HAL open science

Numerical schemes for the simulation of explosion hazards

Nicolas Therme

To cite this version:

Nicolas Therme. Numerical schemes for the simulation of explosion hazards. Numerical Analysis [math.NA]. Aix-Marseille Université, 2015. English. NNT : 2015AIXM4775 . tel-01321731

HAL Id: tel-01321731

https://hal.science/tel-01321731v1

Submitted on 26 May 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

UNIVERSITÉ AIX-MARSEILLE
École Doctorale 184 – Mathématiques et Informatique

Thèse de doctorat

Discipline : Mathématiques appliquées

présentée par

Nicolas Therme

Schémas numériques pour la simulation de l'explosion

---

dirigée par Raphaèle Herbin

Soutenue le 10 décembre 2015 devant le jury composé de :

M. Rémi Abgrall	Université de Zurich	Rapporteur
M. Christophe Chalons	Université de Versailles (UVSQ)	Rapporteur
Mme Laura Gastaldo	IRSN	Encadrant
M. Jean-Marc Hérard	EDF R&D	Examinateur
Mme Raphaèle Herbin	Université Aix-Marseille	Directeur
M. Jean-Claude Latché	IRSN	Examinateur
M. Anthonin Novotny	Université de Toulon	Examinateur
M. Pierre Sagaut	Université Aix-Marseille	Examinateur

Institut de Mathématiques de Marseille
Technopôle Château-Gombert
39, rue Frédéric Joliot-Curie
13453 Marseille Cedex 13

Université Aix-Marseille
Ecole Doctorale de Mathématiques
et Informatique de Marseille
3 Place Victor Hugo
13331 Marseille Cedex 3
case 53, Bat7 - Bloc G1 - 1er étage - Bureau 02

Cette thèse est dédiée à Yves G.
qui m'aura transmis la passion des sciences et de la recherche.

L'essence des mathématiques c'est la liberté.
Georg Cantor

Remerciements

Je tiens tout d'abord à remercier Raphaèle Herbin, ma directrice de thèse, pour ces trois années de conseils et d'aide précieuse, qui m'ont rendu la tâche plus facile. Je remercie aussi mon encadrante Laura qui m'a permis d'aborder le codage dans PRE^{2}MICS de façon plus sereine, et de gagner un temps précieux. Je remercie évidemment Jean-Claude, qui a su tant de fois m'aider à trouver les parades pour débloquer les situations difficiles qui ont jalonné cette thèse.

J'adresse mes remerciements à Rémi Abgrall et Christophe Chalon pour avoir accepté de devenir rapporteurs de ce manuscrit. Je remercie aussi Pierre Sagaut, Jean-Marc Hérard, et Anthonin Novotny pour leur participation à mon jury de thèse.

Je tiens à remercier l'ensemble du LIE pour ces trois années passées dans la joie et la bonne humeur : Jean-Paul, Jean-Marc, Fabrice, Samuel, Thomas, Sylvain, François, Philippe, Laurence, Marc, William, Matthieu, mais aussi Sarah et Pascale. Je remercie aussi Céline pour son aide, sa bonne humeur et les sujets de discussion variés tous les jours à la pause café. Je tiens à remercier Ahmed, Khaled, Germain et Chady avec qui une amitié s'est tissée, source de tant de débats passionnés et passionnants, et de bons moments passés ensemble. Je remercie plus particulièrement Khaled pour son aide précieuse pendant la thèse et ses conseils toujours avisés, un véritable grand frère de thèse.

Je remercie les membres de l'I2M qui ont su m'accueillir chaleureusement durant la fin de ma thèse quand j'ai déménagé de Cadarache à Marseille. Je pense plus particulièrement aux doctorants du labo : Cécile, Flore, Jordan, Damien, Benoit, Vladimiro, Rémi, Thomas. Un remerciement spécial pour Dionysis qui a eu le courage de relire une partie de mon manuscrit à la traque des petites erreurs. Je remercie Thierry, pour sa gentillesse et ses remarques précieuses lors de certaines séances de brainstorming avec Jean-Claude et Raphaèle. Merci aux autres membres du groupe d'AA pour leur sympathie.

Enfin je tiens à remercier mes amis les plus proches pour leur soutien indefectible : Nicolas, Céline, Élodie, Benjamin, et bien evidemment Paul. Mais aussi à ma famille et plus particulièrement mes deux frères Boubou et Fouf, mon père et ma mère, ainsi que mes grands-parents.

Résumé

Résumé

Dans les installations nucléaires, les explosions, qu'elles soient d'origine interne ou externe, peuvent entrainer la rupture du confinement et le rejet de matières radioactives dans l'environnement. Il est donc fondamental, dans un cadre de sûreté de modéliser ce phénomène. La propagation des ondes de choc est modélisée par les équations d'Euler pour un fluide compressible, alors que la phase de déflagration avec propagation du front de flamme est modélisée par les équations de Navier-Stokes avec termes de réaction auxquelles on adjoint une équation de type level-set pour suivre la propagation de la flamme. L'objectif de cette thèse est de contribuer à l'élaboration de schémas numériques performants pour résoudre ces modèles complexes.

Les travaux présentés s'articule autour de deux axes majeurs : le développement de schémas numériques consistants pour les équations d'Euler compressible et celui de schémas performants pour la propagation d'interfaces. On étudie des schémas explicites en temps dans les eux cas, ainsi qu'un schéma de type correction de pression concernant les équations d'Euler. La discrétisation spatiale est de type mailles décalées. Elle se base sur la formulation en énergie interne du système d'Euler, ce qui permet d'en assurer la positivité et évite la discrétisation plutôt difficile de l'énergie totale sur mailles décalées. Un bilan d'énergie cinétique discret est obtenu et un terme source est ajouté dans le bilan d'énergie interne pour permettre de retrouver un bilan d'énergie totale à la limite. Des techniques de montée en ordre de type MUSCL sont utilisées pour la discrétisation des opérateurs convectifs discrets. Elles se basent uniquement sur la vitesse matérielle, et permettent de garantir, sous condition de CFL, la positivité de la masse volumique et de l'énergie interne. On s'assure ainsi que l'énergie totale ne peut croître et on obtient en plus une inégalité d'entropie discrète. Sous des hypothèses de stabilité en normes L^∞ et BV on démontre que si les solutions discrètes du schéma convergent, alors elles le font nécessairement vers la solution faible des équations d'Euler. De plus elles vérifient une inégalité d'entropie faible à la limite.

Concernant la propagation d'interface, on transforme l'équation d'évolution de cette dernière (la "G-equation"), qui est une équation de type Hamilton-Jacobi particulière, en une équation de transport et on utilise les outils déjà introduits pour les équations d'Euler. Il est nécessaire de discrétiser de façon consistante le gradient aux faces. Pour les maillages non réguliers, une construction de type schéma "SUSHI" est utilisée. Cette dernière est modifiée pour les maillages cartésiens afin de pouvoir récupérer des propriétés de monotonie, et de consistance des opérateurs spatiaux discrets à la limite. Ces propriétés permettent de démontrer un résultat de convergence uniforme pour le schéma décentré amont cartésien. Des tests numériques permettent de plus de s'assurer que le schéma converge sur des maillages plus irréguliers.

Mots-clefs

Volumes finis, Équations d'Euler, Hamilton-Jacobi, MUSCL, Maillage décalé, Stabilité, Analyse, fluides compressibles.

Numerical schemes for explosion hazards

Abstract

In nuclear facilities, internal or external explosions can cause confinement breaches and radioactive materials release in the environment. Hence, modeling such phenomena is crucial for safety matters. Blast waves resulting from explosions are modeled by the system of Euler equations for compressible flows, whereas Naviers-Stokes equations with reactive source terms and level set techniques are used to simulate the propagation of flame front during the deflagration phase. The purpose of this thesis is to contribute to the creation of efficient numerical schemes to solve these complex models.

The work presented here focuses on two major aspects: first, the development of consistent schemes for the Euler equations, then the buildup of reliable schemes for the front propagation. In both cases, explicit in time schemes are used, but we also introduce a pressure correction scheme for the Euler equations. Staggered discretization is used in space. It is based on the internal energy formulation of the Euler system, which insures its positivity and avoids tedious discretization of the total energy over staggered grids. A discrete kinetic energy balance is derived from the scheme and a source term is added in the discrete internal energy balance equation to preserve the exact total energy balance at the limit. High order methods of MUSCL type are used in the discrete convective operators, based solely on material velocity. They lead to positivity of density and internal energy under CFL conditions. This ensures that the total energy cannot grow and we can furthermore derive a discrete entropy inequality. Under stability assumptions of the discrete L^∞ and BV norms of the scheme's solutions one can prove that a sequence of converging discrete solutions necessarily converges towards the weak solution of the Euler system. Besides it satisfies a weak entropy inequality at the limit.

Concerning the front propagation, we transform the flame front evolution equation (the so called "G-equation"), which is a particular Hamilton-Jacobi equation, into a transport equation so we can use the methods developed for the Euler system. A consistent gradient discretization at the faces of the mesh is needed though. For irregular meshing a "SUSHI-scheme" technique is used. It is then adapted to cartesian grids in order to get monotonicity of the scheme alongside with the strong consistency of the discrete spatial operators. These joint properties insure a uniform convergence result for the upwind scheme on cartesian grids. Numerical experiments allow to check the convergence of the scheme on more irregular meshes.

Keywords

Finite volumes, Euler equations, Hamilton-Jacobi, Compressible flows, Staggered discretization, MUSCL, Analysis, Stability.

Table des matières

1 Synthèse générale **11**
1.1 Introduction . . . 11
1.2 Modèles physiques . . . 12
1.3 Discrétisation spatiale et temporelle . . . 14
1.4 Équations d'Euler compressible . . . 16
1.5 Equation de propagation du front de flamme . . . 37
1.6 Résultats numériques . . . 45
1.7 Conclusion . . . 55

2 MUSCL-type stable explicit staggered schemes for the compressible Euler equations **57**
2.1 Introduction . . . 57
2.2 Meshes and unknowns . . . 58
2.3 The numerical scheme . . . 61
2.4 Numerical results . . . 75

3 Consistency results of a class of staggered schemes for the compressible Euler equations **93**
3.1 Introduction . . . 93
3.2 Meshes and discretization spaces . . . 94
3.3 Pressure correction and decoupled schemes . . . 96
3.4 Stability properties . . . 103
3.5 Consistency of the schemes . . . 107

4 A class of finite volume schemes for the G-equation **139**
4.1 Introduction . . . 139
4.2 Spatial discretization . . . 140
4.3 The scheme . . . 141
4.4 Properties of the scheme . . . 144
4.5 A convergence result . . . 151
4.6 Numerical results . . . 152

A Euler equations **159**
A.1 Some results concerning explicit finite volume convection operators . . . 159
A.2 Explicit formulas of the WLRs in the MAC case . . . 162
A.3 2D Riemann problems . . . 164

B G-equation **185**
B.1 Viscosity solutions of the eikonal equation . . . 185
B.2 Additional Properties of the scheme . . . 188

Bibliographie **191**

Chapitre 1

Synthèse générale

1.1 Introduction

L'Institut de Radioprotection et de Sûreté Nucléaire (IRSN) a pour vocation première de réaliser des expertises scientifiques pour l'Agence de Sûreté Nucléaire (ASN) française sur des problématiques de sûrété d'installations nucléaires, protection contre les rayonnements ionisants, contrôle des matières nucléaires et prévention des actes de malveillance. Dans cette optique, l'explosion de gaz (hydrogène en particulier) constitue un risque majeur pour les installations nucléaires et de stockage de déchets. L'accident de Fukushima a remis cette préoccupation de sûreté au premier plan et pour y répondre, l'IRSN a décidé de développer un outil dédié à la simulation de l'explosion dénommé P^2REMICS.

La problèmatique de l'explosion peut être découpée en trois thématiques distinctes : la formation de l'atmosphère explosive, l'explosion elle-même, et la propagation d'ondes de souffle qui en résultent. Concernant le premier point, *i.e.* la dispersion du gaz explosif, dans la plupart des cas, la vitesse reste en deçà de quelques (dizaine de) m/s, valeurs pour lesquelles le modèle asymptotique pour les écoulemens à faible nombre de Mach s'applique. Les travaux effectués au cours de cette thèse se concentrent plus particulièrement sur les deux dernières phases. La propagation des ondes de souffle générées par une explosion est étudiée à travers une classe de schémas pour les équations d'Euler compressible. La propagation du front de flamme est quant à elle représentée à l'aide de schémas permettant la résolution d'une équation de type "level-set" : la G-equation.

La classe de schémas présentés dans cette thèse poursuit les développements récents de schémas numériques modélisant des écoulements à tout nombre de Mach à l'IRSN [25, 35, 36, 38]. La discrétisation temporelle est de type explicite en temps pour des écoulements à nombre de Mach élevés (propagation d'ondes de chocs) ou de type correction de pression. Les schémas à correction de pression ont été introduits pour la première fois par Chorin [18] et Temam [60] à la fin des années 60 pour les équations de Navier-Stokes incompressible. Ils s'appuient sur des discrétisations à mailles décalées où les variables scalaires sont discrétisées au centre des mailles et la vitesse aux faces. Dans ce travail deux types de discrétisation sont considérées : l'une est une extension du célèbre schéma MAC développé dans [34, 32, 33] pour le compressible et concerne les maillages cartésiens, l'autre est fondée sur les éléments finis non conformes de type Ranacher-Turek (quadrangles et hexaèdres) [54] ou Crouzeix-Raviart (simplexes) [22]. Les techniques d'interpolation utilisées dans les flux convectifs s'appuient uniquement sur la vitesse matérielle du fluide et non sur la structure des ondes avec la résolution de problèmes de Riemann, technique habituellement utilisée pour les équations hyperboliques ([62, 28, 12, 17]). Néanmoins ils sont utilisés en pratique ([48, 47]), de par la simplicité des flux et leur efficacité. Ils se prêtent particulièrement bien au calcul parallèle.

Ce chapitre s'articule en 5 points. On présente tout d'abord la modélisation physique de l'explosion qui sert de cadre général à ces travaux de thèse avant d'introduire la description

détaillée des discrétisations temporelle et spatiale utilisées. On étudie ensuite la propagation des ondes de souffle générées par une explosion à travers une classe de schémas pour les équations d'Euler compressible. La propagation du front de flamme est quant à elle représentée à l'aide de schémas permettant la résolution d'une équation de type "level-set" : la G-equation. On terminera par quelques résultats numériques illustratifs.

1.2 Modèles physiques

La problèmatique de l'explosion peut être découpée en trois thématiques distinctes : la formation de l'atmosphère explosive, l'explosion elle-même, et la propagation d'ondes de souffle qui en résultent. Ces travaux de thèse se contentrent plus particulièrement sur les deux derniers points. On présente dans les paragraphes qui suivent les modèles physiques mis en jeux et à partir desquels on va développer des schémas numériques.

1.2.1 Propagation des ondes de souffle

L'objectif est ici d'évaluer les conséquences d'une explosion tout en s'affranchissant du calcul de cette dernière. On se donne simplement des conditions initiales représentatives de l'état de l'atmosphère à l'issue de l'explosion, avec une zone limitée de forte surpression, et l'on calcule la propagation de l'onde de souffle résultante et son interaction avec les structures avoisinantes. Pour ce calcul de propagation on se place dans des situations où la simulation numérique directe est possible. Ceci présuppose que l'on s'intéresse aux effets sur un milieu comportant un nombre limité de structures, et en temps court, *i.e.* avant l'apparition de réflexions multiples et de phénomènes turbulents. Il s'agit alors de résoudre les équations d'Euler pour un fluide compressible, en l'absence totale de loi de fermeture empirique. Le système d'équations considéré est donc le suivant :

$$\partial_t \rho + \operatorname{div}(\rho\, \boldsymbol{u}) = 0, \tag{1.1a}$$

$$\partial_t (\rho\, \boldsymbol{u}) + \operatorname{div}(\rho\, \boldsymbol{u} \otimes \boldsymbol{u}) + \boldsymbol{\nabla} p = 0, \tag{1.1b}$$

$$\partial_t (\rho\, E) + \operatorname{div}(\rho\, E\, \boldsymbol{u}) + \operatorname{div}(p\, \boldsymbol{u}) = 0, \tag{1.1c}$$

$$p = (\gamma - 1)\, \rho\, e, \qquad E = \frac{1}{2}|\boldsymbol{u}|^2 + e, \tag{1.1d}$$

où t désigne le temps, ρ,$\boldsymbol{u}$, p, E et e la masse volumique, la vitesse, la pression, l'énergie totale et l'énergie interne respectivement, et $\gamma > 1$ est le rapport des capacités thermiques massiques à pression et à volume constants du fluide considéré. La loi d'état qui relie la pression, la masse volumique et l'énergie est de type gaz parfait. On suppose que le problème est posé sur $\Omega \times (0, T)$, avec Ω ouvert borné connexe de $\mathbb{R}^d$, $1 \leq d \leq 3$, et $(0, T)$ un intervalle de temps fini. Le système (1.1) est complété par des données initiales pour ρ, e et $\boldsymbol{u}$ que l'on note ρ_0, $\boldsymbol{u}_0$ et e_0 respectivement, avec $\rho_0 > 0$ et $e_0 > 0$. La condition au bord est de type Dirichlet *i.e.* , $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ presque partout sur $\partial\Omega$ et pour tout $t \in [0, T)$, $\boldsymbol{n}$ étant le vecteur normal à la frontière $\partial\Omega$.

Supposons maintenant que la solution du problème précédent soit régulière. On introduit l'énergie cinétique du système $E_k = \frac{1}{2}|\boldsymbol{u}|^2$. En prenant le produit scalaire de (1.1b) et $\boldsymbol{u}$ on obtient, en utilisant l'équation de bilan de masse (1.1a) :

$$\partial_t (\rho E_k) + \operatorname{div}\Big(\rho\, E_k\, \boldsymbol{u}\Big) + \boldsymbol{\nabla} p \cdot \boldsymbol{u} = 0. \tag{1.2}$$

Cette équation correspond au bilan d'énergie cinétique. Si on soustrait ce bilan au bilan d'énergie totale (1.1c), on aboutit à une équation de conservation de l'énergie interne :

$$\partial_t (\rho e) + \operatorname{div}(\rho e \boldsymbol{u}) + p \operatorname{div} \boldsymbol{u} = 0. \tag{1.3}$$

On peut observer que

— grâce au bilan de masse, les deux premiers termes de l'équation d'énergie interne (1.3) peuvent se réécrire de la façon suivante :

$$\partial_t(\rho e) + \mathrm{div}(\rho e \boldsymbol{u}) = \rho\left[\partial_t e + \boldsymbol{u}\cdot\nabla e\right],$$

soit sous forme d'un opérateur de transport qui vérifie un principe du maximum,

— et que par l'intermédiaire de la loi d'état on a $p = 0$ quand $e = 0$.

L'équation (1.3) nous assure donc, grâce à la donnée initiale strictement positive, et à des conditions aux limites adaptées, que l'énergie interne reste positive pour tout $t \in [0, T)$. Le bilan de masse assure quant à lui la positivité de la masse volumique.

1.2.2 Explosion : phase de déflagration

Dans les scénarios d'intérêt pour l'analyse de sûreté, les sources d'ignition potentielles ne sont pas, en général, suffisamment énergétiques pour déclencher directement une détonation. L'explosion survient donc sous la forme d'une déflagration, susceptible de transiter vers une détonation. On choisit donc de caractériser la propagation d'un front de flamme subsonique dans un milieu partiellement prémélangé au repos ou turbulent.

Le système complet des équations décrivant le phénomène de la déflagration étant relativement complexe et en dehors du sujet principal de cette thèse, on en présentera simplement les grandes lignes. L'hydrodynamique de l'écoulement est décrite par le système des équations de Navier-Stokes compressible instationnaire avec des termes de turbulence. À ces équations s'ajoutent les équations décrivant le transport des différentes espèces chimiques intervenant dans la réaction explosive. Des termes sources sont ajoutés afin de modéliser la réaction chimique. Enfin le système global est fermé grâce à l'équation d'énergie.

Ce modèle est complexe à mettre en oeuvre et les calculs peuvent s'avérer couteux et fastidieux. Comme la réaction de combustion est très raide et que le front est localisé, un modèle de vitesse de flamme est utilisé pour la modélisation de la combustion. Les modèles de vitesse de flamme sont usuellement écrits pour des systèmes parfaitement prémélangés. Ils permettent de condenser l'ensemble de la chimie du phénomène en une variable simple qu'est la vitesse de flamme. Ils supposent l'écoulement composé d'une phase brûlée et d'une phase imbrûlée séparées entre elles par une interface de flamme. Cette dernière est localisée en espace grâce à une variable d'avancement de la combustion qui représente à la fois la température adimensionnée et les fractions massiques (de combustible, d'oxydant ou de produits de combustion) adimensionnées, qui obéissent à la même équation de bilan et aux mêmes conditions initiales et aux limites. Ce modèle peut être conceptuellement étendu en introduisant une variable caractéristique de l'écoulement, souvent notée G et en supposant que la composition locale peut s'en déduire. Le modèle le plus simple consiste alors à se donner une valeur G_0 et à supposer qu'on se situe dans les gaz brûlés si $G < G_0$ et les gaz frais si $G > G_0$.

La variable G obéit à une équation de transport, dite « G-équation » :

$$\partial_t(\rho G) + \mathrm{div}(\rho G \boldsymbol{u}) + \rho_u u_f |\nabla G| = 0,$$

où ρ_u désigne la masse volumique des gaz frais au voisinage de la zone de réaction et u_f la vitesse de flamme. Sous forme non-conservative, cette relation devient :

$$\partial_t G + \left(\frac{\rho_u}{\rho} u_f \frac{\nabla G}{|\nabla G|} + \boldsymbol{u}\right)\cdot\nabla G = 0.$$

L'un des objectifs de cette thèse étant de développer un schéma pour résoudre la G-équation, on remarque grâce à la relation précédente que l'on peut se réduire à écrire un schéma volumes finis pour une équation de type Hamilton-Jacobi particulière :

$$\partial_t G + H(\nabla G) = 0, \qquad \text{avec} \quad H(\boldsymbol{x}) = \boldsymbol{u}\cdot\boldsymbol{x} + \frac{\rho_u}{\rho} u_f |\boldsymbol{x}|.$$

Le terme original par rapport à une simple équation de transport réside dans le terme $\frac{\rho_u}{\rho} u_f |\boldsymbol{x}|$. De ce fait, la suite de l'étude portera sur le problème canonique suivant :

$$\partial_t G + |\nabla G| = 0,$$

problème défini sur un ouvert borné connexe de $\mathbb{R}^d$ (ou $\mathbb{R}^d$ tout entier), avec une donnée initiale $G_0 \in [0, 1]$.

1.3 Discrétisation spatiale et temporelle

On suppose que le problème est posé sur un domaine Ω, ouvert borné connexe de $\mathbb{R}^d$, $d \in [\![1, 3]\!]$, et sur l'intervalle de temps $(0, T)$ discrétisé de manière uniforme avec un pas de temps δt. Soit $\mathcal{M}$ une décomposition du domaine Ω que l'on suppose régulière au sens usuel des éléments finis. Les mailles peuvent être :

— pour un domaine quelconque Ω, soit des simplexes, soit des quadrilatères ($d = 2$) ou des héxaèdres ($d = 3$) convexes, soit une combinaison des deux,

— pour un domaine dont les frontières sont des hyperplans orthogonaux à un des vecteurs de la base canonique, des rectangles ($d = 2$) ou des parallélépipèdes rectangles ($d = 3$).

On définit $\mathcal{E}$ et $\mathcal{E}(K)$ l'ensemble des faces du maillage $\mathcal{M}$ et de la maille $K \in \mathcal{M}$ respectivement. L'ensemble des faces à la frontière du domaine est noté $\mathcal{E}_{\text{ext}}$, et celui des faces internes $\mathcal{E}_{\text{int}}$. La face interne $\sigma \in \mathcal{E}_{\text{int}}$ séparant les cellules K et L est notée $\sigma = K|L$. On note $\boldsymbol{n}_{K,\sigma}$ le vecteur normal à la face σ sortant de la cellule K. Pour $K \in \mathcal{M}$ et $\sigma \in \mathcal{E}$, on note $|K|$ et $|\sigma|$ les mesures respectives de K et σ, et on note h_K le diamètre de la maille K. Pour $i \in [\![1, d]\!]$, $\mathcal{E}^{(i)} \subset \mathcal{E}$ et $\mathcal{E}^{(i)}_{\text{ext}} \subset \mathcal{E}_{\text{ext}}$ sont les sous ensembles des faces de $\mathcal{E}$ et $\mathcal{E}_{\text{ext}}$ perpendiculaires au $i^{\text{ème}}$ vecteur de la base canonique de $\mathbb{R}^d$. On définit enfin la mesure caractéristique du maillage par $h_{\mathcal{M}} = \max_{K \in \mathcal{M}} h_K$.

La discrétisation spatiale est de type « à mailles décalées ». Quand le maillage est cartésien on utilise un schéma de type MAC (Marker-And Cell). Lorsque le maillage est quelconque, les degrés de liberté sont placés de façon similaire aux éléments finis de type Ranacher-Turek pour les maillages quadrilatéraux ou héxaédriques, ou aux éléments finis de type Crouzeix-Raviart pour les maillages simplectiques.

Pour toutes ces discrétisations, les degrés de liberté des inconnues scalaires (pression, masse volumique, énergie interne, indicatrice de flamme) sont associés aux cellules du maillage $\mathcal{M}$ et sont notés :

$$\{\rho_K, p_K, e_K, G_K,\ K \in \mathcal{M}\}.$$

Les degrés de liberté des inconnues vectorielles (ici la vitesse $\boldsymbol{u}$) dépendent du type de schéma utilisé.

— Discrétisations de type **Rannacher-Turek** (RT) et **Crouzeix-Raviart** (CR) : les inconnues discrètes de la vitesse sont situées au centre des faces du maillage. Les conditions de type Dirichlet sont prises en compte en annulant les inconnues des faces à la frontière du domaine. L'ensemble des inconnues vitesse est noté :

$$\{u_{\sigma,i},\ \sigma \in \mathcal{E}_{\text{int}},\ 1 \leq i \leq d\}.$$

— Discrétisation de type **MAC** : les degrés de liberté pour la $i^{\text{ème}}$ composante de la vitesse sont situés au centre des faces $\sigma \in \mathcal{E}^{(i)}_{\text{int}}$, et l'ensemble des inconnues discrètes est donné par :

$$\left\{u_{\sigma,i},\ \sigma \in \mathcal{E}^{(i)}_{\text{int}},\ 1 \leq i \leq d\right\}.$$

On décrit maintenant le maillage dual, qui sera utilisé pour l'approximation volumes finis du terme convectif et de la dérivée temporelle dans le bilan de quantité de mouvement.

— Discrétisations de type **Rannacher-Turek** (RT) et **Crouzeix-Raviart** (CR) :
le maillage dual est le même pour toutes les composantes de la vitesse. Pour $K \in \mathcal{M}$ simplexe, rectangle ou cuboïde, on désigne par $D_{K,\sigma}$, pour $\sigma \in \mathcal{E}(K)$, le cône de base σ et de sommet le centre de gravité de K. On obtient ainsi une partition de K en m sous volumes, avec m le nombre de faces, telle que chaque sous volume est de mesure identique $|D_{K,\sigma}| = \frac{|K|}{m}$. On généralise cette définition aux quadrangles et hexaèdres quelconques en supposant qu'on a construit une partition de mesure et connectivités identiques. On appelle $D_{K,\sigma}$ la demi-maille diamant associée à K et σ. Pour $\sigma = K|L \in \mathcal{E}_{\text{int}}$, on définit la maille diamant D_σ associée à σ par $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$.
— Discrétisation de type **MAC** : le maillage dual dépend de la composante de la vitesse. Pour chaque composante, le maillage dual diffère du précédent dans le choix des demi-mailles diamant. Pour $K \in \mathcal{M}$ et $\sigma \in \mathcal{E}(K)$, $D_{K,\sigma}$ désigne le rectangle ou le parallélépipède rectangle de base σ et de mesure $|D_{K,\sigma}| = |K|/2$.

On note $|D_\sigma|$ la mesure de la maille diamant D_σ, et par $\epsilon = D_\sigma|D_{\sigma'}$ la face séparant les deux mailles diamant D_σ et $D_{\sigma'}$. L'ensemble des faces de D_σ est noté $\tilde{\mathcal{E}}(D_\sigma)$.

Dans une optique d'unification des notations entre les différentes discrétisations, on introduit l'ensemble des faces $\mathcal{E}_{\mathcal{S}}^{(i)}$ associées aux degrés de liberté de la $i^{\text{ème}}$ composante de la vitesse ($\mathcal{S}$ désigne le schéma) :

$$\mathcal{E}_{\mathcal{S}}^{(i)} = \left| \begin{array}{l} \mathcal{E}^{(i)} \setminus \mathcal{E}_{\text{ext}}^{(i)} \text{ pour le schéma MAC,} \\ \mathcal{E} \setminus \mathcal{E}_{\text{ext}}^{(i)} \text{ pour les schémas RT et CR.} \end{array} \right.$$

En procédant de la même façon pour les notations du maillage dual, on a :

$$\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)} = \left| \begin{array}{l} \tilde{\mathcal{E}}^{(i)} \setminus \tilde{\mathcal{E}}_{\text{ext}}^{(i)} \text{ pour le schéma MAC,} \\ \tilde{\mathcal{E}} \setminus \tilde{\mathcal{E}}_{\text{ext}}^{(i)} \text{ pour les schémas RT et CR.} \end{array} \right.$$

Afin de traiter les conditions d'imperméabilité (*i.e.* $\boldsymbol{u} \cdot \boldsymbol{n} = 0$), on suppose, dans un souci de simplicité, que les faces extérieures sont toutes orthogonales à une des directions spatiales, ce qui permet d'imposer la nullité des inconnues de vitesse correspondantes :

$$\text{pour } i = 1, \ldots, d, \ \forall \sigma \in \mathcal{E}_{\text{ext}}^{(i)}, \qquad u_{\sigma,i} = 0. \tag{1.4}$$

Dans la cas de maillages plus généraux, il est toujours possible de redéfinir, par combinaisons linéaires, les degrés de liberté au niveau des faces externes, de manière à introduire la vitesse normale comme nouveau degré de liberté.

On donne maintenant la définition d'un maillage admissible. Soit $\mathcal{P} = (\boldsymbol{x}_K)_{K \in \mathcal{M}}$ un ensemble de points de Ω tel que, $\forall K \in \mathcal{M}$, $\boldsymbol{x}_K \in K$. Pour tout $\sigma \in \mathcal{E}_{\text{int}}$, on note $\boldsymbol{x}_\sigma$ le centre de gravité de la face σ, et on suppose qu'il existe un ensemble de mailles voisines V_σ tel que :

$$\boldsymbol{x}_\sigma = \sum_{K \in V_\sigma} \alpha_{K,\sigma} \boldsymbol{x}_K \qquad \sum_{K \in V_\sigma} \alpha_{K,\sigma} = 1.$$

Définition 1.1 (*Maillage admissible*)

L'ensemble $(\mathcal{M}, \mathcal{E}, \mathcal{P})$ est dit admissible si et seulement si :
— pour tout $\sigma = K|L \in \mathcal{E}_{\text{int}}$, $\overrightarrow{\boldsymbol{x}_K \boldsymbol{x}_L}$ est perpendiculaire à la face σ.

Pour tout $\sigma = K|L \in \mathcal{E}_{\text{int}}$, D_σ est donc le diamant de base σ et de sommets respectifs $\boldsymbol{x}_K$ et $\boldsymbol{x}_L$. On note enfin $d_\sigma = |\boldsymbol{x}_L - \boldsymbol{x}_K|$.

Ceci permet de définir le gradient normal à la face grâce à un schéma à deux points. Pour plus de détails voir [24].

1.4 Équations d'Euler compressible

L'objectif de cette section est de présenter une classe de schémas volume finis à mailles décalées pour les équations d'Euler compressible (1.1) qui permet :

— d'obtenir un bilan d'énergie cinétique discret, analogue discret de (1.2),
— d'augmenter la précision des résultats numériques via l'utilisation d'interpolations spatiales d'ordre élevé avec le schéma explicite,
— de vérifier la consistance au sens de Lax des schémas explicite et semi-implicite (présentés dans les sections suivantes) avec la version faible du systeme d'Euler,
— de vérifier une inégalité d'entropie faible à la limite.

1.4.1 Schémas numériques

Considérons une partition $0 = t_0 < t_1 < \cdots < t_N = T$ de l'intervalle de temps $(0, T)$ que l'on suppose uniforme et notons $\delta t = t_1 - t_0$ le pas de temps. Pour résoudre le système (1.1), on introduit deux algorithmes. Le premier est purement découplé et verifiera des hypothèses de stabilité sous conditions de CFL comme nous le verrons par la suite. Le second est semi-implicite, et inconditionnellement stable.

Schéma explicite

Le schéma explicite est le suivant :

$$\forall K \in \mathcal{M}, \quad \frac{1}{\delta t}(\rho_K^{n+1} - \rho_K^n) + \mathrm{div}(\rho^n \boldsymbol{u}^n)_K = 0, \tag{1.5a}$$

$$\forall K \in \mathcal{M}, \quad \frac{1}{\delta t}(\rho_K^{n+1} e_K^{n+1} - \rho_K^n e_K^n) + \mathrm{div}(\rho^n e^n \boldsymbol{u}^n)_K + p_K^n(\mathrm{div}(\boldsymbol{u}^n))_K = S_K^n, \tag{1.5b}$$

$$\text{Pour } 1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}, \quad \frac{1}{\delta t}\left(\rho_{D_\sigma}^{n+1} u_{\sigma,i}^{n+1} - \rho_{D_\sigma}^n u_{\sigma,i}^n\right) + \mathrm{div}(\rho^n u_i^n \boldsymbol{u}^n)_\sigma + (\nabla p)_{\sigma,i}^{n+1} + \mathcal{D}(u_i^n)_{\sigma,i} = 0, \tag{1.5c}$$

$$\forall K \in \mathcal{M}, \quad p_K^{n+1} = (\gamma - 1)\, \rho^{n+1}\, e_K^{n+1}. \tag{1.5d}$$

Schéma de correction de pression

Le schéma semi-implicite considéré ici entre dans le cadre des schémas à correction de pression et consiste en plusieurs étapes distinctes :

1- Scaling du gradient de pression :

$$\forall \sigma \in \mathcal{E}_{\mathrm{int}}, \quad (\overline{\nabla p})_\sigma^{n+1} = \left(\frac{\rho_{D_\sigma}^n}{\rho_{D_\sigma}^{n-1}}\right)^{1/2} (\nabla p^n)_\sigma. \tag{1.6a}$$

2- Étape de prédiction – Résoudre en $\tilde{\boldsymbol{u}}^{n+1}$:

$$\begin{gathered}\text{Pour } 1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}, \\ \frac{1}{\delta t}\left(\rho_{D_\sigma}^n \tilde{u}_{\sigma,i}^{n+1} - \rho_{D_\sigma}^{n-1} u_{\sigma,i}^n\right) + \mathrm{div}(\rho^n \tilde{u}_i^{n+1} \boldsymbol{u}^n)_\sigma + (\overline{\nabla p})_{\sigma,i}^{n+1} + \mathcal{D}(\tilde{u}_i^{n+1})_{\sigma,i} = 0.\end{gathered} \tag{1.6b}$$

3- Étape de correction – Résoudre en p^{n+1}, e^{n+1}, ρ^{n+1} et $\boldsymbol{u}^{n+1}$:

$$\text{Pour } 1 \leq i \leq d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)},\quad \frac{1}{\delta t}\,\rho^n_{D_\sigma}\,(u^{n+1}_{\sigma,i} - \tilde{u}^{n+1}_{\sigma,i}) + (\boldsymbol{\nabla} p^{n+1})_{\sigma,i} - (\overline{\boldsymbol{\nabla} p})^{n+1}_{\sigma,i} = 0, \tag{1.6c}$$

$$\forall K \in \mathcal{M},\quad \frac{1}{\delta t}(\rho^{n+1}_K - \rho^n_K) + \operatorname{div}(\rho^{n+1}\boldsymbol{u}^{n+1})_K = 0, \tag{1.6d}$$

$$\forall K \in \mathcal{M},\quad \frac{1}{\delta t}(\rho^{n+1}_K e^{n+1}_K - \rho^n_K e^n_K) + \operatorname{div}(\rho^{n+1} e^{n+1}\boldsymbol{u}^{n+1})_K + p^{n+1}_K \operatorname{div}(\boldsymbol{u}^{n+1})_K = S^{n+1}_K, \tag{1.6e}$$

$$\forall K \in \mathcal{M},\quad p^{n+1}_K = (\gamma - 1)\,\rho^{n+1}\,e^{n+1}_K. \tag{1.6f}$$

L'étape de prédiction est, de façon classique, une résolution semi-implicite du bilan de quantité de mouvement permettant d'obtenir une prédiction de vitesse. L'étape suivante est une étape non linéaire de correction de pression, qui couple le bilan de masse et le bilan d'énergie interne pour des raisons de stabilité. De plus le schéma va ainsi préserver les discontinuités de contact 1D. On détaille dans la section suivante la discrétisation spatiale employée.

1.4.2 Discrétisation spatiale

On détaille les opérateurs spatiaux équation par équation.

Bilan de masse – Les équations (1.6d) et (1.5a) correspondent à une discrétisation volumes finis du bilan de masse sur les mailles primales du maillage. Pour des champs discrets ρ et $\boldsymbol{u}$, on note

$$\operatorname{div}(\rho\boldsymbol{u})_K = \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}(\rho\boldsymbol{u}),$$

avec $F_{K,\sigma}(\rho\boldsymbol{u})$ le flux de masse sortant de K à travers σ. Il s'écrit de la façon suivante :

$$\forall \sigma = K|L \in \mathcal{E}_{\text{int}},\qquad F_{K,\sigma}(\rho\boldsymbol{u}) = |\sigma|\,\rho_\sigma\,u_{K,\sigma}, \tag{1.7}$$

avec $u_{K,\sigma}$ approximation de la vitesse normale à σ sortante de K. Elle est définie par :

$$u_{K,\sigma} = \left|\begin{array}{l} u^n_{\sigma,i}\,\boldsymbol{e}^{(i)}\cdot\boldsymbol{n}_{K,\sigma} \text{ pour } \sigma \in \mathcal{E}^{(i)} \text{ avec le schéma MAC,} \\ \boldsymbol{u}^n_\sigma\cdot\boldsymbol{n}_{K,\sigma} \text{ pour } \sigma \in \mathcal{E} \text{ avec les schémas RT/CR,} \end{array}\right. \tag{1.8}$$

avec $\boldsymbol{e}^{(i)}$ l'$i^{\text{ème}}$ vecteur de la base orthonormale de $\mathbb{R}^d$. A noter que, de par les conditions d'imperméabilité, $u_{K,\sigma}$ est nul sur toutes les faces externes.

La masse volumique interpolée à la face interne $\sigma = K|L$ est supposée satisfaire la propriété suivante :

$$\begin{array}{l} \forall K \in \mathcal{M},\ \forall \sigma = K|L \in \mathcal{E}(K)\cap\mathcal{E}_{\text{int}}, \\ \qquad\qquad \text{il existe } \alpha_{K,\sigma} \in [0,1] \text{ et une cellule voisine } M^K_\sigma \text{ de } K \text{ tels que :} \\ \qquad\qquad\qquad \rho_\sigma - \rho_K = \left|\begin{array}{ll} \alpha_{K,\sigma}(\rho_K - \rho_{M^K_\sigma}) & \text{si } u_{K,\sigma} \geq 0, \\ \alpha_{K,\sigma}(\rho_L - \rho_K) & \text{sinon .} \end{array}\right. \end{array} \tag{1.9}$$

Le choix upwind correspond au cas $\alpha_{K,\sigma} = 0$ si $u_{K,\sigma} \geq 0$ et $\alpha_{K,\sigma} = 1$ sinon. Dans le paragraphe 1.4.3, nous présenterons une technique de type MUSCL qui satisfait cette relation.
A noter que de cette propriété découle naturellement le fait que ρ_σ est une combinaison convexe de ρ_K et ρ_L :

$$\exists\,\alpha_\sigma \in [0,1] \text{ tel que,}\quad \rho_\sigma = \alpha_\sigma\rho_K + (1-\alpha_\sigma)\rho_L. \tag{1.10}$$

Équation de conservation de l'énergie interne – Les équations (1.6e) et (1.5b) sont discrétisées en volumes finis sur les mailles primales du maillage. On a tout naturellement pour le terme de convection :

$$\mathrm{div}(\rho e \boldsymbol{u})_K = \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}(\rho \boldsymbol{u})\, e_\sigma,$$

avec l'interpolée de l'énergie interne e_σ qui vérifie un analogue de la relation (1.9) :

$$\text{Pour } \sigma = K|L \in \mathcal{E}_{\text{int}}, \qquad e_\sigma - e_K = \left| \begin{array}{ll} \beta_{K,\sigma}(e_K - e_{M_\sigma^K}) & \text{si } F_{K,\sigma} \geq 0, \\ \beta_{K,\sigma}(e_L - e_K) & \text{sinon,} \end{array} \right. \tag{1.11}$$

avec la même cellule M_σ^K que pour ρ_σ mais un coefficient $\beta_{K,\sigma}$ différent. On suppose de plus que ce choix de coefficient $\beta_{K,\sigma}$ permette d'écrire :

$$\rho_\sigma e_\sigma = \alpha_\sigma \rho_K e_K + (1 - \alpha_\sigma)\rho_L e_L, \tag{1.12}$$

où $\alpha_{K,\sigma}$ le même coefficient que dans (1.10). Cette construction particulière permet de montrer que le schéma préserve les discontinuités de contact (1D), comme on le verra dans la section suivante.

La divergence discrète de la vitesse est définie de la façon suivante :

$$\text{Pour } K \in \mathcal{M}, \qquad (\mathrm{div}\boldsymbol{u})_K = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} |\sigma|\, u_{K,\sigma}. \tag{1.13}$$

Le terme source S_K sera déterminé par des arguments de consistance dans la section 1.4.5. Notons simplement qu'il s'agit d'un terme correctif issu du bilan d'énergie cinétique discret que nous détaillerons plus tard et qui permet d'obtenir les bonnes vitesses de choc, en s'assurant en particulier que sous des estimations suffisantes, un passage à la limite du schéma donne bien une solution faible des équations d'Euler.

Bilan de quantité de mouvement – Les bilans de quantité de mouvement discrets (1.6b) et (1.5c) sont obtenus par discrétisation volumes finis sur le maillage dual. Le terme de convection s'écrit :

$$\mathrm{div}(\rho \tilde{u}_i \boldsymbol{u})_\sigma = \sum_{\epsilon \in \tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}(\rho, \boldsymbol{u})\, (u_i)_\epsilon,$$

avec $F_{\sigma,\epsilon}(\rho, \boldsymbol{u})$ qui correspond au flux de masse à travers les faces duales du maillage, et dont la définition dépend du type de discrétisation utilisé.

— *Discrétisations Rannacher-Turek et Crouzeix-Raviart* – Pour $K \in \mathcal{M}$ et $\sigma \in \mathcal{E}(K)$, on définit ξ_K^σ de la façon suivante :

$$\xi_K^\sigma = \frac{|D_{K,\sigma}|}{|K|}.$$

Avec la définition de maillage dual donnée dans la section 1.4.4, ξ_K^σ est indépendant de K et σ. En effet, pour les éléments de type RT, $\xi_K^\sigma = \frac{1}{2d}$, et pour les éléments de type CR $\xi_K^\sigma = \frac{1}{d+1}$. On suppose que les flux duaux soient nuls sur les faces duales externes et qu'ils vérifient les propriétés suivantes sur les faces duales internes.

Définition 1.2 (*Définition des flux de masse duaux*)

Les flux de masse à travers les faces du maillage dual doivent vérifier les trois contraintes suivantes :

(H1) le bilan de masse est vérifié à travers les demi-mailles diamant dans le sens suivant : pour toute cellule $K \in \mathcal{M}$, l'ensemble des flux duaux inclus dans K, $(F_{\sigma,\epsilon})_{\epsilon \subset K}$ vérifie

le système linéaire

$$F_{K,\sigma} + \sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma),\ \epsilon\subset K} F_{\sigma,\epsilon} = \sum_{\sigma'\in\mathcal{E}(K)} F_{K,\sigma'}, \qquad \sigma\in\mathcal{E}(K)\,; \tag{1.14}$$

(H2) les flux sont conservatifs, *i.e.* pour $\epsilon = D_\sigma|D_{\sigma'}$ on a $F_{\sigma,\epsilon} = -F_{\sigma',\epsilon}$;
(H3) les flux duaux sont bornés par les flux de masse primaux dans le sens suivant :

$$|F_{\sigma,\epsilon}| \le C\ \max\left\{|F_{K,\sigma}|,\ \sigma\in\mathcal{E}(K)\right\}, \quad K\in\mathcal{M},\ \sigma\in\mathcal{E}(K),\ \epsilon\in\tilde{\mathcal{E}}(D_\sigma),\ \epsilon\subset K. \tag{1.15}$$

La définition 1.2 est incomplète, le systeme linéaire (1.14) ayant une infinité de solutions. Elle est néanmoins suffisante pour les développements théoriques qui seront effectués au cours de ce travail.

— *Maillage MAC* – Pour $\sigma\in\mathcal{E}^{(i)}$, $\sigma = K|L$, on distingue deux cas :
  — le vecteur $\boldsymbol{e}^i$ est orthogonal à ϵ, ainsi la maille primale K contient ϵ. Soit σ' la seconde face de K orthogonale à $\boldsymbol{e}^i$, de sorte que $\epsilon = D_\sigma|D_{\sigma'}$. On a alors

$$F_{\sigma,\epsilon}(\rho,\boldsymbol{u}) = \frac{1}{2}\left[F_{K,\sigma}(\rho,\boldsymbol{u})\,\boldsymbol{n}_{D_\sigma,\epsilon}\cdot\boldsymbol{n}_{K,\sigma} + F_{K,\sigma'}(\rho,\boldsymbol{u})\,\boldsymbol{n}_{D_\sigma,\epsilon}\cdot\boldsymbol{n}_{K,\sigma'}\right]; \tag{1.16}$$

  — le vecteur $\boldsymbol{e}^i$ est tangent à ϵ qui est alors l'union de deux demi-faces $\tau\in\mathcal{E}(K)$ et $\tau'\in\mathcal{E}(K)$. Dans ce cas

$$F_{\sigma,\epsilon}(\rho,\boldsymbol{u}) = \frac{1}{2}\left[F_{K,\tau}(\boldsymbol{u}) + F_{L,\tau'}(\boldsymbol{u})\right]. \tag{1.17}$$

La masse volumique sur le maillage dual est quant à elle donnée par :

$$\begin{aligned} &\text{Pour } \sigma\in\mathcal{E}_{\text{int}},\ \sigma = K|L && |D_\sigma|\,\rho_{D_\sigma} = |D_{K,\sigma}|\,\rho_K + |D_{L,\sigma}|\,\rho_L,\\ &\text{Pour } \sigma\in\mathcal{E}_{\text{ext}},\ \sigma\in\mathcal{E}(K), && \rho_{D_\sigma} = \rho_K. \end{aligned} \tag{1.18}$$

Quelle que soit la discrétisation utilisée, le bilan de masse sur les mailles diamants est vérifié pour les deux types de discrétisations temporelles :

$$\forall\sigma\in\mathcal{E}, \qquad \frac{|D_\sigma|}{\delta t}\left(\rho^{n+1}_{D_\sigma} - \rho^n_{D_\sigma}\right) + \sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)} F^n_{\sigma,\epsilon} = 0, \tag{1.19}$$

pour le schéma découplé et

$$\forall\sigma\in\mathcal{E}, \qquad \frac{|D_\sigma|}{\delta t}\left(\rho^{n}_{D_\sigma} - \rho^{n-1}_{D_\sigma}\right) + \sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)} F^n_{\sigma,\epsilon} = 0, \tag{1.20}$$

pour le schéma de correction de pression. Pour une explication plus détaillée sur la façon de construire les flux duaux, on peut consulter [3, 27].

Le flux de masse étant nul sur les faces duales à la frontière, on a seulement besoin de définir $u^n_{\epsilon,i}$ sur les faces duales internes. La discrétisation choisie est centrée, *i.e.* , pour $\epsilon = D_\sigma|D_{\sigma'}$, $u_{\epsilon,i} = (u_{\sigma,i} + u_{\sigma',i})/2$.

Le terme $(\boldsymbol{\nabla} p)_{\sigma,i}$ correspond à la $i^{\text{ème}}$ composante du gradient discret de p sur la face σ. Ce gradient est construit comme l'opérateur dual de la divergence discrète, *i.e.* de sorte à obtenir la relation duale pour le produit scalaire L^2 :

$$\sum_{K\in\mathcal{M}} |K|\,p_K\,(\mathrm{div}\boldsymbol{u})_K + \sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}^{(i)}_S} |D_\sigma|\,u_{\sigma,i}\,(\boldsymbol{\nabla} p)_{\sigma,i} = 0, \tag{1.21}$$

ce qui conduit à l'expression suivante :

$$\text{Pour } \sigma = K|L \in \mathcal{E}_{\text{int}}, \qquad (\nabla p)_{\sigma,i} = \frac{|\sigma|}{|D_\sigma|}\,(p_L - p_K)\,\boldsymbol{n}_{K,\sigma}\cdot \boldsymbol{e}^{(i)}. \tag{1.22}$$

Le gradient de pression n'a pas besoin d'être défini sur les faces externes, de par les conditions d'imperméabilité.

Le dernier terme dans le bilan de quantité de mouvement est un terme de diffusion. Il comprend à la fois les termes de diffusion dûs à une interpolation de type upwind, mais aussi une éventuelle diffusion numérique additionnelle de type viscosité non linéaire, calculée à partir de la régularité de la solution au pas de temps précédent :

$$\mathcal{D}(u_i)_{\sigma,i} = \sum_{\epsilon = D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \mu_\epsilon\,(u_{\sigma,i} - u_{\sigma',i}).$$

Pour assurer la consistance du schéma, il est nécessaire que ce terme se comporte comme un terme de diffusion avec une viscosité évanescente, ce qui est possible si le terme μ_ϵ se comporte comme $|\sigma|\,h_\epsilon^{\zeta-1}$ avec h_ϵ distance caractéristique de la maille diamant D_σ et ζ réel strictement positif. Ceci est vérifié dans le cas de la viscosité upwind, $\mu_\epsilon^U = \frac{|F_{\sigma,\epsilon}|}{2}$, compte tenu de l'hypothèse (H3) de la définition 1.2. On verra par la suite, dans le paragraphe 1.4.4, deux méthodes pour construire une viscosité artificielle pour le schéma découplé. Pour la suite, afin de clarifier les notations, on va supposer que cette viscosité est de la forme :

$$\mu_\epsilon = \frac{|F_{\sigma,\epsilon}|}{2} + \nu_\epsilon, \tag{1.23}$$

avec ν_ϵ, une viscosité aditionnelle éventuellement nulle. Le paragraphe 1.4.5 est quant à lui consacré à la construction du terme source S_K, dont la principale fonction est de retrouver, au niveau discret, un bilan d'énergie totale. Il vient compenser une dissipation numérique créée par le schéma lorsque l'on essaye d'obtenir un bilan d'énergie cinétique discret, équivalent du bilan continu (1.2).

Conditions initiales – Les données discrètes au temps initial sont obtenues comme la moyenne des conditions initiales sur les mailles primales pour les variables scalaires et sur les mailles duales pour les inconnues vectorielles. Pour le schéma découplé, les différentes variables sont donc discrétisées au temps initial de la façon suivante :

$$\begin{aligned} &\forall K \in \mathcal{M}, \quad \rho_K^0 = \frac{1}{|K|}\int_K \rho_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \text{et } e_K^0 = \frac{1}{|K|}\int_K e_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \\ &\text{pour } 1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}, \quad u_{\sigma,i}^0 = \frac{1}{|D_\sigma|}\int_{D_\sigma} (\boldsymbol{u}_0(\boldsymbol{x}))_i\,\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{1.24}$$

En ce qui concerne le schéma à correction de pression, ρ^{-1} et $\boldsymbol{u}^0$ sont obtenues de manière analogue :

$$\begin{aligned} &\forall K \in \mathcal{M}, \quad \rho_K^{-1} = \frac{1}{|K|}\int_K \rho_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \\ &\text{pour } 1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}, \quad u_{\sigma,i}^0 = \frac{1}{|D_\sigma|}\int_{D_\sigma} (\boldsymbol{u}_0(\boldsymbol{x}))_i\,\mathrm{d}\boldsymbol{x} \end{aligned} \tag{1.25}$$

ρ^0 et p^0 étant obtenus respectivement en résolvant le bilan de masse et le bilan d'énergie interne. Cette façon de procéder permet d'effectuer la première étape de prédiction avec $(\rho_{D_\sigma}^{-1})_{\sigma\in\mathcal{E}}$, $(\rho_{D_\sigma}^0)_{\sigma\in\mathcal{E}}$ et le flux dual satisfaisant un bilan de masse sur maillage dual.

1.4.3 Interpolation MUSCL

Le but de cette section est de construire des interpolations d'ordre élevé de la masse volumique et de l'énergie interne aux faces qui vérifient les propriétés (1.11) et (1.9). Le principe général de la méthode est proche de celle développée dans [53], *i.e.* construire une approximation de l'operateur de convection qui soit d'ordre deux dans les zones de régularité de la solution et préserve les plages de variations des inconnues aux discontinuités grâce à une procédure de limitation adéquate des flux. L'algorithme que l'on présente est donc une extension de celui développé dans [53]. En particulier, contrairement aux interpolations MUSCL usuelles qui utilisent des estimations de pente et des limitations, e.g. [8, 62] pour les reviews et [45, 14, 15] pour les travaux récents, la limitation est ici obtenue directement à partir de conditions de stabilité purement algébriques (dans le sens où elle ne réclame aucun calcul géométrique), ce qui lui permet d'être utilisée avec n'importe quel type de maillage.

L'algorithme est cependant plus complexe que celui développé dans [53], car on veut construire un schéma qui préserve les zones de pression constante, afin d'éviter les instabilités au niveau des discontinuités de contact (en 1D seulement, le problème des glissements en 2D et 3D étant plus complexe, la vitesse n'étant pas nécessairement constante à travers la discontinuité). Une manière de réaliser cet impératif est d'imposer que la pression à la face soit une combinaison convexe des pressions des deux cellules voisines. Cette condition introduit une corrélation dans les interpolations de la masse volumique et l'énergie interne et on perd donc le caractère purement découplé du schéma.

L'algorithme d'interpolation MUSCL se découpe en deux phases distinctes : tout d'abord on construit une première interpolation d'ordre deux (ici seulement pour la masse volumique), puis on applique une procédure de limitation.

Interpolation de la masse volumique – Pour une face $\sigma \in \mathcal{E}_{\text{int}}$ et une cellule $K \in \mathcal{M}$, on note $\boldsymbol{x}_\sigma$ et $\boldsymbol{x}_K$ les centres de masse de σ et K respectivement. Soit $\sigma \in \mathcal{E}_{\text{int}}$ une face interne du maillage. On suppose que l'on a construit un ensemble de réels (ζ_σ^L) tel que :

$$\boldsymbol{x}_\sigma = \sum_{L\in\mathcal{M}} \zeta_\sigma^L \boldsymbol{x}_L, \qquad \sum_{L\in\mathcal{M}} \zeta_\sigma^L = 1. \tag{1.26}$$

Alors, connaissant $\rho_\mathcal{M} = (\rho_K)_{K\in\mathcal{M}}$, on construit une interpolation de la masse volumique à la face $\tilde{\rho}_\sigma$ par la formule suivante :

$$\forall \sigma \in \mathcal{E}_{\text{int}}, \quad \tilde{\rho}_\sigma = \sum_{L\in\mathcal{M}} \zeta_\sigma^L \rho_L. \tag{1.27}$$

En pratique les cellules qui interviennent dans l'interpolation (1.26), sont choisies les plus proches possible de la face σ et, quand cela est possible, la combinaison est convexe (les coefficients (ζ_σ^L) sont positifs). Cette interpolation est simplifiée pour les maillages structurés car elle est obtenue, pour $\sigma = K|L$, par une moyenne pondérée de la valeur en ρ_K et ρ_L (les centres de masse sont alignés). Dans le cas général, l'algorithme permettant de calculer (ζ_σ^L) est le suivant.

— On considère d'abord toutes les familles (ζ_σ^M) qui vérifient la condition (1.26). Pour une face interne $\sigma = K|L$, on considère ensuite toutes celles dont tous les coefficients sont nuls sauf éventuellement en K, L et sur une cellule (2D) ou deux cellules (3D) voisines de K ou L. Pour les faces externes on choisit les familles où tous les coefficients sont nuls sauf en K et sur 2 (2D) ou 3 (3D) mailles voisines.

— On trie parmi les familles restantes celles dont la combinaison est convexe. Parmi celles-ci, on prend, si elle existe, celle dont seulement deux coefficients sont non nuls ($\boldsymbol{x}_\sigma$ est aligné avec les centroides de deux cellules). Sinon on choisit celle qui minimise la quantité $\zeta = \max_{\zeta_\sigma^K \neq 0} |\zeta_\sigma^K - 0.5|$. Ceci signifie intuitivement que l'on choisit la combinaison où $\boldsymbol{x}_\sigma$ est le plus proche possible du centre de gravité de l'ensemble des cellules considérées. Enfin

dans le cas où il n'existerait aucune combinaison convexe dans les familles admissibles, on choisit la combinaison minimisant cette même quantité ζ.

Procédure de limitation – Soit $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = K|L$, et supposons que l'écoulement va de K vers L, *i.e.* $F_{K,\sigma} \geq 0$. On rappelle maintenant les conditions (1.11) et (1.9) qui sont nécessaires pour obtenir la positivité de la masse volumique et de l'énergie interne. Pour la masse volumique, il existe $\alpha_\sigma^\rho \in [0,1]$, $\beta_\sigma^\rho \in [0,1]$ et $M_\sigma^\rho \in \mathcal{M}$ tels que

$$\left| \begin{array}{l} \rho_\sigma - \rho_K = \alpha_\sigma^\rho \, (\rho_K - \rho_{M_\sigma^\rho}), \\ \rho_\sigma - \rho_L = \beta_\sigma^\rho \, (\rho_K - \rho_L). \end{array} \right. \tag{1.28}$$

De façon similaire, pour l'énergie interne, il existe $\alpha_\sigma^e \in [0,1]$, $\beta_\sigma^e \in [0,1]$ et $M_\sigma^e \in \mathcal{M}$ tels que

$$\left| \begin{array}{l} e_\sigma - e_K = \alpha_\sigma^e \, (e_K - e_{M_\sigma^\rho}), \\ e_\sigma - e_L = \beta_\sigma^e \, (e_K - e_L). \end{array} \right. \tag{1.29}$$

Pour simplifier, on suppose que les cellules "upstream" M_σ^ρ et M_σ^e sont les mêmes et on la note M_σ. Dans [53], les équations (1.28) et (1.29) permettent d'obtenir un intervalle d'admissibilité pour ρ_σ et e_σ, et on obtient une procédure de limitation en projetant simplement les interpolations géométriques d'ordre deux sur les intervalles d'admissibilité. Ici la situation est plus complexe car nous voulons en plus préserver les contacts 1D. Comme l'équation d'état fait uniquement dépendre la pression du produit ρe, il faut que $\rho_\sigma e_\sigma$ soit égal à $\rho_K e_K$ et $\rho_L e_L$ dès que ces quantités sont égales. Ici nous allons utiliser une condition encore plus restrictive en supposant que $\rho_\sigma e_\sigma$ est une combinaison convexe de $\rho_K e_K$ et $\rho_L e_L$, *i.e.* il existe $\kappa_\sigma \in [0,1]$ tel que :

$$\rho_\sigma \, e_\sigma = \kappa_\sigma \, \rho_K \, e_K + (1 - \kappa_\sigma) \, \rho_L \, e_L. \tag{1.30}$$

Le but est donc de construire une procédure de limitation qui permette que les conditions (1.28), (1.29) et (1.30) soient vérifiées.

De (1.28), on tire, en combinant les deux équations, la relation suivante :

$$\beta_\sigma^\rho = 1 - \frac{\alpha_\sigma^\rho}{r_\sigma^\rho}, \qquad \text{avec } r_\sigma^\rho = \frac{\rho_L - \rho_K}{\rho_K - \rho_{M_\sigma}}. \tag{1.31}$$

Cette relation montre que (1.28) est vérifiée ($\alpha_\sigma^\rho \in [0,1]$ et $\beta_\sigma^\rho \in [0,1]$) si α_σ^ρ satisfait les inégalités :

$$0 \leq \alpha_\sigma^\rho \leq \min\left(1, \, r_\sigma^\rho\right)^+,$$

avec la notation $a^+ = \max(a, 0)$, $a \in \mathbb{R}$. Ceci suggère la procédure suivante : grâce au lien entre ρ_σ et e_σ qui est induit par la condition (1.30), on exprime les coefficients α_σ^e et β_σ^e comme une fonction de α_σ^ρ, et on transforme les limitations induites par (1.29) en limitations pour le coefficient α_σ^ρ. Dans cette optique, on remarque tout d'abord que (1.28) conduit à $\rho_\sigma = \beta_\sigma^\rho \, \rho_K + (1 - \beta_\sigma^\rho) \, \rho_L$, et on va supposer arbitrairement que $\rho_\sigma \, e_\sigma$ est donnée par la même combinaison :

$$\rho_\sigma \, e_\sigma = \beta_\sigma^\rho \, \rho_K \, e_K + (1 - \beta_\sigma^\rho) \, \rho_L \, e_L,$$

i.e. on prend $\kappa = \beta_\sigma^\rho$ dans (1.30). Ce choix n'est clairement pas univoque. On aurait parfaitement pu prendre $\kappa = \beta_\sigma^e$ par exemple. En divisant par ρ_σ on obtient

$$e_\sigma = \frac{\beta_\sigma^\rho \, \rho_K}{\rho_\sigma} \, e_K + \frac{(1 - \beta_\sigma^\rho) \, \rho_L}{\rho_\sigma} \, e_L.$$

Comme le membre de droite peut être vu comme une combinaison convexe de e_K et e_L, on a :

$$\beta_\sigma^e = \frac{\rho_K}{\rho_\sigma} \, \beta_\sigma^\rho, \tag{1.32}$$

et on en déduit que $\beta^e_\sigma \in [0,1]$ (qui provient aussi du fait que $\rho_\sigma = \beta^\rho_\sigma\,\rho_K + (1-\beta^\rho_\sigma)\,\rho_L \geq \beta^\rho_\sigma\,\rho_K$). En utilisant (1.29), on obtient les relations suivantes, équivalentes pour l'énergie interne de (1.31) :

$$\beta^e_\sigma = 1 - \frac{\alpha^e_\sigma}{r^e_\sigma}, \qquad \text{avec } r^e_\sigma = \frac{e_L - e_K}{e_K - e_{M_\sigma}}. \tag{1.33}$$

Dès lors $\alpha^e_\sigma = (1-\beta^e_\sigma)\,r^e_\sigma$, et en remplaçant β^e_σ par son expression dans (1.32), puis en exprimant β^ρ_σ comme une fonction de α^ρ_σ grâce à (1.31) on obtient après quelques manipulations algébriques :

$$\alpha^e_\sigma = \frac{\rho_L}{\rho_\sigma}\,\frac{r^e_\sigma}{r^\rho_\sigma}\,\alpha^\rho_\sigma. \tag{1.34}$$

De cette expression, on vérifie que (1.29) (où de façon équivalente, $\alpha^e_\sigma \in [0,1]$, comme le fait que $\beta^e_\sigma \in [0,1]$ est déjà vérifié) sera satisfait (ainsi que (1.28)) si α^ρ_σ satisfait :

$$0 \leq \alpha^\rho_\sigma \leq \min\Big(1,\ r^\rho_\sigma,\ \frac{\rho_\sigma}{\rho_L}\,\frac{r^\rho_\sigma}{r^e_\sigma}\Big)^+.$$

Cette relation n'est pas suffisante pour construire α^ρ_σ, parcequ'elle fait intervenir ρ_σ dont l'expression elle même fait intervenir α^ρ_σ. Il suffit de remplacer ρ_σ par une borne inférieure explicite. Comme on l'a déjà vu, $\alpha^\rho_\sigma = 0$ est toujours une valeur admissible (interpolation UPWIND), et donc ρ_K est aussi une valeur possible pour ρ_σ. Par conséquent ρ_σ peut être obtenu par projection de $\tilde\rho_\sigma$ sur un intervalle contenant ρ_K, ce qui conduit nécessairement à $\rho_\sigma \geq \min(\rho_K,\ \tilde\rho_\sigma)$. Finalement on choisit l'intervalle de limitation de α^ρ_σ, noté $\mathcal{I}_\alpha$, et donné par :

$$\mathcal{I}_\alpha = \Big[0,\ \min\Big(1,\ r^\rho_\sigma,\ \frac{\min(\rho_K,\ \tilde\rho_\sigma)}{\rho_L}\,\frac{r^\rho_\sigma}{r^e_\sigma}\Big)^+\Big]. \tag{1.35}$$

L'intervalle d'admissibilité de la masse volumique est donc $\mathcal{I}_\rho$ avec

$$\mathcal{I}_\rho = \Big\{\rho_K + \alpha\,(\rho_K - \rho_{M^\rho_\sigma}),\ \alpha \in \mathcal{I}_\alpha\Big\}. \tag{1.36}$$

L'algorithme de limitation, connaissant $\tilde\rho_\sigma$, consiste à calculer ρ_σ en projetant $\tilde\rho_\sigma$ sur l'intervalle $\mathcal{I}_\rho$, ce qui donne une valeur de α^ρ_σ. Le coefficient α^e_σ est alors donné par (1.34) et e_σ s'en déduit grâce à (1.29).

Il est important de noter que la précision de l'algorithme dépend de la variable considérée :

- L'approximation de ρ, sans limitation, est d'ordre deux en espace.
- On obtient alors une pression à la face en utilisant la même pondération que pour la masse volumique. Pour une discrétisation structurée, cela conduit à une interpolation identique à la première et donc d'ordre deux pour la pression. Au contraire, pour une discrétisation non structurée, où l'interpolation (1.27) (plus précisément l'interpolation (1.27) écrite pour p) et la seconde égalité de (1.28) (en remplaçant ρ par p) sont différentes, le second ordre est perdu.
- L'énergie interne, dont l'interpolation est déduite de la masse volumique et de la pression sera d'ordre 1 comme on peut seulement garantir que sa valeur sera quelquepart entre la valeur des deux cellules voisines. L'énergie interne ne fait qu'ajouter des limitations de flux supplémentaires. En particulier, la définition 1.35 de $\mathcal{I}_\alpha$ suppose que α^ρ_σ s'annule des que l'une des quantités r^ρ_σ ou r^e_σ est négative, *i.e.* dès que ρ ou e admettent des extrema locaux.

Plusieurs variantes de l'algorithme présenté précédemment existent et on en présente quelques unes :

- Comme dit auparavant, on peut intervertir les rôles de ρ et e *i.e.* construire les interpolations d'ordre deux de e et p, et en déduire ρ ; dans ce cas, on choisit $\kappa = \beta^e_\sigma$, et on construit une interpolation géométrique de e.

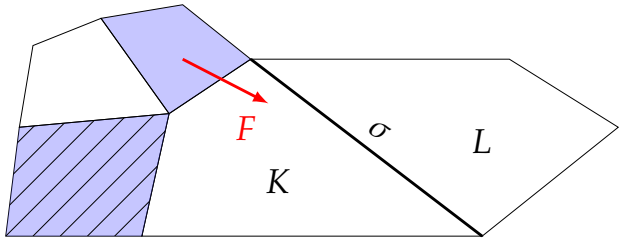


Figure 1.1 – Notations pour la procédure de limitation.
Cellules de l'ensemble V_K pour $\sigma = K|L$, avec un champ d'advection constant $\mathbf{F}$: en bleu cellules upstream – en haché l'unique cellule opposée.

- L'algorithme ci-présent ne permet pas de s'assurer que l'interpolation de e reste bien entre l'upwind et l'interpolation d'ordre deux. Dans le cas 1D ou structuré, on peut s'en assurer en limitant l'intervalle d'admissibilité de β_σ^e à $\beta_\sigma^e \in [\tilde{\beta}_\sigma, 1]$, avec $\tilde{\beta}_\sigma$ le poids permettant de construire l'interpolation d'ordre deux de e_σ à partir de e_K et e_L. Pour des maillages uniformes cet intervalle se réduit à $\beta_\sigma^e \in [1/2, 1]$. Si l'on veut que cela soit vérifié, on ajoute une limitation supplémentaire dans notre procédure.

- Enfin, la cellule commune "upstream" M_σ que nous avons choisie pour (1.28) et (1.29) peut être selectionnée de façon arbitraire, mais de façon raisonnable, on choisit une cellule dans un voisinnage proche de σ. Il existe deux options pour construire V_K dans lequel on choisit M_σ :

  (a) V_K peut être considéré comme l'ensemble des cellules "upstream" à K, *i.e.* $V_K = \{L \in \mathcal{M},\ L$ et K partagent une face σ et $F_{K,\sigma} < 0\}$,

  (b) quand cela a un sens (*i.e.* avec un maillage obtenu par Q_1 mappings de l'élément de référence $(0,1)^d$), V_K peut être constitué de l'unique cellule opposée à σ partageant une face avec K.

Dans les tests numériques effectués on se placera toujours dans le second cas. (voir la Figure 1.1).

1.4.4 Viscosité artificielle

Les résultats numériques effectués dans la dernière section montrent que la dissipation introduite par l'interpolation upwind sur la vitesse est parfois insuffisante, probablement parce que la dissipation numérique est uniquement liée à la vitesse matérielle et non la vitesse des ondes. On voit alors apparaître des oscillations parasites et des overshoots au niveau des chocs. L'interpolation MUSCL des inconnues scalaires a tendance à renforcer ces phénomènes purement numériques, étant donné qu'elle réduit la dissipation numérique pour monter en précision. Il est donc intéréssant d'introduire un surplus de viscosité artificielle dans le bilan de quantité de mouvement afin de pallier ces problèmes. Néanmoins l'ajout de cette viscosité peut dégrader la convergence là où la solution est régulière. L'idée est donc de rajouter de la viscosité uniquement dans les zones ou c'est nécessaire *i.e.* aux chocs. On s'inspire alors des travaux développés dans [31] et [42], où la diffusion est calculée par une analyse *a posteriori* de la solution.

L'objectif de cette section est le calcul de cette viscosité artificielle, *i.e.* le terme ν_ϵ^{n+1} dans (1.23). Le processus consiste à calculer un paramètre de diffusion par maille ζ_K^{n+1}, pour chaque maille primale et d'en déduire une viscosité pour chaque face du maillage dual. Dans cette dernière étape, il existe deux cas de figure :

- La face duale ϵ est strictement incluse dans la maille K ; dans ce cas, on pose $\nu_\epsilon^{n+1} = |\epsilon|\, \zeta_K^{n+1}$.

- La face duale ϵ est à cheval sur 4 cellules primales (cas MAC uniquement) ; alors on pose :

$$\nu_\epsilon^{n+1} = |\epsilon| \, \frac{1}{4} \sum_{K \in \mathcal{N}(\epsilon)} \zeta_K^{n+1},$$

  où $\mathcal{N}(\epsilon)$ est l'ensemble des mailles adjacentes à ϵ.

Le reste de cette section consiste en le calcul des coefficients de diffusions $(\zeta_K^{n+1})_{K \in \mathcal{M}}$. Ce calcul montre que le comportement de ces coefficients est en h (ou en δt, comme la CFL est inférieure à 1 et bornée loin de 0), dans le sens où la quantité ζ_K^{n+1}/h (formellement) ne tend ni vers zéro ni vers l'infini quand h tend vers zéro. Par conséquent le terme de diffusion dans (1.5c) produit bien une viscosité en h^2 dans les zones de régularité de la solution, comme dans [31] et [42]. Néanmoins il existe deux différences notables. Tout d'abord on n'ajoute de la viscosité qu'au bilan de quantité de mouvement et non dans toutes les équations. Ensuite, on garde l'interpolation upwind dans ce même bilan, on n'effectue pas une interpolation centrée.

Viscosité Entropique Cette méthode fondée sur les travaux de J.L. Guermond dans [31, 30] est construite sur l'inégalité d'entropie satisfaite par les solutions physiques du système d'Euler. Elle s'écrit :

$$\partial_t \eta + \operatorname{div}(\eta u) \leq 0, \tag{1.37}$$

l' égalité étant vérifiée lorsque la solution est régulière et aux discontinuités de contact. Dans le cas des équations d'Euler l'entropie physique est définie par :

$$\eta(p, \rho) = \frac{\rho}{\gamma - 1} \log\left(\frac{p}{\rho^\gamma}\right). \tag{1.38}$$

La première étape de cette technique consiste dans la discrétisation volumes finis du résidu d'entropie sur le maillage primal :

$$\mathcal{R}_K^{n+1} = \frac{1}{\delta t}\left(\eta_K^{n+1} - \eta_K^n\right) + \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} |\sigma| \, \eta_\sigma^n \, u_{K,\sigma}^n,$$

où η_σ^n est l'approximation centrée à la face σ. On construit alors un premier paramètre de diffusion

$$\tilde{\zeta}_K^{n+1} = c_E \, \rho_K^{n+1} \, h_K \, |\mathcal{R}_K^{n+1}|. \tag{1.39}$$

où h_K est le diamètre de la maille K et c_E un paramètre de calibration. Il faut noter que $\mathcal{R}_K^{n+1}$ est une discrétisation formelle de $\partial_t \eta + \operatorname{div}(\eta u)$, et donc cette quantité est indépendante du pas de temps et d'espace ; par conséquent, $\tilde{\zeta}_K^{n+1}$ se comporte comme h_K. On limite alors ce paramètre pour qu'il soit dans l'ordre de grandeur de la diffusion numérique upwind de l'operateur de convection discret. Pour une face σ du maillage primal adjacente à une maille L, cette diffusion upwind est égale à $\zeta_\sigma^{n+1} = |\rho_\sigma^n \, u_{L,\sigma}^n|/2$, où ρ_σ^n est la masse volumique à la face utilisée dans le bilan de masse. On définit alors une valeur max de la diffusion par :

$$\zeta_{max,K}^{n+1} = c_{max} \, \max\left((\zeta_\sigma^{n+1})_{\sigma \in \overline{\mathcal{E}}(K)}\right),$$

où c_{max} est un paramètre de calibration et $\overline{\mathcal{E}}(K)$ représente un ensemble de faces dans le voisinnage de K, comprenant au moins $\mathcal{E}(K)$. Pour les calculs effectués dans cette thèse, on choisit un patch de 7 cellules centré sur K en 1D, et en 2D cartésien, on choisit un patch de 7×7 cellules centré sur K. On obtient alors un coefficient limité par :

$$\tilde{\zeta}_K^{n+1} = \min\left(\tilde{\zeta}_K^{n+1}, \zeta_{max,K}^{n+1}\right).$$

Finalement, ζ_K^{n+1} est calculé par une moyenne pondérée des coefficients $(\tilde{\zeta}_L^{n+1})_{L \in \mathcal{M}}$ sur un patch autour de K. En 1D, ce patch est constitué de 3 cellules, K et ses deux cellules adjacentes, avec un poids de 2/3 pour K et 1/3 pour les autres cellules. En 2D avec une grille cartésienne, on utilise un patch 3×3 centré en K, avec un poids de 8/9 pour K et 1/9 pour les autres cellules.

Viscosité WLR La seconde méthode employée pour créer une viscosité artificielle s'appuie sur les travaux [42]. On en rappelle les grandes lignes, pour une loi de conservation générique d'inconnue w et de flux $\boldsymbol{f}$:

$$\partial_t w + \operatorname{div} \boldsymbol{f}(w) = 0. \tag{1.40}$$

Une solution faible de (1.40) est définie par :

$$\mathcal{W}(w,\phi) = \int_0^T \int_\Omega \Big[w(\boldsymbol{x},t)\,\partial_t\phi(\boldsymbol{x},t) + \boldsymbol{f}(\boldsymbol{x},t)\cdot\nabla\phi(\boldsymbol{x},t)\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega w(\boldsymbol{x},0)\,\phi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0,$$

pour toute fonction test $\phi \in C_0^1(\Omega\times[0,T))$. Cette identité est utilisée dans [42] pour construire un instrument de mesure de la régularité locale de la solution, à partir d'une solution discrète w_h obtenue par une méthode de type différences finies. La solution discrète est identifiée comme fonction de l'espace et du temps, des fonctions test spécifiques ϕ (une par cellule, que l'on note $(\phi_K)_{K\in\mathcal{M}}$ pour garder une cohérence avec le reste du manuscrit) sont définies, et les quantités $(\mathcal{W}(w_h,\phi_K))_{K\in\mathcal{M}}$ sont utilisées pour repérer les discontinuités de la solution. On construit alors les paramètres de diffusion à partir de ces valeurs.

On adapte maintenant la stratégie aux équations d'Euler et à un schéma de type volume fini. Premièrement on va calculer le résidu $\mathcal{W}$ du bilan de masse uniquement :

$$\mathcal{W}(\rho,\boldsymbol{u},\phi) = \int_0^T \int_\Omega \Big[\rho(\boldsymbol{x},t)\,\partial_t\phi_t(\boldsymbol{x},t) + \rho(\boldsymbol{x},t)\,\boldsymbol{u}(\boldsymbol{x},t)\cdot\nabla\phi(\boldsymbol{x},t)\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega \rho(\boldsymbol{x},0)\,\phi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}.$$

De façon similaire au schéma différences finies traité dans [42], on identifie les solutions discrètes du schéma à des fonctions constantes par morceaux sur le maillage. On définit donc :

$$\rho_\Delta(\boldsymbol{x},t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} \rho_K^n\,\mathcal{X}_K(\boldsymbol{x})\mathcal{X}_{(t_n,t_{n+1})}(t),$$
$$\boldsymbol{u}_\Delta(\boldsymbol{x},t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} \boldsymbol{u}_K^n\,\mathcal{X}_K(\boldsymbol{x})\mathcal{X}_{(t_n,t_{n+1})}(t),$$

où $\mathcal{X}_K$ et $\mathcal{X}_{(t_n,t_{n+1})}$ désignent les fonctions caractéristiques de K et de l'intervalle (t_n,t_{n+1}), et $\boldsymbol{u}_K$ une interpolée de la vitesse sur le maillage primal :

$$\forall K\in\mathcal{M},\quad \boldsymbol{u}_K = \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} |D_{K,\sigma}|\,\boldsymbol{u}_\sigma.$$

L'étape suivante consiste à se doter d'un ensemble de fonctions polynômiales (ϕ_K^n), pour tout $K\in\mathcal{M}$ et $0\le n\le N-1$, qui feront office de fonctions tests. On précisera plus tard une façon de choisir ces polynômes, on précise simplement leur propriété d'approximation de l'espace des fonctions tests. Pour tout $\phi\in C_0^1(\Omega\times[0,T])$, on suppose qu'il existe $(\beta_K^n)_{K\in\mathcal{M},\,0\le n\le N-1}\subset\mathbb{R}$ tels que :

$$\phi(\boldsymbol{x},t) = \sum_{K\in\mathcal{M}}\sum_{n=0}^{N-1} \beta_K^n\,\phi_K^n(\boldsymbol{x},t) + \mathcal{O}(\Delta^2), \tag{1.41}$$

où $\Delta = \max(h,\delta t)$. Si on suppose que ces fonctions test (ϕ_K^n) sont locales, *i.e.* se comportent en $\delta t\,h^d$ (comme la mesure de leur support se comporte aussi en $\delta t\,h^d$), alors on s'assure que

$$\mathcal{W}(\rho_\Delta,\boldsymbol{u}_\Delta,\phi) = \sum_{K\in\mathcal{M}}\sum_{n=0}^{N-1}\Big(\beta_K^n\mathcal{W}_K^n + \mathcal{O}(\Delta^{d+3})\Big),$$

où les résidus faibles locaux (WLR), $\mathcal{W}_K^n$, prennent la forme suivante :

$$\mathcal{W}_K^n = \int_0^T \int_\Omega \rho_\Delta(\boldsymbol{x},t)\,\partial_t \phi_K^n(\boldsymbol{x},t) + \rho_\Delta(\boldsymbol{x},t)\,\boldsymbol{u}_\Delta(\boldsymbol{x},t)\cdot\nabla\phi_K^n(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \tag{1.42}$$

Sous l'hypothèse (1.41), on sait d'apres [42] que les WLR ont la propriété suivante :

$$|\mathcal{W}_K^n| \text{ se comporte comme } \left|\begin{array}{ll} \Delta, & \text{près des chocs,} \\ \Delta^\alpha, & 1<\alpha\le 2, \text{ près des discontinuités de contact,} \\ \Delta^p & \text{dans les zones de régularité,} \end{array}\right.$$

où $p = \min(r+2,4)$, r ordre de convergence du schéma. Ici, 4 est lié au choix des polynômes (ϕ_K^n) et de la précision avec laquelle ils approximent les fonctions tests. Ces résultats sont démontrés en 1D et seulement observés numériquement sur la quantité $|\mathcal{W}_K^n|/\Delta^{d-1}$ pour les autres dimensions. Ces résidus sont alors utilisés pour construire des coefficients de diffusion :

$$\tilde{\zeta}_K^{n+1} = c_m\,\frac{1}{\delta t\,\Delta^{d-1}}\,|\mathcal{W}_K^{n+1}|, \tag{1.43}$$

où c_m est un paramètre de calibration. Finalement, comme pour la viscosité entropique, les coefficients ζ_K^{n+1} sont des moyennes pondérées des coefficients $(\tilde{\zeta}_L^{n+1})_{L\in\mathcal{M}}$ avec les mêmes patchs et les mêmes poids en 1D et en 2D que pour la viscosité entropique.

Remarque 1.1 (*Cas des maillages cartésiens*)

Dans le cas des schémas MAC, l'utilisation de grilles cartésiennes permet d'obtenir une formule générique des WLR. Il est en effet possible de définir un ensemble univoque de polynômes approximants que l'on appelle les B-splines. On donne en Annexe (A.2) un exemple de calcul explicite des WLR en 1D et 2D. Ces B-splines seront utilisés pour les résultats numériques de la section (1.6)

1.4.5 Bilan d'énergie cinétique discret et terme source

Grâce au choix des différents opérateurs spatiaux discrets, on est capable d'obtenir, quel que soit le type de schéma considéré, un équivalent discret du bilan d'énergie cinétique continu (1.2). On regroupe les résultats dans deux lemmes suivant le type de discrétisation temporelle utilisée. Pour le schéma découplé on a le lemme suivant :

Lemme 1.1 (*Bilan d'énergie cinétique discret explicite*)

Une solution du système (1.5) satisfait, pour $1\le i\le d$, $\sigma\in\mathcal{E}_S^{(i)}$ et $0\le n\le N-1$:

$$\begin{aligned}\frac{1}{2}\frac{|D_\sigma|}{\delta t}\Big[\rho_{D_\sigma}^{n+1}(u_{\sigma,i}^{n+1})^2 - \rho_{D_\sigma}^n(u_{\sigma,i}^n)^2\Big] + \frac{1}{2}\sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n\,(u_{\sigma,i}^n u_{\sigma',i}^n)\\ + |D_\sigma|\,(\nabla p)_{\sigma,i}^{n+1}\,u_{\sigma,i}^{n+1} + \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \frac{1}{2}\mu_\epsilon^n(u_{\sigma,i}^n - u_{\sigma',i}^n)(u_{\sigma',i}^n + u_{\sigma,i}^n) = -R_{\sigma,i}^{n+1},\end{aligned} \tag{1.44}$$

avec :

$$\begin{aligned}R_{\sigma,i}^{n+1} = \frac{1}{2}\frac{|D_\sigma|}{\delta t}\rho_{D_\sigma}^{n+1}(u_{\sigma,i}^{n+1} - u_{\sigma,i}^n)^2 + \frac{1}{2}\sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \mu_\epsilon^n(u_{\sigma',i}^n - u_{\sigma,i}^n)^2\\ + \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} (\mu_\epsilon^n - \frac{F_{\sigma,\epsilon}^n}{2})(u_{\sigma,i}^{n+1} - u_{\sigma,i}^n)\,(u_{\sigma,i}^n - u_{\sigma',i}^n).\end{aligned} \tag{1.45}$$

En ce qui concerne le schéma de correction de pression, le bilan d'énergie cinétique prend la forme suivante :

Lemme 1.2 (*Bilan d'énergie cinétique discret, correction de pression*)

Une solution du système (1.6) satisfait l'égalité suivante, pour $1 \leq i \leq d$, $\sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}$ *et* $0 \leq n \leq N-1$ *:*

$$\frac{1}{2}\frac{|D_\sigma|}{\delta t}\Big[\rho_{D_\sigma}^n (u_{\sigma,i}^{n+1})^2 - \rho_{D_\sigma}^{n-1}(u_{\sigma,i}^n)^2\Big] + \frac{1}{2}\sum_{\epsilon=D_\sigma|D_{\sigma'}} F_{\sigma,\epsilon}^n\, \tilde{u}_{\sigma,i}^{n+1}\, \tilde{u}_{\sigma',i}^{n+1} + |D_\sigma|\,(\nabla p)_{\sigma,i}^{n+1}\, u_{\sigma,i}^{n+1} = -R_{\sigma,i}^{n+1} - P_{\sigma,i}^{n+1}, \quad (1.46)$$

où

$$\begin{aligned} R_{\sigma,i}^{n+1} &= \frac{|D_\sigma|}{2\,\delta t}\,\rho_{D_\sigma}^{n-1}\left(\tilde{u}_{\sigma,i}^{n+1} - u_{\sigma,i}^n\right)^2 + \Big[\sum_{\epsilon=D_\sigma|D_{\sigma'}} \mu_\epsilon^{n+1}\,(\tilde{u}_{\sigma,i}^{n+1} - \tilde{u}_{\sigma',i}^{n+1})\Big]\,\tilde{u}_{\sigma,i}^{n+1}, \\ P_{\sigma,i}^{n+1} &= \frac{|D_\sigma|\,\delta t}{2\,\rho_{D_\sigma}^n}\Big[\big((\nabla p)_{\sigma,i}^{n+1}\big)^2 - \big((\widetilde{\nabla p})_{\sigma,i}^{n+1}\big)^2\Big]. \end{aligned} \quad (1.47)$$

Le résidu P_σ peut être vu comme un terme d'erreur en temps dû à l'algorithme de correction de pression. Lors de l'étude de consistance des schémas avec la formulation faible du modèle continu, on verra que ce terme disparaît quand le pas d'espace tend vers 0. Quant au terme R_σ, celui-ci correspond à une dissipation créée par la diffusion numérique du schéma. De par l'existence de solutions discontinues aux équations d'Euler, ce terme ne va pas converger nécessairement vers 0 avec le pas de temps et d'espace, mais au contraire converger vers des mesures au niveau des chocs. La consistance avec le bilan d'énergie totale est donc compromise et on aboutit à de mauvaises conditions de Rankine-Hugoniot. Pour résoudre ce problème, il est nécessaire de compenser cette dissipation au niveau de l'énergie interne en ajoutant un terme correctif S_K. De par l'utilisation de grilles décalées (les deux équations d'énergie n'étant pas discrétisées sur le même maillage), il est impossible de compenser terme à terme le résidu. L'idée est alors de distribuer le terme R_σ pour $\sigma = K|L \in \mathcal{E}_{\text{int}}$ sur les deux mailles primales voisines K et L.

On commence par le schéma de correction de pression dont le terme source est immédiat :

$$\forall K \in \mathcal{M},\ S_K^{n+1} = \sum_{i=1}^d S_{K,i}^{n+1},$$

avec :

$$S_{K,i}^{n+1} = \frac{1}{2}\,\rho_K^{n-1}\sum_{\sigma\in\mathcal{E}(K)\cap\mathcal{E}_{\mathcal{S}}^{(i)}} \frac{|D_{K,\sigma}|}{\delta t}\left(\tilde{u}_{\sigma,i}^{n+1} - u_{\sigma,i}^n\right)^2 + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset,\\ \epsilon=D_\sigma|D_{\sigma'}}} \alpha_{K,\epsilon}\,\mu_\epsilon^{n+1}(\tilde{u}_{\sigma,i}^{n+1} - \tilde{u}_{\sigma',i}^{n+1})^2. \quad (1.48)$$

Lorsque ϵ est strictement inclus dans la maille K, alors $\alpha_{K,\epsilon} = 1$. C'est l'unique possibilité pour les schémas RT et CR. Dans le cas du schéma MAC, certaines faces duales peuvent être situées à cheval sur deux faces primales. Elles sont donc communes à 4 mailles primales. On répartit alors cette contribution de la manière suivante sur chaque maille :

$$\alpha_{K,\epsilon} = \frac{|K|}{\sum_{M\in\mathcal{N}_\epsilon}|M|}, \quad (1.49)$$

soit pour une grille uniforme, $\alpha_{K,\epsilon} = \frac{1}{4}$.

Concernant le schéma découplé, le travail est plus fastidieux, le terme résiduel étant plus complexe. On décide de décomposer S_K^{n+1} de la façon suivante :

$$\forall K \in \mathcal{M},\ S_K^{n+1} = \sum_{i=1}^d S_{K,i}^{n+1},$$

avec :

$$S_{K,i}^{n+1} = \frac{1}{2}\rho_K^{n+1} \sum_{\sigma\in\mathcal{E}(K)\cap\mathcal{E}_S^{(i)}} \frac{|D_{K,\sigma}|}{\delta t}(u_{\sigma,i}^{n+1} - u_{\sigma,i}^n)^2 + \sum_{\epsilon\in\tilde{\mathcal{E}}_S^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset} S_{K,\epsilon,i}^{n+1},$$

où $S_{K,\epsilon,i}^{n+1}$ représente la contribution sur la face duale ϵ du terme source. Elle est définie comme suit.

Première étape Pour une face interne ϵ, on regroupe tout d'abord les contributions visqueuses des résidus d'énergie cinétique des deux cellules duales qui ont ϵ pour face commune. Pour ce faire, on note σ_ϵ^U et σ_ϵ^D les deux faces primales telles que $\epsilon = D_{\sigma_\epsilon^U}|D_{\sigma_\epsilon^D}$ et $F^n_{\sigma_\epsilon^D,\epsilon} \leq 0$ (*i.e.* la cellule $D_{\sigma_\epsilon^D}$ est à l'aval de ϵ par rapport à l'écoulement), voir la Figure (1.2) pour illustration. Le résidu visqueux de la cellule upwind s'écrit :

$$(R_{\epsilon,i}^U)^{n+1} = \frac{\mu_\epsilon^n}{2}(u^n_{\sigma_\epsilon^D,i} - u^n_{\sigma_\epsilon^U,i})^2 + (\mu_\epsilon^n - \frac{|F^n_{\sigma_\epsilon^D,\epsilon}|}{2})(u^{n+1}_{\sigma_\epsilon^U,i} - u^n_{\sigma_\epsilon^U,i})(u^n_{\sigma_\epsilon^U,i} - u^n_{\sigma_\epsilon^D,i}).$$

Celui de la cellule downwind vaut quant à lui :

$$(R_{\epsilon,i}^D)^{n+1} = \frac{\mu_\epsilon^n}{2}(u^n_{\sigma_\epsilon^D,i} - u^n_{\sigma_\epsilon^U,i})^2 + (\mu_\epsilon^n + \frac{|F^n_{\sigma_\epsilon^D,\epsilon}|}{2})(u^{n+1}_{\sigma_\epsilon^D,i} - u^n_{\sigma_\epsilon^D,i})(u^n_{\sigma_\epsilon^D,i} - u^n_{\sigma_\epsilon^U,i}).$$

On obtient ainsi une contribution totale qui est égale à $R_{\epsilon,i}^{n+1} = (R_{\epsilon,i}^U)^{n+1} + (R_{\epsilon,i}^D)^{n+1}$.

Seconde étape On compense le terme $R_{\epsilon,i}^{n+1}$ avec le terme $S_{K,\epsilon,i}^{n+1}$. Il existe plusieurs cas de figure. Si la face duale ϵ est strictement incluse dans la cellule K, on pose $R_{\epsilon,i}^{n+1} = S_{K,\epsilon,i}^{n+1}$. Cette situation est l'unique possible pour les schémas de type RT ou CR. Pour les schémas MAC, certaines faces duales sont à cheval sur deux faces primales. Si K est en amont de ϵ par rapport à l'écoulement ($\sigma_\epsilon^U \in \mathcal{E}(K)$), soit L la deuxième cellule amont ($\sigma_\epsilon^U = K|L$), alors :

$$S_{K,\epsilon,i}^{n+1} = \frac{|K|}{|K|+|L|}\left[(R_{\epsilon,i}^U)^{n+1} - \frac{|F^n_{\sigma_\epsilon^U,\epsilon}|}{4}(u^n_{\sigma_\epsilon^D,i} - u^n_{\sigma_\epsilon^U,i})^2\right].$$

De façon analogue, si K est en aval :

$$S_{K,\epsilon,i}^{n+1} = \frac{|K|}{|K|+|L|}\left[(R_{\epsilon,i}^D)^{n+1} + \frac{|F^n_{\sigma_\epsilon^D,\epsilon}|}{4}(u^n_{\sigma_\epsilon^D,i} - u^n_{\sigma_\epsilon^U,i})^2\right],$$

où $\sigma_\epsilon^D = K|L$.

Grâce à cette procédure, on obtient l'identité recherchée :

$$\sum_{K\in\mathcal{M}} S_K^{n+1} - \sum_{i=1}^d \sum_{\sigma\in\mathcal{E}_S^{(i)}} R_{\sigma,i}^{n+1} = 0, \tag{1.50}$$

qui permet de garantir une conservation de l'énergie totale discrète.

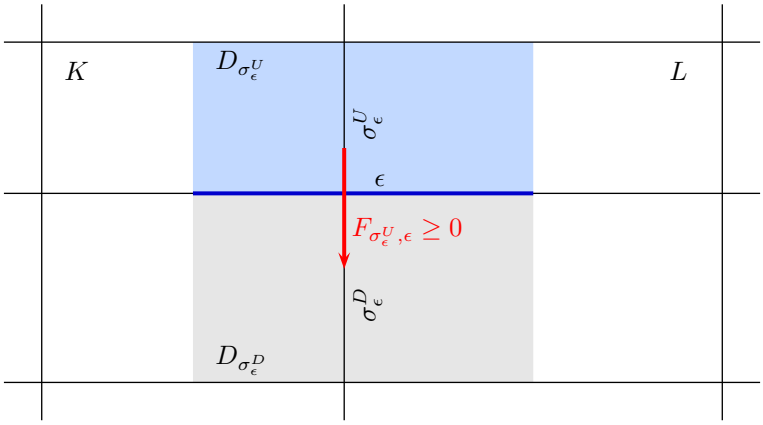


Figure 1.2 – Notations Pour la construction du terme source $S_{K,1}$, dans le cas MAC, pour une face duale à cheval sur deux faces primales. ϵ : face duale considérée. $D_{\sigma_\epsilon^U}$: cellule duale amont. $D_{\sigma_\epsilon^D}$: cellule duale aval. K, L : cellules amonts.

1.4.6 Propriétés des schémas

On peut découpler les propriétés du schéma en deux catégories : les propriétés de stabilité de la solution discrète et les propriétés de consistance du schéma. L'ensemble des preuves de ces résultats fera l'object du troisière chapitre de cette thèse.

Propriétés de stabilité On présente l'ensemble des propriétés de stabilité dans les trois lemmes suivants : Le premier lemme permet d'assurer la positivité de l'énergie interne et de la masse volumique au niveau discret.

Lemme 1.3 (*Positivité de la masse volumique et de l'énergie*)

Soit $n \in \mathbb{N}$, soit $(\rho_K^n,\ \boldsymbol{u}_K^n,\ e_K^n)_{K\in\mathcal{M}} \in (\mathbb{R}^{\mathrm{card}\mathcal{M}} \times (\mathbb{R}^{\mathrm{card}\mathcal{E}})^d \times \mathbb{R}^{\mathrm{card}\mathcal{M}})$, et supposons que e_K^n et ρ_K^n soient positifs, $\forall K \in \mathcal{M}$; si $(\rho_K^n,\ \boldsymbol{u}_K^n,\ e_K^n)_{K\in\mathcal{M}}$ satisfont (1.6a)-(1.6f) et (1.5a)-(1.5c) ainsi que la condition de CFL

$$\delta t \leq \min\Big(\frac{|K|}{\sum\limits_{\sigma\in\mathcal{E}(K)} |\sigma|\ (1+(\alpha_{K,\sigma})^n)\,(u_{K,\sigma}^n)^+},\ \frac{|K|\rho_K^n}{\sum\limits_{\sigma\in\mathcal{E}(K)} (\gamma-1)|\sigma|\ \rho_K^n(u_{K,\sigma}^n)^+ + (F_{K,\sigma}^n)^+ + \alpha_{K,\sigma}|F_{K,\sigma}^n|},\ \frac{|D_{K,\sigma}|\ \rho_K^n}{\sum\limits_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma),\ \epsilon\cap\bar{K}\neq\emptyset} \nu_\epsilon^n + |F_{\sigma,\epsilon}^n|}\Big), \tag{1.51}$$

alors $e_K^{n+1} \geq 0$ et $\rho_K^{n+1} \geq 0$, pour tout $K \in \mathcal{M}$.

Le théorème suivant permet de s'assurer de l'existence d'une solution discrète ainsi que de la conservation de l'énergie totale discrète pour les deux types de discrétisations temporelles.

Theorème 1.4 (*Existence et conservation de l'énergie totale*)

Supposons que pour tout $K \in \mathcal{M}$, $e_K^0 > 0$, $\rho_K^0 > 0$ et $\rho_K^{-1} > 0$. Alors il existe une solution pour chacun des schémas qui vérifie, $\forall n \in \mathbb{N}$ et $\forall K \in \mathcal{M}$:

$$\sum_{K\in\mathcal{M}} |K|\,\rho_K^n e_K^n + \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_S^{(i)}} |D_\sigma|\ \rho_{D_\sigma}^{n-1}\ (u_{\sigma,i}^n)^2 + \mathcal{R}^n \leq \sum_{K\in\mathcal{M}} |K|\,\rho_K^0 e_K^0 + \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_S^{(i)}} |D_\sigma|\ \rho_{D_\sigma}^{-1}\ (u_{\sigma,i}^0)^2 + \mathcal{R}^0,$$

avec :

$$\mathcal{R}^n = \delta t^2 \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \frac{|D_\sigma|}{\rho_{D_\sigma}^{n-1}}\,|(\nabla p)_\sigma^n|^2 = \delta t^2 \sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} \frac{|\sigma|^2}{|D_\sigma|\ \rho_{D_\sigma}^{n-1}}\,(p_K^n - p_L^n)^2,$$

pour le schéma à correction de pression et

$$\sum_{K\in\mathcal{M}} |K|\,\rho_K^{n+1} e_K^{n+1} + \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_S^{(i)}} |D_\sigma|\ \rho_{D_\sigma}^{n}\ (u_{\sigma,i}^n)^2 \leq \sum_{K\in\mathcal{M}} |K|\,\rho_K^1 e_K^1 + \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_S^{(i)}} |D_\sigma|\ \rho_{D_\sigma}^{0}\ (u_{\sigma,i}^0)^2$$

pour le schéma découplé.

Avant de donner la troisième et dernière propriété de stabilité, il est nécessaire d'introduire, grâce au lemme suivant, une notation supplémentaire.

Lemme 1.5 (*Propriété de convexité*)

Soit $g(.)$ une fonction strictement convexe régulière, dérivable sur l'ouvert $I \subset \mathbb{R}$. Soient $x_1 \in I$ et $x_2 \in I$ deux réels distincts. La relation suivante :

$$g(x_1) + (\bar{x} - x_1)\,g'(x_1) = g(x_2) + (\bar{x} - x_2)\,g'(x_2)$$

définit de façon unique le réel $\bar{x}$. De plus, $\bar{x} \in |[x_1, x_2]|$, avec $|[a,b]| = \{\alpha x_1 + (1-\alpha)x_2\}_{\alpha\in[0,1]}$.

Le dernier théorème de stabilité montre que les solutions des schémas vérifient une forme discrète de l'inégalité d'entropie.

Theorème 1.6 (*Inégalité d'entropie discrète*)

Pour $K \in \mathcal{M}$ et $n \in \mathbb{N}$, on définit l'entropie discrète suivante

$$\rho_K^n \eta_K^n = \phi(\rho_K^n) + \rho_K^n \psi(e_K^n), \tag{1.52}$$

avec $\phi(\rho) = \rho ln(\rho)$, $\psi(e) = \dfrac{1}{1-\gamma} ln(e)$. On suppose que les interpolations MUSCL de ρ et e satisfont les limitations addtionnelles suivantes, pour tout $\sigma = K|L \in \mathcal{E}_{\text{int}}$ tel que $F_{K,\sigma} \geq 0$,

$$\rho_\sigma \in |[\rho_K, \bar{\rho}_\sigma]| \qquad \text{et} \qquad e_\sigma \in |[e_K, \bar{e}_\sigma]|, \tag{1.53}$$

avec $\bar{\rho}_\sigma \in |[\rho_K, \rho_L]|$ et $\bar{e}_\sigma \in |[e_K, e_L]|$ sont tels que

$$\phi(\rho_L) + (\bar{\rho}_\sigma - \rho_L)\,\phi'(\rho_L) = \phi(\rho_K) + (\bar{\rho}_\sigma - \rho_K)\,\phi'(\rho_K), \tag{1.54}$$

$$\psi(e_L) + (\bar{e}_\sigma - e_L)\,\psi'(e_L) = \psi(e_K) + (\bar{e}_\sigma - e_K)\,\psi'(e_K). \tag{1.55}$$

L'existence et l'unicité de $\bar{e}_\sigma$ et $\bar{\rho}_\sigma$ est une conséquence directe de (1.5).

Alors les inégalités suivantes sont vérifiées :

$$\frac{|K|}{\delta t}\left(\rho_K^{n+1}\eta_K^{n+1} - \rho_K^n \eta_K^n\right) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^{n+1}\eta_\sigma^{n+1} + T_{conv,K}^{n+1} \leq 0, \quad \forall K \in \mathcal{M} \text{ et } \quad n \in \mathbb{N} \tag{1.56}$$

pour le schéma à correction de pression, et

$$\frac{|K|}{\delta t}\left(\rho_K^{n+1}\eta_K^{n+1} - \rho_K^n \eta_K^n\right) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^{n}\eta_\sigma^{n} + P_K^n + T_{conv,K}^{n} \leq 0, \tag{1.57}$$

pour le schéma découplé, avec, pour $m = n|n+1$:

$$\rho_\sigma^m \eta_\sigma^m = \phi(\rho_\sigma^m) + \rho_\sigma^m \psi(e_\sigma^m),$$

$T_{conv,K}^m = T_{conv,K,\rho}^m + T_{conv,K,e}^m$ *avec*

$$\begin{aligned} T_{conv,K,\rho}^m &= \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}^m P_{\sigma,\rho}^m, \\ T_{conv,K,e}^m &= \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^m P_{\sigma,e}^m, \end{aligned} \tag{1.58}$$

où

$$\begin{aligned} P_{\sigma,\rho}^m &= (\rho_\sigma^m - \bar{\rho}_\sigma^m)\left(\frac{\phi'(\rho_K^m) + \phi'(\rho_L^m)}{2}\right) + (\bar{\rho}_\sigma^m - \rho_K^m)\phi'(\rho_K^m) + \phi(\rho_K^m) - \phi(\rho_\sigma^m), \\ P_{\sigma,e}^m &= (e_\sigma^m - \bar{e}_\sigma^m)\left(\frac{\psi'(e_K^m) + \psi'(e_L^m)}{2}\right) + (\bar{e}_\sigma^m - e_K^m)\psi'(e_K^m) + \psi(e_K^m) - \psi(e_\sigma^m), \end{aligned}$$

et

$$P_K^n = \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n\left(\rho_K^{n+1} - \rho_K^n\right)\phi''(\rho_K^{(1)}) + \sum_{\sigma\in\mathcal{E}(K)} \left[|\sigma| p_K^n u_{K,\sigma}^n + F_{K,\sigma}^n(e_\sigma^n - e_K^n)\right]\left(e_K^{n+1} - e_K^n\right)\psi''(e_K^{(1)}),$$

avec $\rho_K^{(1)} \in |[\rho_K^n, \rho_K^{n+1}]|$ et $e_K^{(1)} \in |[e_K^n, e_K^{n+1}]|$.

Tout comme le bilan discret d'énergie cinétique, l'inégalité d'entropie n'est pas exacte au niveau discret. Il existe des termes résiduels qui sont dûs à l'interpolation MUSCL d'une part et à la discrétisation explicite d'autre part. A noter cependant, que le bilan d'entropie discret est exact pour le schéma de correction de pression avec discrétisation upwind.

Tous les résultats introduits dans ce paragraphe sont nécessaires pour démontrer le théorème de consistance que nous allons énoncer dans la section suivante.

1.4.7 Consistance des schémas numériques

Les schémas qui ont été introduits vérifient une propriété importante de consistance au sens de Lax ; toute suite de solutions discrètes convergente, sous certaines hypothèses de contrôle de normes, va converger vers une limite qui satisfait une formulation faible du système des équations d'Euler.

Une solution faible du système (1.1) satisfait le système suivant, pour tout $\varphi \in \mathrm{C}_c^\infty\big(\Omega\times[0,T)\big)$ (et tout $\boldsymbol{\varphi} \in \mathrm{C}_c^\infty\big(\Omega \times [0,T)\big)^d$) :

$$-\int_0^T \int_\Omega \Big[\rho\, \partial_t\varphi + \rho\, u \cdot \nabla\varphi\Big]\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} = 0, \tag{1.59a}$$

$$-\int_0^T \int_\Omega \Big[\rho\, \boldsymbol{u} \cdot \partial_t\boldsymbol{\varphi} + (\rho\boldsymbol{u} \otimes \boldsymbol{u}) : \underline{\underline{\nabla}}\boldsymbol{\varphi} + p\mathrm{div}(\boldsymbol{\varphi})\Big]\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(\boldsymbol{x})\, \boldsymbol{u}_0(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} = 0, \tag{1.59b}$$

$$-\int_{\Omega\times(0,T)} \Big[\rho\, E\, \partial_t\varphi + (\rho\, E + p)\, \boldsymbol{u} \cdot \nabla\varphi\Big]\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(\boldsymbol{x})\, E_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} = 0, \tag{1.59c}$$

$$p = (\gamma - 1)\rho\, e, \quad E = \frac{1}{2}|\boldsymbol{u}|^2 + e, \quad E_0 = \frac{1}{2}|\boldsymbol{u}_0|^2 + e_0. \tag{1.59d}$$

On complète ces équations par une inégalité d'entropie faible, pour tout $\varphi \in \mathrm{C}_c^\infty\big(\Omega \times [0,T)\big)$, φ positive :

$$-\int_0^T \int_\Omega \Big[\rho\, \eta\, \partial_t\varphi + (\rho\, \eta\, \boldsymbol{u}) \cdot \nabla\varphi\Big]\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(\boldsymbol{x})\, \eta_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} \le 0. \tag{1.60}$$

Rigoureusement, cette formulation n'est pas suffisante pour définir une solution faible du problème car les conditions aux limites ne sont pas prises en compte. Néanmoins, elle permet d'obtenir les conditions de Rankine-Hugoniot. Cela permet donc de s'assurer que si les suites convergentes de solutions discrètes de ces schémas vérifient ces relations à la limite, alors ces derniers peuvent calculer des chocs correctement. Cette propriété fera l'objet des deux théorèmes (un pour chaque discrétisation temporelle) présentés ci-dessous. Avant de ce faire, il est nécessaire d'introduire quelques notations, ainsi que des hypothèses sur le contrôle des normes des solutions discrètes.

Définitions – Dans un soucis de simplicité, on se limite dans un premier temps aux schémas découplés. On présentera par la suite les différences minimes à apporter pour les schémas de correction de pression.

On introduit tout d'abord les notations qui sont communes aux discrétisations de type RT-CR et MAC. Soit $L_{\mathcal{M}}(\Omega \times (0,T))$ l'espace des fonctions constantes sur chaque $K \times (t^n, t^{n+1})$, $K \in \mathcal{M}$, $n \in [\![0, N-1]\!]$. On rappelle que h_K désigne le diamètre de la cellule, et on note r_K le rayon de la plus grosse boule contenue dans K. Pour une face duale interne $\epsilon = D_\sigma|D_{\sigma'} \in \tilde{\mathcal{E}}_{\mathrm{int}}$ on désigne par $d_{\sigma,\epsilon}$ la distance euclidienne entre le centre de masse de σ $\boldsymbol{x}_\sigma$ et la face ϵ. On note $d_\epsilon = d_{\sigma,\epsilon} + d_{\sigma',\epsilon}$. La taille de la discrétisation est caractérisée par la grandeur suivante :

$$h_{\mathcal{M}} = \sup\{h_K,\ K \in \mathcal{M}\}.$$

Pour $q \in L_{\mathcal{M}}(\Omega \times (0,T))$, on définit la norme discrète $\mathrm{L}^1((0,T);\mathrm{BV}(\Omega))$ par :

$$\|q\|_{\mathcal{T},x,\mathrm{BV}} = \sum_{n=0}^{N} \delta t \sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} |\sigma||q_L^n - q_K^n|,$$

et la norme discrète $\mathrm{L}^1(\Omega;\mathrm{BV}((0,T)))$ par :

$$\|q\|_{\mathcal{T},t,\mathrm{BV}} = \sum_{K\in\mathcal{M}} |K| \sum_{n=0}^{N-1} |q_K^{n+1} - q_K^n|.$$

Les autres définitions nécessaires dépendent de la discrétisation spatiale et seront détaillées pour chacune d'entre elles.

Schéma RT-CR – On mesure la régularité du maillage à travers le réel positif θ dont l'expression est la suivante :

$$\theta = \max\left\{\frac{h_K}{r_K},\ K \in \mathcal{M}\right\} \cup \left\{\frac{h_K}{d_\epsilon},\ K \in \mathcal{M},\ \epsilon \subset K\right\} \cup \left\{\frac{d_{\sigma,\epsilon}}{d_{\sigma',\epsilon}},\ \sigma,\sigma' \in \mathcal{E},\ \epsilon = \sigma|\sigma'\right\}. \tag{1.61}$$

La première donnée permet de mesurer l'aplatissement du maillage primal, la seconde de relier les ordres de grandeur des maillages primaux et duaux et la dernière de vérifier qu'il n'y ait aucun aplatissement du maillage dual.

Soient $H_{\mathcal{E}}(\Omega \times (0,T))$ l'espace des fonctions constantes sur chaque maille spatio-temporelle duale $D_\sigma \times (t^n, t^{n+1})$, $\sigma \in \mathcal{E}$, $n \in [\![0, N-1]\!]$ et $H_{\mathcal{E},0}(\Omega \times (0,T))$ le sous espace des fonctions nulles sur chaque D_σ, $\sigma \in \tilde{\mathcal{E}}_{\mathrm{ext}}$.

Soit $v \in H_{\mathcal{E},0}(\Omega \times (0,T))$, on définit la norme BV discrète $\mathrm{L}^1((0,T);\mathrm{BV}(\Omega))$ par :

$$\|v\|_{\mathcal{T},x,\mathrm{BV}} = \sum_{n=0}^{N} \delta t \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}_{\mathrm{int}}} |\epsilon||v_{\sigma'}^n - v_\sigma^n|,$$

et la norme $\mathrm{L}^1(\Omega;\mathrm{BV}((0,T)))$ par :

$$\|v\|_{\mathcal{T},t,\mathrm{BV}} = \sum_{\sigma\in\mathcal{E}} |D_\sigma| \sum_{n=0}^{N-1} |v_\sigma^{n+1} - v_\sigma^n|.$$

Schéma MAC– La régularité du maillage est mesurée de façon plus simple pour le maillage cartésien :

$$\theta = \max\left\{\frac{h_K}{r_K},\ K \in \mathcal{M}\right\} \cup \left\{\frac{d_{\sigma,\epsilon}}{d_{\sigma',\epsilon}},\ \sigma,\sigma' \in \mathcal{E},\ \epsilon = \sigma|\sigma'\right\}. \tag{1.62}$$

On note $H_{\mathcal{E}}^{(i)}(\Omega \times (0,T))$ l'espace des fonctions constantes sur chaque $D_\sigma \times (t^n, t^{n+1})$, $\sigma \in \mathcal{E}^{(i)}$, $i \in [\![1, d]\!]$ et $H_{\mathcal{E},0}^{(i)}(\Omega \times (0,T))$ le sous espace des fonctions nulles au bord.

Les normes BV discrètes sont maintenant définies suivant chaque direction d'espace. Pour $v \in H_{\mathcal{E},0}^{(i)}(\Omega \times (0,T))$, la norme discrète $\mathrm{L}^1(\Omega;\mathrm{BV}((0,T)))$ est définie par :

$$\|v\|_{\mathcal{T},x,\mathrm{BV},(i)} = \sum_{n=0}^{N} \delta t \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}_{\mathrm{int}}^{(i)}} |\epsilon||v_{\sigma'}^n - v_\sigma^n|,$$

et la norme discrète $\mathrm{L}^1(\Omega;\mathrm{BV}((0,T)))$ par :

$$\|v\|_{\mathcal{T},t,\mathrm{BV},(i)} = \sum_{\sigma\in\mathcal{E}^{(i)}} |D_\sigma| \sum_{n=0}^{N-1} |v_\sigma^{n+1} - v_\sigma^n|.$$

Soit $\boldsymbol{v} = (v^{(1)}, \cdots v^{(d)}) \in H^{(1)}_{\mathcal{E},0}(\Omega\times(0,T))\times\cdots\times H^{(d)}_{\mathcal{E},0}(\Omega\times(0,T))$. On définit la norme globale comme suit :

$$\|\boldsymbol{v}\|_{\mathcal{T},t,\mathrm{BV}} = \max_{i\in[\![1,d]\!]} \|v^{(i)}\|_{\mathcal{T},x,\mathrm{BV},(i)} \qquad \|\boldsymbol{v}\|_{\mathcal{T},x,\mathrm{BV}} = \max_{i\in[\![1,d]\!]} \|v^{(i)}\|_{\mathcal{T},x,\mathrm{BV},(i)}.$$

Passons maintenant à la définition des solutions globales des schémas découplés.

Définition 1.3 (*Définition des solutions globales explicites*)

Soit une suite de discrétisations $(\mathcal{M}^{(m)},\ \delta t^{(m)})_{m\in\mathbb{N}}$ donnée. On note $h^{(m)}$ la taille caractéristique du maillage. Soient $\rho^{(m)}$, $p^{(m)}$, $e^{(m)}$ et $\boldsymbol{u}^{(m)}$ les solutions du schéma (1.5) sur le maillage $\mathcal{M}^{(m)}$ avec un pas de temps $\delta t^{(m)}$. Aux inconnues discrètes, on associe des fonctions constantes sur les intervalles de temps et sur les éléments du maillage primal et dual, de telle sorte que la masse volumique $\rho^{(m)}$, la pression $p^{(m)}$, l'énergie interne $e^{(m)}$ et la vitesse $\boldsymbol{u}^{(m)}$ soient définies presque partout sur $\Omega\times(0,T)$ par :

$$\begin{aligned}
\rho^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (\rho^{(m)})^n_K\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t),\\
p^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (p^{(m)})^n_K\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t),\\
e^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (e^{(m)})^n_K\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t),\\
\boldsymbol{u}^{(m)}(\boldsymbol{x},t) &= \begin{cases} \displaystyle\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}} (\boldsymbol{u}^{(m)})^n_\sigma\, \mathcal{X}_{D_\sigma}(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t) & \text{pour les schémas RT et CR}\\ \displaystyle\sum_{i=1}^{d}\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}^{(i)}} (u^{(m)})^n_{\sigma,i}\, \mathcal{X}_{D_\sigma}(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t) & \text{pour le schéma MAC,}\end{cases}
\end{aligned} \tag{1.63}$$

avec $\mathcal{X}_O$ qui désigne la fonction indicatrice de l'ensemble O. De manière analogue on définit $\eta^{(m)} \in L_{\mathcal{M}^{(m)}}(\Omega\times(0,T))$ de la façon suivante :

$$\eta^{(m)}(\boldsymbol{x},t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (\eta^{(m)})^n_K\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t), \tag{1.64}$$

où l'entropie $(\eta^{(m)})^n_K$ est donnée par (1.52).

Hypothèses – Dans le but de pouvoir obtenir les résultats de consistance voulus, il est nécessaire de faire des hypothèses d'estimations sur la solution discrète du schéma. Sous la condition CFL (1.51), une suite de solutions discrètes $(\rho^{(m)}, p^{(m)}, e^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ vérifie $\rho^{(m)} > 0$, $p^{(m)} > 0$ et $e^{(m)} > 0$, $\forall m \in \mathbb{N}$. On suppose qu'elle est uniformément bornée dans $\mathrm{L}^\infty(\Omega\times(0,T))^{d+3}$, *i.e.*, pour $m \in \mathbb{N}$ et $0 \le n \le N^{(m)}$:

$$0 < (\rho^{(m)})^n_K \le C,\ 0 < (p^{(m)})^n_K \le C,\ 0 < (e^{(m)})^n_K \le C, \quad \forall K \in \mathcal{M}^{(m)}, \tag{1.65}$$

et

$$|(\boldsymbol{u}^{(m)})^n_\sigma| \le C, \quad \forall \sigma \in \mathcal{E}^{(m)}, \tag{1.66}$$

où C est une constante réelle positive. Un certain nombre d'hypothèses supplémentaires doivent être faites pour obtenir le caractère entropique des schémas. Supposons donc de plus que $\dfrac{1}{\rho^{(m)}}$ et $\dfrac{1}{e^{(m)}}$ soient dans $\mathrm{L}^\infty(\Omega\times(0,T))$. A noter que ces hypothèses impliquent que ρ_0, e_0, $\boldsymbol{u}_0$

appartiennent à $L^\infty(\Omega)$ et que

$$\rho_0(\boldsymbol{x}) > \rho_{\min} > 0 \qquad e_0 > e_{\min} > 0.$$

Il est aussi nécessaire de supposer que la suite de solutions discrètes vérifie une hypothèse sur les normes BV :

$$\lim_{m\to\infty} \left(h^{(m)} + \delta t^{(m)}\right) \left[\|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|p^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|e^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},t,\mathrm{BV}}\right] = 0. \tag{1.67}$$

A noter que cette hypothèse est plus faible qu'une borne uniforme sur les estimations BV.

Supposons enfin que les hypothèses du théorème (1.6) soient vérifiées, ainsi que la condition de CFL additionnelle suivante :

$$\lim_{m\to+\infty} \frac{\delta t^{(m)}}{\min\limits_{K\in\mathcal{M}^{(m)}} h_K} \left(\|\rho^{(m)}\|_{\mathcal{T},t,\mathrm{BV}} + \|e^{(m)}\|_{\mathcal{T},t,\mathrm{BV}}\right) = 0. \tag{1.68}$$

Théorèmes de consistance Le théorème qui suit est une version multidimensionnelle d'un résultat démontré en 1D dans [38].

Theorem 1.1 (*Consistance du schéma découplé*)

Soit Ω un intervalle ouvert de $\mathbb{R}$. Soient $\rho_0 \in L^\infty(\Omega)$, $p_0 \in \mathrm{BV}(\Omega)$, $e_0 \in L^\infty(\Omega)$ et $\boldsymbol{u}_0 \in L^\infty(\Omega)^d$. Soit $(\mathcal{M}^{(m)}, \delta t^{(m)})_{m\in\mathbb{N}}$ une suite de discrétisations telle que le pas de temps $\delta t^{(m)}$ et le pas $h^{(m)}$ du maillage $\mathcal{M}^{(m)}$ tendent vers zéro quand $m \to \infty$, et soit $(\rho^{(m)}, p^{(m)}, e^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ la suite des solutions discrètes du schéma associé. Supposons que cette suite vérifie les estimations du paragraphe 1.4.7 et converge dans $L^r(\Omega\times(0,T))^3 \times L^r(\Omega\times(0,T))^d$, pour $1 \le r < \infty$, vers $(\bar\rho, \bar p, \bar e, \bar{\boldsymbol{u}}) \in L^\infty(\Omega\times(0,T))^3 \times L^\infty(\Omega\times(0,T))^d$.

Alors la limite de cette suite $(\bar\rho, \bar p, \bar e, \bar{\boldsymbol{u}})$ est solution du système (1.59)–(1.60).

Idée de preuve :

La démonstration complète de ce théorème fait l'objet du chapitre 3 de la thèse. Nous allons simplement donner les idées principales ici.

— On multiplie chaque équation discrète par des interpolées sur le maillage de fonctions test φ et on somme sur tous les éléments de la discrétisation.
— Par des manipulations algébriques (intégration par partie discrète, ...) on transporte les dérivées discrètes sur les fonctions tests et on fait apparaître les intégrales discrètes plus des termes de reste.
— On passe à la limite dans les intégrales discrètes pour obtenir les équations intégrales voulues.
— On utilise les hypothèses sur les normes des solutions discrètes pour faire tendre les termes de reste vers 0.

Un théorème similaire existe pour le schéma à correction de pression, mais nécessite auparavant quelques ajustements. La définition des solutions globales discrètes du maillage est

différente pour les schémas de correction de pression. En effet, on a :

$$
\begin{aligned}
\rho^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (\rho^{(m)})_K^{n+1}\,\mathcal{X}_K(\boldsymbol{x})\,\mathcal{X}_{(n,n+1]}(t),\\
p^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (p^{(m)})_K^{n+1}\,\mathcal{X}_K(\boldsymbol{x})\,\mathcal{X}_{(n,n+1]}(t),\\
e^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (e^{(m)})_K^{n+1}\,\mathcal{X}_K(\boldsymbol{x})\,\mathcal{X}_{(n,n+1]}(t),\\
\boldsymbol{u}^{(m)}(\boldsymbol{x},t) &= \begin{cases} \displaystyle\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}} (\boldsymbol{u}^{(m)})_\sigma^{n+1}\,\mathcal{X}_{D_\sigma}(\boldsymbol{x})\,\mathcal{X}_{(n,n+1]}(t), & \text{schéma RT et CR}\\ \displaystyle\sum_{i=1}^{d}\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}^{(i)}} (u^{(m)})_{\sigma,i}^{n+1}\,\mathcal{X}_{D_\sigma}(\boldsymbol{x})\,\mathcal{X}_{(n,n+1]}(t), & \text{schéma MAC}\end{cases}\\
\tilde{\boldsymbol{u}}^{(m)}(\boldsymbol{x},t) &= \begin{cases} \displaystyle\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}} (\tilde{\boldsymbol{u}}^{(m)})_\sigma^{n+1}\,\mathcal{X}_{D_\sigma}(\boldsymbol{x})\,\mathcal{X}_{(n,n+1]}(t), & \text{schéma RT-CR}\\ \displaystyle\sum_{i=1}^{d}\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}^{(i)}} (\tilde{u}^{(m)})_{\sigma,i}^{n+1}\,\mathcal{X}_{D_\sigma}(\boldsymbol{x})\,\mathcal{X}_{(n,n+1]}(t) & \text{schéma MAC.}\end{cases}
\end{aligned}
\tag{1.69}
$$

Les hypothèses sur le contrôle des normes sont elles aussi légèrement modifiées :

$$
\lim_{m\to\infty}\left(h^{(m)}+\delta t^{(m)}\right)\left[\|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}+\|p^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}+\|e^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}+\|\tilde{\boldsymbol{u}}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}+\|\boldsymbol{u}^{(m)}\|_{\mathcal{T},t,\mathrm{BV}}\right]=0. \tag{1.70}
$$

Nous supposons en outre une borne L^∞ sur $\tilde{\boldsymbol{u}}^{(m)}$. Il est à noter que nous n'avons besoin d'aucune hypothèse sur les dérivées discrètes temporelles de $\tilde{\boldsymbol{u}}^{(m)}$ ou sur les dérivées discrètes spatiales de $\boldsymbol{u}^{(m)}$.

Theorem 1.2 (*Consistance du schéma de correction de pression*)

Soit Ω un ouvert borné de $\mathbb{R}^d$. Supposons que les données initiales vérifient $\rho_0 \in \mathrm{L}^\infty(\Omega)$, $p_0 \in \mathrm{BV}(\Omega)$, $e_0 \in \mathrm{L}^\infty(\Omega)$ et $\boldsymbol{u}_0 \in \mathrm{L}^\infty(\Omega)^d$. Soit $(\mathcal{M}^{(m)}, \delta t^{(m)})_{m\in\mathbb{N}}$ une suite de discrétisations telle que $\delta t^{(m)}$ et $h^{(m)}$ tendent vers zéro quand $m\to\infty$, et soient $(\rho^{(m)}, p^{(m)}, e^{(m)}, \boldsymbol{u}^{(m)}, \tilde{\boldsymbol{u}}^{(m)})_{m\in\mathbb{N}}$ les solutions discrètes du schéma associé. On suppose que cette suite de solutions vérifie les hypothèses (1.69) et (1.70) et qu'elle converge dans $\mathrm{L}^r(\Omega\times(0,T))^3\times(\mathrm{L}^r(\Omega\times(0,T))^d)^2$, pour $1\le r<\infty$, vers $(\bar\rho, \bar p, \bar e, \bar{\boldsymbol{u}}, \bar{\tilde{\boldsymbol{u}}})\in \mathrm{L}^\infty(\Omega\times(0,T))^3\times(\mathrm{L}^\infty(\Omega\times(0,T))^d)^2$.

Alors $\bar{\tilde{\boldsymbol{u}}} = \bar{\boldsymbol{u}}$ et la limite $(\bar\rho, \bar p, \bar e, \bar{\boldsymbol{u}})$ est solution du système faible (1.59).

La démonstration est similaire à celle effectuée pour le schéma découplé. Il est néanmoins nécessaire de démontrer additionnellement que $\boldsymbol{u}^{(m)}$ et $\tilde{\boldsymbol{u}}^{(m)}$ ont la même limite.

1.5 Equation de propagation du front de flamme

Comme on a pu le voir dans le chapitre introductif, on représente la propagation du front de flamme pendant la phase de déflagration par une équation de type level-set appelée « G-équation » :

$$
\partial_t(\rho G) + \mathrm{div}(\rho G\boldsymbol{u}) + \rho_u u_f|\nabla G| = 0,
$$

avec u_f la vitesse de flamme et ρ_u la masse volumique des gaz frais. Cette équation est de type Hamilton-Jacobi. On peut en effet la réécrire, grâce au bilan de masse, sous la forme suivante :

$$
\partial_t G + H(\nabla G) = 0, \qquad \text{avec} \quad H(\boldsymbol{x}) = \boldsymbol{u}\cdot\boldsymbol{x} + \frac{\rho_u}{\rho}\,u_f|\boldsymbol{x}|.
$$

Les schémas développés pour les équations d'Hamilton-Jacobi s'appuient sur la théorie d'existence et d'unicité développée au niveau continu par P.-L. Lions dans [21, 46]. Elle fait intervenir la notion de solution de viscosité comme nous le verrons dans la section suivante. Les premiers schémas différences finies développés pour résoudre numériquement ces équations ont été introduits dans [20], accompagnés d'un résultat de convergence de ces schémas. Ce cadre théorique a été généralisé à tout type d'approximations numériques dans [6, 58]. Des extensions d'ordre élevé ont été introduites par S.Osher et James A. Sethian dans [51]. D'autre part des schémas fondés sur des discrétisations non structurées ont été développés par R. Abgrall [1] et P. Souganidis [41]. Dès lors, de nombreux schémas ont été développés pour résoudre ces équations ; des schémas aux différences finies d'ordre élevé dans [13, 56, 52], mais aussi des schémas sur maillage non structuré comme par exemple [9, 57, 65, 5]. L'idée principale est ici de développer un schéma volumes finis original fondé sur le même type de discrétisations que celles introduites pour les équations d'Euler et sur la forme particulière de la G-équation, en se basant sur la théorie d'existence de solutions au niveau continu comme dans le cas des travaux sur les schémas différences finies pour Hamilton-Jacobi. A partir de maintenant, on se concentrera sur le modèle canonique de l'équation eikonale instationnaire suivant :

$$\begin{cases} \partial_t G + |\nabla G| = 0, \\ G(0,x) = G_0(x) \in \mathrm{BUC}(\mathbb{R}^d), \end{cases} \tag{1.71}$$

avec BUC l'espace des fonctions bornées uniformément continues. Avant de rentrer dans la description des schémas utilisés, on donne le cadre théorique d'existence de solutions à ce problème continu.

1.5.1 Existence de solutions continues

La théorie d'existence de solutions au niveau continu s'appuie sur un passage à la limite d'un problème d'Hamilton-Jacobi régularisé par la méthode de la viscosité évanescente. On aboutit alors à l'existence d'une unique solution d'un problème faible fondé sur le principe du maximum. Afin de donner une idée du résultat qui va suivre, considérons le problème régularisé suivant, pour $\epsilon > 0$ et $\mathbb{R}^d = \mathbb{R}^s$,

$$\begin{aligned} &\partial_t G^\epsilon + |\nabla G^\epsilon| - \epsilon \Delta G^\epsilon = 0, \\ &G^\epsilon(0,x) = G_0(x) \in \mathrm{BUC}(\mathbb{R}^d). \end{aligned} \tag{1.72}$$

On sait qu'il existe une solution unique régulière à ce problème. Supposons que l'on ait une solution régulière de classe C^2 et soit $\varphi \in C^2(\mathbb{R}^d \times (0,T])$ telle que $G^\epsilon - \varphi$ admette un maximum local en $(x_0, t_0) \in \mathbb{R}^d \times (0,T]$. Alors selon le principe du maximum, les dérivées premières sont nulles en (x_0, t_0), *i.e.*

$$\nabla G^\epsilon(t_0,x_0) = \nabla \varphi(t_0,x_0) \qquad \partial_t G^\epsilon(t_0,x_0) = \partial_t \varphi(t_0,x_0),$$

et le laplacien est négatif :

$$\Delta(G^\epsilon - \varphi)(t_0,x_0) \leq 0. \tag{1.73}$$

On a alors

$$\partial_t \varphi(t_0,x_0) + |\nabla \varphi(t_0,x_0)| - \epsilon \Delta \varphi(t_0,x_0) - \epsilon \Delta(G^\epsilon - \varphi)(t_0,x_0) = 0.$$

En utilisant (1.73) on aboutit finalement à :

$$\partial_t \varphi(t_0,x_0) + |\nabla \varphi(t_0,x_0)| - \epsilon \Delta \varphi(t_0,x_0) \leq 0$$

De façon analogue pour un minimum local on a :

$$\partial_t \varphi(t_0,x_0) + |\nabla \varphi(t_0,x_0)| - \epsilon \Delta \varphi(t_0,x_0) \geq 0.$$

On obtient ainsi la formulation faible naturelle qu'on va faire tendre à la limite quand ϵ va tendre vers zéro. De cette même manière, on prouve l'existence et l'unicité de la solution des équations d'Hamilton-Jacobi. Ainsi dans notre cas, on prouve qu'il existe une unique fonction $G \in \mathrm{BUC}([0,T]\times\mathbb{R}^d)$ telle que $G(0,\boldsymbol{x}) = G_0(\boldsymbol{x})$, et, $\forall\varphi \in C^1(\mathbb{R}^d\times(0,\infty))$, si (x_0,t_0) est un maximum local de $G-\varphi$ sur $\mathbb{R}^d\times(0,T]$, alors :

$$\partial_t\varphi(x_0,t_0) + H(\boldsymbol{\nabla}\varphi(x_0,t_0)) \leq 0, \tag{1.74}$$

et $\forall\varphi \in C^1(\mathbb{R}^d\times(0,\infty))$, si (x_0,t_0) est un minimum local de $G-\varphi$ sur $\mathbb{R}^d\times(0,T]$, on a :

$$\partial_t\varphi(x_0,t_0) + H(\boldsymbol{\nabla}\varphi(x_0,t_0)) \geq 0. \tag{1.75}$$

Comme on vient de le voir, on a donné le problème continu sous forme de problème de Cauchy. Néanmoins pour des raisons pratiques (tests numériques), on va être amené à considérer des domaines ouverts bornés Ω. Pour simuler des frontières libres de notre domaine, un choix judicieux des données initiales sera fait ainsi que des conditions aux limites de type Neumann homogène.

1.5.2 Schéma numérique

Considérons la G-équation sous forme non conservative :

$$\partial_t G + \left(\frac{\boldsymbol{\nabla}G}{|\boldsymbol{\nabla}G|}\right)\cdot\boldsymbol{\nabla}G = 0, \tag{1.76}$$

et rappellons l'identité suivante :

$$\boldsymbol{u}.\boldsymbol{\nabla}\phi = \mathrm{div}(\phi\boldsymbol{u}) - \phi\mathrm{div}(\boldsymbol{u}). \tag{1.77}$$

Considérons une partition $0 = t_0 < t_1 < \cdots < t_N = T$ de l'intervalle de temps $(0,T)$ que l'on suppose uniforme en notant $\delta t = t_1 - t_0$ le pas de temps, et un maillage $\mathcal{M}$ de Ω. Le type de schéma utilisé dépend de la régularité du maillage. La classe de schémas construite pour résoudre (1.76) est de type explicite en temps. En utilisant l'expression (1.77) avec $u = \dfrac{\boldsymbol{\nabla}G}{|\boldsymbol{\nabla}G|}$ et $\phi = \boldsymbol{\nabla}G$, on obtient en semi-discret :

$$\frac{G_K^{n+1} - G_K^n}{\delta t} + \mathrm{div}\left(\frac{\boldsymbol{\nabla}_{\mathcal{E}}G^n}{|\boldsymbol{\nabla}_{\mathcal{E}}G^n|}G^n\right)_K - G_K^n\mathrm{div}\left(\frac{\boldsymbol{\nabla}_{\mathcal{E}}G^n}{|\boldsymbol{\nabla}_{\mathcal{E}}G^n|}\right)_K = 0. \tag{1.78}$$

où la divergence discrète est définie d'une façon analogue à (1.13) :

$$\text{pour } K \in \mathcal{M}, \qquad (\mathrm{div}\boldsymbol{u})_K = \frac{1}{|K|}\sum_{\sigma=K|L\in\mathcal{E}(K)} \kappa_{K,\sigma}^{\mathcal{M}}|\sigma|\,\boldsymbol{u}_\sigma.\boldsymbol{n}_{K,\sigma},$$

avec $\kappa_{K,\sigma}^{\mathcal{M}}$ égal à 1 pour un maillage quelconque et à $\frac{|K|}{|D_\sigma|}$ pour un maillage cartésien. De même :

$$(\mathrm{div}\,G\boldsymbol{u})_K = \frac{1}{|K|}\sum_{\sigma=K|L\in\mathcal{E}(K)} \kappa_{K,\sigma}^{\mathcal{M}}|\sigma|\,G_\sigma\boldsymbol{u}_\sigma.\boldsymbol{n}_{K,\sigma}.$$

On utilise une technique de type upwind ou MUSCL pour exprimer l'interpolée de G à la face σ. Afin de prendre en compte les conditions aux limites, pour $\sigma \in \mathcal{E}(K)\cap\mathcal{E}_{\mathrm{ext}}$, on pose $G_\sigma = G_K$ (il est facile de se convaincre que c'est un équivalent discret de $\boldsymbol{\nabla}G\cdot\boldsymbol{n}_{\mathrm{ext}} = 0$). On suppose donc que pour tout $K \in \mathcal{M}$, et pour tout $\sigma \in \mathcal{E}(K)\cap\mathcal{E}_{\mathrm{int}}$, il existe $\beta_{K,\sigma} \in [0,1]$ et un voisin M_σ^K de K tels que :

$$G_\sigma - G_K = \left|\begin{array}{ll} \beta_{K,\sigma}(G_K - G_{M_\sigma^K}) & \text{si } \geq 0, \\ \beta_{K,\sigma}(G_{M_\sigma^K} - G_K) & \text{sinon.} \end{array}\right. \tag{1.79}$$

Il nous reste maintenant à discrétiser le gradient de G sur la face, $(\boldsymbol{\nabla}_{\mathcal{E}}G)_\sigma$, où $\boldsymbol{\nabla}_{\mathcal{E}}$ désigne un gradient discret constant sur chaque maille duale D_σ. En s'inspirant des notations de la section (1.4.7), on appelle $\mathrm{L}_{\mathcal{M}}(\Omega)$ l'espace des fonctions constantes sur chaque maille K du maillage $\mathcal{M}$. On distingue trois cas suivant la régularité du maillage.

Maillage quelconque Quand le maillage ne possède aucune régularité, on définit, pour $\sigma = K|L \in \mathcal{E}_{\text{int}}$ et $G \in \mathrm{L}(\Omega)$:

$$(\nabla_{\mathcal{E}} G)_\sigma = \sum_{\sigma \in \partial(K \cup L)} \frac{|\sigma|}{|K \cup L|} G_\sigma \boldsymbol{n}_{K \cup L, \sigma}. \tag{1.80}$$

Grâce aux hypothèses sur le maillage, il existe un voisinage de mailles V_σ tel que :

$$\exists (\alpha_{K,\sigma})_{K \in V_\sigma}, \quad \boldsymbol{x}_\sigma = \sum_{K \in V_\sigma} \alpha_{K,\sigma} \boldsymbol{x}_K \text{ et } \sum_{K \in V_\sigma} \alpha_{K,\sigma} = 1.$$

On définit alors :

$$\tilde{G}_\sigma = \sum_{K \in V_\sigma} \alpha_{K,\sigma} G_K. \tag{1.81}$$

Il est à noter que cela correspond à l'interpolation de type SUSHI (voir [23]) effectuée lors de la construction des interpolations MUSCL.

Maillage admissible Quand le maillage est admissible, on peut récupérer plus naturellement la composante normale du gradient à la face. On décompose le gradient discret en deux composantes distinctes : une composante normale à la face et une composante colinéaire à la face :

$$\text{pour } \sigma \in \mathcal{E}_{\text{int}}, \qquad (\nabla_{\mathcal{E}} G)_\sigma = \frac{G_L - G_K}{d_\sigma} \boldsymbol{n}_{K,\sigma} + \nabla_{//\sigma} G. \tag{1.82}$$

Pour des raisons de clarté, pour $\sigma = K|L \in \mathcal{E}_{\text{int}}$, on renomme $\nabla_{K \cup L}$ le gradient défini par (1.80). On considère une base orthonormale de la face σ que l'on note $\boldsymbol{n}_{//\sigma}$ dans le cas 2D et $(\boldsymbol{n}^1_{//\sigma}, \boldsymbol{n}^2_{//\sigma})$ dans le cas 3D. On écrit alors :

$$\begin{aligned} \nabla_{//\sigma} G &= \left(\nabla_{K \cup L} G \cdot \boldsymbol{n}_{//\sigma}\right) \boldsymbol{n}_{//\sigma} && \text{(2D)}, \\ \nabla_{//\sigma} G &= \left(\nabla_{K \cup L} G \cdot \boldsymbol{n}^1_{//\sigma}\right) \boldsymbol{n}^1_{//\sigma} + \left(\nabla_{K \cup L} G \cdot \boldsymbol{n}^2_{//\sigma}\right) \boldsymbol{n}^2_{//\sigma} && \text{(3D)}. \end{aligned} \tag{1.83}$$

Maillage cartésien Dans le cadre du maillage cartésien, il est possible de récupérer plus simplement les composantes du gradient suivant chaque direction d'espace. Pour $\sigma = \overrightarrow{K|L}$ (en d'autres termes $G_L - G_K \geq 0$), on définit :

$$\text{pour } \sigma \in \mathcal{E}_{\text{int}}, \qquad (\nabla_{\mathcal{E}} G)_\sigma = \frac{G_L - G_K}{d_\sigma} \boldsymbol{n}_{K,\sigma} + \nabla_{//\sigma} G, \tag{1.84}$$

où $\nabla_{//\sigma}$ est construit comme suit :

$$(\nabla G)^{\mathrm{C}}_{//\sigma} = \sum_{i=1,\ \boldsymbol{e}^{(i)} \cdot \boldsymbol{n}_{K,\sigma} = 0}^{d} \left\{ \frac{(G_{K_i^+} - G_K)^+}{d_{\sigma_i^+}} - \left(1 - \operatorname{sgn}(G_{K_i^+} - G_K)^+\right) \frac{(G_K - G_{K_i^-})^-}{d_{\sigma_i^-}} \right\} \boldsymbol{e}^{(i)}. \tag{1.85}$$

Pour une maille K, on désigne par σ_i^+ et σ_i^- les deux faces de K orthogonales à $\boldsymbol{e}^{(i)}$. Le $-$ et le $+$ désignent la face inférieure et supérieure de la maille K (rangement par ordre croissant de la coordonnée i). On a $\sigma_i^+ = K|K_i^+$ et $\sigma_i^- = K|K_i^-$. On illustre cela en 2D dans la figure suivante (1.3). On rappelle que $a^+ = \max(a, 0)$ et $a^- = \max(-a, 0)$, pour $a \in \mathbb{R}$. Ce choix particulier de la composante colinéaire à la face est importante car elle doit à la fois garantir des propriétés de consistance du gradient discret, ainsi que préserver la monotonie du schéma.

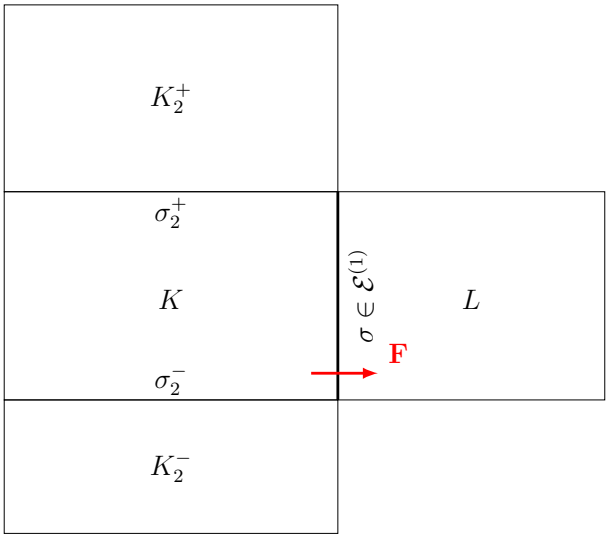


Figure 1.3 – Notations pour la construction du gradient sur maillage cartésien avec $\mathbf{F} = (G_L - G_K)\boldsymbol{n}_{K,\sigma}$.

Enfin la donnée initiale est interpolée naturellement

$$G_K^0 = G_0(\boldsymbol{x}_K), \quad \forall K \in \mathcal{M}. \tag{1.86}$$

Interpolation MUSCL L'interpolation MUSCL est similaire à celle qui a été effectuée pour les équations d'Euler. Seuls les intervalles de limitation changent.

$$\begin{aligned} &(H1) \quad G_\sigma \in |[G_K, G_K + \tfrac{\zeta^+}{2}(G_L - G_K)]| \\ &(H2) \quad \text{il existe } M \in V_K \text{ tel que } G_\sigma \in |[G_K, G_K + \tfrac{\zeta^-}{2}\tfrac{d_\sigma}{d_\epsilon}(G_K - G_M)]|, \end{aligned} \tag{1.87}$$

avec ζ^+ et ζ^- dans l'intervalle $[0, 2]$. Pour le reste on renvoie à la section (1.4.3).

Remarque 1.2 (*Cas cartésien*)

Dans le cadre de grilles cartésiennes on impose $\zeta^+ = \zeta^- = 1$, dans le but d'obtenir des résultats de consistance (voir le théorème 1.10).

Avant de poursuivre plus avant, on va donner une forme génériques du schéma qui sera utile pour formaliser les résultats théoriques qui vont suivre. Pour tout $n \in [0, N-1]$:

$$\eth t G_{\mathcal{M}}^n + F_{\mathcal{M}}(G_{\mathcal{M}}^n) = 0, \tag{1.88}$$

avec,

$$\eth t G_{\mathcal{M}}^n = \sum_{K \in \mathcal{M}} \frac{G_K^{n+1} - G_K^n}{\delta t} \chi_K, \tag{1.89}$$

et

$$F_{\mathcal{M}}(G_{\mathcal{M}}^n) = \sum_{K \in \mathcal{M}} \left[\sum_{\sigma = K|L \in \mathcal{E}(K)} \kappa_{K,\sigma}^{\mathcal{M}} \frac{|\sigma|}{|K|} \frac{\nabla_{\mathcal{E}} G_\sigma^n}{|\nabla_{\mathcal{E}} G_\sigma^n|} \cdot \boldsymbol{n}_{K,\sigma}(G_\sigma^n - G_K^n) \right] \chi_K. \tag{1.90}$$

1.5.3 Propriétés du schéma

Cette section regroupe les propriétés du schéma que nous venons de présenter. Certaines d'entre elles sont fondamentales pour obtenir un résultat de convergence de la solution numérique vers la solution faible visqueuse unique du problème, car ce sont les hypothèse classiques du théorème de Barles-Souganidis [6] que l'on adapte ici.

Stabilité La première propriété est une conséquence directe de la définition de l'opérateur convectif. On la résume dans le lemme suivant :

Lemme 1.7 (*Principe du maximum*)

Soit $G^n_{\mathcal{M}} \in \mathrm{L}_{\mathcal{M}}(\Omega)$, $n \in [0, N]$, la solution du schéma (1.88). Pour tout $K \in \mathcal{M}$ et $n \in [0, N-1]$, on a :

$$\min_{L\in\mathcal{M}} G^n_L \leq G^{n+1}_K \leq \max_{L\in\mathcal{M}} G^n_L,$$

sous la condition de CFL :

$$\delta t \leq \min_{K\in\mathcal{M}} \frac{|K|}{\sum_{\sigma\in\mathcal{E}(K)} |\sigma|}. \tag{1.91}$$

Remarque 1.3 (*Maillage cartésien*)

Dans le cadre d'une discrétisation cartésienne, la condition de CFL est légèrement modifiée :

$$\delta t \leq \min_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} \frac{|D_\sigma|}{|\sigma|}.$$

Invariance par translations Réécrivons le problème de la façon suivante :

$$\forall n \in [0, N-1], \qquad G^{n+1}_{\mathcal{M}} = SCH(G^n_{\mathcal{M}}), \tag{1.92}$$

avec

$$SCH(G^n_{\mathcal{M}}) = G^n_{\mathcal{M}} - \delta t F_{\mathcal{M}}(G^n_{\mathcal{M}}).$$

Le schéma satisfait la propriété suivante :

Lemme 1.8 (*Invariance par translations*)

$\forall \lambda \in \mathbb{R}$, et $\forall \phi^{(m)} \in H_{\mathcal{M}}$,

$$SCH(\phi_{\mathcal{M}} + \lambda) = \lambda + SCH(\phi^{(m)}). \tag{1.93}$$

Consistance Tout d'abord il est nécessaire de définir les interpolations sur le maillage des fonctions test $\phi \in \mathrm{C}^1(\Omega)$:

$$\phi_{\mathcal{M}} = \sum_{K\in\mathcal{M}} \phi_K \mathcal{X}_K, \qquad \text{avec} \quad \phi_K = \phi(\boldsymbol{x}_K). \tag{1.94}$$

On donne ensuite la définition de la consistance, telle qu'elle sera considérée dans cette section. On va voir qu'elle est différente de celle que nous avons abordée dans la partie sur les équations d'Euler.

Définition 1.4

Soit un opérateur $F(G)$ que l'on discrétise par $F_{\mathcal{M}}(G_{\mathcal{M}})$. Soient $h_{\mathcal{M}} = \max_{K\in\mathcal{M}} h_K$ et $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}\right\}$ une suite de discrétisations telle que $h^{(m)}_{\mathcal{M}}$ tend vers zéro quand $m \to \infty$. L'opérateur spatial discret $F_{\mathcal{M}}$ est dit consistant avec F si pour tout $\phi \in \mathrm{C}^1(\Omega)$:

$$\lim_{m\to\infty} \|F_{\mathcal{M}^{(m)}}(\phi_{\mathcal{M}^{(m)}}) - F(\phi)\|_{\mathrm{L}^\infty(\Omega)} = 0.$$

On a alors les propriétés suivantes :

Lemme 1.9 (*Consistance du gradient discret*)

Pour $\phi \in \mathrm{L}_{\mathcal{M}}(\Omega)$, le gradient $\nabla_{\mathcal{E}}$ défini par :

$$\nabla_{\mathcal{E}}\phi = \sum_{\sigma\in\mathcal{E}} (\nabla_{\mathcal{E}}\phi)_\sigma \mathcal{X}_{D_\sigma},$$

est consistant.

Il existe un résultat encore plus fort, mais il ne peut être obtenu que sur maillage cartésien.

Theorème 1.10 (*Consistance du schéma sur un maillage cartésien*)

L'opérateur spatial sur maillage cartésien défini pour $G_{\mathcal{M}} \in \mathrm{L}_{\mathcal{M}}(\Omega)$ par :

$$F_{\mathcal{M}}(G_{\mathcal{M}}) = \sum_{K\in\mathcal{M}} \left[\sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{|\sigma|}{d|D_\sigma|} \frac{(G_L - G_K)}{\sqrt{(G_L - G_K)^2 + d_\sigma^2 |\nabla_{//\sigma} G_{\mathcal{M}}|^2}} (G_\sigma - G_K) \right] \mathcal{X}_K, \tag{1.95}$$

est consistant avec $|\nabla G|$.

Il n'est malheureusement pas possible de prouver ce même résultat pour le schéma sur un maillage non cartésien. On aboutit seulement à une forme moins forte de la consistance qui pourrait être vue comme une consistance faible et qu'on présentera en Annexe.

Monotonie Cette dernière propriété est très forte mais aussi très restrictive. Elle s'applique uniquement au cas cartésien avec une interpolation de G aux faces de type upwind.

Soient $(\phi^{(m)}, \psi^{(m)}) \in \mathrm{L}_{\mathcal{M}^{(m)}}(\Omega)$, on définit l'ordre partiel suivant :

$$\phi^{(m)} \leq \psi^{(m)} \qquad \Longleftrightarrow \qquad \forall K \in \mathcal{M}, \quad \phi_K \leq \psi_K. \tag{1.96}$$

Le schéma cartésien avec interpolation upwind satisfait la propriété suivante :

Lemme 1.11 (*Monotonie*)

Supposons que la condition de cfl suivante est satisfaite,

$$\delta t \leq \frac{1}{\sum_{\sigma\in\mathcal{E}(K)} \frac{1+\frac{1}{2}\sqrt{1+r^2}}{d_\sigma}}, \qquad r = \max_{\sigma,\sigma'\in\mathcal{E}(K)} \frac{d_\sigma}{d_{\sigma'}}. \tag{1.97}$$

Alors on a le résultat suivant :

$$\forall (\phi_{\mathcal{M}}, \psi_{\mathcal{M}}) \in H_{\mathcal{M}}, \qquad \phi_{\mathcal{M}} \leq \psi_{\mathcal{M}} \quad \Longrightarrow \quad F_{\mathcal{M}}(\phi_{\mathcal{M}}) \leq F_{\mathcal{M}}(\psi_{\mathcal{M}}).$$

Remarque 1.4 (*Formulation discrète faible*)

La combinaison des propriétés d'invariance par translation et de monotonie permet d'obtenir une version discrète de (1.74) (et (1.75) respectivement). En effet, soit $G^{(T)}_{\mathcal{M}^{(m)}} = \sum_{n=0}^{N-1} G^{n+1}_{\mathcal{M}^{(m)}} \mathcal{X}_{[t^n,t^{n+1}]}$ la solution du schéma

$$G^{n+1}_{\mathcal{M}} = SCH(G^n_{\mathcal{M}}), \quad \forall n \in [0, N-1],$$

que l'on suppose monotone et invariant par translations. Soit $L_{\mathcal{M}}(\Omega \times [0,T))$ l'espace des fonctions constantes sur chaque $K \times [t^n, t^{n+1})$ sur $\mathcal{M} \times [0,T)$. Soit $\varphi \in L_{\mathcal{M}}(\Omega \times [0,T))$ telle que $G^{(T)}_{\mathcal{M}^{(m)}} - \varphi$ admette un maximum local en (K_0, t^{n_0}). On a alors, $\forall K \in \mathcal{M}$:

$$G^{n_0}_{K_0} - \varphi^{n_0}_{K_0} \geq G^{n_0-1}_K - \varphi^{n_0-1}_K,$$

c'est-à-dire :

$$G_K^{n_0-1} \leq G_{K_0}^{n_0} - \varphi_{K_0}^{n_0} + \varphi_K^{n_0-1}.$$

Grâce à la monotonie du schéma on obtient :

$$SCH(G_K^{n_0-1}) \leq SCH(G_{K_0}^{n_0} - \varphi_{K_0}^{n_0} + \varphi_K^{n_0-1}).$$

Or on a $SCH(G_K^{n_0-1}) = G_K^{n_0}$ et $G_{K_0}^{n_0} - \varphi_{K_0}^{n_0}$ constante donc l'invariance par translation nous permet d'écrire :

$$G_K^{n_0} \leq SCH(\varphi_K^{n_0-1}) + G_{K_0}^{n_0} - \varphi_{K_0}^{n_0}.$$

En prenant la composante K_0 on aboutit à :

$$\varphi_{K_0}^{n_0} \leq SCH(\varphi_K^{n_0-1})_{K_0}.$$

En d'autres termes :

$$\eth t \varphi_{K_0}^{n_0-1} + F_{\mathcal{M}}(\varphi^{n_0-1})_{K_0} \leq 0,$$

que l'on peut voir comme un équivalent discret de (1.74). On obtient un résultat analogue pour le minimum local.

On est maintenant en mesure de donner un résultat de convergence pour le schéma upwind cartésien.

1.5.4 Théorème de convergence

Le théorème que l'on va donner ci-dessous est un résultat démontré initialement par P.-L. Lions et G.Crandall pour les schémas de type différences finies. Un cadre plus général a ensuite été donné par G. Barles et P. Souganidis. Nous nous inspirons ici de ces travaux pour donner une version adaptée aux notations de cette section.

Theorème 1.12 (*Théorème de convergence*)

Soit $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}, \delta t^{(m)}\right\}$ *une suite de discrétisations telle que le pas de temps et d'espace tendent vers zéro quand* $m \to \infty$. *Soit* $\bar{G}$ *la solution de viscosité de* (1.71). *On considère le schéma explicite suivant, pour* $n \in [0, N-1]$ *:*

$$\eth t G_{\mathcal{M}}^n + F_{\mathcal{M}}(G_{\mathcal{M}}^n) = 0,$$

et la solution discrète associée $G_m^{(T)} = \sum_{n=0}^{N-1} G_{\mathcal{M}^{(m)}}^{n+1} \mathcal{X}_{[t^n, t^{n+1}]}$. *On suppose que :*

- *l'opérateur spatial* $F_{\mathcal{M}}$ *est consistant, dans le sens* (1.4), *avec l'opérateur* $G \longmapsto |\nabla G|$,
- *le schéma est invariant par translations :* $F_{\mathcal{M}}(G_{\mathcal{M}} + v) = F_{\mathcal{M}}(G_{\mathcal{M}})$,
- *le schéma est monotone.*

Alors,

$$G_m^{(T)} \longrightarrow \bar{G} \text{ uniformément, en espace et temps, quand } m \to \infty.$$

On a alors le corollaire suivant qui nous permet d'appliquer ce résultat au schéma présenté ici, grâce aux propriétés vérifiées au dessus.

Corollaire 1.13

Supposons qu'il existe $r > 0$, *tel que* $\forall m \in \mathbb{N}$, $\forall \sigma, \sigma' \in \mathcal{M}^{(m)}$,

$$\frac{d_\sigma}{d_{\sigma'}} \leq r.$$

Soit $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}, \delta t^{(m)}\right\}$ *une suite de discrétisations telle que le pas de temps et*

d'espace tendent vers zéro quand $m \to \infty$. Supposons que pour tout $m \in \mathbb{N}$,

$$\delta t^{(m)} \leq \max_{K \in \mathcal{M}^{(m)}} \frac{1}{\sum_{\sigma \in \mathcal{E}(K)} \frac{1+\frac{1}{2}\sqrt{1+r^2}}{d_\sigma}}.$$

Alors la solution du schéma cartésien upwind (1.88)–(1.95) $G_m^{(T)}$ converge uniformément, en espace et temps, vers $\bar{G}$.

Remarque 1.5 (*Estimation de la vitesse de convergence*)

Nous ne donnons pas de vitesse de convergence, puisqu'elle va dépendre de la régularité de la donnée initiale et de l'erreur de consistance au niveau du schéma. Par exemple, en supposant que la donnée initiale est Lipschitzienne, on obtient, avec le schéma upwind, une vitesse de convergence en $\sqrt{h_{\mathcal{M}}}$.

1.6 Résultats numériques

Nous allons donner un aperçu des résultats numériques obtenus pour les différents schémas présentés. Des résultats plus complets seront présentés dans les différents chapitres de cette thèse et en Annexe.

1.6.1 Équations d'Euler

Les résultats présentés ici ont pour but d'illustrer l'efficacité des interpolations d'ordre élevées de type MUSCL effectuées sur ρ et e ainsi que l'utilité de l'ajout de viscosité artificielle pour atténuer l'apparition d'instabilités au niveau des chocs. On commence par présenter des résultats simples en une dimension avant de donner quelques résultats classiques en deux dimensions.

Résultats 1D Nous présentons un cas test supersonique issu de la littérature : le problème de Riemann 1D n° 5 tiré de [62, Chapter 4]. Le problème est défini sur $\Omega = (0, 1)$. Les conditions initiales consistent en deux états constants :

$$\text{état gauche : } \begin{bmatrix} \rho_L = 5.99924 \\ u_L = 19.5975 \\ p_L = 460.894 \end{bmatrix}; \qquad \text{état droit : } \begin{bmatrix} \rho_R = 5.99242 \\ u_R = -6.19633 \\ p_R = 46.0950 \end{bmatrix}.$$

Le temps final de la simulation est $T = 0.035s$. Le pas de discrétisation spatiale est $h = 0.001$ et celui temporel est $\delta t = h/60$. La vitesse du son étant proche de 10, la vitesse des ondes les plus rapides est donc égale à $u + c \approx 30$ ce qui correspond à une CFL acoustique d'environ 0.5. Les résultats obtenus avec une discrétisation de type upwind et une discrétisation de type MUSCL avec ajout de viscosité artificielle sont présentés dans la figure ci-dessous (1.4).

On effectue le calcul avec deux versions du schéma. Une version découplée Upwind, et une version de type MUSCL plus ajout de viscosité WLR et de coefficient $c_m = 2$. Avec l'interpolation upwind, le front est très diffusé au niveau de la discontinuité de contact, alors qu'il est bien plus raide avec l'interpolation MUSCL. On remarque aussi la présence d'un overshoot au niveau du choc dans le cas upwind. A noter qu'un raffinement du maillage ne permet pas de faire disparaître cet overshoot bien qu'il soit borné dans L^∞. Ce phénomène est encore plus prononcé avec une diffusion moindre comme c'est le cas du MUSCL, d'où l'utilité de l'ajout d'une viscosité artificielle au choc uniquement. Afin de quantifier le gain en précision, on effectue une analyse d'erreur sur un cas test similaire : le cas test numéro 3 de Toro [62, Chapter

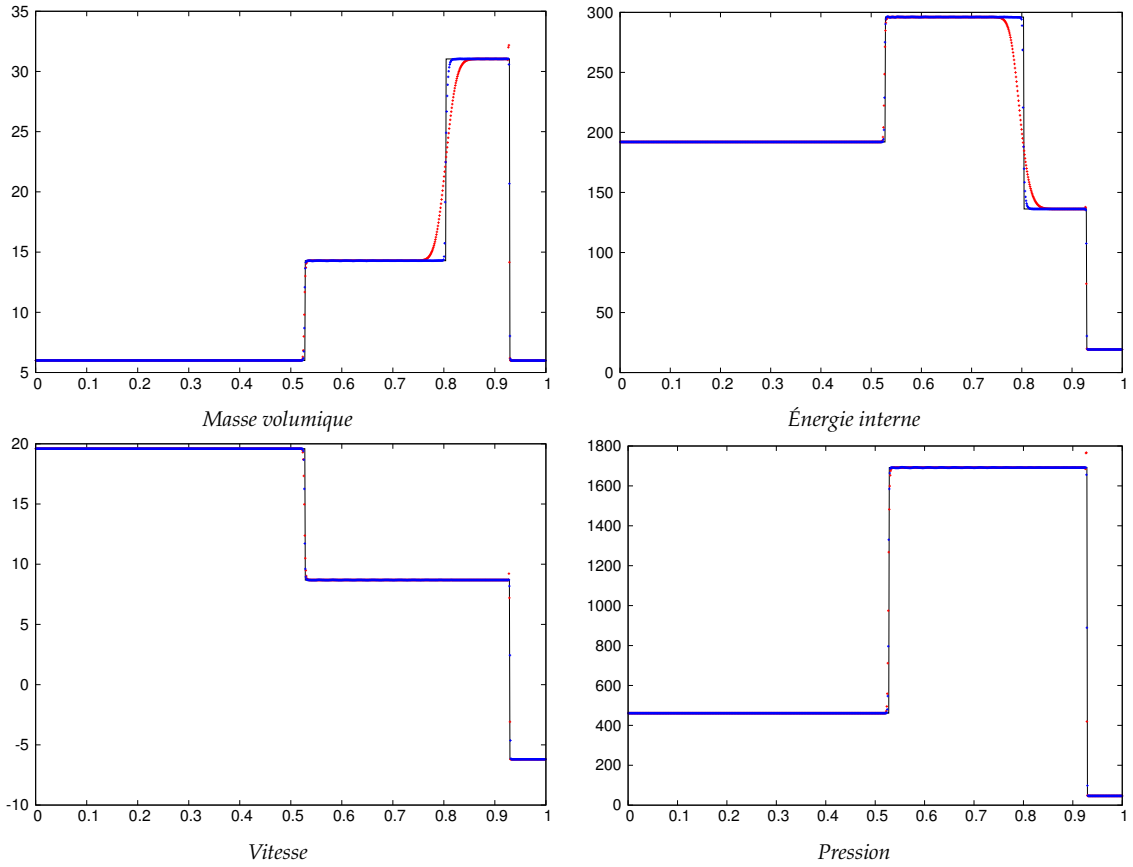


Figure 1.4 – Problème de Riemann (Test 5 de [62, Chapter 4]) – En rouge : Upwind ; En bleu : MUSCL + viscosité WLR – $h = 0.001$ et $\delta t = h/90$ – Résultats à $t = 0.035s$.

4]. Les données initiales sont les suivantes :

$$\text{état gauche : } \begin{bmatrix} \rho_L = 1 \\ u_L = 0 \\ p_L = 1000 \end{bmatrix}; \qquad \text{état droit : } \begin{bmatrix} \rho_R = 1 \\ u_R = 0 \\ p_R = 0.001 \end{bmatrix}.$$

L'analyse d'erreur en norme $L^1(\Omega)$ est effectuée à $t = 0.012s$ et les résultats sont reportés dans le tableau suivant.

		$h_0 = 0.001$	$h_0/2$	$h_0/4$	$h_0/8$	$h_0/16$
$\|\rho - \bar{\rho}\|_{L^1(\Omega)}$	MUSCL	0.0108	0.0058	0.0025	0.0012	0.0007
	UPWIND	0.0651	0.0455	0.0310	0.0217	0.0153
$\|p - \bar{p}\|_{L^1(\Omega)}$	MUSCL	1.2827	0.6734	0.3316	0.1800	0.1044
	UPWIND	1.87	1.05	0.530	0.284	0.164

On voit que la vitesse de convergence est améliorée par l'interpolation MUSCL. La vitesse de convergence des schémas numériques pour les équations d'Euler (et plus globalement pour les systèmes hyperboliques) étant toujours limité par les discontinuités de contact, on observe le gain en vitesse par l'intermédiaire des variables discontinues aux contacts (ici ρ). Le passage upwind/MUSCL approche le schéma de l'ordre deux (qui correspond à une vitesse de convergence de 2/3).

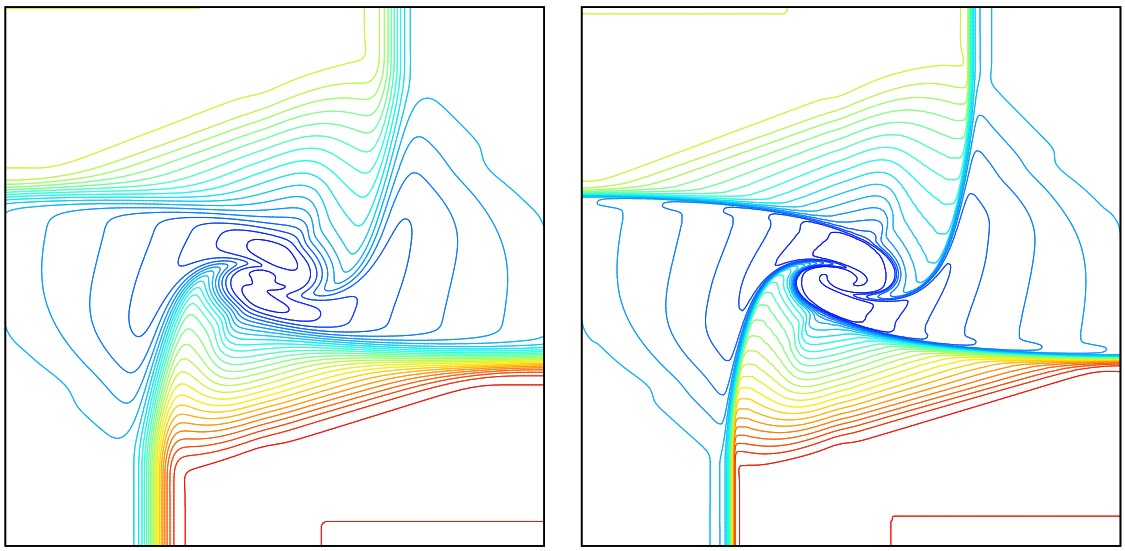
Résultats 2D Discontinuités de contact – Nous présentons tout d'abord un des problèmes de Riemann tiré de [44]. Il est défini sur $\Omega = (-0.5, 0.5)^2$, les conditions initiales consistent

en quatre états constants dans 4 quadrants qui composent le domaine : $\Omega_1 = (0, 0.5)^2$, $\Omega_2 = (-0.5, 0) \times (0, 0.5)$, $\Omega_3 = (-0.5, 0)^2$, $\Omega_4 = (0, 0.5) \times (-0.5, 0)$. Ils sont choisis de telle sorte qu'il ne puisse exister qu'une seule onde au niveau de chaque interface entre les quadrants (les autres ondes fondamentales sont d'amplitude nulle). On aboutit alors à 19 configurations possibles. Etant donné qu'on souhaite mettre en lumière le gain en précision du schéma MUSCL on choisit une configuration qui comporte des discontinuités de contact. On présente donc le cas test numéro 5 qui correspond aux états constants initiaux :

$$\Omega_1 : \begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0.75 \\ v_1 = -0.5 \end{bmatrix} \qquad \Omega_2 : \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = 0.75 \\ v_2 = 0.5 \end{bmatrix}$$

$$\Omega_3 : \begin{bmatrix} \rho_3 = 1 \\ p_3 = 1 \\ u_3 = -0.75 \\ v_3 = 0.5 \end{bmatrix} \qquad \Omega_4 : \begin{bmatrix} \rho_4 = 3 \\ p_4 = 1 \\ u_4 = -0.75 \\ v_4 = -0.5 \end{bmatrix}.$$

Le temps final du calcul est $T = 0.3s$. Les résultats sont obtenus en utilisant un maillage 400×400 avec le schéma sur maillage MAC, et un pas de temps $\delta t = \frac{1}{5 \times 400}$. Ils sont tracés sur la figure (1.5).



Upwind *MUSCL*

Figure 1.5 – Problème de Riemann – (Test 6 de [44]) – comparaison entre les schémas Upwind et MUSCL – $h = 0.0025$ et $\delta t = h/10$ – Isocontours (100 valeurs) de la masse volumique à $t = 0.3s$.

Avec une CFL égale à 1/5, on aboutit à une CFL acoustique de 0.5 pour ce calcul. Les résultats confirment les analyses effectuées en 1D. Les discontinuités de contact sont plus fines avec l'interpolation MUSCL.

Écoulement autour d'un cylindre à Mach 10 – Afin de montrer les résultats d'un calcul avec le schéma de type RT-CR, on prend une configuration géométrique qui nécessite un maillage non cartésien : un écoulement autour d'un cylindre. Il s'agit d'une adaptation pour les équations d'Euler d'un cas test d'un benchmark pour les écoulements de type faiblement compressible tiré de [55]. La géométrie du problème est représentée sur la figure (1.6).

Le fluide entre à gauche du domaine avec une vitesse constante égale à $\boldsymbol{u} = \left(1, 0\right)^t$. Afin d'obtenir un écoulement à Mach 10, on prend une vitesse du son égale à $c = (\gamma p/\rho)^{1/2} = 0.1m/s$.

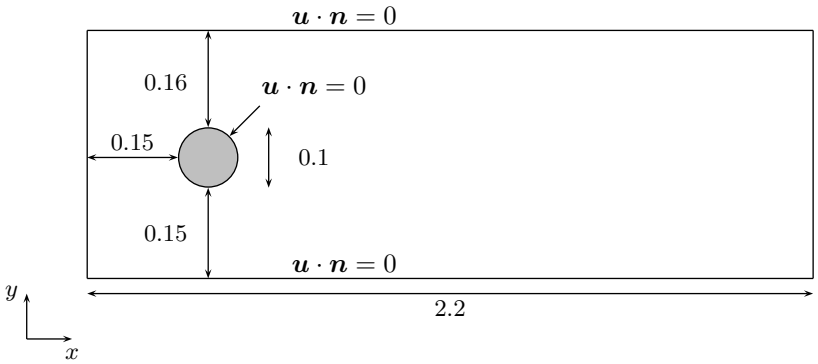


Figure 1.6 – Geométrie du cas test de l'écoulement autour d'un cylindre.

On a donc :

$$\begin{bmatrix} \rho \\ p \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1/140 \end{bmatrix}.$$

Les conditions initiales sont identiques aux données d'entrée sur la frontière gauche du domaine. On laisse l'écoulement sortir librement à la frontière droite. Quant au cylindre, on impose une condition de type glissement parfait sur sa surface, ainsi que sur les frontières supérieures et inférieures.

Une version grossière du maillage utilisé est présentée dans la figure (1.7). Les versions raffinées de ce maillage sont obtenues en réduisant le pas d'espace au niveau des lignes caractéristiques de la géométrie (les frontières et les cercles autour du cylindre).

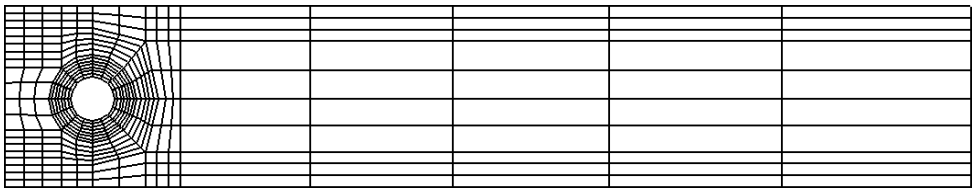


Figure 1.7 – Version grossière du maillage RT pour le cas test cylindre.

Le temps final du calcul est fixé à $T = 5s$. On impose une viscosité additionnelle égale à $\mu = 0.05$ ce qui correspond environ à un dixième de la viscosité upwind. On impose un pas de temps de $\delta t = 10^{-4}$ et un pas d'espace de 10^{-3} (environ 5.3110^5 mailles). Par conséquent, la CFL obtenue pour notre calcul est proche de 0.1, sachant que la vitesse de la plus rapide des ondes acoustiques est égale à $1.1m/s$.

La figure 1.8 présente les résultats obtenus au temps final. On observe un choc fort à l'avant du cylindre. Il se réfléchit sur les parois haute et basse du domaine, en produisant une succession de chocs faibles conduisant à une structure en X pour les champs de pression et de masse volumique. Ils finissent par être dissipés par la diffusion numérique.

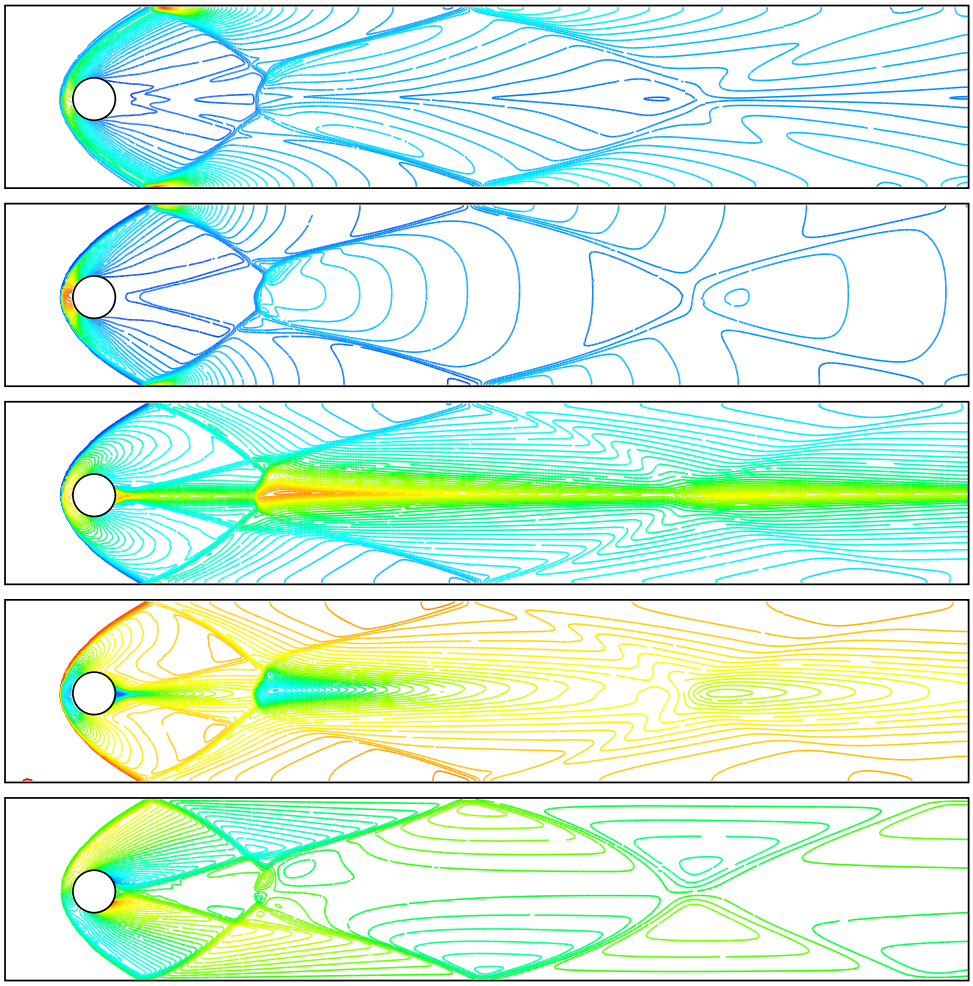


Figure 1.8 – Ecoulement à Mach 10 contre un cylindre – De haut en bas : énergie interne, masse volumique, composante-x de la vitesse, composante-y de la vitesse à $t = 5s$. Les plages de variations des inconnues sont $e \in [0.178,\ 0.536]$, $\rho \in [0.804,\ 12.23]$, $u_1 \in [-0.11,\ 1]$, $u_2 \in [-0.326,\ 0.327]$.

1.6.2 G-équation

On va maintenant donner quelques résultats numériques pour illustrer l'efficacité du schéma pour la propagation du front de flamme.

Résultats 1D – Le problème est posé sur un domaine $\Omega = (0,1)$. Les conditions aux bords sont de type Neumann homogène. Initialement G est défini de la façon suivante :

$$G_0(x) = |\sin(4\pi x)|. \tag{1.98}$$

Il est possible dans le cas 1D de calculer exactement les solutions de viscosité. Dans ce cas, pour $T < \frac{1}{8}s$, la solution s'écrit :

$$G_{\text{visc}}(x,t) = \begin{cases} 0, & \forall x \in [0,T] \cup [\frac{1}{4}-T, \frac{1}{4}+T] \cup [\frac{1}{2}-T, \frac{1}{2}+T] \cup [\frac{3}{4}-T, \frac{3}{4}+T] \cup [1-T, 1], \\ |\sin(4\pi(x-T))|, & \forall x \in [T, \frac{1}{8}] \cup [\frac{1}{4}+T, \frac{3}{8}] \cup [\frac{1}{2}+T, \frac{5}{8}] \cup [\frac{3}{4}+T, \frac{7}{8}], \\ |\sin(4\pi(x+T))|, & \forall x \in [\frac{1}{8}, \frac{1}{4}-T] \cup [\frac{3}{8}, \frac{1}{2}-T] \cup [\frac{5}{8}, \frac{3}{4}-T] \cup [\frac{7}{8}, 1-T]. \end{cases}$$

Plus de détails seront donnés en Annexe. Les calculs présentés ont été réalisés avec le schéma upwind. Le temps final est $T = 0.05s$. Par soucis de simplicité on choisit un pas d'espace et de temps constant. Le pas d'espace est fixé à $h = \dfrac{1}{400}$ et le pas de temps est calculé de manière à obtenir une CFL égale à $\dfrac{\delta t}{h} = 0.1$.

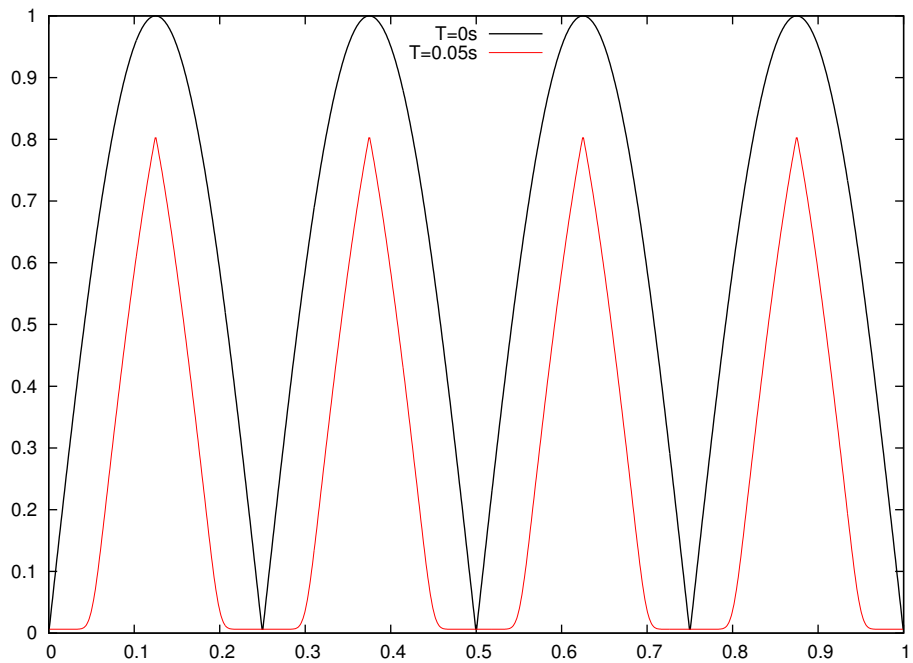


Figure 1.9 – Solution de l'équation eikonale pour $T = 0.05s$ avec G_0 donné par (1.98)

Une analyse d'erreur a été effectuée afin de souligner numériquement le théorème de convergence valable en 1D. Les résultats sont donnés dans la figure ci-dessous.

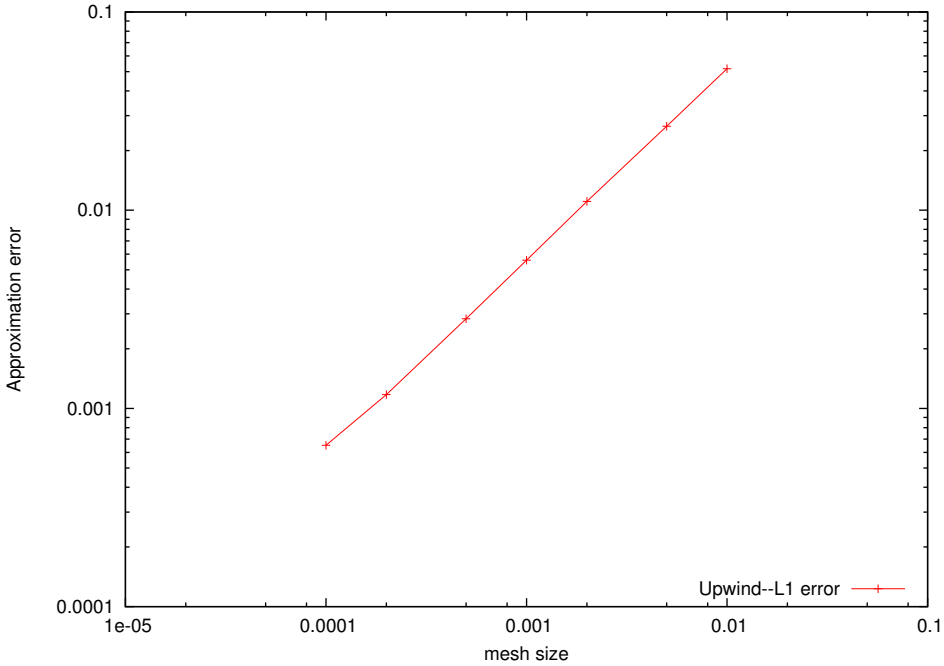


Figure 1.10 – Erreur L^1 à $T = 0.05s$ et une cfl de $\frac{1}{10}$ – Schéma Upwind.

Résultats 2D – L'objectif de cette section est d'effectuer une analyse d'erreur et un comparatif des vitesses de convergence upwind/MUSCL. On s'inspire du cas test 1D introduit précedemment. Le problème est posé sur $\Omega = [-\frac{1}{2}, \frac{1}{2}]^2$. Les conditions aux bords sont de type Neumann homogène. La condition initiale en coordonnées polaires s'écrit :

$$G_0(r, \theta) = |\sin(4\pi r)|. \tag{1.99}$$

La solution analytique de viscosité est donnée par :

$$G_{\text{visc}}(r, \theta, T) = \begin{cases} 0, & \forall r \in [0, T] \cup [\frac{1}{4} - T, \frac{1}{4} + T] \cup [\frac{1}{2} - T, \frac{1}{2} + T] \cup [\frac{3}{4} - T, \frac{3}{4} + T] \cup [1 - T, 1], \\ |\sin(4\pi(r - T))|, & \forall r \in [T, \frac{1}{8}] \cup [\frac{1}{4} + T, \frac{3}{8}] \cup [\frac{1}{2} + T, \frac{5}{8}] \cup [\frac{3}{4} + T, \frac{7}{8}], \\ |\sin(4\pi(r + T))|, & \forall r \in [\frac{1}{8}, \frac{1}{4} - T] \cup [\frac{3}{8}, \frac{1}{2} - T] \cup [\frac{5}{8}, \frac{3}{4} - T] \cup [\frac{7}{8}, 1 - T]. \end{cases}$$

Le temps final est $T = 0.04s$, le pas d'espace égal à $h = \frac{1}{400}$ et le pas de temps est calculé afin d'obtenir une condition de CFL de 0.1. Trois maillages non cartésiens différents sont utilisés :

- un maillage triangulaire obtenu en coupant un maillage cartésien uniforme suivant les diagonales des cellules,
- un maillage composé de parallélogrammes uniformes de grand angle $\frac{2\pi}{3}$,
- un maillage quelconque qui est une déformation d'un maillage cartésien par le principe suivant : soit $\epsilon \in [0, 1]$, chaque sommet interne est déplacé d'une distance ϵh suivant une direction aléatoire. Plus ϵ est important, plus le maillage est déformé. Dans notre cas on fixe $\epsilon = 0.35$. Un exemple de ce type de maillage obtenu à partir d'une grille cartésienne 10×10 est donné dans la figure suivante.

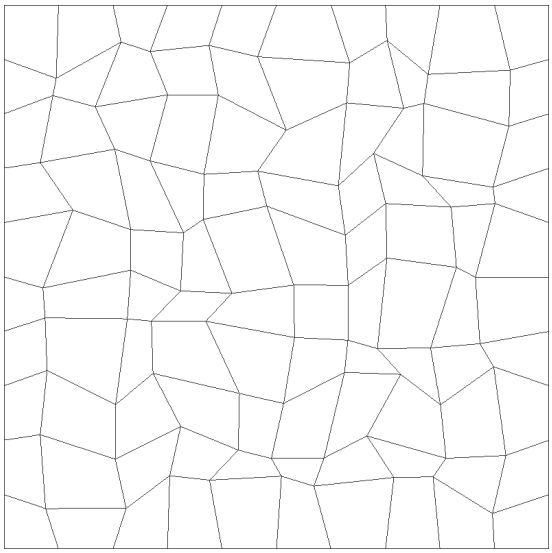


Figure 1.11 – exemple de grille 10×10 non structurée

L'ensemble des résultats est reporté sur la figure suivante.

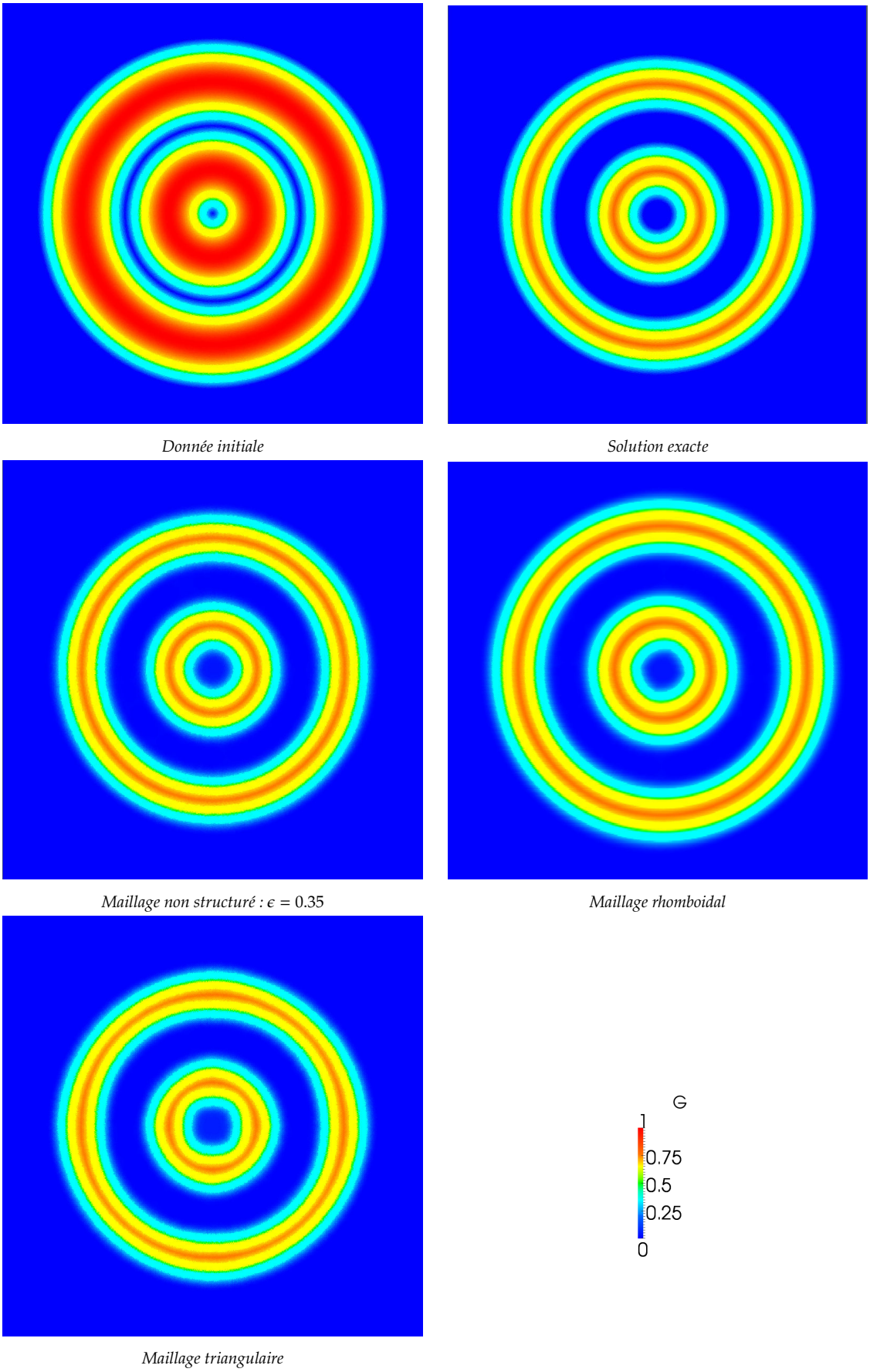


Figure 1.12 – G pour différents maillages avec le schéma Upwind –$T = 0.04$ – $h = \dfrac{1}{400}$ – $c.f.l. = \dfrac{1}{10}$

On illustre ces résultats avec une analyse de convergence pour chacun de ces maillages en prenant $G_{\text{visc}}(r, \theta, T = 0.01)$ comme donnée initiale (Figure 1.13).

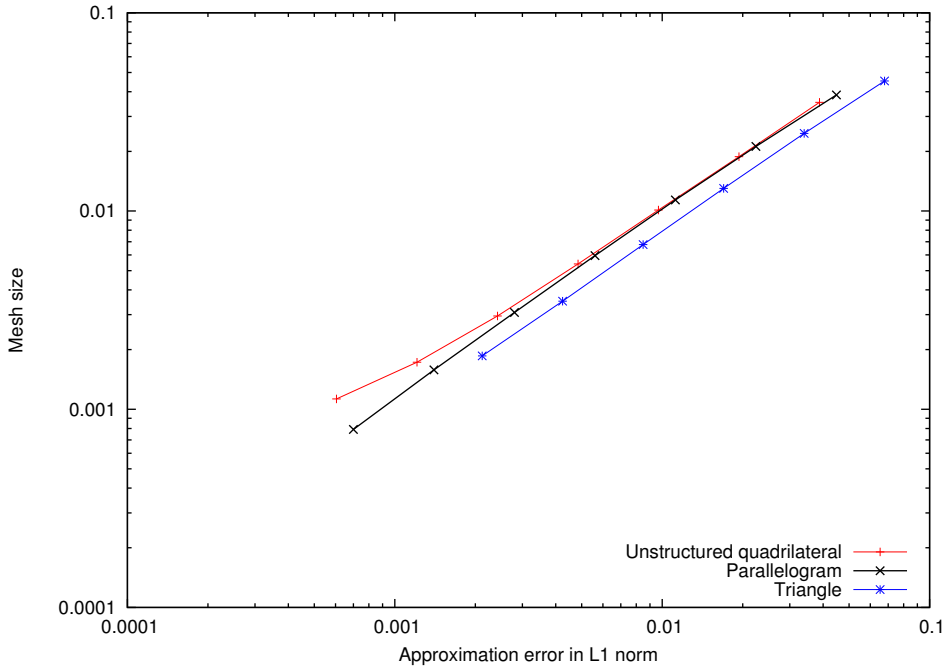


Figure 1.13 – Erreur L^1 pour $T = 0.04$ avec une cfl de 0.1 et le schéma Upwind.

Enfin on compare les vitesses de convergence pour le schéma upwind, le schéma MUSCL avec une discrétisation temporelle RK2 et enfin le schéma différences finies introduit par Lions dans [20]. On obtient les résultats suivants, sur maillage cartésien :

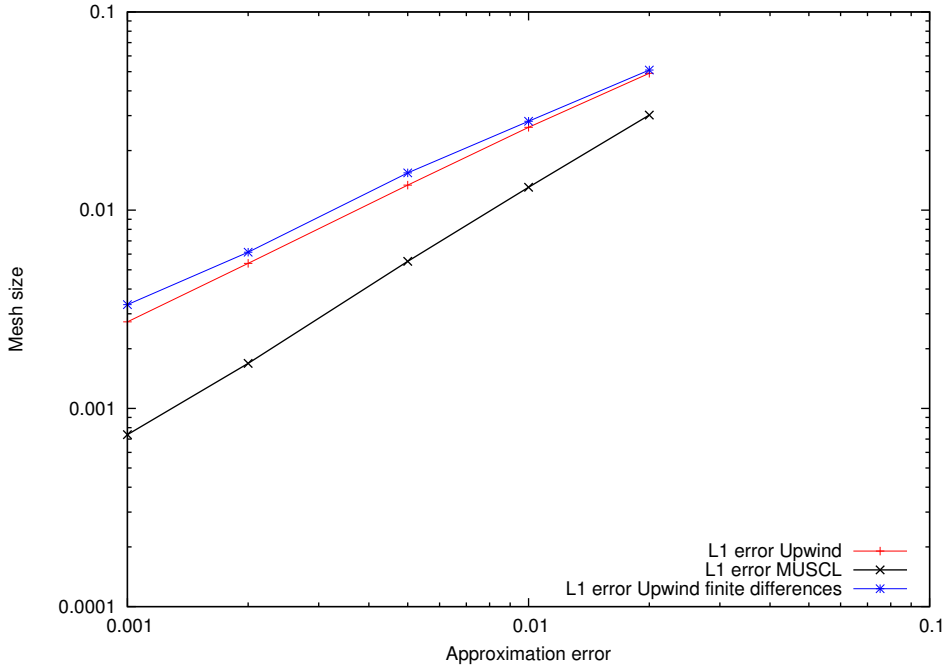


Figure 1.14 – Erreur L^1 pour $T = 0.04$ avec une cfl de 0.1

1.7 Conclusion

Cette thèse contribue à la résolution numérique du phénomène d'explosion en s'articulant autour de deux axes :

— Le premier axe, prépondérant dans cette thèse, consiste à développer une classe de schémas sur mailles décalées pour les équations d'Euler, afin de modéliser le phénomène d'ondes de chocs consécutives à une explosion. Les travaux réalisés se découpent en deux parties :
  - Tout d'abord l'extension à l'ordre élevé des schémas découplés développés dans [39], par des techniques de type MUSCL. Ces interpolations sont effectuées sur les variables discontinues sur les contacts, et permettent ainsi d'améliorer la convergence des schémas. Une corrélation judicieuse de ces interpolations permet en outre de conserver les contacts en 1D. Afin de pallier les éventuels manque de dissipation numérique au niveau des chocs, on utilise un modèle de viscosité selective dans le bilan de quantité de mouvement afin de réduire les oscillations et overshoots parasites sans dégrader la solution dans les zones de régularité. Les propriétés initiales de ces schémas (en upwind), *i.e.* le bilan d'énergie cinétique discret, la conservation de l'énergie totale ainsi que la positivité de l'énergie interne et la masse volumique sont préservées.
  - Ensuite une extension des résultats théoriques pour l'ensemble des schémas développés à l'IRSN pour les équations d'Euler, *i.e.* en incluant les schémas de type correction de pression aux schémas étudiés en première partie. Pour cette classe de schéma, un bilan d'entropie discrèt est obtenu, et un résultat de consistance de Lax (toute suite convergente de solutions des schémas converge vers une solution faible des équations du modèle continu) est obtenu. Enfin on s'assure que les solutions vérifient à la limite une inégalité d'entropie.

— La propagation du front de flamme, modélisée par une équation de type Hamilton-Jacobi appelée la G-équation, est résolue grâce à une classe de schémas qui s'appuie sur les discrétisations existantes pour résoudre les équations du fluide. Il est possible de transformer cette G-équation en une équation de transport suivant la normale du gradient de l'indicatrice de flamme. On est alors capable d'utiliser les techniques de discrétisations des opérateurs de convection précédentes. L'avantage principal est de pouvoir utiliser ces schémas sur des maillages non cartésiens de façon simple, alors que la majorité des développements de schémas pour Hamilton-Jacobi s'appuie sur des différences finies. On obtient un schéma qui vérifie un principe du maximum discret. De plus on construit une adaptation de ces schémas pour les mailles cartésiennes qui possède des propriétés de consistance et de monotonie. Ces propriétés permettent de prouver la convergence des solutions du schéma vers la solution de viscosité du problème continu. Des calculs viennent confirmer ce résultat et montrent la bonne convergence numérique sur maillage non cartésien.

Les schémas développés pour les équations d'Euler sont naturellement applicables aux équations de Saint-Venant et une analyse théorique similaire peut être effectuée. Une version découplée de ces schémas pour les équations de Navier-Stokes est en cours d'implémentation dans les codes de calcul de l'IRSN. Elle permettra à terme l'utilisation de modèles de turbulence de type LES pour la propagation de la flamme. En outre, une extension de ces schémas pour les équations d'Euler multi-espèces non réactives dans un premier temps, puis réactives dans un second est aussi en cours. Enfin une application de ces schémas à mailles décalées pour les équations de Baer-Nunziato est étudiée dans le cadre de la thèse de Sophie Dallet et a fait l'object d'un travail conjoint avec R. Abgrall dans [2].

Chapitre 2

MUSCL-type stable explicit staggered schemes for the compressible Euler equations

2.1 Introduction

The main objective of this paper is to develop and test a numerical scheme for the simulation of non viscous compressible flows modeled by the full Euler equations for an ideal gas :

$$\partial_t \rho + \mathrm{div}(\rho\, \boldsymbol{u}) = 0, \tag{2.1a}$$

$$\partial_t (\rho\, \boldsymbol{u}) + \mathrm{div}(\rho\, \boldsymbol{u} \otimes \boldsymbol{u}) + \nabla p = 0, \tag{2.1b}$$

$$\partial_t (\rho\, E) + \mathrm{div}(\rho\, E\, \boldsymbol{u}) + \mathrm{div}(p\, \boldsymbol{u}) = 0, \tag{2.1c}$$

$$p = (\gamma - 1)\, \rho\, e, \qquad E = \frac{1}{2}|\boldsymbol{u}|^2 + e, \tag{2.1d}$$

where t stands for the time, ρ, $\boldsymbol{u}$, p, E and e are the density, velocity, pressure, total energy and internal energy respectively, and $\gamma > 1$ is a coefficient specific to the considered fluid. The problem is supposed to be posed over $\Omega \times (0, T)$, where Ω is an open bounded connected subset of $\mathbb{R}^d$, $1 \leq d \leq 3$, and $(0, T)$ is a finite time interval. System (2.1) is complemented by initial conditions for ρ, e and $\boldsymbol{u}$, denoted by ρ_0, e_0 and $\boldsymbol{u}_0$ respectively, with $\rho_0 > 0$ and $e_0 > 0$, and by a boundary condition which we suppose to be $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ at any time and *a.e.* on $\partial\Omega$, where $\boldsymbol{n}$ stands for the normal vector to the boundary.

Let us list here the essential features of the proposed numerical scheme :

- First, we use a staggered arrangement of the unknowns, on general simplicial or quadrangular/hexahedral meshes : the so-called scalar variables (density, pressure and thus, to allow a straightforward formulation of the equation of state, the internal energy) are approximated by piecewise constant functions on the cells while the velocity is approximated at the faces of the cells.
- Second, the energy equation solved by the scheme is the internal energy balance (see Equations (2.3)-(2.4) below), which presents two advantages : first of all, it allows to preserve, by construction of the scheme, the positivity of the internal energy ; in addition, it avoids to build an approximation of the total energy which, for staggered discretizations, is a "composite" variable, in the sense that it combines quantities discretized on the cells and at the faces. Note that a blunt discretization of the internal energy balance is known to yield uncorrect shock solutions ; a corrective term is added here to circumvent this problem.
- Third, the positivity of the scheme (in the sense that it keeps positive the density and internal energy, and thus the pressure) is obtained by a very simple way, namely by building for the

mass and internal energy balance a positivity-preserving convection operator. For a first-order scheme, it would amount to use an upwinding with respect to the material velocity only ; here, we rather develop a MUSCL-like procedure, which consists in computing a (formally) second-order in space fluxes and then applying a limitation procedure to obtain positivity under a CFL-like condition, since we use a explicit time discretization. This limitation step is purely algebraic : it does not require any geometric argument and thus works on quite general meshes. It is carefully designed to keep the pressure constant in the zones where it actually should be, and in particular across contact discontinuities. Such a scheme is often referred to in the literature as a "flux splitting scheme", since it may be obtained by splitting the system by a two-steps technique (usually into a "convective" and "acoustic" part), apply a standard scheme to each part (which, for the convection system, indeed yields, at first order, an upwinding with respect to the material velocity) and then sum both steps to obtain the final flux. Works in this direction may be found in [59, 48, 66, 47, 63]. Here, following strictly this line seems difficult, since we work on staggered meshes and with a non-conservative formulation of the system, and obtain some non-standard fluxes ; in particular, the pressure gradient is discretized as the dual of the velocity divergence, and thus essentially centered. However, the scheme used here presents similarities with the above references, and its derivation does not use the ingredients usual in the context of hyperbolic systems, in particular (approximate) Riemann solvers (see *e.g.* [62, 28, 12] for surveys).

- Finally, the limitation procedure introduces a rather low stabilizing viscosity in the scheme : roughly speaking, the numerical viscosity is at most scaled by the material velocity, and this may be not sufficient in the zones where the (local) Mach number is low. To cope with this problem, we add a non-linear viscosity (in the sense that it depends on the solution) in the momentum balance equation, in the spirit of [31, 42].

The work presented here is an extension of [39] in two directions : first of all, the scheme proposed in [39] is only first-order in space and stabilization through a non-linear viscosity is non implemented ; second, we present here an extensive numerical assessment. These tests show that the new scheme is much more accurate that its first-order variant ; in addition, we observe that the straightforward formulation of the fluxes yields a very low cpu-time consuming algorithm. The present work has been or is being complemented in two directions. First, we present in [36, 29] partially implicited variants, under the form of pressure-correction algorithms, for Euler and Navier-Stokes equations respectively, which are shown to be unconditionally stable, *i.e.* stable irrespectively of the time and space steps. In addition, these schemes boil down to usual pressure correction schemes for incompressible flows when the Mach number tends to zero, with *inf-sup* stable discretizations. Second, weak consistency (or Lax-Wendroff type consistencty) results are shown for all this class of schemes (including the one presented here) in [26].

The paper is organized as follows. The space discretization is described in Section 2.2, and the scheme is given in Section 2.3. Numerical experiments are presented in Section 2.4.

2.2 Meshes and unknowns

In this section, we focus on the discretization of a multi-dimensional domain (*i.e.* $d = 2$ or $d = 3$) ; the extension to the one-dimensional case is straightforward.

Let $\mathcal{M}$ be a mesh of the domain Ω, supposed to be regular in the usual sense of the finite element literature (*e.g.* [19]). The cells of the mesh are assumed to be :

- for a general domain Ω, either non-degenerate quadrilaterals ($d = 2$) or hexahedra ($d = 3$), or simplices, both types of cells being possibly combined in a same mesh,

- for a domain the boundaries of which are hyperplanes normal to a coordinate axis, rectangles ($d = 2$) or rectangular parallelepipeds ($d = 3$) (the faces of which, of course, are then also necessarily normal to a coordinate axis).

By $\mathcal{E}$ and $\mathcal{E}(K)$ we denote the set of all $(d-1)$-faces σ of the mesh and of the element $K \in \mathcal{M}$ respectively. The set of faces included in the boundary of Ω is denoted by $\mathcal{E}_{\text{ext}}$ and the set of internal faces (*i.e.* $\mathcal{E} \setminus \mathcal{E}_{\text{ext}}$) is denoted by $\mathcal{E}_{\text{int}}$; a face $\sigma \in \mathcal{E}_{\text{int}}$ separating the cells K and L is denoted by $\sigma = K|L$. The outward normal vector to a face σ of K is denoted by $\boldsymbol{n}_{K,\sigma}$. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, we denote by $|K|$ the measure of K and by $|\sigma|$ the $(d-1)$-measure of the face σ. For $1 \leq i \leq d$, we denote by $\mathcal{E}^{(i)} \subset \mathcal{E}$ and $\mathcal{E}^{(i)}_{\text{ext}} \subset \mathcal{E}_{\text{ext}}$ the subset of the faces of $\mathcal{E}$ and $\mathcal{E}_{\text{ext}}$ respectively which are perpendicular to the ith unit vector of the canonical basis of $\mathbb{R}^d$.

The space discretization is staggered, using either the Marker-And Cell (MAC) scheme [34, 33], or nonconforming low-order finite element approximations, namely the Rannacher and Turek element (RT) [54] for quadrilateral or hexahedric meshes, or the nonconforming P1 [22] for simplicial meshes.

For all these space discretizations, the degrees of freedom for the pressure, the density and the internal energy (*i.e.* the discrete pressure, density and internal energy unknowns) are associated to the cells of the mesh $\mathcal{M}$, and are denoted by :

$$\left\{p_K,\ \rho_K,\ e_K,\ K \in \mathcal{M}\right\}.$$

Let us then turn to the degrees of freedom for the velocity (*i.e.* the discrete velocity unknowns).

- **Rannacher-Turek** or **Crouzeix-Raviart** discretizations – The degrees of freedom for the velocity components are located at the center of the faces of the mesh, and we choose the version of the element where they represent the average of the velocity through a face. The set of degrees of freedom reads :

$$\{u_{\sigma,i},\ \sigma \in \mathcal{E},\ 1 \leq i \leq d\}.$$

- **MAC** discretization – The degrees of freedom for the ith component of the velocity are defined at the centre of the faces $\sigma \in \mathcal{E}^{(i)}$, so the whole set of discrete velocity unknowns reads :

$$\left\{u_{\sigma,i},\ \sigma \in \mathcal{E}^{(i)},\ 1 \leq i \leq d\right\}.$$

We now introduce a dual mesh, which will be used for the finite volume approximation of the time derivative and convection terms in the momentum balance equation.

- **Rannacher-Turek** or **Crouzeix-Raviart** discretizations – For the RT or CR discretizations, the dual mesh is the same for all the velocity components. When $K \in \mathcal{M}$ is a simplex, a rectangle or a cuboid, for $\sigma \in \mathcal{E}(K)$, we define $D_{K,\sigma}$ as the cone with basis σ and with vertex the mass center of K (see Figure 2.1). We thus obtain a partition of K in m sub-volumes, where m is the number of faces of the mesh, each sub-volume having the same measure $|D_{K,\sigma}| = |K|/m$. We extend this definition to general quadrangles and hexahedra, by supposing that we have built a partition still of equal-volume sub-cells, and with the same connectivities. Note that this is of course always possible, but that such a volume $D_{K,\sigma}$ may be no longer a cone ; indeed, if K is far from a parallelogram, it may not be possible to build a cone having σ as basis, the opposite vertex lying in K and a volume equal to $|K|/m$.

  The volume $D_{K,\sigma}$ is referred to as the half-diamond cell associated to K and σ.

  For $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = K|L$, we now define the diamond cell D_σ associated to σ by $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$; for an external face $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}(K)$, D_σ is just the same volume as $D_{K,\sigma}$.

- **MAC** discretization – For the MAC scheme, the dual mesh depends on the component of the velocity. For each component, the MAC dual mesh only differs from the RT or CR dual mesh by the choice of the half-diamond cell, which, for $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$, is now the rectangle or rectangular parallelepiped of basis σ and of measure $|D_{K,\sigma}| = |K|/2$.

We denote by $|D_\sigma|$ the measure of the dual cell D_σ, and by $\epsilon = D_\sigma|D_{\sigma'}$ the face separating two diamond cells D_σ and $D_{\sigma'}$. The set of the faces of a dual cell D_σ is denoted by $\tilde{\mathcal{E}}(D_\sigma)$.

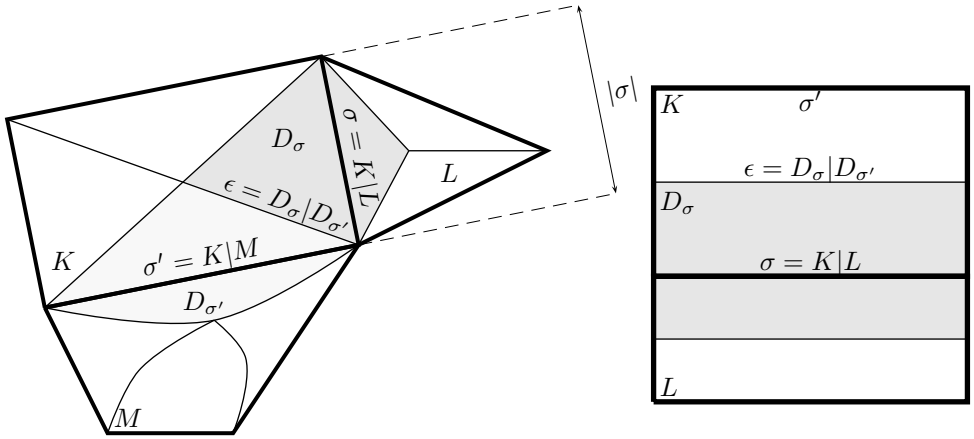


Figure 2.1 – Notations for control volumes and dual cells – Left : Finite Elements (the present sketch illustrates the possibility, implemented in the software ISIS [40], of mixing simplicial (Crouzeix-Raviart) and quadrangular (Rannacher-Turek) cells) – Right : MAC discretization, dual cell for the y-component of the velocity.

Finally, we need to deal with the impermeability (*i.e.* $\boldsymbol{u} \cdot \boldsymbol{n} = 0$) boundary condition. Since the velocity unknowns lie on the boundary (and not inside the cells), these conditions are taken into account in the definition of the discrete spaces. To avoid technicalities in the expression of the schemes, we suppose throughout this paper that the boundary is *a.e.* normal to a coordinate axis, (even in the case of the RT or CR discretizations), which allows to simply set to zero the corresponding velocity unknowns :

$$\text{for } i = 1, \dots, d, \ \forall \sigma \in \mathcal{E}^{(i)}_{\text{ext}}, \qquad u_{\sigma,i} = 0. \tag{2.2}$$

Therefore, there are no degrees of freedom for the velocity on the boundary for the MAC scheme, and there are only $d-1$ degrees of freedom on each boundary face for the CR and RT discretizations, which depend on the orientation of the face. In order to be able to write a unique expression of the discrete equations for both MAC and CR/RT schemes, we introduce the set of faces $\mathcal{E}^{(i)}_{\mathcal{S}}$ associated to the degrees of freedom of each component of the velocity ($\mathcal{S}$ stands for "scheme") :

$$\mathcal{E}^{(i)}_{\mathcal{S}} = \left| \begin{array}{l} \mathcal{E}^{(i)} \setminus \mathcal{E}^{(i)}_{\text{ext}} \text{ for the MAC scheme,} \\ \mathcal{E} \setminus \mathcal{E}^{(i)}_{\text{ext}} \text{ for the CR or RT schemes.} \end{array} \right.$$

For both schemes, we define $\tilde{\mathcal{E}}^{(i)}$, for $1 \le i \le d$, as the set of faces of the dual mesh associated to the ith component of the velocity. For the RT or CR discretizations, the sets $\tilde{\mathcal{E}}^{(i)}$ does not depend on the component (*i.e.* of i), up to the elimination of some unknowns (and so some dual cells and, finally, some external faces) to take the boundary conditions into account. For the MAC scheme, $\tilde{\mathcal{E}}^{(i)}$ depends on i ; note that each face of $\tilde{\mathcal{E}}^{(i)}$ is perpendicular to a unit vector of the canonical basis of $\mathbb{R}^d$, but not necessarily to the ith one.

General domains can be addressed (of course, with the CR or RT discretizations) by re-defining, through linear combinations, the degrees of freedom at the external faces, so as to introduce the normal velocity as a new degree of freedom.

2.3 The numerical scheme

We build in this section a scheme for the Euler equations (2.1). We recall that the conservative energy equation of the system is the total energy equation :

$$\partial_t(\rho E) + \mathrm{div}\Big(\rho\, E\, \boldsymbol{u}\Big) + \mathrm{div}(p\, \boldsymbol{u}) = 0.$$

Let us suppose that the solution is regular, and let E_k be the kinetic energy, defined by $E_k = \frac{1}{2}\,|\boldsymbol{u}|^2$. Taking the inner product of (2.1b) by $\boldsymbol{u}$ yields, after formal compositions of partial derivatives and using the mass balance (2.1a) :

$$\partial_t(\rho E_k) + \mathrm{div}\Big(\rho\, E_k\, \boldsymbol{u}\Big) + \boldsymbol{\nabla} p \cdot \boldsymbol{u} = 0. \tag{2.3}$$

This relation is referred to as the kinetic energy balance. Substracting this relation from the total energy balance (2.1c), we obtain the internal energy balance equation :

$$\partial_t(\rho e) + \mathrm{div}(\rho e \boldsymbol{u}) + p\, \mathrm{div}\boldsymbol{u} = 0. \tag{2.4}$$

Since,

- thanks to the mass balance equation, the first two terms in the left-hand side of (2.4) may be recast as a transport operator : $\partial_t(\rho e) + \mathrm{div}(\rho e \boldsymbol{u}) = \rho\,[\partial_t e + \boldsymbol{u} \cdot \boldsymbol{\nabla} e]$,
- and, from the equation of state, the pressure vanishes when $e = 0$,

this equation implies, if $e \geq 0$ at $t = 0$ and with suitable boundary conditions, that e remains non-negative at all times. Solving this equation instead of the total energy equation seems appealing to preserve the positivity of the internal energy by construction of the scheme. Furthermore it avoids to introduce a discrete approximation for the total energy which would not be straightforward since the internal energy and the kinetic energy are not discretized on the same grid. We thus choose here to design a scheme solving the internal energy balance. However, the internal energy being a non-conservative variable, a raw discretization of (2.4) can lead to non-consistent solutions (wrong shock predictions for example). We overstep this difficulty by adding, as in [39], a corrective term in the discrete internal energy balance equation ; this point is discussed in Section 2.3.2 below.

In adddition, the proposed scheme features two ingredients :

- The algorithm of [39] used an elementary first-order upwinding "equation-by-equation" of the convection terms with respect to the material velocity, while the approximation of the pressure gradient is basically centered (more precisely speaking, the gradient is built as the transposed of the natural divergence). We keep here the same philosophy, but build a more accurate scheme by adopting a MUSCL-like aproximation for the convection in the mass and energy balance equations, while the discretization of the momentum balance is still first-order (and even more diffusive, see next item). Indeed, our goal is here to get a better approximation of the (1D) contact discontinuity, where the density and internal energy are discontinuous while the velocity is constant, and which is known to be the part of the solution where the scheme diffusion essentially spoils the solution. Since we deal with each equation separately, the MUSCL technique may be directly inspired from the work on the transport operator presented in [53].
- In [39], we observed that the numerical solution presented oscillations in the zones where the fluid was at rest. This may be explained by the fact that, because of the particular upwinding used here, when the velocity vanishes, no stabilizing diffusion remains. An alternative numerical diffusion for a staggered scheme, obtained thanks to a kinetic approach and which does not vanish with the velocity, may be found in [10]. Here we follow a different line, in the spirit of [30, 31, 42], which consists in introducing in the momentum balance equation (only) an artificial viscosity estimated a posteriori thanks to the solution at the previous time step.

The presentation of the scheme is organized as follows. We first give the general form of the scheme (Section 2.3.1). Then we detail the construction of the corrective terms in the energy balance (Section 2.3.2). The next section (Section 2.3.3) is devoted to the stability analysis of the scheme ; we prove that, under a CFL condition, the convex of admissible states is preserved (so, in other words, $\rho > 0, e > 0$ and $p > 0$) and show that the velocity and pressure are kept constant at the contact discontinuity. These results are obtained thanks to some abstract assumptions on the approximation of the density and internal energy at the face, in the discretization of the mass and internal energy convection term, respectively. We build in Section 2.3.4 a MUSCL algorithm (more specifically, a limitation procedure) which allows to satisfy these assumptions, so the density and energy convection operator is fully specified. Finally, Section 2.3.5 is devoted to the design of the artificial viscosity.

2.3.1 General form of the scheme

Let us consider a partition $0 = t_0 < t_1 < \ldots < t_N = T$ of the time interval $(0, T)$, which we suppose uniform for the sake of simplicity, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \ldots, N-1$ be the (constant) time step. We consider an explicit-in-time scheme, which reads in its fully discrete form, for $0 \le n \le N-1$:

$$\forall K \in \mathcal{M},\ \frac{|K|}{\delta t}(\rho_K^{n+1} - \rho_K^n) + \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^n = 0, \tag{2.5a}$$

$$\forall K \in \mathcal{M},\ \frac{|K|}{\delta t}(\rho_K^{n+1} e_K^{n+1} - \rho_K^n e_K^n) + \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^n e_\sigma^n + |K|\, p_K^n\, (\mathrm{div}\boldsymbol{u})_K^n = S_K^n, \tag{2.5b}$$

$$\forall K \in \mathcal{M},\ p_K^{n+1} = (\gamma - 1)\, \rho_K^{n+1}\, e_K^{n+1}, \tag{2.5c}$$

For $1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}$,

$$\begin{aligned} \frac{|D_\sigma|}{\delta t}(\rho_{D_\sigma}^{n+1} u_{\sigma,i}^{n+1} - \rho_{D_\sigma}^n u_{\sigma,i}^n) + \sum_{\epsilon \in \tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n u_{\epsilon,i}^n \\ + \sum_{\epsilon \in \tilde{\mathcal{E}}(D_\sigma)} \nu_\epsilon^{n+1}(u_{\sigma,i}^n - u_{\sigma',i}^n) + |D_\sigma|\, (\boldsymbol{\nabla} p)_{\sigma,i}^{n+1} = 0, \end{aligned} \tag{2.5d}$$

where the terms introduced for each discrete equation are defined hereafter.

Equation (2.5a) is obtained by the discretization of the mass balance equation (2.1a) over the primal mesh, and $F_{K,\sigma}^n$ stands for the mass flux across σ outward K, which, because of the impermeability condition, vanishes on external faces and is given on the internal faces by :

$$\forall \sigma = K|L \in \mathcal{E}_{\mathrm{int}}, \qquad F_{K,\sigma}^n = |\sigma|\, \rho_\sigma^n\, u_{K,\sigma}^n, \tag{2.6}$$

where $u_{K,\sigma}^n$ is an approximation of the normal velocity to the face σ outward K. This latter quantity is defined by :

$$u_{K,\sigma}^n = \left| \begin{array}{l} u_{\sigma,i}^n\, \boldsymbol{e}^{(i)} \cdot \boldsymbol{n}_{K,\sigma} \text{ for } \sigma \in \mathcal{E}^{(i)} \text{ in the MAC case,} \\ \boldsymbol{u}_\sigma^n \cdot \boldsymbol{n}_{K,\sigma} \text{ in the CR and RT cases,} \end{array} \right. \tag{2.7}$$

where $\boldsymbol{e}^{(i)}$ denotes the i-th vector of the orthonormal basis of $\mathbb{R}^d$. The density at the face $\sigma = K|L$ is approximated by a MUSCL technique, detailed in Section 2.3.4. We only state here the algebraic condition which we require to this reconstruction, which is that for any $K \in \mathcal{M}$ and for any $\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\mathrm{int}}$, there exists $\alpha_{K,\sigma} \in [0,1]$ and $M_\sigma^K \in \mathcal{M}$ such that :

$$\rho_\sigma^n - \rho_K^n = \left| \begin{array}{ll} \alpha_{K,\sigma}(\rho_K^n - \rho_{M_\sigma^K}) & \text{if } u_{K,\sigma}^n \ge 0, \\ \alpha_{K,\sigma}(\rho_{M_\sigma^K} - \rho_K^n) & \text{otherwise.} \end{array} \right. \tag{2.8}$$

We now turn to the discrete momentum balance (2.5d), which is obtained by discretizing the momentum balance equation (2.1b) on the dual cells associated to the faces of the mesh. Up to the addition of a viscosity term, this equation is the same as in [39], and we refer to this work for details. The first task is to define the values $\rho^{n+1}_{D_\sigma}$ and $\rho^n_{D_\sigma}$, which approximate the density over the dual cell D_σ at time t^{n+1} and t^n respectively, and the discrete mass flux through the dual face ϵ outward D_σ, denoted by $F^n_{\sigma,\epsilon}$; the guideline for their construction is that a finite volume discretization of the mass balance equation over the diamond cells, of the form

$$\forall \sigma \in \mathcal{E}, \qquad \frac{|D_\sigma|}{\delta t}\,(\rho^{n+1}_{D_\sigma} - \rho^n_{D_\sigma}) + \sum_{\epsilon \in \tilde{\mathcal{E}}(D_\sigma)} F^n_{\sigma,\epsilon} = 0, \tag{2.9}$$

must hold in order to be able to derive a discrete kinetic energy balance (see Section 2.3.2 below). The density on the dual cells is given by the following weighted average :

$$\text{for } \sigma = K|L \in \mathcal{E}_{\text{int}}, \text{ for } k = n \text{ and } k = n+1, \qquad |D_\sigma|\,\rho^k_{D_\sigma} = |D_{K,\sigma}|\,\rho^k_K + |D_{L,\sigma}|\,\rho^k_L. \tag{2.10}$$

For the MAC scheme, the flux on a dual face which is located on two primal faces is the mean value of the sum of fluxes on the two primal faces, and the flux of a dual face located between two primal faces is again the mean value of the sum of fluxes on the two primal faces [37]. In the case of the CR and RT schemes, for a dual face ϵ included in the primal cell K, this flux is computed as a linear combination (with constant coefficients, *i.e.* independent of the cell) of the mass fluxes through the faces of K, *i.e.* the quantities $(F^n_{K,\sigma})_{\sigma\in\mathcal{E}(K)}$ appearing in the discrete mass balance (2.5a). We refer to [4, 25] for a detailed construction of this approximation. Let us remark that a dual face lying on the boundary is then also a primal face, and the flux across this face is zero. Therefore, the values $u^n_{\epsilon,i}$ are only needed at the internal dual faces, and we make the upwind choice for their discretization :

$$\text{for } \epsilon = D_\sigma|D_{\sigma'}, \quad u^n_{\epsilon,i} = \left| \begin{array}{ll} u^n_{\sigma,i} & \text{if } F^n_{\sigma,\epsilon} \geq 0, \\ u^n_{\sigma',i} & \text{otherwise.} \end{array} \right. \tag{2.11}$$

The last term $(\nabla p)^{n+1}_{\sigma,i}$ stands for the i-th component of the discrete pressure gradient at the face σ. The gradient operator is built as the transpose of the discrete operator for the divergence of the velocity, the discretization of which is based on the primal mesh. Let us denote the divergence of $\boldsymbol{u}^{n+1}$ over $K \in \mathcal{M}$ by $(\text{div}\boldsymbol{u})^{n+1}_K$; its natural approximation reads :

$$\text{for } K \in \mathcal{M}, \qquad (\text{div}\boldsymbol{u})^{n+1}_K = \frac{1}{|K|} \sum_{\sigma\in\mathcal{E}(K)} |\sigma|\, u^{n+1}_{K,\sigma}. \tag{2.12}$$

Consequently, the components of the pressure gradient are given by :

$$\text{for } \sigma = K|L \in \mathcal{E}_{\text{int}}, \qquad (\nabla p)^{n+1}_{\sigma,i} = \frac{|\sigma|}{|D_\sigma|}\,(p^{n+1}_L - p^{n+1}_K)\,\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{e}^{(i)}, \tag{2.13}$$

this expression being derived thanks to the following duality relation with respect to the L^2 inner product :

$$\sum_{K\in\mathcal{M}} |K|\, p^{n+1}_K\, (\text{div}\boldsymbol{u})^{n+1}_K + \sum_{i=1}^{d} \sum_{\sigma\in\mathcal{E}^{(i)}_{\mathcal{S}}} |D_\sigma|\, u^{n+1}_{\sigma,i}\, (\nabla p)^{n+1}_{\sigma,i} = 0. \tag{2.14}$$

Note that, because of the impermeability boundary conditions, the discrete gradient is not defined at the external faces.

Equation (2.5b) is an approximation of the internal energy balance over the primal cell K. For the discretization of the internal energy at the primal faces we use the same MUSCL technique as for the density to ensure the positivity of the convection operator, see Section 2.3.4 ; hence we have For any $K \in \mathcal{M}$, and for any $\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\text{int}}$, there exists $\alpha_{K,\sigma} \in \mathbb{R}$ such that :

$$\text{for } \sigma = K|L \in \mathcal{E}_{\text{int}},\ e^n_\sigma - e^n_K = \left| \begin{array}{l} \alpha_{K,\sigma}(e^n_K - e^n_{M^K_\sigma}),\ \alpha_{K,\sigma} \in [0,1],\ \text{if } F^n_{K,\sigma} \geq 0, \\ \alpha_{K,\sigma}(e^n_{M^K_\sigma} - e^n_K),\ \alpha_{K,\sigma} \geq 0,\ \text{otherwise.} \end{array} \right. \tag{2.15}$$

The discrete divergence of the velocity, $(\text{div}\boldsymbol{u})^n_K$, is defined by (2.12). The right-hand side, S^n_K, is derived using consistency arguments in the next section ; at the first time step, it is simply set to zero :

$$\forall K \in \mathcal{M}, \qquad S^0_K = 0.$$

Finally, the initial approximations for ρ, e and $\boldsymbol{u}$ are given by the average of the initial conditions ρ_0 and e_0 on the primal cells and of $\boldsymbol{u}_0$ on the dual cells :

$$\begin{aligned} &\forall K \in \mathcal{M}, \quad \rho^0_K = \frac{1}{|K|}\int_K \rho_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \text{and } e^0_K = \frac{1}{|K|}\int_K e_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \\ &\text{for } 1 \leq i \leq d,\ \forall \sigma \in \mathcal{E}^{(i)}_{\mathcal{S}}, \quad u^0_{\sigma,i} = \frac{1}{|D_\sigma|}\int_{D_\sigma} (\boldsymbol{u}_0(\boldsymbol{x}))_i\,\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.16}$$

2.3.2 Discrete kinetic energy balance and corrective source terms

Equation (2.17) below is a discrete analogue of the kinetic energy balance equation (2.3), with some additional terms due to the artificial viscosity terms implemented in the scheme.

At the continuous level, the kinetic energy convection term is obtained by taking the inner product of the momentum balance equation by the velocity and using twice the mass balance equation. At the discrete level, the computation is essentially the same, provided that a momentum balance and a mass balance hold on the same cell, which we ensured thanks to the definition of the dual densities and fluxes which entail the discrete dual mass balance (2.9). The obtained kinetic energy convection flux is upwind with respect to the mass flux.

For the diffusion term, the algebraic manipulation performed at the discrete level are reminiscent of the continuous identity $-\mu u_i \Delta u_i = -\text{div}(\mu u_i \boldsymbol{\nabla} u_i) + \mu|\boldsymbol{\nabla} u_i|^2$ (valid for a constant viscosity). The conservative term is left at the left-hand side of the equation, while the dissipation term is considered as a residual term.

Lemma 2.1 (*Discrete kinetic energy balance*)

A solution to the system (2.5) satisfies the following equality, for $1 \leq i \leq d$, $\sigma \in \mathcal{E}^{(i)}_{\mathcal{S}}$ and $0 \leq n \leq N-1$:

$$\begin{aligned} &\frac{1}{2}\frac{|D_\sigma|}{\delta t}\Big[\rho^{n+1}_{D_\sigma}(u^{n+1}_{\sigma,i})^2 - \rho^n_{D_\sigma}(u^n_{\sigma,i})^2\Big] \\ &\qquad + \frac{1}{2}\sum_{\epsilon = D_\sigma|D_{\sigma'} \in \tilde{\mathcal{E}}(D_\sigma)} F^n_{\sigma,\epsilon}\, u^n_{\sigma,i} u^n_{\sigma',i} + |D_\sigma|\,(\boldsymbol{\nabla} p)^{n+1}_{\sigma,i}\, u^{n+1}_{\sigma,i} \\ &\qquad\qquad + \frac{1}{2}\sum_{\epsilon = D_\sigma|D_{\sigma'} \in \tilde{\mathcal{E}}(D_\sigma)} \mu^n_\epsilon\,(u^n_{\sigma,i} - u^n_{\sigma',i})\,(u^n_{\sigma,i} + u^n_{\sigma',i}) = -R^{n+1}_{\sigma,i}, \end{aligned} \tag{2.17}$$

with

$$R^{n+1}_{\sigma,i} = \frac{1}{2}\frac{|D_\sigma|}{\delta t}\rho^{n+1}_{D_\sigma}(u^{n+1}_{\sigma,i} - u^n_{\sigma,i})^2 + \frac{1}{2}\sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \mu^n_\epsilon\,(u^n_{\sigma',i} - u^n_{\sigma,i})^2 + \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \left(\mu^n_\epsilon - \frac{F_{\sigma,\epsilon}}{2}\right)(u^{n+1}_{\sigma,i} - u^n_{\sigma,i})\,(u^n_{\sigma,i} - u^n_{\sigma',i}), \quad (2.18)$$

and where $\mu^n_\epsilon = |F^n_{\sigma,\epsilon}|/2 + \nu^{n+1}_\epsilon$.

Proof : The proof is similar to that of [39, Lemma 4.1]. Multiplying the i-th component of the momentum balance equation (2.5d) associated to the face σ by $u^{n+1}_{\sigma,i}$, using the dual mass balance equation (2.9) and invoking [39, Lemma A.2] yields the terms associated the convection operators. For the diffusion term, we just use the following elementary computation :

$$(u^n_{\sigma,i} - u^n_{\sigma',i})\,u^{n+1}_{\sigma,i} = (u^n_{\sigma,i} - u^n_{\sigma',i})\,(u^{n+1}_{\sigma,i} - u^n_{\sigma,i}) + \frac{1}{2}(u^n_{\sigma,i} - u^n_{\sigma',i})^2 + \frac{1}{2}(u^n_{\sigma,i} - u^n_{\sigma',i})(u^n_{\sigma,i} + u^n_{\sigma',i}).$$

■

The residual terms $R^{n+1}_{\sigma,i}$ may be seen as a numerical dissipation generated by the numerical diffusions. Because of the discontinuous solutions that exist in the case of the inviscid Euler equations which we are dealing with here, these terms do not tend to zero with the mesh and time steps, but subsist as measures borne by the shocks (see [29, Remark 4.1]). In order for the scheme to be consistent with the total energy balance, we thus need to compensate this dissipation in the internal energy balance by adding the corrective terms S^n_K in (2.5b). Because of the staggered discretization, or, in other terms, since the kinetic energy balance is associated to the dual mesh while the internal energy balance is discretized on the primal mesh, we are not able to recover a local total energy balance, and a direct term-to-term compensation is not possible. We thus are lead to build the quantities (S^{n+1}_K) by dispatching the terms $(R^{n+1}_{\sigma,i})$ given by (2.18) on the neighbouring primal cells. For $K \in \mathcal{M}$, S^{n+1}_K is computed as $S^{n+1}_K = \sum_{i=1}^d S^{n+1}_{K,i}$ with :

$$S^{n+1}_{K,i} = \frac{1}{2}\rho^{n+1}_K \sum_{\sigma\in\mathcal{E}(K)\cap\mathcal{E}^{(i)}_S} \frac{|D_{K,\sigma}|}{\delta t}\left(u^{n+1}_{\sigma,i} - u^n_{\sigma,i}\right)^2 + \sum_{\epsilon\in\tilde{\mathcal{E}}^{(i)}_S,\ \epsilon\cap\bar{K}\neq\emptyset} S^{n+1}_{K,\epsilon,i}, \quad (2.19)$$

where $S^{n+1}_{K,\epsilon,i}$ stands for the contribution of ϵ to $S^{n+1}_{K,i}$, which we now define.

Step 1 - To this purpose, our first task is, for a given dual face ϵ, to gather the remainders issued from the kinetic energy balances associated to the two neighbour dual cells. Let us begin with the terms issued from the upwinding of the convection. Let σ^U_ϵ and σ^D_ϵ be the two primal faces such that $\epsilon = D_{\sigma^D_\epsilon}|D_{\sigma^U_\epsilon}$ and $F^n_{\sigma^D_\epsilon,\epsilon} \leq 0$ (*i.e.* $D_{\sigma^D_\epsilon}$ is the dual cell located downstream ϵ, see Figure 2.2). Then, we get for ϵ the following contribution from the upwind dual cell :

$$(R^U_{\epsilon,i})^{n+1} = \frac{1}{2}\mu^n_\epsilon\,(u^n_{\sigma^D_\epsilon,i} - u^n_{\sigma^U_\epsilon,i})^2 + \left(\mu^n_\epsilon - \frac{|F^n_{\sigma^U_\epsilon,\epsilon}|}{2}\right)\left(u^{n+1}_{\sigma^U_\epsilon,i} - u^n_{\sigma^U_\epsilon,i}\right)(u^n_{\sigma^U_\epsilon,i} - u^n_{\sigma^D_\epsilon,i}).$$

The contribution of the downwind cell reads :

$$(R^D_{\epsilon,i})^{n+1} = \frac{1}{2}\mu^n_\epsilon\,(u^n_{\sigma^D_\epsilon,i} - u^n_{\sigma^U_\epsilon,i})^2 + \left(\mu^n_\epsilon + \frac{|F^n_{\sigma^D_\epsilon,\epsilon}|}{2}\right)\left(u^{n+1}_{\sigma^U_\epsilon,i} - u^n_{\sigma^U_\epsilon,i}\right)(u^n_{\sigma^U_\epsilon,i} - u^n_{\sigma^D_\epsilon,i}).$$

Gathering both terms, we obtain that $R^{n+1}_{\epsilon,i} = (R^D_{\epsilon,i})^{n+1} + (R^U_{\epsilon,i})^{n+1}$.

Step 2 - Let us now distribute $R^{n+1}_{\epsilon,i}$ in $S^{n+1}_{K,\epsilon,i}$.

There are two different cases. First, if ϵ is included in K, we just set $S^{n+1}_{K,\epsilon,i} = R^{n+1}_{\epsilon,i}$; this is the only situation to consider for the RT and CR discretizations, and it happens for some dual faces for the MAC scheme (precisely speaking, the dual faces which are normal to $\boldsymbol{e}^{(i)}$).

For the MAC scheme and for the dual faces which coincide (for one part) with ∂K. Let us consider the case where K is upstream to ϵ (or, in other words, the case where σ^U_ϵ is a face of K). Then, le L be the other upstream primal cell to ϵ (or, in other words, the cell such as $\sigma^U_\epsilon = K|L$). Then we set :

$$S^{n+1}_{K,\epsilon,i} = \frac{|K|}{|K|+|L|}\left((R^U_{\epsilon,i})^{n+1} - \frac{|F^n_{\sigma^U_\epsilon,\epsilon}|}{4}(u^n_{\sigma^D_\epsilon,i} - u^n_{\sigma^U_\epsilon,i})^2\right).$$

If K is a dowstream cell, we set :

$$S^{n+1}_{K,\epsilon,i} = \frac{|K|}{|K|+|L|}\left((R^D_{\epsilon,i})^{n+1} + \frac{|F^n_{\sigma^U_\epsilon,\epsilon}|}{4}(u^n_{\sigma^D_\epsilon,i} - u^n_{\sigma^U_\epsilon,i})^2\right).$$

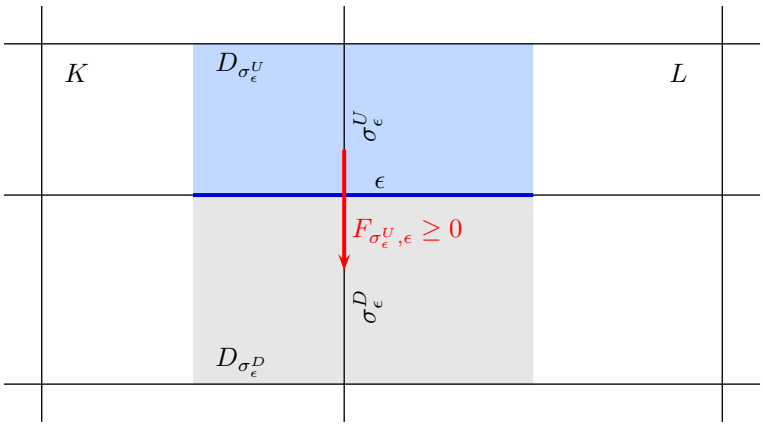


Figure 2.2 – Notations the construction of the corrective term $S_{K,1}$, in the MAC case, for a dual face lying on the primal cells boundaries. ϵ : considered dual edge. $D_{\sigma^U_\epsilon}$: upstream dual cell. $D_{\sigma^D_\epsilon}$: downstream dual cell. K, L : upstream cells.

The expression of the terms $(S^{n+1}_K)_{K\in\mathcal{M}}$ may be justified by showing that with this choice, under some compactness assumptions, we may pass to the limit in the scheme to show that any possible limit of approximate solutions is indeed a weak solution to the Euler equations[26]. We may already note here that :

$$\sum_{K\in\mathcal{M}} S^{n+1}_K - \sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}^{(i)}_S} R^{n+1}_{\sigma,i} = 0, \tag{2.20}$$

and so, by summing the kinetic energy balance over the component and faces and the internal energy balance over the cells, we observe that the integral of the total energy over the domaine is conserved.

2.3.3 Stability results

The following positivity result is a consequence of the MUSCL interpolation of the density in (2.5a).

Lemma 2.2 (*Positivity of the density*)

Let ρ^0 be given by (2.16). Then, since ρ_0 is assumed to be a positive function, $\rho^0 > 0$ and, under the CFL *condition :*

$$\delta t \leq \frac{|K|}{\sum_{\sigma\in\mathcal{E}(K)} |\sigma| \, (1+\alpha_{K,\sigma})(u^n_{K,\sigma})^+}, \quad \forall K \in \mathcal{M}, \text{ for } 0 \leq n \leq N-1, \tag{2.21}$$

where, for $a \in \mathbb{R}$, $a^+ \geq 0$ is defined by $a^+ = \max(a, 0)$, the solution to the scheme satisfies $\rho^n > 0$, for $1 \leq n \leq N$.

The definition (2.19) of $(S^{n+1}_K)_{K\in\mathcal{M}}$ allows to prove that, under a CFL condition, the scheme also preserves the positivity of e.

Lemma 2.3 (*Positivity of the internal energy*)

We assume that the CFL condition (2.21) holds, and we furthermore assume that, for $0 \leq n \leq N-1$, for all $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$, we have :

$$\delta t \leq \min\Big(\frac{|K|\rho^n_K}{\displaystyle\sum_{\sigma\in\mathcal{E}(K)} (\gamma-1)|\sigma| \, \rho^n_K (u^n_{K,\sigma})^+ + (F^n_{K,\sigma})^+ + \alpha_{K,\sigma}|F^n_{K,\sigma}|}, \\ \frac{|D_{K,\sigma}| \, \rho^n_K}{\displaystyle\sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma),\, \epsilon\cap\bar{K}\neq\emptyset} \nu^{n+1}_\epsilon + |F^n_{\sigma,\epsilon}|} \Big). \tag{2.22}$$

Then the internal energy $(e^n)_{1\leq n\leq N}$ given by the scheme (2.5) is positive.

Proof : Let n such that $0 < n \leq N-1$ be given, and let us assume in a first step that $e^n_K \geq 0$ and $S^n_K \geq 0$ for all $K \in \mathcal{M}$. Because (2.21) is satisfied we have $\rho^n_K \geq 0$ and $\rho^{n+1}_K \geq 0$. In the internal energy equation (2.5b), let us express the pressure thanks to the equation of state (2.5c) to obtain :

$$\frac{|K|}{\delta t}\rho^{n+1}_K e^{n+1}_K = \left[\frac{|K|}{\delta t}\rho^n_K - \sum_{\sigma\in\mathcal{E}(K)} \left[(F^n_{K,\sigma})^+ + \alpha_{K,\sigma}|F^n_{K,\sigma}| - (\gamma-1)\,\rho^n_K|\sigma|(u^n_{K,\sigma})^+\right]\right] \\ e^n_K + \sum_{\sigma\in\mathcal{E}(K)} \alpha_{K,\sigma}|F^n_{K,\sigma}|e^n_{M^K_\sigma} + \sum_{\sigma\in\mathcal{E}(K)} (F^n_{K,\sigma})^- e^n_K \\ + (\gamma-1)\rho^n_K e^n_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma|(u^n_{K,\sigma})^- + S^n_K. \tag{2.23}$$

Then we get $e^{n+1}_K > 0$ under the following CFL condition :

$$\delta t \leq \frac{|K|\rho^n_K}{\displaystyle\sum_{\sigma\in\mathcal{E}(K)} (\gamma-1)|\sigma| \, \rho^n_K (u^n_{K,\sigma})^+ + (F^n_{K,\sigma})^+ + \alpha_{K,\sigma}|F^n_{K,\sigma}|}.$$

Let us now derive a condition for the non-negativity of the source term S^n_K. For $i \in [\![1, d]\!]$ and $\epsilon = D_\sigma|D_{\sigma'} \in \tilde{\mathcal{E}}^{(i)}_S$, $\epsilon \cap \bar{K} \neq \emptyset$, we have, using a Taylor-Young inequality :

$$S^n_{K,\epsilon,i} \geq -\frac{1}{2} \sum_{\sigma\in\mathcal{E}(K),\, \epsilon\cap D_\sigma\neq\emptyset} \left(\nu^{n+1}_\epsilon + |F_{\sigma,\epsilon}|^n\right)\left(u^{n+1}_{\sigma,i} - u^n_{\sigma,i}\right)^2 \tag{2.24}$$

Recalling that

$$S^n_{K,i} = \sum_{\sigma\in\mathcal{E}(K)\cap\mathcal{E}^{(i)}_S} \frac{1}{2}\rho^{n+1}_K \frac{|D_{K,\sigma}|}{\delta t}\left(u^{n+1}_{\sigma,i} - u^n_{\sigma,i}\right)^2 + \sum_{\epsilon\in\tilde{\mathcal{E}}^{(i)}_S,\, \epsilon\cap\bar{K}\neq\emptyset} S^{n+1}_{K,\epsilon,i},$$

we get, using (2.24) and reordering the terms :

$$S_{K,i}^n \geq \sum_{\sigma\in\mathcal{E}(K)\cap\mathcal{E}_S^{(i)}} \left\{ \frac{1}{2}\rho_K^{n+1}\frac{|D_{K,\sigma}|}{\delta t} - \frac{1}{2}\sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma),\epsilon\cap\bar{K}\neq\emptyset} \nu_\epsilon^{n+1} + |F_{\sigma,\epsilon}|^n \right\} \left(u_{\sigma,i}^{n+1} - u_{\sigma,i}^n\right)^2.$$

The positivity of $S_{K,i}^n$ is then ensured, provided that :

$$\delta t \leq \min_{\sigma\in\mathcal{E}(K)} \frac{|D_{K,\sigma}|\rho_K^{n+1}}{\displaystyle\sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma),\ \epsilon\cap\bar{K}\neq\emptyset} \nu_\epsilon^{n+1} + |F_{\sigma,\epsilon}^n|}.$$

The positivity of S_K^n and e_K^{n+1} follow directly, under the condition (2.22). ■

The upwind version of the scheme studied in [39] preserves the contact discontinuities if the pressure is a function of the product ρe, which is the case of the perfect gas EOS (2.1d) considered here ; indeed, if the pressure and velocity are constant through a contact discontinuity at time t_n, then they remain so at time t_{n+1}. We show in the proposition below that under a condition which correlates the MUSCL reconstructions of the face values e_σ and ρ_σ, the scheme (2.1a)-(2.1b) also preserves 1D contact discontinuities.

Proposition 2.4 (*Preservation of the contact discontinuities*)

Let us suppose that $u_0 = u$ and $p_0 = p$, u and p constant. Additionally assume that

$$\forall n \in [\![1,N]\!], \forall\sigma\in\mathcal{E}_{\text{int}}, \sigma = K|L, \exists\, \kappa_\sigma^n \in [0,1]\ ;\ \rho_\sigma^n e_\sigma^n = \kappa_\sigma^n \rho_K^n e_K^n + (1-\kappa_\sigma^n)\rho_L^n e_L^n, \tag{2.25}$$

then $\forall n \in [\![1,N]\!]$ and $\forall K \in \mathcal{M}$, $u_K^n = u$ and $p_K^n = p$.

Proof : Without loss of generality, we restrict ourselves to the one-dimensional case. A cell $K \in \mathcal{M}$ is then denoted $K = [\sigma', \sigma]$, where σ' and σ are the two interfaces of K. Assume that the proposition is true for all $k \in [\![0,n]\!]$ and for all $K = [\sigma',\sigma] \in \mathcal{M}$. It is easy to see that $S_K^n = 0$ and $(\text{div}u)_K^n = 0$. The internal energy equation (2.5b) for $K = [\sigma',\sigma]$ then reads :

$$\frac{|K|}{\delta t}\left(\rho_K^{n+1}e_K^{n+1} - \rho_K^n e_K^n\right) + u\left(\rho_\sigma^n e_\sigma^n - \rho_{\sigma'}^n e_{\sigma'}^n\right) = 0.$$

From the EOS (2.5c), we get that $\rho_K^n e_K^n = \dfrac{p}{\gamma-1}$, $\quad\forall K \in \mathcal{M}$, and so

$$\rho_\sigma^n e_\sigma^n = \kappa_\sigma^n \rho_K^n e_K^n + (1-\kappa_\sigma^n)\rho_L^n e_L^n = \frac{p}{\gamma-1}, \quad \forall\sigma\in\mathcal{E}_{\text{int}},\ \sigma = K|L.$$

Thus : $\rho_K^{n+1}e_K^{n+1} = \rho_K^n e_K^n$ and $p_K^{n+1} = p$, $\quad\forall K \in \mathcal{M}$, and $\nabla p_\sigma^{n+1} = 0 \quad \forall\sigma\in\mathcal{E}_{\text{int}},\ \sigma = K|L$. Denoting by F_K^n and F_L^n the numerical fluxes $F_{\sigma,\epsilon}^n$ on the dual interfaces ϵ included in K and L respectively, and noting that $u_\epsilon^n = u$ for both interfaces, the momentum equation (2.5d) then reads :

$$\frac{|D_\sigma|}{\delta t}\left(\rho_{D_\sigma}^{n+1}u_{D_\sigma}^{n+1} - \rho_{D_\sigma}^n u_{D_\sigma}^n\right) + \left(F_K^n - F_L^n\right)u = 0.$$

Together with the discrete dual mass balance (2.9) which reads

$$\frac{|D_\sigma|}{\delta t}\left(\rho_{D_\sigma}^{n+1} - \rho_{D_\sigma}^n\right) + \left(F_K^n - F_L^n\right) = 0,$$

we obtain that

$$\frac{|D_\sigma|}{\delta t}\rho_{D_\sigma}^{n+1}\left(u_\sigma^{n+1} - u\right) = 0,$$

and therefore $u_\sigma^{n+1} = u\ \forall\sigma\in\mathcal{E}_{\text{int}}$, which concludes the proof of the proposition. ■

Finally we mention that for both MAC and CR-RT discretizations, one may show that the scheme is weakly consistent, or consistent in the Lax-Wendroff sense : a sequence of converging discrete solutions of the scheme necessarily converges to the solution of the weak formulation of (2.1) when the time step and the space step of the mesh tend to 0. The proof of this result is quite technical and out of the scope of this paper and is the object of ongoing work.

2.3.4 MUSCL interpolation

As already mentioned when introducing the general form of the scheme (2.5), the upwinding process is performed equation-per-equation, on the basis of the material velocity only ; a MUSCL-like strategy is applied only for the density and internal energy balance equations. The objective of this section is to detail this algorithm, thus, precisely speaking, the approximation of the density and internal energy at the primal face in equations (2.5a) and (2.5b) respectively.

As a consequence of this equation-per-equation process, the problem that we face is close to the program realized in [53], namely to built an approximation for a convection operator (satisfying a maximum principle) which is formally second order in space when the solution is regular, and preserves the range of variation of the unknowns even in case of shocks, by an adequate flux limitation procedure. The algorithm presented here is thus an extension of the scheme developped in [53] ; in particular, contrary to most MUSCL reconstructions which use slope estimation and limitation, see e.g. [8, 62] for reviews and [45, 11, 14, 15] for recent works, the limitation is here directly derived from stability conditions which are purely algebraic (in the sense that they do not require any geometric computation), and thus work with arbitrary meshes.

Compared to [53], the algorithm is however complicated by the requirement that the scheme should preserve pressure-constant zones, to avoid to destabilize the computation of contact discontinuities (more precisely, of the one-dimensional contact discontinuity, across which the velocity is constant, the difficult problem posed by slip interfaces in 2D or 3D being out of the scope of this study). *In fine*, this is realized by imposing to the face pressure (*i.e.* the pressure obtained by applying the equation of state to the face density and internal energy) to be a convex interpolation of the pressure in the two neighbour cells. This condition leads to a limitation procedure which takes into account both mass and internal energy equations, so that we somehow loose here our equation decoupling strategy.

As often in MUSCL techniques, the algorithm consists in two steps : first compute a tentative second-order approximation (here for the density only) and then apply a limitation procedure. We describe these two steps successively in the following. For the sake of clarity, we omit in this section all the superscripst relative to the time step number.

Computation of a tentative value for the density – For an edge $\sigma \in \mathcal{E}_{\text{int}}$ and $K \in \mathcal{M}$, let us call $\boldsymbol{x}_\sigma$ and $\boldsymbol{x}_K$ the mass center of σ and K respectively. Let $\sigma \in \mathcal{E}_{\text{int}}$ be a given internal face. We suppose that we have computed a set of real coefficients (ζ_σ^L) such that :

$$\boldsymbol{x}_\sigma = \sum_{L\in\mathcal{M}} \zeta_\sigma^L \boldsymbol{x}_L, \qquad \sum_{L\in\mathcal{M}} \zeta_\sigma^L = 1. \tag{2.26}$$

Then, $\rho_\mathcal{M} = (\rho_K)_{K\in\mathcal{M}}$ being known, we define the interpolate of the density at the face $\tilde{\rho}_\sigma$ by :

$$\forall \sigma \in \mathcal{E}_{\text{int}}, \quad \tilde{\rho}_\sigma = \sum_{L\in\mathcal{M}} \zeta_\sigma^L \rho_L \tag{2.27}$$

In practice, the cells used in Relation (2.26) are chosen as close as possible to σ, and a convex interpolation (*i.e.* positive reals $\left(\zeta_\sigma^L\right)$) is preferred each time it is possible. For structured discretization, the value at the internal face $\sigma = K|L$ is obtained as a weighted average of ρ_K and ρ_L.

In the general case, the computation of the coefficients (ζ_σ^K) is performed as follows :

- We first consider several possible families $(\zeta_\sigma^M)_{M\in\mathcal{M}}$ such that (2.26) holds : for an internal face $\sigma = K|L$, we consider all the families $(\zeta_\sigma^M)_{M\in\mathcal{M}}$ which satisfy (2.26) and are such that $\zeta_\sigma^M = 0$ except for $M = K$, $M = L$, and for one (in 2D) or two (in 3D) cell(s) M which share a face with K or L ; for an external face of a cell K, we consider all the families $(\zeta_\sigma^M)_{M\in\mathcal{M}}$ which satisfy (2.26) and are such that $\zeta_\sigma^M = 0$ except for $M = K$ and for two (in 2D) or three (en 3D) other cells M sharing a face with K.

- Then we have to choose among the obtained families. We first choose among the families which yield a convex combination in (2.26) (*i.e.* which satisfy $\zeta_\sigma^K \geq 0,\ \forall K \in \mathcal{M}$), if any. If, for one of these convex combinations, only two coefficients differ from zero (which means that the center of mass of the face $\boldsymbol{x}_\sigma$ is aligned with the centroids of two cells), then it is chosen for the computations. Otherwise, for each combination, we compute the real number $\zeta = \max_{\zeta_\sigma^K \neq 0} |\zeta_\sigma^K - 0.5|$ and choose the combination which leads to the minimum value for ζ ; loosely speaking, we thus pick the configuration where $\boldsymbol{x}_\sigma$ is best located "at the center" of the convex set. If there is no convex combination, we turn to non-convex ones (which is almost always the case for an external face), and choose once again the one which is characterized by the lowest parameter ζ.

Limitation precedure – Let $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = K|L$, and let us suppose that the flow goes from K to L, *i.e.* $F_{K,\sigma} \geq 0$. We now recall the conditions which were used to prove that the density and the internal energy remain positive, gathering the condition used for the cell K and the condition used for L. For the density, we get that there exists $\alpha_\sigma^\rho \in [0,1]$, $\beta_\sigma^\rho \in [0,1]$ and $M_\sigma^\rho \in \mathcal{M}$ such that

$$\left| \begin{array}{l} \rho_\sigma - \rho_K = \alpha_\sigma^\rho\,(\rho_K - \rho_{M_\sigma^\rho}), \\ \rho_\sigma - \rho_L = \beta_\sigma^\rho\,(\rho_K - \rho_L). \end{array} \right. \tag{2.28}$$

Similarly, we have for the internal energy that there exists $\alpha_\sigma^e \in [0,1]$, $\beta_\sigma^e \in [0,1]$ and $M_\sigma^e \in \mathcal{M}$ such that

$$\left| \begin{array}{l} e_\sigma - e_K = \alpha_\sigma^e\,(e_K - e_{M_\sigma^\rho}), \\ e_\sigma - e_L = \beta_\sigma^e\,(e_K - e_L). \end{array} \right. \tag{2.29}$$

For the sake of simplicity, we suppose that the "upstream cells" M_σ^ρ and M_σ^e are the same and, from now on, we denote this cell by M_σ. We have shown in [53] that Equation (2.28) (respectively Equation (2.29)) define an admissible interval for ρ_σ (resp. e_σ), and that a limitation procedure may be obtained by just projecting the tentavive value for the density at the face $\tilde{\rho}_\sigma$ (resp. $\tilde{e}_\sigma$) on this interval. Here, the situation is more complicated, since we also need to comply with the condition required for the scheme to keep the pressure constant at contact discontinuities, which states that the product $\rho_\sigma\, e_\sigma$ must be equal to $\rho_K\, e_K$ and $\rho_L\, e_L$, of course as soon as these quantities are the same (recall that we use here the fact that the equation of state is such that the pressure only depends on the product $\rho\, e$). In fact, we use here the more restrictive assumption that $\rho_\sigma\, e_\sigma$ is a convex combination of $\rho_K\, e_K$ and $\rho_L\, e_L$, *i.e.* that there exists $\kappa_\sigma \in [0,1]$ so that :

$$\rho_\sigma\, e_\sigma = \kappa_\sigma\, \rho_K\, e_K + (1 - \kappa_\sigma)\, \rho_L\, e_L. \tag{2.30}$$

Our aim is now to find an admissible interval for ρ_σ and e_σ such that (2.28), (2.29) and (2.30) hold.

Let us first have a look on (2.28). Combining both relations, we obtain that α_σ^ρ and β_σ^ρ satisfy :

$$\beta_\sigma^\rho = 1 - \frac{\alpha_\sigma^\rho}{r_\sigma^\rho}, \qquad \text{with } r_\sigma^\rho = \frac{\rho_L - \rho_K}{\rho_K - \rho_{M_\sigma}}. \tag{2.31}$$

From this relation, it appears that (2.28) is satisfied (or, in other words, $\alpha_\sigma^\rho \in [0,1]$ and $\beta_\sigma^\rho \in [0,1]$) provided that α_σ^ρ satisfies :

$$0 \leq \alpha_\sigma^\rho \leq \min\left(1,\ r_\sigma^\rho\right)^+,$$

with still the notation $a^+ = \max(a,0)$, for $a \in \mathbb{R}$. This observation suggests the following strategy : thanks to the link between the value of ρ_σ and e_σ induced by Equation (2.30), try to express the coefficients α_σ^e and β_σ^e as a fonction of α_σ^ρ, and then express the limitations produced by (2.29) as limitations for α_σ^ρ. To this purpose, we remark that the second relation of (2.28) reads

$\rho_\sigma = \beta^\rho_\sigma \rho_K + (1-\beta^\rho_\sigma)\rho_L$, and arbitrarily suppose that the product $\rho_\sigma e_\sigma$ is given by the same interpolation between neighbouring cells values :

$$\rho_\sigma e_\sigma = \beta^\rho_\sigma \rho_K e_K + (1-\beta^\rho_\sigma)\rho_L e_L,$$

i.e. we take $\kappa = \beta^\rho_\sigma$ in (2.30). Note that many other choices would be possible, as, for instance, $\kappa = \beta^e_\sigma$. Dividing by ρ_σ yields :

$$e_\sigma = \frac{\beta^\rho_\sigma \rho_K}{\rho_\sigma} e_K + \frac{(1-\beta^\rho_\sigma)\rho_L}{\rho_\sigma} e_L.$$

Since the right hand side may be seen as a convex interpolation between e_K and e_L, we get :

$$\beta^e_\sigma = \frac{\rho_K}{\rho_\sigma}\beta^\rho_\sigma, \tag{2.32}$$

and also the fact that $\beta^e_\sigma \in [0,1]$ (which may also be inferred directly from the fact that $\rho_\sigma = \beta^\rho_\sigma \rho_K + (1-\beta^\rho_\sigma)\rho_L \geq \beta^\rho_\sigma \rho_K$). From (2.29), we derive the following relation, which is the analogue of (2.31) :

$$\beta^e_\sigma = 1 - \frac{\alpha^e_\sigma}{r^e_\sigma}, \qquad \text{with } r^e_\sigma = \frac{e_L - e_K}{e_K - e_{M_\sigma}}. \tag{2.33}$$

So $\alpha^e_\sigma = (1-\beta^e_\sigma) r^e_\sigma$, and substituting β^e_σ by its expression (2.32) and then expressing β^ρ_σ as a function of α^ρ_σ thanks to (2.31) yields, after some agebraic manipulations :

$$\alpha^e_\sigma = \frac{\rho_L}{\rho_\sigma}\frac{r^e_\sigma}{r^\rho_\sigma}\alpha^\rho_\sigma. \tag{2.34}$$

From this expression, we get that (2.29) (or, more precisely speaking, $\alpha^e_\sigma \in [0,1]$, since the fact that $\beta^e_\sigma \in [0,1]$ is already known) will be satisfied (together with (2.28)) if α^ρ_σ satisfies :

$$0 \leq \alpha^\rho_\sigma \leq \min\Big(1,\ r^\rho_\sigma,\ \frac{\rho_\sigma}{\rho_L}\frac{r^\rho_\sigma}{r^e_\sigma}\Big)^+.$$

This relation still does not provide an interval for α^ρ_σ, since it involves ρ_σ which expression itself involves α^ρ_σ. But we just need now to replace ρ_σ by an explicit lower bound. As we already remarked, $\alpha^\rho_\sigma = 0$ is always an admissible value, and so ρ_K is also an admissible value for ρ_σ. Thus ρ_σ will be obtained by a projection of the tentative value $\tilde\rho_\sigma$ on an interval containing ρ_K, which ensures that $\rho_\sigma \geq \min(\rho_K,\ \tilde\rho_\sigma)$. Consequently, we finally choose for admissible interval for α^ρ_σ the interval $\mathcal{I}_\alpha$ given by :

$$\mathcal{I}_\alpha = \Big[0,\ \min\Big(1,\ r^\rho_\sigma,\ \frac{\min(\rho_K,\ \tilde\rho_\sigma)}{\rho_L}\frac{r^\rho_\sigma}{r^e_\sigma}\Big)^+\Big]. \tag{2.35}$$

The admissible interval for the density is thus $\mathcal{I}_\rho$ with

$$\mathcal{I}_\rho = \Big\{\rho_K + \alpha\,(\rho_K - \rho_{M^\rho_\sigma}),\ \alpha \in \mathcal{I}_\alpha\Big\}. \tag{2.36}$$

The limitation algorithm is, knowing $\tilde\rho_\sigma$, to compute ρ_σ by projection on $\mathcal{I}_\rho$, which yields α^ρ_σ. The coefficient α^e_σ is given by (2.34) and e_σ is computed from the first relation of (2.29).

We should note that the accuracy of this algorithm depends on the considered variable :

- The approximation for ρ, in the absence of limitation, is second order in space.
- Then we derive from this approximation a value for the pressure, using the same wheighted average between the neighbouring cells values. In a structured discretization, without limitation, this averaging formula is also the interpolation one, and thus the face pressure is also given by a second-order formula. On the opposite, for unstructured discretizations, where the interpolation formula (2.27) (more exactly, the analogue of (2.27) written for p) and the second relation of (2.28) (still replacing ρ by p) are not the same, the second-order accuracy is lost.

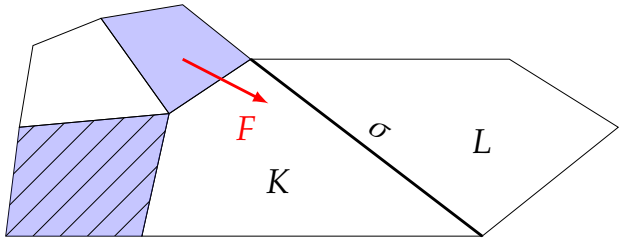


Figure 2.3 – Notations for the definition of the limitation process.
Control volumes of the set V_K for $\sigma = K|L$, with a constant advection field $\mathbf{F}$: in blue upwind cells – hatched unique opposite cell

- Finally, the internal energy is obtained from the density and the pressure (so its approximation is, in general, only first order, since it only satisfies that the face value lies between the values in the two neighbour cells), and potentially generates limitations of the fluxes. In particular, the definition (2.35) of $\mathcal{I}_\alpha$ implies that α_σ^ρ vanishes as soon as either r_σ^ρ or r_σ^e is non-positive, *i.e.* as soon as either ρ or e presents a local extrema.

A lot of variants of the present scheme may be designed, among which the following ones :

- As mentioned above, the roles of ρ and e may be switched, in the sense that one may choose a "limited second order" interpolation for e and p, and deduce ρ from these values ; to this purpose, one must choose $\kappa = \beta_\sigma^e$, and start from the non-limited approximation of e instead of the one for ρ.
- The present algorithm does not ensure that the value taken for e at the face will lie in-between the second-order approximation and the upwind value. In the case of one-dimensional or structured discretizations, it may be done by restricting the admissible range for β_σ^e to $\beta_\sigma^e \in [\tilde{\beta}_\sigma, 1]$, where $\tilde{\beta}_\sigma$ is the weight which yields for e_σ the second-order average between e_K and e_L. For uniform meshes, the admissible interval is thus $\beta_\sigma^e \in [1/2, 1]$. The results of such a choice would just be an additional limitation of the algorithm.
- Finally, from a theoretical point of view, the upstream cell M_σ used in the first relation of (2.28) and of (2.29) may be chosen arbitrarily in the mesh, but any reasonable implementation of the algorithm should restrict this choice to the vicinity of the face σ.

Two different choices of are implemented for the choice of the set of cells V_K in which M_σ is searched for :

(a) V_K is defined as the set of "upstream cells" to K, *i.e.* $V_K = \{L \in \mathcal{M},\ L$ shares a face σ with K and $F_{K,\sigma} < 0\}$,

(b) when this makes sense (*i.e.* with a mesh obtained by Q_1 mappings from the $(0,1)^d$ reference element), V_K may be chosen as the opposite cells to σ in K.

In the tests performed here in the remaining of this paper, M_σ is always the opposite neighbour of the upwind cell K (see Figure 2.3).

2.3.5 Artificial viscosity

Numerical experiments (see Section 2.4) show some oscillations at shocks with the Upwind scheme developed in [39], probably due to the fact that the aartificial viscosity brought by the upwinding behaves as the material velocity only, and not as the celerity of waves ; with the MUSCL algorithm, this phenomenon is even enhanced since the numerical diffusion is reduced. To cure this problem, we add some viscosity in the discrete momentum balance equation (while the numerical diffusion in the other equations is left unchanged) and only where it is needed, that is at the shocks. To this purpose, we test here two different methods, inspired from the works [31] and [42] respectively, where the diffusion is evaluated thanks to an *a posteriori* analysis of the solution.

The aim of this section is to describe the computation of this artificial viscosity, *i.e.* the parameter ν_ϵ^{n+1} in Equation (2.5d). The process followed for this computation is to first define a "cell diffusion parameter" ζ_K^{n+1} on each primal cell K, and then to deduce the "dual face viscosity" from these cell values. For this latter step, two situations may be encountered :

- The dual face ϵ is strictly included in a primal cell K ; in this case, we take $\nu_\epsilon^{n+1} = |\epsilon|\, \zeta_K^{n+1}$.
- The dual face ϵ lies on the boundary of four primal cells (in the MAC case) ; then we take :

$$\nu_\epsilon^{n+1} = |\epsilon|\, \frac{1}{4} \sum_{K \in \mathcal{N}(\epsilon)} \zeta_K^{n+1},$$

  where $\mathcal{N}(\epsilon)$ is the set of cells adjacent to ϵ.

The remaining of this section is devoted to the description of the computation of the $(\zeta_K^{n+1})_{K\in\mathcal{M}}$. According to this computation, these parameters are "homogeneous to the space step h" (or, equivalently, the time step, since the CFL number is bounded away from zero and lower than 1), in the sense that ζ_K^{n+1}/h (formally) does not tend neither to zero or infinity when the space and time steps tend to zero. Consequently, the artificial diffusion term in (2.5d) produces a viscosity which scales as h^2 in smooth zones of the solution, as in [31, 42]. However, the scheme proposed here presents two essential differences with these previous works : first, artificial diffusion is added only in the momentum balance equation (while it is introduced in all the equations in [31, 42]) ; second, a first-order upwind discretization is kept in the convection term of the momentum balance equation.

Entropic viscosity

The method, developed in [31], is based on the entropy inequality satisfied by the weak solutions of the system, which reads :

$$\partial_t \eta + \mathrm{div}(\eta u) \leq 0,$$

this inequality becoming an equality in the zones where the solution is smooth and at contact discontinuities. The idea is to compute the numerical diffusion in the momentum balance equation as a function of the entropy production, to introduce an additional numerical dissipation at shocks. We use here the usual physical definition of the entropy :

$$\eta(p, \rho) = \frac{\rho}{\gamma - 1} \log\Big(\frac{p}{\rho^\gamma}\Big).$$

The first step consists in computing the residual of the discrete entropy equations in every element K of the mesh :

$$\mathcal{R}_K^{n+1} = \frac{1}{\delta t}\Big(\eta_K^{n+1} - \eta_K^n\Big) + \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} |\sigma|\, \eta_\sigma^n\, u_{K,\sigma}^n,$$

where η_σ^n stands for a centered approximation of the entropy at the faces σ. Then we compute a tentative diffusion parameter by :

$$\tilde{\zeta}_K^{n+1} = c_E\, \rho_K^{n+1}\, h_K\, |\mathcal{R}_K^{n+1}|. \tag{2.37}$$

where h_K is the diameter of the cell K and c_E is a calibration parameter. Note that $\mathcal{R}_K^{n+1}$ is a formal discretization of $\partial_t \eta + \mathrm{div}(\eta u)$, and thus is a quantity formally independent of the space and time steps ; consequently, $\tilde{\zeta}_K^{n+1}$ scales as h_K. Then this parameter is limited to the (range of the) diffusion generated by the first-order upwinding of the convection operator. For any face σ of the primal mesh adjacent to a cell L, this latter reads $\zeta_\sigma^{n+1} = |\rho_\sigma^n\, u_{L,\sigma}^n|/2$, with ρ_σ^n the face

density used in the mass balance equation. We then define a maximum value for the diffusion parameter by :

$$\zeta_{max,K}^{n+1} = c_{max}\ \max\Big((\zeta_\sigma^{n+1})_{\sigma\in\overline{\mathcal{E}}(K)}\Big),$$

where c_{max} is once again a calibration parameter and $\overline{\mathcal{E}}(K)$ stands for a set of faces located in the vicinity of K, whichs includes at least $\mathcal{E}(K)$. In applications realized here, this set is in fact much larger, since it is composed of the faces of the 3 left and 3 right cells to K in one dimension, and for structured 2D discretizations, the faces of the cells of a 7×7 patch centered on K. Then we obtain a second tentative diffusion parameter by :

$$\tilde{\zeta}_K^{n+1} = \min\Big(\tilde{\zeta}_K^{n+1}, \zeta_{max,K}^{n+1}\Big).$$

Finally, ζ_K^{n+1} is computed as a weighted average of the parameters $(\tilde{\zeta}_L^{n+1})_{L\in\mathcal{M}}$ over a patch around K. In one dimension, this patch includes the left and right cells of K and K itself, and the weight is $2/3$ for K and $1/3$ for the other cells. For structured discretizations in two dimensions, we use a 3×3 patch centered on K, the weight is $8/9$ for K and $1/9$ for the other cells.

WLR viscosity

The second method is based on [42]. We first briefly recall the ideas developped in this work, for a generic conservation law of unknown w and flux $\boldsymbol{f}$:

$$\partial_t w + \mathrm{div}\boldsymbol{f}(w) = 0. \tag{2.38}$$

A weak solution of (2.38) is defined by :

$$\mathcal{W}(w,\phi) = \int_0^T\int_\Omega\Big[w(\boldsymbol{x},t)\,\partial_t\phi(\boldsymbol{x},t) + \boldsymbol{f}(\boldsymbol{x},t)\cdot\nabla\phi(\boldsymbol{x},t)\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega w(\boldsymbol{x},0)\,\phi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0,$$

for all test functions $\phi\in C_0^1(\Omega\times\ [0,T))$. This identity is used in [42] to build, on the basis of a discrete solution w_h obtained by a finite difference method, a measure of the local regularity of the solution. The discrete solution is identified to a function of time and space, specific test functions ϕ (one per cell, let us say $(\phi_K)_{K\in\mathcal{M}}$ to keep notations consistent with the rest of the present paper) are defined, and the quantities $(\mathcal{W}(w_h,\phi_K))_{K\in\mathcal{M}}$ are used to track the discontinuities. On their basis, a stabilizing diffusion is then introduced in the scheme.

Here, we use an adaptation of this strategy for Euler equations and a finite volume scheme. First, we do not compute the residual $\mathcal{W}$ for each equation, but just for the mass balance :

$$\mathcal{W}(\rho,\boldsymbol{u},\phi) = \int_0^T\int_\Omega\Big[\rho(\boldsymbol{x},t)\,\partial_t\phi_t(\boldsymbol{x},t) + \rho(\boldsymbol{x},t)\,\boldsymbol{u}(\boldsymbol{x},t)\cdot\nabla\phi(\boldsymbol{x},t)\Big]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega \rho(\boldsymbol{x},0)\,\phi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}.$$

As for the finite difference scheme treated in [42], we identify the discrete solution to piecewise functions. We thus define :

$$\rho_\Delta(\boldsymbol{x},t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} \rho_K^n\,\mathcal{X}_K(\boldsymbol{x})\mathcal{X}_{(t_n,t_{n+1})}(t),$$

$$\boldsymbol{u}_\Delta(\boldsymbol{x},t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} \boldsymbol{u}_K^n\,\mathcal{X}_K(\boldsymbol{x})\mathcal{X}_{(t_n,t_{n+1})}(t),$$

where $\mathcal{X}_K$ and $\mathcal{X}_{(t_n,t_{n+1})}$ stand for the characteristic functions of the cell K and the interval (t_n, t_{n+1}) and $\boldsymbol{u}_K$ is an interpolate of the velocity on the primal mesh :

$$\forall K \in \mathcal{M}, \quad \boldsymbol{u}_K = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} |D_{K,\sigma}| \, \boldsymbol{u}_\sigma.$$

The next step is to introduce a set of local polynomials (ϕ_K^n), for every $K \in \mathcal{M}$ and for $0 \le n \le N-1$, to be used as test functions. We postpone the exact definition of these polynomials for a while, and only state here the approximation property that they have to satisfy for the subsequent theory to hold. For any $\phi \in C_0^1(\Omega \times [0,T])$, we suppose that there exists $(\beta_K^n)_{K\in\mathcal{M},\, 0\le n\le N-1} \subset \mathbb{R}$ such that :

$$\phi(\boldsymbol{x}, t) = \sum_{K\in\mathcal{M}} \sum_{n=0}^{N-1} \beta_K^n \, \phi_K^n(\boldsymbol{x}, t) + O(\Delta^2), \tag{2.39}$$

where $\Delta = \max(h, \delta t)$. If we suppose that the test functions (ϕ_K^n) are local in the sense that their integral behaves like $\delta t \, h^d$ (since the measure of their support also behaves like $\delta t \, h^d$), this condition ensures that

$$\mathcal{W}(\rho_\Delta, \boldsymbol{u}_\Delta, \phi) = \sum_{K\in\mathcal{M}} \sum_{n=0}^{N-1} \Big(\beta_K^n \mathcal{W}_K^n + O(\Delta^{d+3})\Big),$$

where the weak local residual (WLR), $\mathcal{W}_K^n$, takes the following expression :

$$\mathcal{W}_K^n = \int_0^T \int_\Omega \rho_\Delta(\boldsymbol{x}, t) \, \partial_t \phi_K^n(\boldsymbol{x}, t) + \rho_\Delta(\boldsymbol{x}, t) \, \boldsymbol{u}_\Delta(\boldsymbol{x}, t) \cdot \nabla \phi_K^n(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t. \tag{2.40}$$

In the one-dimensional case, it is proven in [42] that, under the assumption (2.39), these weak local residuals have the following properties :

$$|\mathcal{W}_K^n| \text{ behaves as } \left| \begin{array}{ll} \Delta, & \text{near shock waves,} \\ \Delta^\alpha, & 1 < \alpha \le 2, \text{ near contact waves,} \\ \Delta^3, & \text{in smooth regions.} \end{array} \right.$$

These results are not proven in two and three dimensions ; however the same behaviour for $|\mathcal{W}_K^n|/\Delta^{d-1}$ is observed on numerical tests. These residuals are used to define a tentative diffusion coefficient by :

$$\tilde{\zeta}_K^{n+1} = c_m \, \frac{1}{\delta t \, \Delta^{d-1}} \, |\mathcal{W}_K^{n+1}|, \tag{2.41}$$

with c_m a calibration parameter. Finally, as in the previous section, ζ_K^{n+1} is computed as a weighted average of the parameters $(\tilde{\zeta}_L^{n+1})_{L\in\mathcal{M}}$ over the same patch around K : in one dimension, this patch includes the left and right cells of K and K itself, and the weight is 2/3 for K and 1/3 for the other cells ; for structured discretizations in two dimensions, we use a 3×3 patch centered on K, the weight is 8/9 for K and 1/9 for the other cells.

In the applications presented in Section 2.4 below, we use for the polynomials $(\phi_K^n)_{K\in\mathcal{M},\, 0\le n\le N-1}$ the same definition based on B-splines as in [42]. The definition of these polynomials, together with the expression of the residuals for structured grids, is given in appendix.

2.4 Numerical results

We present in this section numerical tests to assess the behaviour of the scheme. We first address the accuracy of the MUSCL interpolation and artificial viscosity techniques on several 1D numerical test cases. A convergence rate analysis is performed, to compare Upwind and

MUSCL interpolations. We then deal with computation of high speed inviscid flows, using classical benchmarks for Euler solvers. Since these 2D computations are performed using the MAC space discretization, we complete the study by computing a high speed inviscid flow around a cylinder with a Rannacher-Turek space discretization. For all the computations, the fluid obeys the equation of state (2.1d) with $\gamma = 1.4$.

2.4.1 One Dimension

2.4.2 One dimensional tests

This section is devoted to the computation of one-dimensional Rieman problems. In all the tests, the computational domain is $\Omega = (0, 1)$.

Single contact discontinuity wave – First of all we give a numerical evidence of the necessity of a correlation between the density and the internal energy. To this purpose, we compute a Riemann problem consisting in a single contact discontinuity wave travelling to the right of the domain. It corresponds to the following initial conditions :

$$\text{left state : } \begin{bmatrix} \rho_L = 14.282 \\ u_L = 8.6898 \\ p_L = 1691.6 \end{bmatrix}; \qquad \text{right state : } \begin{bmatrix} \rho_R = 31.043 \\ u_R = 8.6898 \\ p_R = 1691.6 \end{bmatrix}.$$

The pressure fields obtained, at $t = 0.02$, respectively with and without a correlation between approximation of the density and the internal energy at faces of the cells, are shown on Figure 2.4. The computation referred to as "non-correlated approximation" is performed by applying the interpolation/limitation procedure used for the density to the internal energy also, thus without imposing to the product $\rho\, e$ at the face to be an interpolation of $\rho\, e$ at the two neighbour cells. As one can see, this approximation generates oscillations of the pressure at the contact discontinuity ; we even observe in our computations that these oscillations tend to get worse and worse with time. In addition, pressure variations appear at the locations of zero-amplitude 1-shock and 3-shock waves. On the opposite, the proposed scheme yields a constant pressure with respect to time and space, as in the continuous solution.

Two classical Riemann problems – We will now compare the Upwind and the MUSCL schemes on two Riemann problems classically used in the literature, namely Test 4 and Test 5 from [62, Chapter 4]. In Test 4, the left and right states are :

$$\text{left state : } \begin{bmatrix} \rho_L = 1 \\ u_L = 0 \\ p_L = 0.01 \end{bmatrix}; \qquad \text{right state : } \begin{bmatrix} \rho_R = 1 \\ u_R = 0 \\ p_R = 100 \end{bmatrix}.$$

The solution consists of a shock travelling to the left and a rarefaction wave travelling to the right, separated by the contact discontinuity. We first evaluate the stability of the scheme, by performing computations with a (constant) time step larger and larger, until obtaining a blow-up of the computation ; for $h = 0.001$, strong oscillations are observed for $\delta t = h/17$ and the computation fails for $\delta t = h/17$ (to be related to a maximal celerity of waves close to 17 also). This stability limit is obtained without artificial viscosity ; adding such a term reduces the stability domain. Note hovever that, since the artificial diffusion is limited by the viscosity generated by the first-order Upwind scheme, the stability domain still keeps the form $\delta t \leq C\, h$ (and not $\delta t \leq C\, h^2$, which would be characteristic of a viscosity constant (in order of magnitude) with respect to the space step). Results obtained at $t = 0.035$ with $h = 0.001$ and $\delta t = h/30$ are reported on Figure 2.5. As seen on the internal energy and density profiles, the numerical diffusion at the contact discontinuity is drastically reduced by the MUSCL approximation. At the shock, the results of the Upwind and the MUSCL scheme look similar : on one hand, the compressive effect of the shock prevents the Upwind scheme to be too dissipative, and, on the other hand,

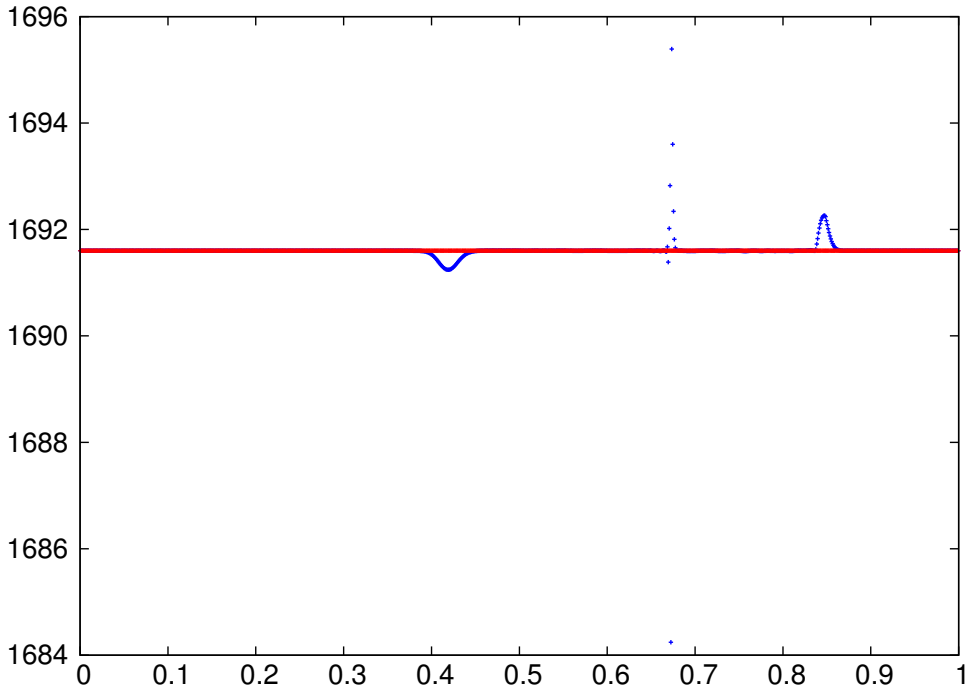


Figure 2.4 – A 1D Riemann problem : single contact discontinuity – $h = 0.001$ and $\delta t = h/40$ – Pressure at $t = 0.02$. Results obtained with a correlated (resp. non-correlated) face approximation for ρ and e are drawn in red (resp. in blue). The analytical solution is the same as the discrete solution obtained with the correlated approximation (non-visible black line).

the numerical dissipation introduced by the limitation procedure in the MUSCL scheme seems to be sufficient.

In Test 5, the initial conditions are :

$$\text{état gauche : } \begin{bmatrix} \rho_L = 5.99924 \\ u_L = 19.5975 \\ p_L = 460.894 \end{bmatrix}; \qquad \text{état droit : } \begin{bmatrix} \rho_R = 5.99242 \\ u_R = -6.19633 \\ p_R = 46.0950 \end{bmatrix}.$$

In this test, the genuinely non-linear waves are two shocks travelling to the left. The numerical stability analysis shows that, without artificial viscosity and for $h = 0.001$, the scheme blows up for $\delta t \simeq h/29$ (while the greatest wave celerity is close to 30 in the left state). Results obtained at $t = 0.035$ with $h = 0.001$ and $\delta t = h/90$ are reported on Figure 2.6. One may observe on the density and the pressure some overshoots at the 3-shock with the Upwind scheme ; this phenomenon is strengthened with the MUSCL algorithm (results are not shown here). This problem is completely cured by the WLR viscosity introduced in Section 2.3.5, with $c_m = 2$ in Relation (2.41).

A convergence study – In addition, we perform a convergence study, successively dividing by two the space and time steps (so keeping the CFL number constant). We use the same test as in [39], *i.e.* Test 3 in [62, Chapter 4]. The left and right states are given by :

$$\text{left state : } \begin{bmatrix} \rho_L = 1 \\ u_L = 0 \\ p_L = 1000 \end{bmatrix}; \qquad \text{right state : } \begin{bmatrix} \rho_R = 1 \\ u_R = 0 \\ p_R = 0.001 \end{bmatrix}.$$

The differences between the computed and analytical solution at $t = 0.012$, measured in $L^1(\Omega)$ norm, are reported in the following table.

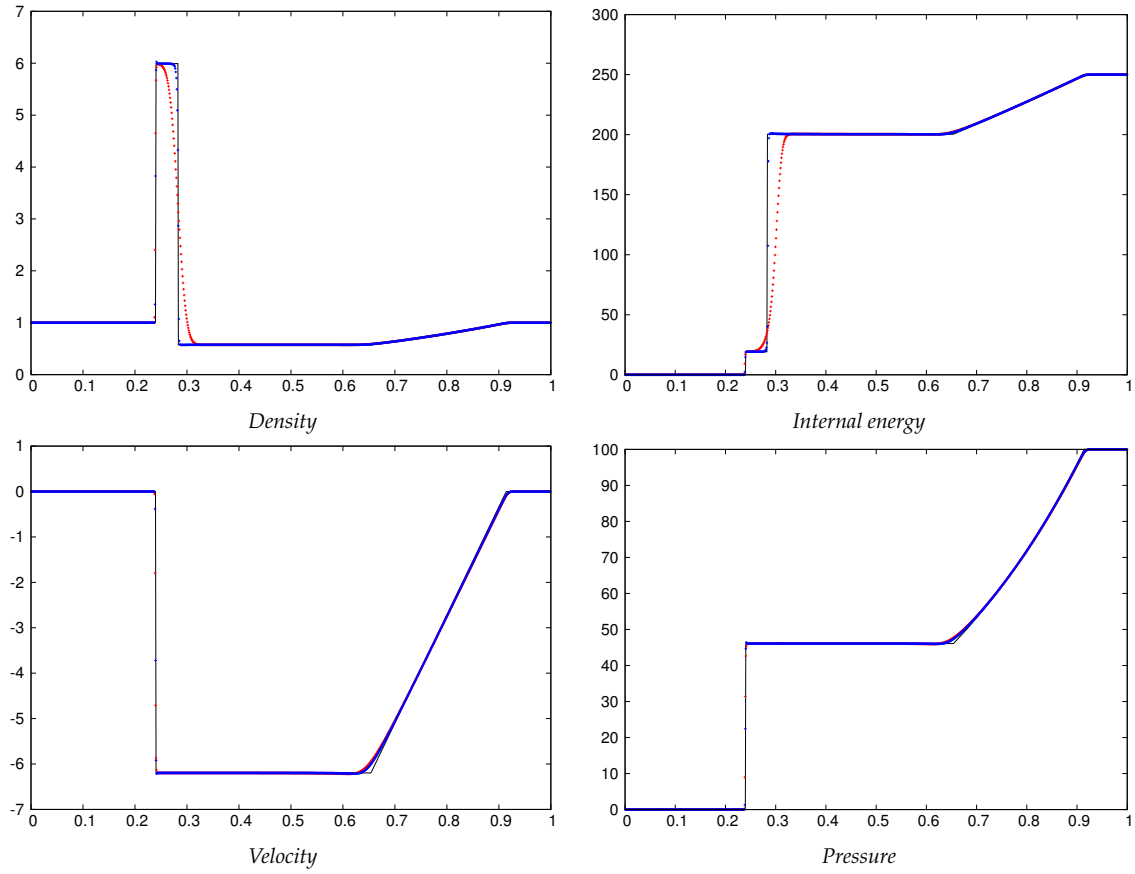


Figure 2.5 – A 1D Riemann problem : Test 4 of [62, Chapter 4]) – $h = 0.001$ and $\delta t = h/30$ – Results at $t = 0.035$. The Upwind and MUSCL solutions are drawn in red and blue respectively, the analytical solution corresponds to the black line.

space step	$h_0 = 0.001$	$h_0/2$	$h_0/4$	$h_0/8$	$h_0/16$
$\|\rho - \bar{\rho}\|_{L^1(\Omega)}$	0.0108	0.0058	0.0025	0.0012	0.0007
$\|p - \bar{p}\|_{L^1(\Omega)}$	1.2827	0.6734	0.3316	0.1800	0.1044

As a reminder, we give the results obtained with the Upwind scheme.

space step	$h_0 = 0.001$	$h_0/2$	$h_0/4$	$h_0/8$	$h_0/16$
$\|\rho - \bar{\rho}\|_{L^1(\Omega)}$	0.0651	0.0455	0.0310	0.0217	0.0153
$\|p - \bar{p}\|_{L^1(\Omega)}$	1.87	1.05	0.530	0.284	0.164

As one can see, the convergence rate is improved by the MUSCL interpolation. Indeed, for variables which are not constant through contact discontinuities, the convergence rate is now close to 2/3. For the other variables, it is slightly improved.

The symetrical double-shock - To conclude this one dimension part, we introduce a pathological case, where the initial data consists in opposite initial velocities, the density and pressure being constant all over Ω. Precisely speaking, we take :

$$\text{left state : } \begin{bmatrix} \rho_L = 5.99924 \\ u_L = 19.5975 \\ p_L = 460.894 \end{bmatrix}; \qquad \text{right state : } \begin{bmatrix} \rho_R = 5.99924 \\ u_R = -19.5975 \\ p_R = 460.894 \end{bmatrix}.$$

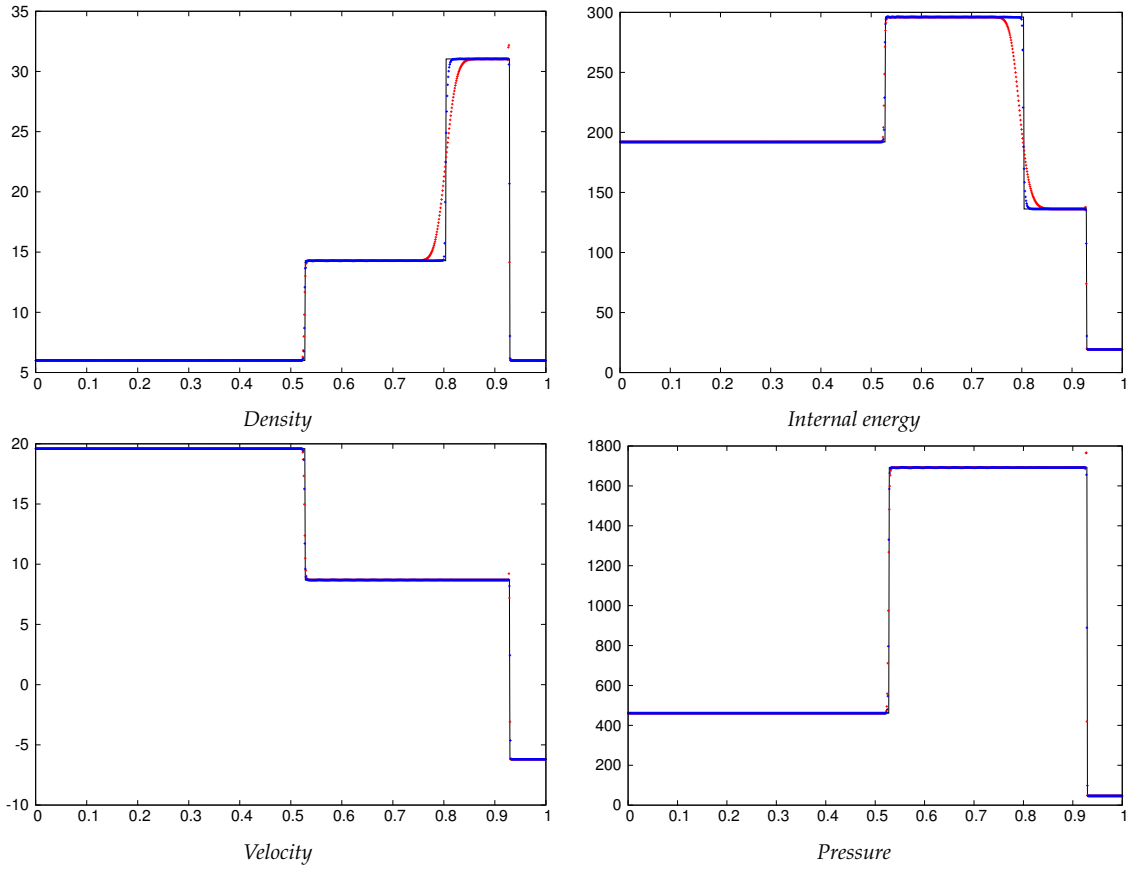


Figure 2.6 – A 1D Riemann problem : Test 5 of [62, Chapter 4] – with artificial viscosity – $h = 0.001$ and $\delta t = h/90$ – Results at $t = 0.035$. The Upwind and MUSCL solutions are drawn in red and blue respectively, the analytical solution corresponds to the black line.

The analytical solution consists in two shocks, travelling with the same velocity to the left and the right respectively, which separate the left and right initial states from a constant state, where the fluid is at rest ; the contact discontinuity is stationary, located in $x = 0.5$ and of zero amplitude. Results obtained at $t = 0.035$ with $h = 0.001$ and $\delta t = h/60$ are reported on Figure 2.7. This test case is particulary interesting because the dual convection fluxes vanish in the intermediate state. Consequently, the Upwind scheme (solution in red on the figure) does not bring any numerical viscosity at the shocks (since this viscosity is proportional to $|F_{\sigma,\epsilon}|/2$), and spurious oscillations appear in the central zone. The WLR viscosity, with $c_m = 3$, allows to drastically reduce this phenomenon. However, it also generates artificial variations at the contact discontinuity for the (possibly) discontinuous variables ; the other ones are not affected.

2.4.3 Two dimensions

Two-dimensional Riemann problems

subsectionTwo-dimensional Riemann problems

We address in this section two-dimensional Riemann problem introduced in [44]. The computational domain is $\Omega = (-0.5, 0.5)^2$ and the initial data consists in 4 quadrants in which initial data are constant. These quadrants are $\Omega_1 = (0, 0.5)^2$, $\Omega_2 = (-0.5, 0) \times (0, 0.5)$, $\Omega_3 = (-0.5, 0)^2$, $\Omega_4 = (0, 0.5) \times (-0.5, 0)$. The constant states are chosen so that the solution to the four Riemann problems associated with each interface of the quadrants consist in a single wave. There exists 19 posible configurations. All the computations of this section are performed with the MAC space discretization.

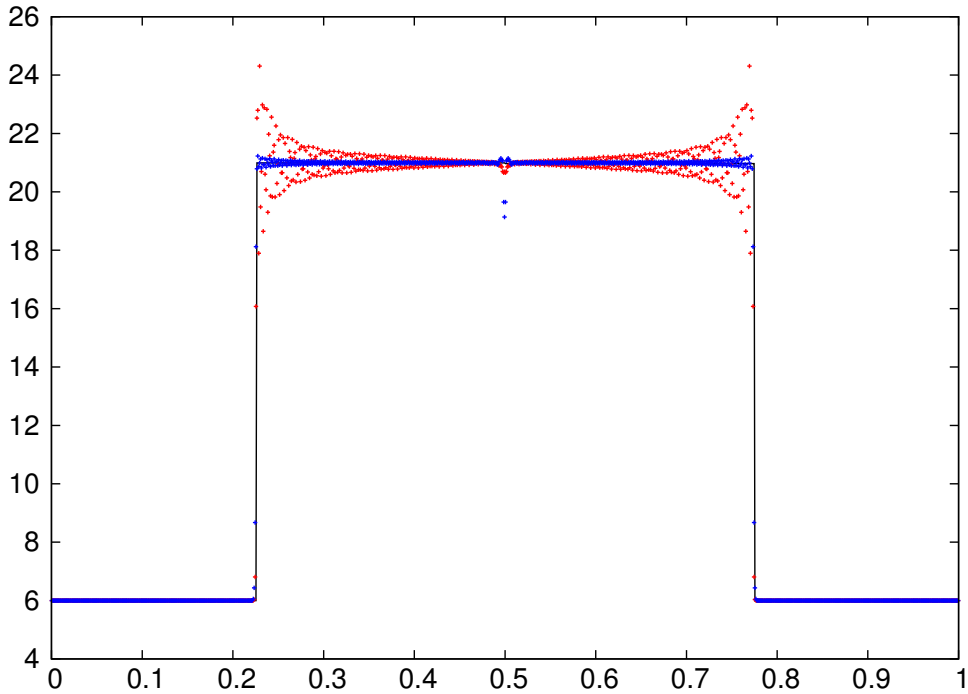


Figure 2.7 – A 1D Riemann problem : the "symetrical double shock" – $h = 0.001$ and $\delta t = h/60$ – Density at $t = 0.035$. Upwind solution without (red) and with (blue) WLR viscosity

Configurations 5 and 6 – MUSCL interpolation is primarily used to improve precision at contact discontinuity lines. To illustrate this effect, we address Configurations referred to as 5 and 6 in [44]. The initial condition is, for Configuration 5 :

$$\Omega_1 : \begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = -0.75 \\ v_1 = -0.5 \end{bmatrix} \qquad \Omega_2 : \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = -0.75 \\ v_2 = 0.5 \end{bmatrix}$$

$$\Omega_3 : \begin{bmatrix} \rho_3 = 1 \\ p_3 = 1 \\ u_3 = 0.75 \\ v_3 = 0.5 \end{bmatrix} \qquad \Omega_4 : \begin{bmatrix} \rho_4 = 3 \\ p_4 = 1 \\ u_4 = 0.75 \\ v_4 = -0.5 \end{bmatrix}.$$

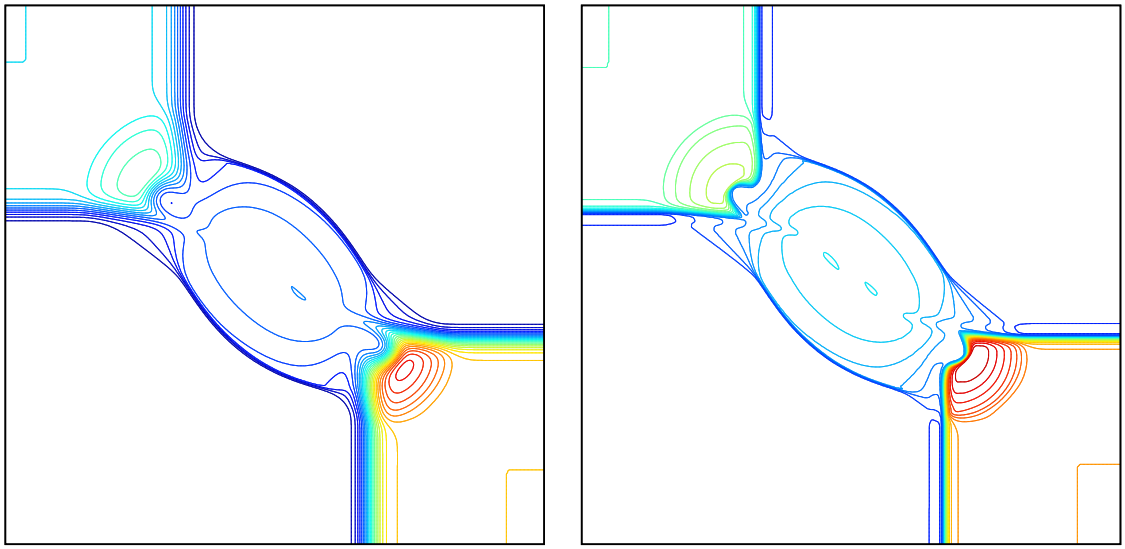
For Configuration 6, we have :

$$\Omega_1 : \begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0.75 \\ v_1 = -0.5 \end{bmatrix} \qquad \Omega_2 : \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = 0.75 \\ v_2 = 0.5 \end{bmatrix}$$

$$\Omega_3 : \begin{bmatrix} \rho_3 = 1 \\ p_3 = 1 \\ u_3 = -0.75 \\ v_3 = 0.5 \end{bmatrix} \qquad \Omega_4 : \begin{bmatrix} \rho_4 = 3 \\ p_4 = 1 \\ u_4 = -0.75 \\ v_4 = -0.5 \end{bmatrix}.$$

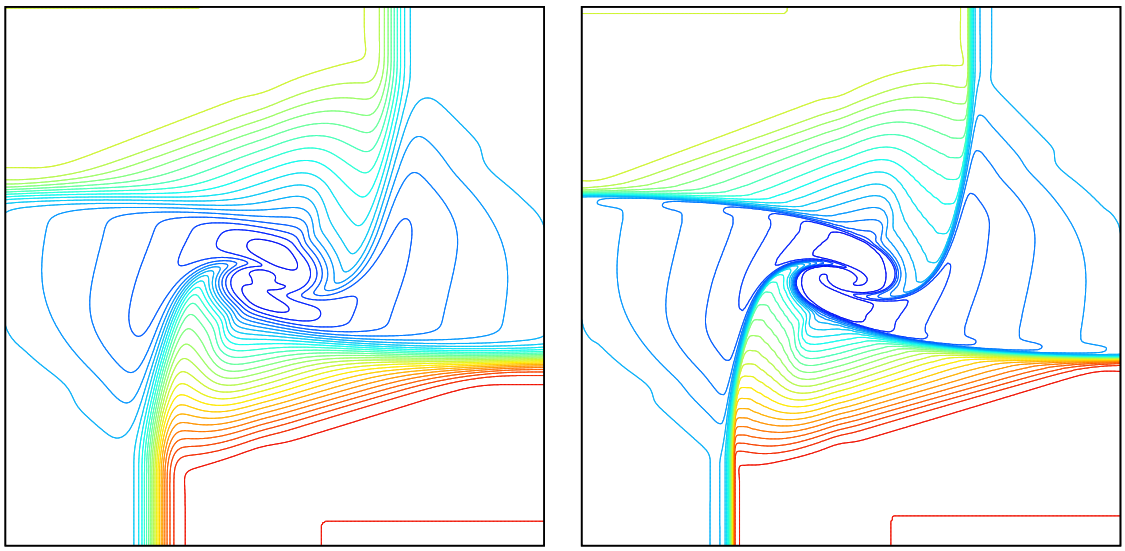
The final time is $t = 0.23$ for Configuration 5 and $t = 0.3$ for Configuration 6. In both cases, the solution results from the combination of four contact discontinuities (precisely speaking, "1D contact discontinuity", in the sense that the discontinuity line is normal to the velocity).

Results obtained at the end of the computation, with a 400×400 grid and with $\delta t = 1/(10 \times 400)$, are reported on Figures 2.8 and 2.9 respectively. As we expect, the contact discontinuities are sharper with the MUSCL interpolation. These two Riemann problems thus comfort the results obtained in 1D.



Upwind *MUSCL*

Figure 2.8 – A two-dimensional Riemann problem : Configuration 5 in [44] – Comparison of the Upwind and MUSCL schemes – $h = 1/400$ and $\delta t = h/10$ – density at $t = 0.23$.



Upwind *MUSCL*

Figure 2.9 – A two-dimensional Riemann problem : Configuration 6 in [44] – Comparison of the Upwind and MUSCL schemes – $h = 1/400$ and $\delta t = h/10$ – density at $t = 0.3$.

Configuration 4 –

We now turn to a test case composed with shockwaves to evidence the properties of entropic and WLR viscosities. The initial states are now given by :

$$\Omega_1 : \begin{bmatrix} \rho_1 = 1.1 \\ p_1 = 1.1 \\ u_1 = 0 \\ v_1 = 0 \end{bmatrix}; \qquad \Omega_2 : \begin{bmatrix} \rho_2 = 0.5065 \\ p_2 = 0.35 \\ u_2 = 0.8939 \\ v_2 = 0 \end{bmatrix};$$

$$\Omega_3 : \begin{bmatrix} \rho_3 = 1.1 \\ p_3 = 1.1 \\ u_3 = 0.8939 \\ v_3 = 0.8939 \end{bmatrix}; \qquad \Omega_4 : \begin{bmatrix} \rho_4 = 0.5065 \\ p_4 = 0.35 \\ u_4 = 0 \\ v_4 = 0.8939 \end{bmatrix}.$$

Results obtained at $t = 0.3$ on a 400×400 grid with $\delta t = 1/(10 \times 400)$ are reported on Figure 2.10. This test case is composed of 4 simple shocks. The first-order Upwind scheme yields a solution (Figure 2.10, top-right) with spurious oscillations in the downstream section of the top and right shock, in the area where the fluid is at rest. This is caused by the lack of numerical dissipation of our scheme, because the dissipation produced by the upwind interpolation vanishes with the velocity. To cure this problem, we first add a constant artificial viscosity (Figure 2.10, top-right) equal to 1/10 of the maximum upwind viscosity (3.510^{-4}) . We also plot the results obtained using the WLR and entropic viscosities (Figure 2.10, middle and bottom line, respectively). Concerning the calibration parameters, we have $c_m = 1$ for the WLR viscosity and $c_{max} = 3$, $c_E = 0.4$ for the entropic viscosity. As one can see they correctly pick up shocks and reduce oscillations inside the subsonic area surrounded by the sonic shocks. Furthermore a cutline (2.11) shows that oscillations in the downstream section of the top shock are greatly reduced by WLR viscosity (idem for the entropic viscosity but not shown here).

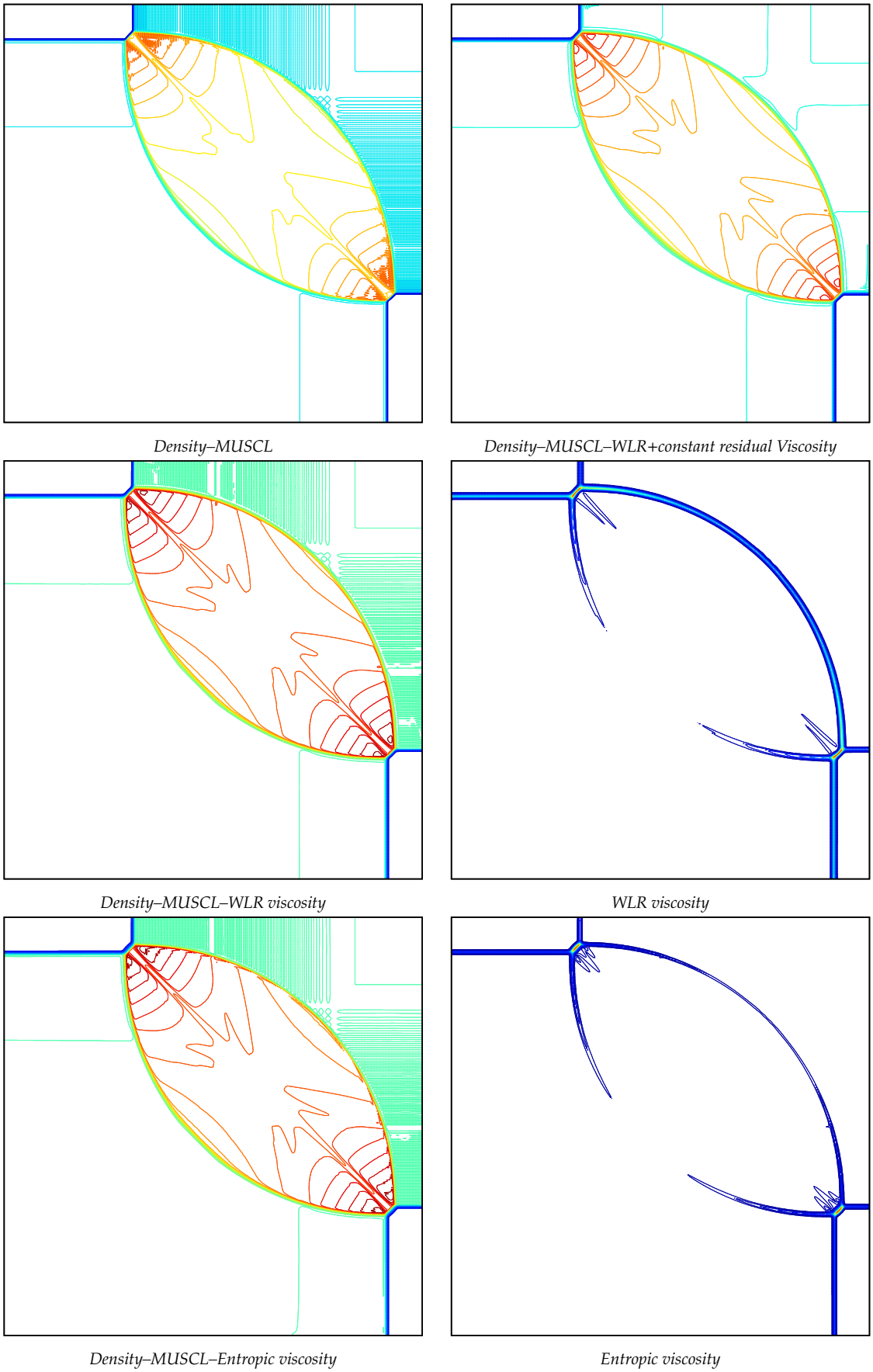


Figure 2.10 – A two-dimensional Riemann problem : Configuration 4 in [44] – $h = 1/400$ and $\delta t = h/10$ – Results at $t = 0.3$ – Both viscosities lie in $[0, 0.002]$.

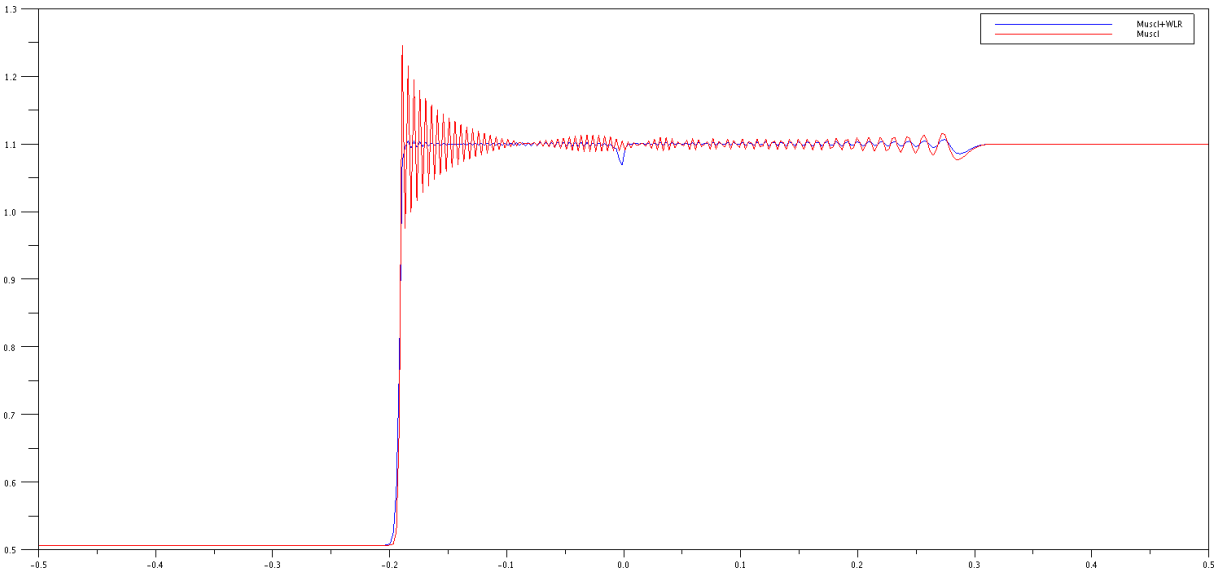


Figure 2.11 – Configuration 4 –Density– Cutline at $y = 0.48$ – Solution without artificial viscosity in red and with WLR viscosity in blue.

The Mach 3 facing step

This benchmark has been popularized in [64]. The computational domain is $\Omega = \overline{\Omega} \backslash \mathcal{S}$, with $\overline{\Omega} = (0,3) \times (0,1)$, and $\mathcal{S} = (0.6,3) \times (0,0.2)$. The time interval is $(0,4)$. A Mach 3 flow is coming from the left boundary $\{0\} \times (0,1)$ with the following properties :

$$\begin{bmatrix} \rho \\ \boldsymbol{u} \\ p \end{bmatrix} \left((0,x_2)^t, t \right) = \begin{bmatrix} 1.4 \\ (3,0)^t \\ 1 \end{bmatrix}, \qquad \forall x_2 \in (0,1),\ \forall t \in (0,4).$$

The initial data is the same as the inflow conditions :

$$\begin{bmatrix} \rho \\ \boldsymbol{u} \\ p \end{bmatrix} (\boldsymbol{x},0) = \begin{bmatrix} 1.4 \\ (3,0)^t \\ 1 \end{bmatrix}, \qquad \forall \boldsymbol{x} \in \Omega.$$

The right boundary is free, since the flow leaves the domain at a velocity greater than the sound speed. Finally we prescribe a perfect slip condition ($\boldsymbol{u} \cdot \boldsymbol{n} = 0$, where $\boldsymbol{n}$ is the unit outward normal on $\partial\Omega$).

We display on figure 2.12 the results obtained with the MAC space discretization, using the MUSCL interpolation. The mesh is a 4800×1600 uniform grid where we remove the cells included in $\mathcal{S}$. The time step is set to $t = \frac{h}{10} = 6.25e-5$, which corresponds to a CFL number approximatively equal to 0.5 with respect to the celerity of the fastest wave (equal to 4 at the inlet boundary).

Results are comparable to those presented in recent literature (see [42]). The scheme seems rather diffusive. Kevin-Helmoltz instability is often observed at the contact discontinuity issued from the Mach triple point which is not the case here. Adding some numerical dissipation does not affect the slip line (even if it is unstable). A spurious Mach reflection at the bottom boundary is observed on coarser versions of the mesh, but is greatly softened here.

Double Mach reflexion

This section is devoted to an other classical test case which consists in a Mach 10 shock impacting a wall with a 60° slope. The right state (pre-shock) initial conditions correspond to a immobile fluid and we complete the left state thanks to the Rankine-Hugoniot conditions, supposing the shock velocity is equal to $\omega = 10$ and the speed of sound in the right state is equal to 1 :

$$\begin{bmatrix} \rho_R \\ \boldsymbol{u}_R \\ p_R \end{bmatrix} = \begin{bmatrix} 1.4 \\ (0,0)^t \\ 1 \end{bmatrix}, \qquad \begin{bmatrix} \rho_L \\ \boldsymbol{u}_L \\ p_L \end{bmatrix} = \begin{bmatrix} 8 \\ 8.25\ (\sqrt{3}/2, 1/2)^t \\ 116.5 \end{bmatrix}.$$

The computational domain is $\Omega = (0,4) \times (0,1)$, and we suppose that the wall lies in the bottom of the domain, more precisely $\partial\Omega_w = (1/6,4) \times \{0\}$. At $t = 0$, the shock impinges the reflecting wall (at $x_1 = 1/6$), so the fluid is in the left state for $x_1 \leq 1/6 + x_2/\sqrt{3}$ and in the right state in the rest of the domain. Then, in the zones of Ω which are not perturbed by the reflections, the shock moves with a velocity equal to $\omega\ (\sqrt{3}/2, -1/2)^t$. The external pressure at the outflow boundary $\partial\Omega_o$ is thus prescribed throughout the transient to $p_L = 116.5$. On the top of the domain $(0,4) \times \{1\}$, the boundary condition is consistent to the undisturbed shock wave, thus the unknowns ρ, $\boldsymbol{u}$ and p are prescribed to the left state values for $x_1 \leq 1/6 + 1/\sqrt{3} + (2 * \omega/\sqrt{3})\, t$ and to the right state values on the other part of the boundary. Finally, on $\{4\} \times (0,1)$, the velocity is prescribed to $\boldsymbol{u}_R = (0,0)^t$. These results strengthen the previous one as they are comparable to those presented in recent literature (see [42]).

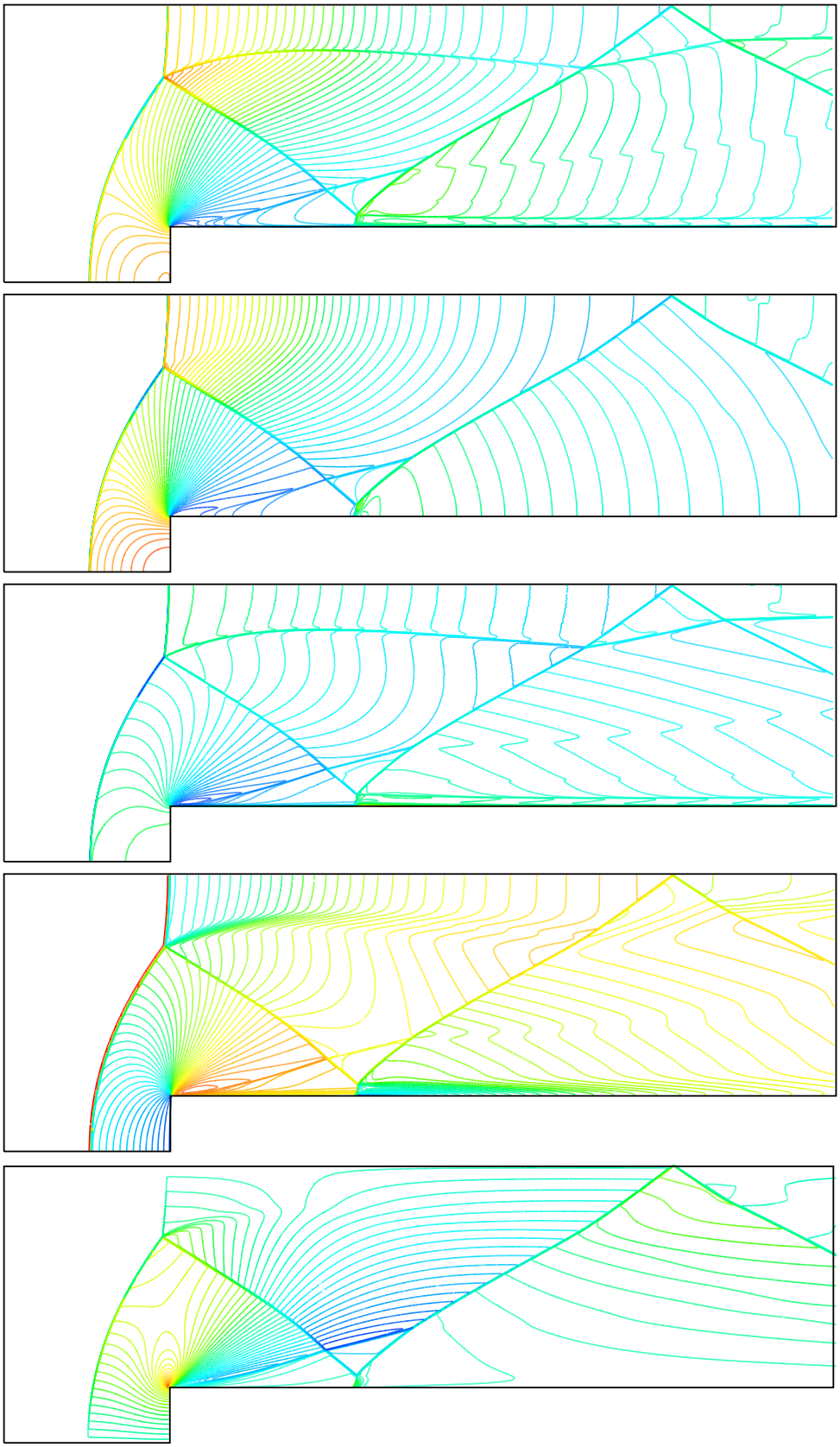


Figure 2.12 – Mach 3 step – From top to bottom : density, pressure, internal energy, first and second component of the velocity at $t = 4$, obtained with $h = 2.5 \times 10^{-3}$, $\delta t = 10^{-3}$ and $\mu = 10^{-3}$. The variation intervals of the unknowns are $\rho \in [0.235,\ 6.4]$, $p \in [0.216,\ 12.04]$, $\mathcal{H} \in [2.46, 8.11]$, $u_1 \in [0.,\ 3.046]$, and $u_2 \in [-0.92,\ 1.82]$.

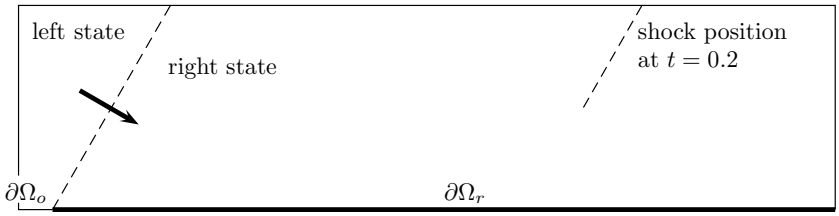


Figure 2.13 – Double Mach reflection – Geometry and initial conditions.

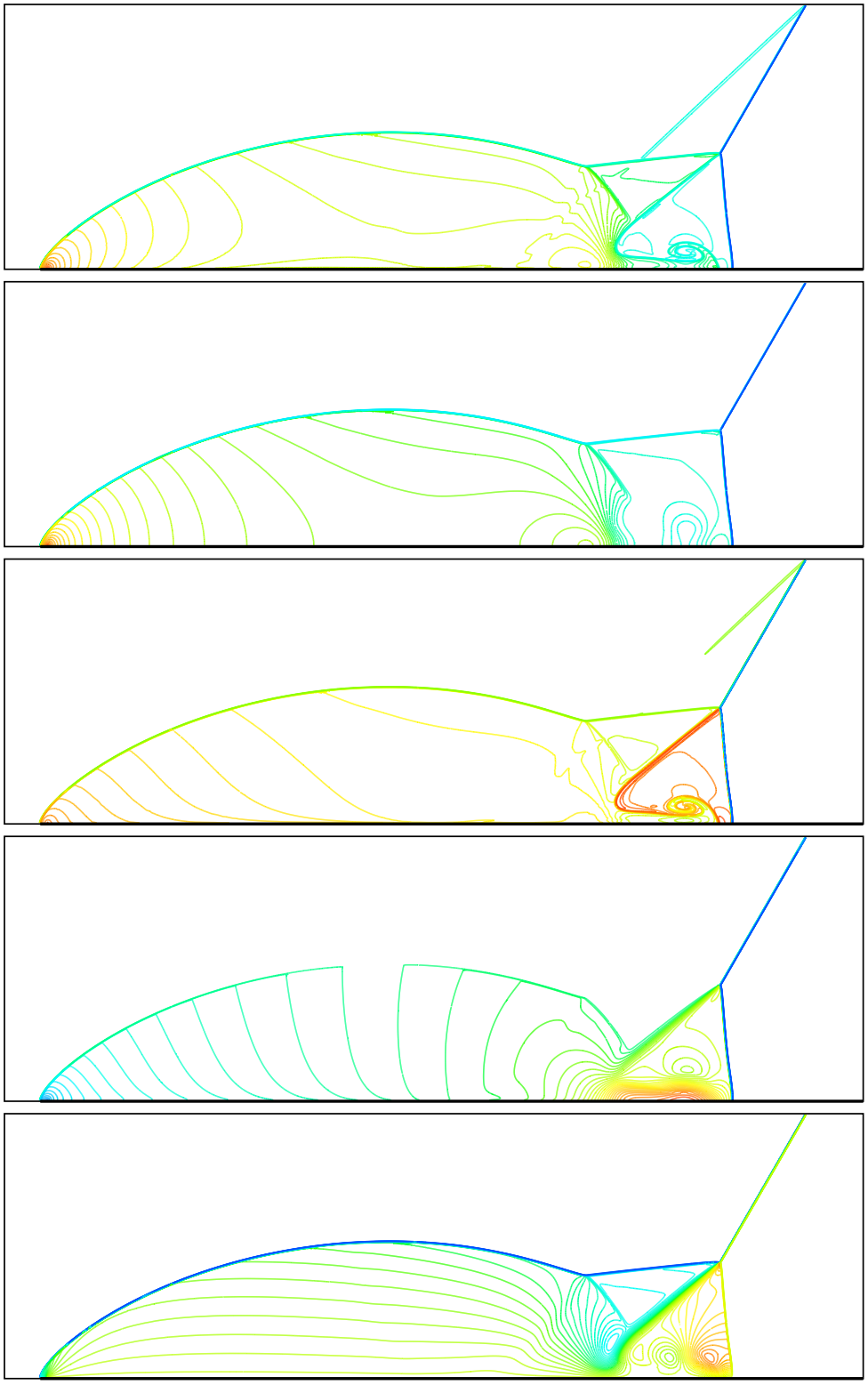


Figure 2.14 – Double Mach reflection – From top left to bottom right : density, pressure, internal energy and first and second component of the velocity at $t = 0.2$, obtained with $h = 2.5\ 10^{-3}$, $\delta t = 2.5\ 10^{-5}$ and $\mu = 0.01$. The variation ranges of the unknowns are $\rho \in [1.4,\ 22.4]$, $p \in [1,\ 559]$, $\mathcal{H} \in [2.5,\ 87.8]$, $u_1 \in [-1.74,\ 15.9]$, and $u_2 \in [-5.53,\ 1.74]$. A right part of the domain, where the solution is constant, is not drawn.

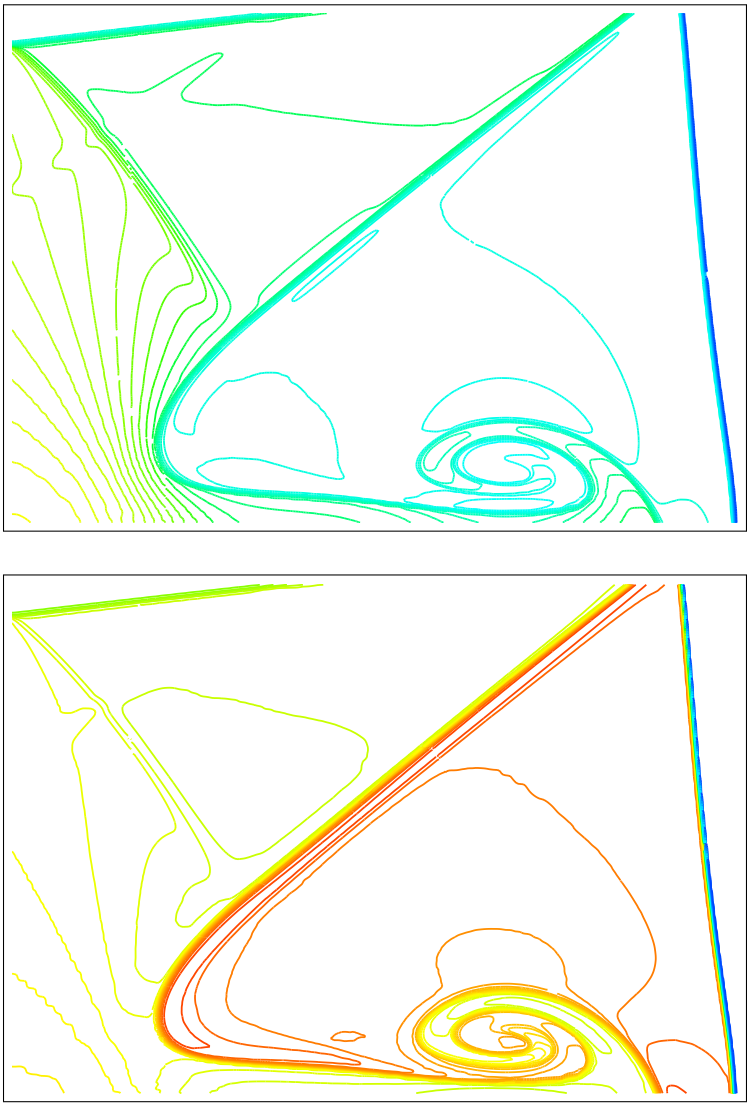


Figure 2.15 – Double Mach reflection, zoom of the density (top) and internal energy (bottom) fields

Mach 10 flow past a cylinder

The last test case is a compressible version of a benchmark originally developed for incompressible Navier-Stokes solvers in [55]. The geometry of the problem is described in Figure 2.16. The fluid enters the domain on the left boundary with a constant velocity :

$$\boldsymbol{u} = \left(1,\, 0\right)^t.$$

We want a Mach 10 flow entering the domain so we set $c = (\gamma p/\rho)^{1/2} = 0.1$.

$$\begin{bmatrix} \rho \\ p \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1/140 \end{bmatrix},$$

A coarse version of the meshes used for this computation is presented in Figure 2.17. Refined versions of this mesh are obtained by reducig the space step along the characteristic lines (the boundaries and the circles around the cylinder). We consequently use the RT discretization. Initial conditions are the same as inlet values.

The right boundary condition is free. We impose a perfect slip conditions on the cylinder, the top and the bottom boundaries. The computationnal time interval is set to $(0, 5)$. We impose

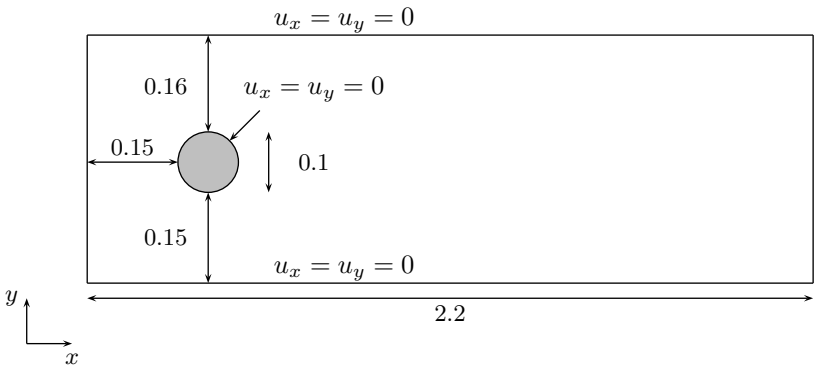


Figure 2.16 – Low Mach flow past a cylinder – Geometry.

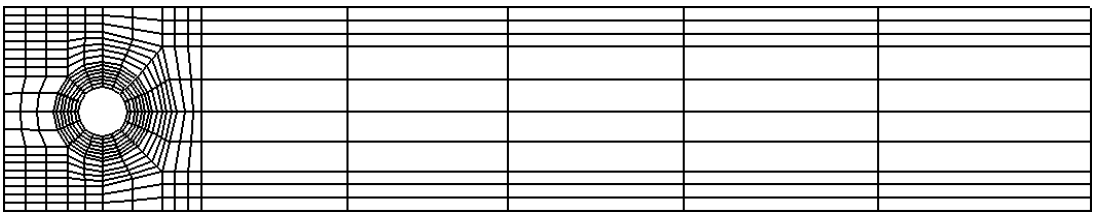


Figure 2.17 – A "coarse version" of the mesh.

a residual viscosity $\mu = 0.05$ which roughly corresponds to 1/10 of the upwind dissipation. The time step is set to $\delta t = 10^{-4}$ and the computations are performed on a mesh with $5.31e5$ cells which corresponds approximately to a space step of $1e-3$. Consequently the value of the fastest wave being 1.1 the acoustic CFL is close to 0.1.

We present in Figure 2.18 results obtained at $t = 5$. We observe a strong shock in front of the cylinder. Subsequent weak shock reflections yield the X-structure for the pressure and density fields. They are progressively damped by the scheme diffusion.

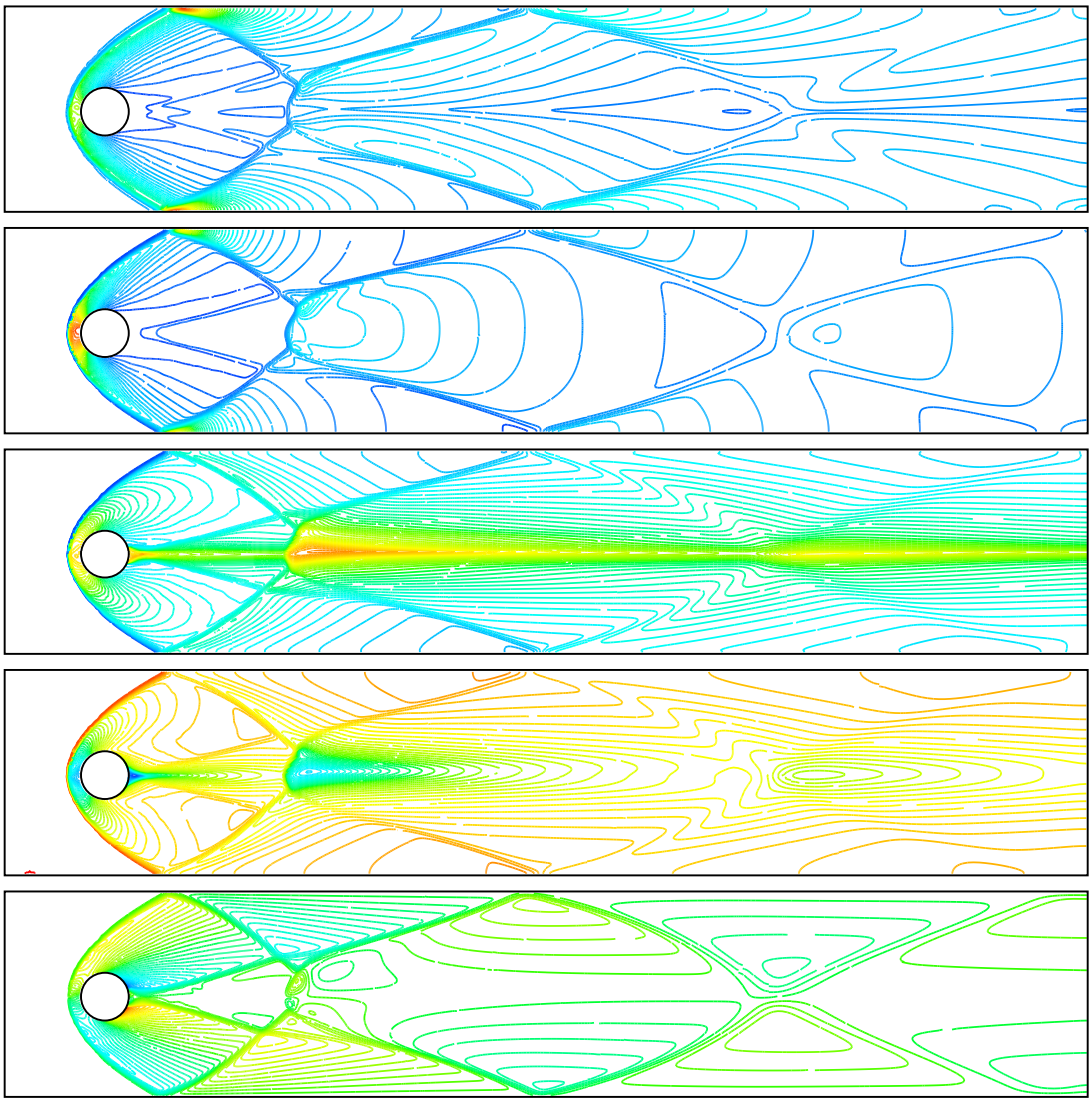


Figure 2.18 – Mach=10 flow past a cylinder – From top to bottom : internal energy, density, x-component of velocity, y-component of velocity at $t = 5$. The variation ranges of the unknowns are $e \in [0.178,\ 0.536]$, $\rho \in [0.804,\ 12.23]$, $u_1 \in [-0.11,\ 1]$, and the value $u_1 = 0$ corresponds to the fourth iso-line, $u_2 \in [-0.326,\ 0.327]$.

Chapitre 3

Consistency results of a class of staggered schemes for the compressible Euler equations

3.1 Introduction

Let Ω be an open bounded connected subset of $\mathbb{R}^d$, with $d \in \{2,3\}$. Let $T \in \mathbb{R}^+$. We address in this paper the system of unstationnary compressible Euler equations :

$$\partial_t \rho + \mathrm{div}(\rho\, \boldsymbol{u}) = 0, \tag{3.1a}$$

$$\partial_t(\rho\, \boldsymbol{u}) + \mathrm{div}(\rho\, \boldsymbol{u} \otimes \boldsymbol{u}) + \boldsymbol{\nabla} p = 0, \tag{3.1b}$$

$$\partial_t(\rho\, E) + \mathrm{div}(\rho\, E\, \boldsymbol{u}) + \mathrm{div}(p\, \boldsymbol{u}) = 0, \tag{3.1c}$$

$$p = (\gamma - 1)\, \rho\, e, \qquad E = \frac{1}{2}|\boldsymbol{u}|^2 + e, \tag{3.1d}$$

where t stands for the time, ρ, $\boldsymbol{u}$, p, E and e are the density, velocity, pressure, total energy and internal energy respectively, and $\gamma > 1$ is a coefficient specific to the considered fluid. The problem is supposed to be posed over $\Omega \times (0, T)$. System (3.1) is complemented by initial conditions for ρ, e and $\boldsymbol{u}$, denoted by ρ_0, e_0 and $\boldsymbol{u}_0$ respectively, with $\rho_0 > 0$ and $e_0 > 0$, and by a boundary condition which we suppose to be $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ at any time and *a.e.* on $\partial\Omega$, where $\boldsymbol{n}$ stands for the normal vector to the boundary.

This paper falls in with a research program undertaken to develop staggered schemes for all-Mach flows satisfying a kinetic energy balance [25, 35, 36, 38]. We recall that a combination of (3.1a) and (3.1b) leads to :

$$\partial_t(\rho E_k) + \mathrm{div}\Big(\rho\, E_k\, \boldsymbol{u}\Big) + \boldsymbol{\nabla} p \cdot \boldsymbol{u} = 0, \tag{3.2}$$

with $E_k = \frac{1}{2}\,|\boldsymbol{u}|^2$. the kinetic energy of the fluid. Subtracting this relation from the total energy balance (3.1c), we obtain the internal energy balance equation :

$$\partial_t(\rho e) + \mathrm{div}(\rho e \boldsymbol{u}) + p\, \mathrm{div}\boldsymbol{u} = 0. \tag{3.3}$$

Since,

- thanks to the mass balance equation, the first two terms in the left-hand side of (3.3) may be recast as a transport operator : $\partial_t(\rho e) + \mathrm{div}(\rho e \boldsymbol{u}) = \rho\, [\partial_t e + \boldsymbol{u} \cdot \boldsymbol{\nabla} e]$,
- and, from the equation of state, the pressure vanishes when $e = 0$,

this equation implies, if $e \geq 0$ at $t = 0$ and with suitable boundary conditions, that e remains non-negative at all times. The key point is to obtain such energy balances at a discrete level.

This paper takes over the work developed in [39, 36], which answers the above questions, and extends the consistency results obtained in 1D to higher dimensions. It is organized as follows : We first present the meshes and the spatial discretisation, then we introduce the decoupled scheme and we finish by the main result of this paper and its proof.

3.2 Meshes and discretization spaces

Let $\mathcal{M}$ be a mesh of the domain Ω, supposed to be regular in the usual sense of the finite element literature (*e.g.* [19]). The cells of the mesh are assumed to be :

- for a general domain Ω, either non-degenerate quadrilaterals ($d = 2$) or hexahedra ($d = 3$) or simplices, both type of cells being possibly combined in a same mesh,
- for a domain the boundaries of which are hyperplanes normal to a coordinate axis, rectangles ($d = 2$) or rectangular parallelepipeds ($d = 3$) (the faces of which, of course, are then also necessarily normal to a coordinate axis).

By $\mathcal{E}$ and $\mathcal{E}(K)$ we denote the set of all $(d-1)$-faces σ of the mesh and of the element $K \in \mathcal{M}$ respectively. The set of faces included in the boundary of Ω is denoted by $\mathcal{E}_{\text{ext}}$ and the set of internal faces (*i.e.* $\mathcal{E} \setminus \mathcal{E}_{\text{ext}}$) is denoted by $\mathcal{E}_{\text{int}}$; a face $\sigma \in \mathcal{E}_{\text{int}}$ separating the cells K and L is denoted by $\sigma = K|L$. The outward normal vector to a face σ of K is denoted by $\boldsymbol{n}_{K,\sigma}$. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, we denote by $|K|$ the measure of K and by $|\sigma|$ the $(d-1)$-measure of the face σ. For $1 \leq i \leq d$, we denote by $\mathcal{E}^{(i)} \subset \mathcal{E}$ and $\mathcal{E}^{(i)}_{\text{ext}} \subset \mathcal{E}_{\text{ext}}$ the subset of the faces of $\mathcal{E}$ and $\mathcal{E}_{\text{ext}}$ respectively which are perpendicular to the ith unit vector of the canonical basis of $\mathbb{R}^d$.

The space discretization is staggered, using either the Marker-And Cell (MAC) scheme [34, 33], or nonconforming low-order finite element approximations, namely the Rannacher and Turek element (RT) [54] for quadrilateral or hexahedric meshes, or the lowest degree Crouzeix-Raviart element (CR) [22] for simplicial meshes.

For all these space discretizations, the degrees of freedom for the pressure, the density and the internal energy (*i.e.* the discrete pressure, density and internal energy unknowns) are associated to the cells of the mesh $\mathcal{M}$, and are denoted by :

$$\left\{p_K,\ \rho_K,\ e_K,\ K \in \mathcal{M}\right\}.$$

Let us then turn to the degrees of freedom for the velocity (*i.e.* the discrete velocity unknowns).

- **Rannacher-Turek** or **Crouzeix-Raviart** discretizations – The degrees of freedom for the velocity components are located at the center of the faces of the mesh, and we choose the version of the element where they represent the average of the velocity through a face. The set of degrees of freedom reads :

$$\{u_{\sigma,i},\ \sigma \in \mathcal{E},\ 1 \leq i \leq d\}.$$

- **MAC** discretization – The degrees of freedom for the ith component of the velocity are defined at the centre of the faces $\sigma \in \mathcal{E}^{(i)}$, so the whole set of discrete velocity unknowns reads :

$$\left\{u_{\sigma,i},\ \sigma \in \mathcal{E}^{(i)},\ 1 \leq i \leq d\right\}.$$

We now introduce a dual mesh, which will be used for the finite volume approximation of the time derivative and convection terms in the momentum balance equation.

- **Rannacher-Turek** or **Crouzeix-Raviart** discretizations – For the RT or CR discretizations, the dual mesh is the same for all the velocity components. When $K \in \mathcal{M}$ is a simplex, a rectangle or a cuboid, for $\sigma \in \mathcal{E}(K)$, we define $D_{K,\sigma}$ as the cone with basis σ and with vertex the mass center of K (see Figure 3.1). We thus obtain a partition of K in m sub-volumes, where m is the number of faces of the mesh, each sub-volume having the same

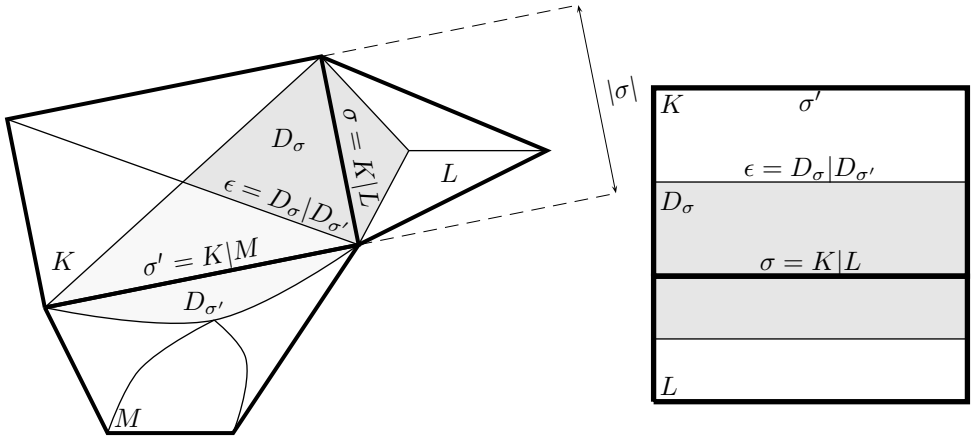


Figure 3.1 – Notations for control volumes and dual cells – Left : Finite Elements (the present sketch illustrates the possibility, implemented in our software (CALIF^{3}S [16]), of mixing simplicial (Crouzeix-Raviart) and quadrangular (Rannacher-Turek) cells) – Right : MAC discretization, dual cell for the y-component of the velocity.

measure $|D_{K,\sigma}| = |K|/m$. We extend this definition to general quadrangles and hexahedra, by supposing that we have built a partition still of equal-volume sub-cells, and with the same connectivities. Note that this is of course always possible, but that such a volume $D_{K,\sigma}$ may be no longer a cone ; indeed, if K is far from a parallelogram, it may not be possible to build a cone having σ as basis, the opposite vertex lying in K and a volume equal to $|K|/m$.

The volume $D_{K,\sigma}$ is referred to as the half-diamond cell associated to K and σ.

For $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = K|L$, we now define the diamond cell D_σ associated to σ by $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$; for an external face $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}(K)$, D_σ is just the same volume as $D_{K,\sigma}$.

- **MAC** discretization – For the MAC scheme, the dual mesh depends on the component of the velocity. For each component, the MAC dual mesh only differs from the RT or CR dual mesh by the choice of the half-diamond cell, which, for $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$, is now the rectangle or rectangular parallelepiped of basis σ and of measure $|D_{K,\sigma}| = |K|/2$.

We denote by $|D_\sigma|$ the measure of the dual cell D_σ, and by $\epsilon = D_\sigma|D_{\sigma'}$ the face separating two diamond cells D_σ and $D_{\sigma'}$. The set of the faces of a dual cell D_σ is denoted by $\tilde{\mathcal{E}}(D_\sigma)$.

Finally, we need to deal with the impermeability (*i.e.* $\boldsymbol{u} \cdot \boldsymbol{n} = 0$) boundary condition. Since the velocity unknowns lie on the boundary (and not inside the cells), these conditions are taken into account in the definition of the discrete spaces. To avoid technicalities in the expression of the schemes, we suppose throughout this paper that the boundary is *a.e.* normal to a coordinate axis, (even in the case of the RT or CR discretizations), which allows to simply set to zero the corresponding velocity unknowns :

$$\text{for } i = 1, \dots, d, \ \forall \sigma \in \mathcal{E}^{(i)}_{\text{ext}}, \qquad u_{\sigma,i} = 0. \tag{3.4}$$

Therefore, there are no discrete velocity unknowns on the boundary for the MAC scheme, and there are only $d-1$ discrete velocity unknowns on each boundary face for the CR and RT discretizations, which depend on the orientation of the face. In order to be able to write a unique expression of the discrete equations for both MAC and CR/RT schemes, we introduce the set of faces $\mathcal{E}^{(i)}_{\mathcal{S}}$ associated with the degrees of freedom of each component of the velocity ($\mathcal{S}$ stands for "scheme") :

$$\mathcal{E}^{(i)}_{\mathcal{S}} = \left| \begin{array}{l} \mathcal{E}^{(i)} \setminus \mathcal{E}^{(i)}_{\text{ext}} \text{ for the MAC scheme,} \\ \mathcal{E} \setminus \mathcal{E}^{(i)}_{\text{ext}} \text{ for the CR or RT schemes.} \end{array} \right.$$

Similarly, we unify the notation for the set of dual faces for both schemes by defining :

$$\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)} = \left| \begin{array}{l} \tilde{\mathcal{E}}^{(i)} \setminus \tilde{\mathcal{E}}_{\text{ext}}^{(i)} \text{ for the MAC scheme,} \\ \tilde{\mathcal{E}} \setminus \tilde{\mathcal{E}}_{\text{ext}}^{(i)} \text{ for the CR or RT schemes,} \end{array} \right.$$

where the symbol ˜ refers to the dual mesh ; for instance, $\tilde{\mathcal{E}}^{(i)}$ is thus the set of faces of the dual mesh associated with the ith component of the velocity, and $\tilde{\mathcal{E}}_{\text{ext}}^{(i)}$ stands for the subset of these dual faces included in the boundary. Note that, for the MAC scheme, the faces of $\tilde{\mathcal{E}}^{(i)}$ are perpendicular to a unit vector of the canonical basis of $\mathbb{R}^d$, but not necessarily to the ith one.

Note that general domains can easily be addressed (of course, with the CR or RT discretizations) by redefining, through linear combinations, the degrees of freedom at the external faces, so as to introduce the normal velocity as a new degree of freedom.

3.3 Pressure correction and decoupled schemes

3.3.1 General form of the schemes

Let us consider a partition $0 = t_0 < t_1 < \ldots < t_N = T$ of the time interval $(0, T)$, which we suppose uniform, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \ldots, N-1$ be the (constant) time step. For the solution of the system (3.1), we introduce two algorithms. The first one is purely explicit in time and, as we will show later, is stable under some CFL condition ; the second one is semi-implicit, and is unconditionally stable. A first order (upwind) version of the first algorithm was introduced in [39] ; then formally second-order (MUSCL) expressions of the convection fluxes were implemented, and the scheme was tested numerically in [61], showing much better accuracy properties. The pressure correction algorithm was introduced in [36] and extended to the Navier-Stokes equations in [29].

The decoupled scheme reads :

$$\forall K \in \mathcal{M}, \quad \frac{1}{\delta t}(\rho_K^{n+1} - \rho_K^n) + \operatorname{div}(\rho^n \boldsymbol{u}^n)_K = 0, \tag{3.5a}$$

$$\forall K \in \mathcal{M}, \quad \frac{1}{\delta t}(\rho_K^{n+1} e_K^{n+1} - \rho_K^n e_K^n) + \operatorname{div}(\rho^n e^n \boldsymbol{u}^n)_K + p_K^n (\operatorname{div}(\boldsymbol{u}^n))_K = S_K^n, \tag{3.5b}$$

$$\forall K \in \mathcal{M}, \quad p_K^{n+1} = (\gamma - 1)\, \rho^{n+1}\, e_K^{n+1}. \tag{3.5c}$$

$$\text{For } 1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}, \quad \frac{1}{\delta t}\left(\rho_{D_\sigma}^{n+1} u_{\sigma,i}^{n+1} - \rho_{D_\sigma}^n u_{\sigma,i}^n\right) + \operatorname{div}(\rho^n u_i^n \boldsymbol{u}^n)_\sigma + (\nabla p)_{\sigma,i}^{n+1} + \mathcal{D}(u_i^n)_{\sigma,i} = 0. \tag{3.5d}$$

The pressure correction algorithm falls in the class of pressure correction schemes, and consists in the two following steps :

Pressure gradient scaling step :

$$\forall \sigma \in \mathcal{E}_{\text{int}}, \quad (\overline{\nabla p})_\sigma^{n+1} = \left(\frac{\rho_{D_\sigma}^n}{\rho_{D_\sigma}^{n-1}}\right)^{1/2} (\nabla p^n)_\sigma. \tag{3.6a}$$

Prediction step – Solve for $\tilde{\boldsymbol{u}}^{n+1}$:

$$\begin{gathered} \text{For } 1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}, \\ \frac{1}{\delta t}\left(\rho_{D_\sigma}^n \tilde{u}_{\sigma,i}^{n+1} - \rho_{D_\sigma}^{n-1} u_{\sigma,i}^n\right) + \operatorname{div}(\rho^n \tilde{u}_i^{n+1} \boldsymbol{u}^n)_\sigma + (\overline{\nabla p})_{\sigma,i}^{n+1} + \mathcal{D}(\tilde{u}_i^{n+1})_{\sigma,i} = 0. \end{gathered} \tag{3.6b}$$

Correction step – Solve for p^{n+1}, e^{n+1}, ρ^{n+1} and $\boldsymbol{u}^{n+1}$:

$$\text{For } 1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)},\quad \frac{1}{\delta t}\,\rho^n_{D_\sigma}\,(u^{n+1}_{\sigma,i} - \tilde{u}^{n+1}_{\sigma,i}) + (\boldsymbol{\nabla} p^{n+1})_{\sigma,i} - (\overline{\boldsymbol{\nabla} p})^{n+1}_{\sigma,i} = 0, \tag{3.6c}$$

$$\forall K \in \mathcal{M},\quad \frac{1}{\delta t}(\rho^{n+1}_K - \rho^n_K) + \operatorname{div}(\rho^{n+1}\boldsymbol{u}^{n+1})_K = 0, \tag{3.6d}$$

$$\forall K \in \mathcal{M},\quad \frac{1}{\delta t}(\rho^{n+1}_K e^{n+1}_K - \rho^n_K e^n_K) + \operatorname{div}(\rho^{n+1} e^{n+1}\boldsymbol{u}^{n+1})_K + p^{n+1}_K \operatorname{div}(\boldsymbol{u}^{n+1})_K = S^{n+1}_K, \tag{3.6e}$$

$$\forall K \in \mathcal{M},\quad p^{n+1}_K = (\gamma - 1)\,\rho^{n+1}\,e^{n+1}_K. \tag{3.6f}$$

The first step is a classical pressure correction solution of the momentum balance equation to obtain a tentative velocity field. The second step is a nonlinear pressure correction step, which couples, for stability reasons, the mass balance equation with the internal energy balance equation (see [49, 50, 29]). In addition, it also allows the scheme to keep the velocity and pressure constant across (1D) contact discontinuities [29].

The right hand side S_K is a correction term to the internal energy which is required in order for the approximate solutions to converge to an entropy weak solution in presence of shock discontinuities. We shall describe this term later. Note that it does not tend to zero (in a natural L^1 norm) as the mesh size and time step tend to zero.

We now give the space discretization of the terms involved in these algorithms. For details on their construction, the reader is referred to [39, 61, 36].

3.3.2 Mass balance equation

Equations (3.5a) and (3.6d) are a finite volume discretization of the mass balance over the primal mesh. For a discrete density field ρ and a discrete velocity field $\boldsymbol{u}$, we write :

$$|K| \operatorname{div}(\rho\boldsymbol{u})_K = \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}(\rho\boldsymbol{u}),$$

where $F_{K,\sigma}(\rho\boldsymbol{u})$ stands for the mass flux across σ outward K. This quantity is given by :

$$\forall \sigma = K|L \in \mathcal{E}_{\text{int}},\qquad F_{K,\sigma}(\rho\boldsymbol{u}) = |\sigma|\,\rho_\sigma\,u_{K,\sigma}, \tag{3.7}$$

where $u_{K,\sigma}$ is an approximation of the normal velocity to the face σ outward K, defined by :

$$u_{K,\sigma} = \left| \begin{array}{l} u^n_{\sigma,i}\,\boldsymbol{e}^{(i)} \cdot \boldsymbol{n}_{K,\sigma} \text{ for } \sigma \in \mathcal{E}^{(i)} \text{ in the MAC case,} \\ \boldsymbol{u}^n_\sigma \cdot \boldsymbol{n}_{K,\sigma} \text{ for } \sigma \in \mathcal{E} \text{ in the CR and RT cases,} \end{array} \right. \tag{3.8}$$

with $\boldsymbol{e}^{(i)}$ the i-th vector of the orthonormal basis of $\mathbb{R}^d$. Thanks to the boundary conditions, $u_{K,\sigma}$ vanishes for any external face σ. The density at the internal face $\sigma = K|L$ is supposed to satisfy the following property :

$$\forall K \in \mathcal{M},\ \forall \sigma = K|L \in \mathcal{E}(K) \cap \mathcal{E}_{\text{int}},\ \text{ there exists } \alpha_{K,\sigma} \in [0,1] \text{ and a neighbour cell } M^K_\sigma \text{ of } K$$
$$\text{such that}\quad \rho_\sigma - \rho_K = \left| \begin{array}{ll} \alpha_{K,\sigma}(\rho_K - \rho_{M^K_\sigma}) & \text{if } u_{K,\sigma} \ge 0, \\ \alpha_{K,\sigma}(\rho_L - \rho_K) & \text{otherwise.} \end{array} \right. \tag{3.9}$$

The upwind choice corresponds to $\alpha_{K,\sigma} = 0$ if $u_{K,\sigma} \ge 0$ and $\alpha_{K,\sigma} = 1$, $M^K_\sigma = L$ if $u_{K,\sigma} \le 0$. A computation for $\alpha_{K,\sigma}$ derived from a MUSCL technique developed for the scalar transport equation [53] is given in [61]. An important property derived from the MUSCL interpolation is the existence of a unique $\alpha_\sigma \in [0,1]$ (depending on the different $\alpha_{K,\sigma}$) such that, for $\sigma = K|L$,

$$\rho_\sigma = \alpha_\sigma \rho_K + (1 - \alpha_\sigma)\rho_L \tag{3.10}$$

3.3.3 The discrete internal energy balance equation

Equations (3.5b) and (3.6e) are a finite volume discretization of the internal energy balance over the primal cell K. The convection term reads :

$$\mathrm{div}(\rho e \boldsymbol{u})_K = \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}(\rho \boldsymbol{u})\, e_\sigma,$$

and the approximation of the internal energy at the face e_σ has the same properties as the density approximation, *i.e.*

$$\text{for } \sigma = K|L \in \mathcal{E}_{\text{int}}, \qquad e_\sigma - e_K = \left| \begin{array}{ll} \beta_{K,\sigma}(e_K - e_{M_\sigma^K}) & \text{if } F_{K,\sigma} \geq 0, \\ \beta_{K,\sigma}(e_L - e_K) & \text{otherwise,} \end{array} \right. \tag{3.11}$$

with the same cell M_σ^K but a different coefficient $\beta_{K,\sigma}$. A particular coupled choice of these coefficients $\alpha_{K,\sigma}$ for the density and energy approximation ensures that, for the present equation of state (more generally, for any equation of state giving the pressure p as a function of the product ρe), the internal energy can be see as an equation for the pressure which leaves p unchanged whenever the velocity $\boldsymbol{u}$ is constant in space, and thus, in particular, across contact discontinuities (see [61] for more details). This particular choice of coefficients $\beta_{K,\sigma}$ leads to the same convex combination as for ρ (3.10) for the product ρe :

$$\rho_\sigma e_\sigma = \alpha_\sigma \rho_K e_K + (1 - \alpha_\sigma)\rho_L e_L. \tag{3.12}$$

The discrete divergence of the velocity has a natural approximation :

$$\text{for } K \in \mathcal{M}, \qquad (\mathrm{div}\boldsymbol{u})_K = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} |\sigma|\, u_{K,\sigma}. \tag{3.13}$$

Finally, the right-hand side, S_K, is derived using consistency (in the Lax sense) arguments in the next section ; at the first time step, it is simply set to zero.

3.3.4 The discrete momentum balance equation

We now turn to the discrete momentum balances (3.5d) and (3.6b), which are obtained by discretizing the momentum balance equation (3.1b) on the dual cells associated to the faces of the mesh. The convection operator reads :

$$\mathrm{div}(\rho \tilde{u}_i \boldsymbol{u})_\sigma = \sum_{\epsilon \in \tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}(\rho, \boldsymbol{u})\, (u_i)_\epsilon,$$

where $F_{\sigma,\epsilon}(\rho, \boldsymbol{u})$ stands for a mass flux through the dual faces of the mesh, the definition of which differs in the Rannacher-Turek/Crouzeix-Raviart case and in the MAC case. In both cases though these definitions ensures that a finite volume discretization of the mass balance equation over the diamond cells holds, whatever the time discretization be :

$$\forall \sigma \in \mathcal{E}, \qquad \frac{|D_\sigma|}{\delta t}\,(\rho_{D_\sigma}^{n+1} - \rho_{D_\sigma}^n) + \sum_{\epsilon \in \tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n = 0, \tag{3.14}$$

in the decoupled case and

$$\forall \sigma \in \mathcal{E}, \qquad \frac{|D_\sigma|}{\delta t}\,(\rho_{D_\sigma}^n - \rho_{D_\sigma}^{n-1}) + \sum_{\epsilon \in \tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n = 0, \tag{3.15}$$

in the pressure correction case. This is a necessary condition to be able to derive a discrete kinetic energy balance in both cases.

The density on a dual cell is the same for each type of discretization, namely :

$$\begin{aligned} &\text{for } \sigma \in \mathcal{E}_{\text{int}},\ \sigma = K|L && |D_\sigma|\, \rho_{D_\sigma} = |D_{K,\sigma}|\, \rho_K + |D_{L,\sigma}|\, \rho_L, \\ &\text{for } \sigma \in \mathcal{E}_{\text{ext}},\ \sigma \in \mathcal{E}(K), && \rho_{D_\sigma} = \rho_K. \end{aligned} \tag{3.16}$$

We now turn to the definition of the dual fluxes.

Rannacher-Turek and Crouzeix-Raviart cases – For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$, let ξ_K^σ be given by :

$$\xi_K^\sigma = \frac{|D_{K,\sigma}|}{|K|}.$$

With the definition of the dual mesh adopted here, the value of the coefficients ξ_K^σ is independent of the cell and the face. For the Rannacher-Turek elements, we have $\xi_K^\sigma = 1/(2d)$ and, for the Crouzeix-Raviart elements, $\xi_K^\sigma = 1/(d+1)$. We suppose first that the flux through the external dual faces, which are also faces of the primal mesh, is equal to zero. Then the mass fluxes through the inner dual faces are supposed to satisfy the following properties.

Définition 3.1 (*Definition of the dual fluxes CR-RT*)

The fluxes through the faces of the dual mesh are defined so as to satisfy the following three constraints :

(H1) The discrete mass balance over the half-diamond cells is satisfied, in the following sense. For all primal cell K in $\mathcal{M}$, the set $(F_{\sigma,\epsilon})_{\epsilon\subset K}$ of dual fluxes included in K solves the following linear system

$$F_{K,\sigma} + \sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma),\ \epsilon\subset K} F_{\sigma,\epsilon} = \xi_K^\sigma \sum_{\sigma'\in\mathcal{E}(K)} F_{K,\sigma'}, \qquad \sigma \in \mathcal{E}(K). \tag{3.17}$$

(H2) The dual fluxes are conservative, *i.e.* for any dual face $\epsilon = D_\sigma|D'_\sigma$, we have $F_{\sigma,\epsilon} = -F_{\sigma',\epsilon}$.

(H3) The dual fluxes are bounded with respect to the primal fluxes $(F_{K,\sigma})_{\sigma\in\mathcal{E}(K)}$, in the sense that there exists a constant real number C such that :

$$|F_{\sigma,\epsilon}| \le C \max\left\{|F_{K,\sigma}|,\ \sigma \in \mathcal{E}(K)\right\}, \quad K \in \mathcal{M},\ \sigma \in \mathcal{E}(K),\ \epsilon \in \tilde{\mathcal{E}}(D_\sigma),\ \epsilon \subset K. \tag{3.18}$$

In fact, definition 3.1 is not complete, since the system of equations (3.17) has an infinite number of solutions, which makes necessary to impose in addition the constraint (3.18) ; however, assumptions (H1)-(H3) are sufficient for the subsequent developments of this paper (and thus, in particular, imply the consistency of the discrete convection operator). A detailed process of the dual fluxes construction can be found in [3, 25].

MAC case – We define the dual flux $F_{\sigma,\epsilon}$ for the MAC case : For $\sigma \in \mathcal{E}^{(i)}$, $\sigma = K|L$, we have to distinguish two cases :

- First case – The vector $\boldsymbol{e}^i$ is normal to ϵ, so ϵ is included in a primal cell K, and we denote by σ' the second face of K which, in addition to σ, is normal to $\boldsymbol{e}^i$. We thus have $\epsilon = D_\sigma|D_{\sigma'}$. Then the mass flux through ϵ is given by :

$$F_{\sigma,\epsilon}(\rho,\boldsymbol{u}) = \frac{1}{2}\left[F_{K,\sigma}(\rho,\boldsymbol{u})\, \boldsymbol{n}_{D_\sigma,\epsilon}\cdot\boldsymbol{n}_{K,\sigma} + F_{K,\sigma'}(\rho,\boldsymbol{u})\, \boldsymbol{n}_{D_\sigma,\epsilon}\cdot\boldsymbol{n}_{K,\sigma'}\right]. \tag{3.19}$$

- Second case – The vector $\boldsymbol{e}_i$ is tangent to ϵ, and ϵ is the union of the halves of two primal faces τ and τ' such that $\tau \in \mathcal{E}(K)$ and $\tau' \in \mathcal{E}(L)$. The mass flux through ϵ is then given by :

$$F_{\sigma,\epsilon}(\rho,\boldsymbol{u}) = \frac{1}{2}\left[F_{K,\tau}(\boldsymbol{u}) + F_{L,\tau'}(\boldsymbol{u})\right]. \tag{3.20}$$

Note that, with this definition, we have the usual finite volume property of local conservativity of the flux through an a dual face $D_\sigma|D_{\sigma'}$ (*i.e.* $F_{\sigma,\epsilon}(\rho,\boldsymbol{u}) = -F_{\sigma',\epsilon}(\rho,\boldsymbol{u})$), and that the flux through a dual face included in the boundary still vanishes.

Since the flux across a dual face lying on the boundary is zero, the values $u^n_{\epsilon,i}$ are only needed at the internal dual faces, and we make the centered choice for their discretization, *i.e.*, for $\epsilon = D_\sigma|D_{\sigma'}$, $u_{\epsilon,i} = (u_{\sigma,i} + u_{\sigma',i})/2$.

The term $(\boldsymbol{\nabla} p_{\sigma,i})$ stands for the i-th component of the discrete pressure gradient at the face σ. This gradient operator is built as the transpose of the discrete operator for the divergence of the velocity, *i.e.* in such a way that the following duality relation with respect to the L^2 inner product holds :

$$\sum_{K\in\mathcal{M}} |K|\, p_K\, (\mathrm{div}\boldsymbol{u})_K + \sum_{i=1}^{d} \sum_{\sigma\in\mathcal{E}_{\mathcal{S}}^{(i)}} |D_\sigma|\, u_{\sigma,i}\, (\boldsymbol{\nabla} p)_{\sigma,i} = 0. \tag{3.21}$$

This yields to the following expression :

$$\text{for } \sigma = K|L \in \mathcal{E}_{\mathrm{int}}, \qquad (\boldsymbol{\nabla} p)_{\sigma,i} = \frac{|\sigma|}{|D_\sigma|}\,(p_L - p_K)\,\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{e}^{(i)}. \tag{3.22}$$

Note that, because of the impermeability boundary conditions, the discrete gradient is not defined at the external faces.

Finally, the last term is a diffusion-like stabilization term, which may account for the use of a so-called "non-linear viscosity" [30, 31, 42], computed from the regularity of the solution at the previous time step, but also for the numerical diffusion produced by the up winding of the convection term :

$$\mathcal{D}(u_i)_{\sigma,i} = \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \mu_\epsilon\,(u_{\sigma,i} - u_{\sigma',i}).$$

An upwind discrimination of $\mathrm{div}(\rho\tilde{u}_i\boldsymbol{u})_\sigma$ yields $\mu_\epsilon = \mu^{\mathrm{u}}_\epsilon = |F_{\sigma,\epsilon}(\rho,\boldsymbol{u})|/2$. For the Lax consistency of the scheme, we need this stabilization term to behave as a diffusion term with a vanishing viscosity, which is realized for instance if μ_ϵ behaves as $\sigma\, h_\epsilon^{\zeta-1}$, where h_ϵ stands for a characteristic dimension of the face ϵ and ζ is a positive real number. This is indeed true for $\mu^{\mathrm{u}}_\epsilon$ (with $\zeta = 1$), if the density and the velocity are supposed to be uniformly bounded.

3.3.5 Discrete initial conditions

Finally, the initial approximations for ρ, e and $\boldsymbol{u}$ are given by the average of the initial conditions ρ_0 and e_0 on the primal cells and of $\boldsymbol{u}_0$ on the dual cells :

$$\begin{aligned} &\forall K\in\mathcal{M}, \quad \rho^0_K = \frac{1}{|K|}\int_K \rho_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \text{and } e^0_K = \frac{1}{|K|}\int_K e_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &\text{for } 1\le i\le d,\ \forall\sigma\in\mathcal{E}_{\mathcal{S}}^{(i)}, \quad u^0_{\sigma,i} = \frac{1}{|D_\sigma|}\int_{D_\sigma} (\boldsymbol{u}_0(\boldsymbol{x}))_i\,\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{3.23}$$

for the decoupled scheme. Concerning the pressure correction scheme, ρ^{-1} and $\boldsymbol{u}^0$ are given by the same process as for the decoupled scheme :

$$\begin{aligned} &\forall K\in\mathcal{M}, \quad \rho^{-1}_K = \frac{1}{|K|}\int_K \rho_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &\text{for } 1\le i\le d,\ \forall\sigma\in\mathcal{E}_{\mathcal{S}}^{(i)}, \quad u^0_{\sigma,i} = \frac{1}{|D_\sigma|}\int_{D_\sigma} (\boldsymbol{u}_0(\boldsymbol{x}))_i\,\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.24}$$

Then we compute ρ^0 by solving the mass balance equation (3.6d). The initial pressure p^0 is obtained from the internal energy which is computed as in the decoupled case. Finally, this procedure allows to perform the first prediction step with $(\rho_{D_\sigma}^{-1})_{\sigma\in\mathcal{E}}$, $(\rho_{D_\sigma}^{0})_{\sigma\in\mathcal{E}}$ and the dual mass fluxes satisfying the mass balance.

3.3.6 Discrete kinetic energy balance and corrective source term

Thanks to the above definition of the spatial operators it is possible to derive, for each temporal discretization, a discrete analogue of the kinetic energy balance equation (3.2).

For the decoupled discretization it takes the form of the equation (3.25) just below.

Lemma 3.1 (*Discrete kinetic energy balance*)

A solution to the system (3.5) satisfies the following equality, for $1 \le i \le d$, $\sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}$ and $0 \le n \le N-1$:

$$\frac{1}{2}\frac{|D_\sigma|}{\delta t}\Big[\rho_{D_\sigma}^{n+1}(u_{\sigma,i}^{n+1})^2 - \rho_{D_\sigma}^{n}(u_{\sigma,i}^{n})^2\Big] + \frac{1}{2}\sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n\,(u_{\sigma,i}^n u_{\sigma',i}^n)$$
$$+ |D_\sigma|\,(\nabla p)_{\sigma,i}^{n+1}\,u_{\sigma,i}^{n+1} + \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \frac{1}{2}\mu_\epsilon^n(u_{\sigma,i}^n - u_{\sigma',i}^n)(u_{\sigma',i}^n + u_{\sigma,i}^n) = -R_{\sigma,i}^{n+1}, \quad (3.25)$$

with :

$$R_{\sigma,i}^{n+1} = \frac{1}{2}\frac{|D_\sigma|}{\delta t}\rho_{D_\sigma}^{n+1}(u_{\sigma,i}^{n+1} - u_{\sigma,i}^n)^2 + \frac{1}{2}\sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \mu_\epsilon^n(u_{\sigma',i}^n - u_{\sigma,i}^n)^2$$
$$+ \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} (\mu_\epsilon^n - \frac{F_{\sigma,\epsilon}^n}{2})(u_{\sigma,i}^{n+1} - u_{\sigma,i}^n)\,(u_{\sigma,i}^n - u_{\sigma',i}^n). \quad (3.26)$$

Its proof may be found in [61].

The same reasoning can be applied to the pressure correction scheme, and it leads to the following result :

Lemma 3.2 (*Discrete kinetic energy balance, pressure correction scheme*)

A solution to the system (3.6) satisfies the following equality, for $1 \le i \le d$, $\sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}$ and $0 \le n \le N-1$:

$$\frac{1}{2}\frac{|D_\sigma|}{\delta t}\Big[\rho_{D_\sigma}^{n}(u_{\sigma,i}^{n+1})^2 - \rho_{D_\sigma}^{n-1}(u_{\sigma,i}^{n})^2\Big] + \frac{1}{2}\sum_{\epsilon=D_\sigma|D_{\sigma'}} F_{\sigma,\epsilon}^n\,\tilde{u}_{\sigma,i}^{n+1}\,\tilde{u}_{\sigma',i}^{n+1}$$
$$+ |D_\sigma|\,(\nabla p)_{\sigma,i}^{n+1}\,u_{\sigma,i}^{n+1} = -R_{\sigma,i}^{n+1} - P_{\sigma,i}^{n+1}, \quad (3.27)$$

where

$$R_{\sigma,i}^{n+1} = \frac{|D_\sigma|}{2\,\delta t}\,\rho_{D_\sigma}^{n-1}\left(\tilde{u}_{\sigma,i}^{n+1} - u_{\sigma,i}^n\right)^2 + \Big[\sum_{\epsilon=D_\sigma|D_{\sigma'}} \nu\, h_\epsilon^{d-2}\,(\tilde{u}_{\sigma,i}^{n+1} - \tilde{u}_{\sigma',i}^{n+1})\Big]\,\tilde{u}_{\sigma,i}^{n+1},$$
$$P_{\sigma,i}^{n+1} = \frac{|D_\sigma|\,\delta t}{2\,\rho_{D_\sigma}^n}\Big[\big((\nabla p)_{\sigma,i}^{n+1}\big)^2 - \big((\widetilde{\nabla p})_{\sigma,i}^{n+1}\big)^2\Big]. \quad (3.28)$$

Its proof can be found in [36].

In the presence of shock discontinuities, the kinetic energy equation is not satisfied at the continuous level. It is therefore natural to expect the the residual R does not tend to zero in the

weak sense, as the time and space step of the discretization vanish. We wish to obtain a discrete total energy balance by summing the discrete internal energy balance and the discrete kinetic energy balance. This is obtained by choosing ad hoc term S in the discrete internal energy equation to somewhat compensate the residual R. An other issue lies in the fact that both energy equations are discretized on different meshes which means we cannot easily recover an exact discrete total energy balance. The truth is that it is not really needed. We only want to recover a weak form of the total energy balance equation for vanishing space and time steps of the discretization. In particular, we also obtain a global compensation of the residual on the domain, which reads :

$$\sum_{K\in\mathcal{M}} S_K^{n+1} - \sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_{\mathcal{S}}^{(i)}} R_{\sigma,i}^{n+1} = 0. \tag{3.29}$$

Furthermore in order to ensure the positivity of the internal energy, we choose S_K to be positive (unconditionally with the pressure correction scheme and under a C.F.L. condition concerning the decoupled scheme). A possible choice is the following expressions :

$$\forall K \in \mathcal{M},\ S_K^{n+1} = \sum_{i=1}^{d} S_{K,i}^{n+1},$$

with :

$$S_{K,i}^{n+1} = \frac{1}{2}\rho_K^{n-1}\sum_{\sigma\in\mathcal{E}(K)\cap\mathcal{E}_{\mathcal{S}}^{(i)}} \frac{|D_{K,\sigma}|}{\delta t}\left(\tilde{u}_{\sigma,i}^{n+1} - u_{\sigma,i}^{n}\right)^2 + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset,\\ \epsilon=D_\sigma|D_{\sigma'}}} \alpha_{K,\epsilon}\,\frac{|F_{\sigma,\epsilon}^n|}{2}(\tilde{u}_{\sigma,i}^{n+1} - \tilde{u}_{\sigma',i}^{n+1})^2. \tag{3.30}$$

for the pressure correction scheme, with $\alpha_{K,\epsilon}$ equal to 1 when ϵ is strictly included in K which is always the case for RT-CR discretizations. For the MAC scheme, some dual faces are included in the primal cells, but some lie on their boundary ; for such a boundary edge ϵ,we have $\beta_{K,\epsilon} = 0$ and we denote by $\mathcal{N}_\epsilon$ the set of cells M such that $\bar{M}\cap\epsilon \neq \emptyset$ (the cardinal of this set is always 4, except for boundary edges through which, anyway, the mass flux vanishes). We compute $\alpha_{K,\epsilon}$ by :

$$\alpha_{K,\epsilon} = \frac{|K|}{\sum_{M\in\mathcal{N}_\epsilon}|M|}.$$

We notice for uniform grids that $\alpha_{K,\epsilon} = 1/4$.

For the decoupled scheme, one can write the source term as follows :

$$S_{K,i}^{n+1} = \frac{1}{2}\rho_K^{n+1}\sum_{\sigma\in\mathcal{E}(K)\cap\mathcal{E}_{\mathcal{S}}^{(i)}} \frac{|D_{K,\sigma}|}{\delta t}(u_{\sigma,i}^{n+1} - u_{\sigma,i}^{n})^2 + \sum_{\epsilon\in\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset} S_{K,\epsilon,i}^{n+1}, \tag{3.31}$$

with $S_{K,\epsilon,i}^{n+1}$ defined as follows :

We denote by σ_ϵ^U and σ_ϵ^D the two primal faces such that $\epsilon = D_{\sigma_\epsilon^U}|D_{\sigma_\epsilon^D}$ and $F_{\sigma_\epsilon^D,\epsilon}^n \leq 0$ (*i.e.* the cell $D_{\sigma_\epsilon^D}$ is downstream to ϵ). The viscous residual of the upstream cell is equal to

$$(R_{\epsilon,i}^U)^{n+1} = \frac{\mu_\epsilon^n}{2}(u_{\sigma_\epsilon^D,i}^n - u_{\sigma_\epsilon^U,i}^n)^2 + (\mu_\epsilon^n - \frac{|F_{\sigma_\epsilon^D,\epsilon}^n|}{2})(u_{\sigma_\epsilon^U,i}^{n+1} - u_{\sigma_\epsilon^U,i}^n)(u_{\sigma_\epsilon^U,i}^n - u_{\sigma_\epsilon^D,i}^n).$$

The downstream cell residual is equal to

$$(R_{\epsilon,i}^D)^{n+1} = \frac{\mu_\epsilon^n}{2}(u_{\sigma_\epsilon^D,i}^n - u_{\sigma_\epsilon^U,i}^n)^2 + (\mu_\epsilon^n + \frac{|F_{\sigma_\epsilon^D,\epsilon}^n|}{2})(u_{\sigma_\epsilon^D,i}^{n+1} - u_{\sigma_\epsilon^D,i}^n)(u_{\sigma_\epsilon^D,i}^n - u_{\sigma_\epsilon^U,i}^n),$$

so that the total viscous residual on the dual face ϵ is equal to $R^{n+1}_{\epsilon,i} = (R^D_{\epsilon,i})^{n+1} + (R^U_{\epsilon,i})^{n+1}$. If ϵ is strictly included in K one simply take $S^{n+1}_{K,\epsilon,i} = R^{n+1}_{\epsilon,i}$. Ii is the only possible case for the RT-CR discretization. Concerning the MAC discretization, some dual faces are lying on two primal faces. In this case if K is upstream to ϵ (or equivalently $\sigma^U_\epsilon \in \mathcal{E}(K)$) then, denoting by L the other upstream cell ($\sigma^U_\epsilon = K|L$), we take

$$S^{n+1}_{K,\epsilon,i} = \frac{|K|}{|K|+|L|}\left[(R^U_{\epsilon,i})^{n+1} - \frac{|F^n_{\sigma^U_\epsilon,\epsilon}|}{4}(u^n_{\sigma^D_\epsilon,i} - u^n_{\sigma^U_\epsilon,i})^2\right].$$

Likewise, if K is downstream to ϵ, we have

$$S^{n+1}_{K,\epsilon,i} = \frac{|K|}{|K|+|L|}\left[(R^D_{\epsilon,i})^{n+1} + \frac{|F^n_{\sigma^D_\epsilon,\epsilon}|}{4}(u^n_{\sigma^D_\epsilon,i} - u^n_{\sigma^U_\epsilon,i})^2\right],$$

with $\sigma^D_\epsilon = K|L$.

3.4 Stability properties

The joint properties of the schemes are gathered in the three lemmas presented below. The first lemma will ensure that the scheme preserve the positivity of the density and the internal energy.

Lemma 3.3 (*Positivity on the internal energy and the density*)

Let $n \in \mathbb{N}$, let $(\rho^n_K, \boldsymbol{u}^n_K, e^n_K)_{K\in\mathcal{M}} \in (\mathbb{R}^{\mathrm{card}\mathcal{M}} \times (\mathbb{R}^{\mathrm{card}\mathcal{E}})^d \times \mathbb{R}^{\mathrm{card}\mathcal{M}})$, and assume that e^n_K and ρ^n_K are positive, $\forall K \in \mathcal{M}$; let $(\rho^n_K, \boldsymbol{u}^n_K, e^n_K)_{K\in\mathcal{M}}$ satisfy (3.6a)-(3.6f) or (3.5a)-(3.5d) plus the following C.F.L. condition

$$\delta t \le \min\Bigg(\frac{|K|}{\sum_{\sigma\in\mathcal{E}(K)} |\sigma| \left(1+\alpha^n_{K,\sigma}\right)(u^n_{K,\sigma})^+}, \frac{|K|\,\rho^n_K}{\sum_{\sigma\in\mathcal{E}(K)} |\sigma| \left\{\left(1+\beta^n_{K,\sigma}\right)|F^n_{K,\sigma}| + (\gamma-1)\,\rho^n_K u^n_{K,\sigma}\right\}}, \frac{\rho^{n+1}_K |D_{K,\sigma}|}{\sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma),\ \epsilon\cap\tilde{K}\neq\emptyset} \alpha_{K,\epsilon}\,(F^n_{\sigma,\epsilon})^-}\Bigg). \tag{3.32}$$

then $e^{n+1}_K \ge 0$ and $\rho^{n+1}_K \ge 0$, for all $K \in \mathcal{M}$.

The proofs of this lemma can be found in [39] and [36] for the decoupled scheme and pressure correction scheme respectively. The second lemma ensures the existence of a discrete solution of the pressure correction scheme and guarantees, for both schemes, a discrete conservation of the total energy.

Lemma 3.4 (*Existence and stability*)

Assume that for all $K \in \mathcal{M}$, $e^0_K > 0$, $\rho^0_K > 0$ and $\rho^{-1}_K > 0$. Then there exists a solution for each scheme which satisfies, $\forall n \in \mathbb{N}$ and $\forall K \in \mathcal{M}$:

$$\sum_{K\in\mathcal{M}} |K|\,\rho^n_K e^n_K + \frac{1}{2}\sum_{i=1}^d \sum_{\sigma\in\mathcal{E}^{(i)}_S} |D_\sigma|\,\rho^{n-1}_{D_\sigma}\,(u^n_{\sigma,i})^2 + \mathcal{R}^n \le \sum_{K\in\mathcal{M}} |K|\,\rho^0_K e^0_K + \frac{1}{2}\sum_{i=1}^d \sum_{\sigma\in\mathcal{E}^{(i)}_S} |D_\sigma|\,\rho^{-1}_{D_\sigma}\,(u^0_{\sigma,i})^2 + \mathcal{R}^0,$$

where :

$$\mathcal{R}^n = \delta t^2 \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \frac{|D_\sigma|}{\rho^{n-1}_{D_\sigma}} |(\nabla p)^n_\sigma|^2 = \delta t^2 \sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} \frac{|\sigma|^2}{|D_\sigma|\,\rho^{n-1}_{D_\sigma}} (p^n_K - p^n_L)^2.$$

for the pressure correction scheme, and

$$\sum_{K\in\mathcal{M}} |K|\,\rho_K^{n+1}e_K^{n+1} + \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_S^{(i)}} |D_\sigma|\,\rho_{D_\sigma}^n\,(u_{\sigma,i}^n)^2 \le \sum_{K\in\mathcal{M}} |K|\,\rho_K^1 e_K^1 + \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_S^{(i)}} |D_\sigma|\,\rho_{D_\sigma}^0\,(u_{\sigma,i}^0)^2$$

for the decoupled scheme.

The proof of this lemma can be found in [36] for the pressure correction scheme. Its adaption to the decoupled case concerning the total energy inequality is straightforward.

Finally the third lemma proves that the solution of the scheme satisfies a discrete analogue of the continuous entropy inequality. But before stating the main result, we need a preliminary lemma.

Lemma 3.5 (*Convexity property*)

Let $g(.)$ be a strictly convex and once continuously derivable function over an open interval $I \subset \mathbb{R}$. Let $x_1 \in I$ and $x_2 \in I$ two distinct real numbers. Then the following relation :

$$g(x_1) + (\bar{x} - x_1)\,g'(x_1) = g(x_2) + (\bar{x} - x_2)\,g'(x_2)$$

uniquely defines the real number $\bar{x}$. In addition, we have $\bar{x} \in |[x_1, x_2]|$, where $|[a,b]| = (\alpha x_1 + (1-\alpha)x_2)_{\alpha\in[0,1]}$.

Theorem 3.6 (*Discrete entropy inequality*)

For $K \in \mathcal{M}$ and $n \in \mathbb{N}$, let us define the following discrete entropy

$$\rho_K^n\eta_K^n = \phi(\rho_K^n) + \rho_K^n\psi(e_K^n), \tag{3.33}$$

with $\phi(\rho) = \rho ln(\rho)$, $\psi(e) = \dfrac{1}{1-\gamma}ln(e)$. Suppose that the MUSCL interpolations on ρ and e satisfy the additional limitations, for all $\sigma = K|L \in \mathcal{E}_{\text{int}}$ such that $F_{K,\sigma} \ge 0$,

$$\rho_\sigma \in |[\rho_K, \bar{\rho}_\sigma]| \qquad \text{and} \qquad e_\sigma \in |[e_K, \bar{e}_\sigma]|, \tag{3.34}$$

with $\bar{\rho}_\sigma \in |[\rho_K, \rho_L]|$ and $\bar{e}_\sigma \in |[e_K, e_L]|$ are such that

$$\phi(\rho_L) + (\bar{\rho}_\sigma - \rho_L)\,\phi'(\rho_L) = \phi(\rho_K) + (\bar{\rho}_\sigma - \rho_K)\,\phi'(\rho_K), \tag{3.35}$$

$$\psi(e_L) + (\bar{e}_\sigma - e_L)\,\psi'(e_L) = \psi(e_K) + (\bar{e}_\sigma - e_K)\,\psi'(e_K). \tag{3.36}$$

The existence and uniqueness of $\bar{e}_\sigma$ and $\bar{\rho}_\sigma$ is a direct consequence of (3.5).

Then the following inequality holds :

$$\frac{|K|}{\delta t}\left(\rho_K^{n+1}\eta_K^{n+1} - \rho_K^n\eta_K^n\right) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^{n+1}\eta_\sigma^{n+1} + T_{\text{conv},K}^{n+1} \le 0, \quad \forall K \in \mathcal{M} \text{ and } \quad n \in \mathbb{N} \tag{3.37}$$

for the pressure correction scheme, and

$$\frac{|K|}{\delta t}\left(\rho_K^{n+1}\eta_K^{n+1} - \rho_K^n\eta_K^n\right) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^{n}\eta_\sigma^{n} + P_K^n + T_{\text{conv},K}^{n} \le 0, \tag{3.38}$$

for the decoupled scheme, with, for $m = n|n+1$:

$$\rho_\sigma^m\eta_\sigma^m = \phi(\rho_\sigma^m) + \rho_\sigma^m\psi(e_\sigma^m),$$

$T^m_{conv,K} = T^m_{conv,K,\rho} + T^m_{conv,K,e}$ with

$$T^m_{conv,K,\rho} = \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u^m_{K,\sigma} P^m_{\sigma,\rho},$$
$$T^m_{conv,K,e} = \sum_{\sigma\in\mathcal{E}(K)} F^m_{K,\sigma} P^m_{\sigma,e}, \tag{3.39}$$

where

$$P^m_{\sigma,\rho} = (\rho^m_\sigma - \bar{\rho}^m_\sigma)(\frac{\phi'(\rho^m_K) + \phi'(\rho^m_L)}{2}) + (\bar{\rho}^m_\sigma - \rho^m_K)\phi'(\rho^m_K) + \phi(\rho^m_K) - \phi(\rho^m_\sigma),$$
$$P^m_{\sigma,e} = (e^m_\sigma - \bar{e}^m_\sigma)(\frac{\psi'(e^m_K) + \psi'(e^m_L)}{2}) + (\bar{e}^m_\sigma - e^m_K)\psi'(e^m_K) + \psi(e^m_K) - \psi(e^m_\sigma),$$

and

$$P^n_K = \sum_{\sigma\in\mathcal{E}(K)} F^n_{K,\sigma}\left(\rho^{n+1}_K - \rho^n_K\right)\phi''(\rho^{(1)}_K) + \sum_{\sigma\in\mathcal{E}(K)} \left[|\sigma| p^n_K u^n_{K,\sigma} + F^n_{K,\sigma}(e^n_\sigma - e^n_K)\right]\left(e^{n+1}_K - e^n_K\right)\psi''(e^{(1)}_K),$$

where $\rho^{(1)}_K \in |[\rho^n_K, \rho^{n+1}_K]|$ and $e^{(1)}_K \in |[e^n_K, e^{n+1}_K]|$.

Proof : We start the proof for the pressure correction scheme. For the sake of readability we drop the temporal superscript and denote by ρ the quantity at time $n+1$ and $\bar{\rho}$ at time n. Let ϕ be a regular convex function from $\mathbb{R}^*_+$ to $\mathbb{R}$. Multiplying (3.6d) by $\phi'(\rho^{n+1}_K)$ and reordering, we get that :

$$\frac{|K|}{\delta t}(\rho_K - \bar{\rho}_K)\phi'(\rho_K) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\phi(\rho_\sigma) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\left[\rho_K\phi'(\rho_K) - \phi(\rho_K)\right] + T_{K,\rho} = 0.$$

with

$$T_{K,\rho} = \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\left\{(\rho_\sigma - \rho_K)\phi'(\rho_K) + \phi(\rho_K) - \phi(\rho_\sigma)\right\}.$$

Thanks to the convexity of ϕ, the following inequality holds :

$$(\rho_K - \bar{\rho}_K)\phi'(\rho_K) \geq \phi(\rho_K) - \phi(\bar{\rho}_K),$$

so we can deduce that :

$$\frac{|K|}{\delta t}(\phi(\rho_K) - \phi(\bar{\rho}_K)) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\phi(\rho_\sigma) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\left(\rho_K\phi'(\rho_K) - \phi(\rho_K)\right) + T_{K,\rho} \leq 0.$$

The term $T_{K,\rho}$ can also be written :

$$T_{K,\rho} = \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\left(\phi(\rho_K) + (\bar{\rho}_\sigma - \rho_K)\,\phi'(\rho_K) - \phi(\rho_\sigma)\right) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\,(\rho_\sigma - \bar{\rho}_\sigma)\,\phi'(\rho_K),$$

where $\bar{\rho}_\sigma$ is defined by (3.35). Noticing that

$$(\rho_\sigma - \bar{\rho}_\sigma)\,\phi'(\rho_K) = \frac{1}{2}(\rho_\sigma - \bar{\rho}_\sigma)\left(\phi'(\rho_K) - \phi'(\rho_L)\right) + \frac{1}{2}(\rho_\sigma - \bar{\rho}_\sigma)\left(\phi'(\rho_K) + \phi'(\rho_L)\right),$$

we have

$$T_{K,\rho} = T_{\text{conv},K,\rho} + \frac{1}{2}\sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\,(\rho_\sigma - \bar{\rho}_\sigma)\left(\phi'(\rho_K) - \phi'(\rho_L)\right) \geq T_{\text{conv},K,\rho},$$

thanks to the MUSCL limitation (3.34) and the convexity of ϕ. Therefore,

$$\frac{|K|}{\delta t}(\phi(\rho_K) - \phi(\bar{\rho}_K)) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\phi(\rho_\sigma) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\left(\rho_K\phi'(\rho_K) - \phi(\rho_K)\right) + T_{\text{conv},K,\rho} \leq 0. \tag{3.40}$$

Now consider a regular nonincreasing convex function ψ from $\mathbb{R}_+^*$ to $\mathbb{R}$. Multiplying equation (3.6e) by $\psi'(e_K)$ leads to

$$\frac{|K|}{\delta t}(\rho_K e_K - \bar{\rho}_K \bar{e}_K)\psi'(e_K) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}(e_\sigma - e_K)\psi'(e_K) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma} e_K \psi'(e_K) + p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\psi'(e_K) = S_K\psi'(e_K) \leq 0.$$

Using the mass balance equation (3.6d), we get that $\sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma} e_K \psi'(e_K) = \frac{|K|}{\delta t}(\bar{\rho}_K - \rho_K)e_K\psi'(e_K)$. This leads to

$$\frac{|K|}{\delta t}\bar{\rho}_K(e_K - \bar{e}_K)\psi'(e_K) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}(e_\sigma - e_K)\psi'(e_K) + p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\psi'(e_K) \leq 0$$

The convexity of the function ψ implies that

$$(e_K - \bar{e}_K)\psi'(e_K) \leq \psi(e_K) - \psi(\bar{e}_K)$$

so we have

$$\frac{|K|}{\delta t}\bar{\rho}_K(\psi(e_K) - \psi(\bar{e}_K)) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}(e_\sigma - e_K)\psi'(e_K) + p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\psi'(e_K) \leq 0.$$

Proceeding as for the density and using once again the mass balance equation to get that $\sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\psi(e_K) = \frac{|K|}{\delta t}(\bar{\rho}_K - \rho_K)\psi(e_K)$, we obtain

$$\frac{|K|}{\delta t}(\rho_K\psi(e_K) - \bar{\rho}_K\psi(\bar{e}_K)) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\psi(e_\sigma) + p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\psi'(e_K) + T_{K,e} \leq 0,$$

with

$$T_{K,e} = \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\left\{(e_\sigma - e_K)\psi'(e_K) + \psi(e_K) - \psi(e_\sigma)\right\}.$$

Applying the exact same process as for the density leads directly to :

$$\frac{|K|}{\delta t}(\rho_K\psi(e_K) - \bar{\rho}_K\psi(\bar{e}_K)) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\psi(e_\sigma) + p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\psi'(e_K) + T_{\text{conv},K,e} \leq 0, \tag{3.41}$$

where $T_{\text{conv},K,e}$ is defined in (3.39). In order to conclude, we only need to take $\phi(\rho) = \rho\ln(\rho)$, $\psi(e) = \frac{1}{1-\gamma}\ln(e)$ and sum the inequalities :

$$\frac{|K|}{\delta t}(\rho_K\eta_K - \bar{\rho}_K\bar{\eta}_K) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\eta_\sigma + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}(\rho_K + \rho_K\ln(\rho_K) - \rho_K\ln(\rho_K)) + p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\frac{1}{(1-\gamma)e_K} + T_{\text{conv},K} \leq 0.$$

Recalling that $p = (\gamma - 1)\rho e$ we directly have

$$\frac{|K|}{\delta t}(\rho_K\eta_K - \bar{\rho}_K\bar{\eta}_K) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\eta_\sigma + T_{\text{conv},K} \leq 0,$$

with $\eta_\sigma = \phi(\rho_\sigma) + \rho_\sigma\psi(e_\sigma)$, which concludes the proof for the pressure correction scheme.

For the decoupled scheme, we take the same notations and functions except that we consider that ρ refers to the time n and $\bar{\rho}$ refers to the time $n+1$. Let us multiply the decoupled mass balance equation (3.5a) by $\phi'(\bar{\rho}_K)$

$$\frac{|K|}{\delta t}(\bar{\rho}_K - \rho_K)\phi'(\bar{\rho}_K) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\rho_\sigma\phi'(\rho_K) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\rho_\sigma(\phi'(\bar{\rho}_K) - \phi'(\rho_K)) = 0$$

The first two terms are already studied for the pressure correction scheme. A simple Taylor expansion implies

$$\sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\rho_\sigma(\phi'(\bar\rho_K)-\phi'(\rho_K)) = \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\rho_\sigma(\bar\rho_K-\rho_K)\phi''(\rho_K^{(1)}),$$

with $\rho_K^{(1)} \in |[\rho_K, \bar\rho_K]|$ so that

$$\frac{|K|}{\delta t}(\phi(\bar\rho_K)-\phi(\rho_K)) + \sum_{\sigma\in\mathcal{E}(K)} |\sigma|\phi(\rho_\sigma)u_{K,\sigma} + \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\left(\rho_K\phi'(\rho_K)-\phi(\rho_K)\right)$$
$$+ \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\rho_\sigma(\bar\rho_K-\rho_K)\phi''(\rho_K^{(1)}) + T_{\mathrm{conv},K,\rho} \le 0. \quad (3.42)$$

We now multiply the internal energy balance equation (3.5b) by $\psi'(\bar e_K)$. Reordering and using the mass balance equation we directly get that

$$\frac{|K|}{\delta t}\bar\rho_K(\bar e_K - e_K)\psi'(\bar e_K) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}(e_\sigma - e_K)\psi'(e_K) + p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\psi'(e_K)$$
$$+ p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}(\psi'(\bar e_K)-\psi'(e_K)) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}(e_\sigma - e_K)(\psi'(\bar e_K)-\psi'(e_K)) \le 0$$

Thanks to the convexity of the function ψ we have :

$$(\bar e_K - e_K)\psi'(\bar e_K) \ge \psi(\bar e_K) - \psi(e_K),$$

and combined with the mass balance equation we get :

$$\frac{|K|}{\delta t}\bar\rho_K(\bar e_K - e_K)\psi'(\bar e_K) \ge \frac{|K|}{\delta t}\left(\bar\rho_K\psi(\bar e_K) - \rho_K\psi(e_K)\right) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\psi(e_K).$$

As a result, performing the same computations as in the pressure correction case, we have :

$$\frac{|K|}{\delta t}\bar\rho_K(\bar e_K - e_K)\psi'(\bar e_K) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}(e_\sigma - e_K)\psi'(e_K) \ge \frac{|K|}{\delta t}\left(\bar\rho_K\psi(\bar e_K) - \rho_K\psi(e_K)\right) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\psi(e_\sigma) + T_{\mathrm{conv},K,e}.$$

Finally we get that :

$$\frac{|K|}{\delta t}(\rho_K\psi(e_K) - \bar\rho_K\psi(\bar e_K)) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\psi(e_\sigma) + p_K \sum_{\sigma\in\mathcal{E}(K)} |\sigma| u_{K,\sigma}\psi'(e_K)$$
$$+ \sum_{\sigma\in\mathcal{E}(K)} \left[|\sigma| p_K u_{K,\sigma} + F_{K,\sigma}(e_\sigma - e_K)\right](\bar e_K - e_K)\psi''(e_K^{(1)}) + T_{\mathrm{conv},K,e} \le 0, \quad (3.43)$$

with $e_K^{(1)} \in |[e_K, \bar e_K]|$ Summing (3.42) and (3.43) leads directly to the desired result. ■

Remark 3.1

The previous theorem can be seen as a discrete version of the continuous entropy inequality of the Euler system. As in the case of the discrete version of the kinetic energy balance, one can notice the presence of remainder terms. These terms, under some assumed uniform estimates on the discrete solutions, will tend to zero as time and space step vanish so the limit of a converging sequence of discrete solutions will satisfy a weak entropy inequality at the limit.

3.5 Consistency of the schemes

The objective of this section is to show that if a sequence of solutions is controlled in suitable norms and converges to a limit, this latter necessarily satisfies a weak formulation of

the continuous problem. This is the so called Lax consistency. For the sake of clarity we focus on the results for the decoupled scheme and we point out the main differences with the pressure correction scheme at a later stage.

A weak solution to the continuous problem satisfies, for any $\varphi \in \mathrm{C}_c^\infty\big(\Omega \times [0,T)\big)$ ($\boldsymbol{\varphi} \in \mathrm{C}_c^\infty\big(\Omega \times [0,T)\big)^d$) :

$$-\int_0^T \int_\Omega \Big[\rho\, \partial_t \varphi + \rho\, u \cdot \nabla\varphi\Big]\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} = 0, \tag{3.44a}$$

$$-\int_0^T \int_\Omega \Big[\rho\, \boldsymbol{u} \cdot \partial_t \boldsymbol{\varphi} + (\rho \boldsymbol{u} \otimes \boldsymbol{u}) : \underline{\underline{\nabla}}\boldsymbol{\varphi} + p \mathrm{div}(\boldsymbol{\varphi})\Big]\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(\boldsymbol{x})\, \boldsymbol{u}_0(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} = 0, \tag{3.44b}$$

$$-\int_{\Omega\times(0,T)} \Big[\rho\, E\, \partial_t \varphi + (\rho\, E + p)\, \boldsymbol{u} \cdot \nabla\varphi\Big]\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(\boldsymbol{x})\, E_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} = 0, \tag{3.44c}$$

$$p = (\gamma - 1)\rho\, e, \quad E = \frac{1}{2}|\boldsymbol{u}|^2 + e, \quad E_0 = \frac{1}{2}|\boldsymbol{u}_0|^2 + e_0. \tag{3.44d}$$

This weak system is completed with a weak entropy inequality, for any $\varphi \in \mathrm{C}_c^\infty\big(\Omega \times [0,T),\ \mathbb{R}_+\big)$:

$$-\int_0^T \int_\Omega \Big[\rho\, \eta\, \partial_t \varphi + (\rho\, \eta\, \boldsymbol{u}) \cdot \nabla\varphi\Big]\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(\boldsymbol{x})\, \eta_0(\boldsymbol{x})\, \varphi(\boldsymbol{x},0)\, \mathrm{d}\boldsymbol{x} \le 0, \tag{3.45}$$

Note that these relations are not sufficient to define a weak solution to the problem, since they do not imply anything about the boundary conditions. However, they allow to derive the Rankine-Hugoniot conditions ; hence if we show that they are satisfied by the limit of a sequence of solutions to the discrete problem, this implies, loosely speaking, that *the scheme computes correct entropic shocks* (*i.e.* shocks where the jumps of the unknowns and of the fluxes are linked to the shock speed by the Rankine-Hugoniot conditions and for which an entropy condition is satisfied). This is the result we are seeking and which we state in Theorems 3.9-3.10. In order to prove these theorems, we need some definitions of interpolates of regular test functions on the primal and dual meshes, along with some assumed uniform estimates on the approximate solutions.

3.5.1 Definitions and assumptions

Some notations and definitions are common between the MAC and RT discretization. Let us denote by $L_{\mathcal{M}}(\Omega \times (0,T))$ the space of piecewise constant functions on each $K \times (t^n, t^{n+1})$, $K \in \mathcal{M}$, $n \in [\![0, N-1]\!]$. We also define the natural interpolation operator $\mathcal{P}_{\mathcal{M}}$

$$\mathcal{P}_{\mathcal{M}} : \left| \begin{array}{lcl} \mathrm{C}((0,T); H_0^1(\Omega)) & \longrightarrow & L_{\mathcal{M}}(\Omega \times (0,T)) \\ \varphi & \mapsto & \mathcal{P}_{\mathcal{M}}\varphi(\boldsymbol{x}) = \displaystyle\sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}} \varphi_K^{n+1}\, \mathcal{X}_K\, \mathcal{X}_{(t^n, t^{n+1}]}, \end{array} \right.$$

with $\varphi_K^n = \dfrac{1}{|K|} \displaystyle\int_K \varphi(\boldsymbol{x}, t^n) \mathbf{d}\boldsymbol{x}$ and $\mathcal{X}_P$ is the indicator function of P.

We define by h_K the diameter of a cell $K \in \mathcal{M}$ and by r_K the radius of the largest ball included in K. For $\sigma \in \mathcal{E}$ we denote by $\boldsymbol{x}_\sigma$ its mass center. For a dual edge $\epsilon = D_\sigma | D_{\sigma'} \in \tilde{\mathcal{E}}_{\mathrm{int}}$, we denote by $d_{\sigma,\epsilon}$ the Euclidean distance between ϵ and $\boldsymbol{x}_\sigma$ and $d_\epsilon = d_{\sigma,\epsilon} + d_{\sigma',\epsilon}$. The size of the mesh is measured through the quantity

$$h_{\mathcal{M}} = \sup \{h_K,\ K \in \mathcal{M}\}.$$

For a function $q \in L_{\mathcal{M}}(\Omega \times (0,T))$ we define a discrete $\mathrm{L}^1((0,T); \mathrm{BV}(\Omega))$ norm by :

$$\|q\|_{\mathcal{T},x,\mathrm{BV}} = \sum_{n=0}^{N} \delta t \sum_{\sigma = K|L \in \mathcal{E}_{\mathrm{int}}} |\sigma| |q_L^n - q_K^n|,$$

and a discrete $\mathrm{L}^1(\Omega;\mathrm{BV}((0,T)))$ norm by :

$$\|q\|_{\mathcal{T},t,\mathrm{BV}} = \sum_{K\in\mathcal{M}} |K| \sum_{n=0}^{N-1} |q_K^{n+1} - q_K^n|.$$

RT scheme

We measure the regularity of the mesh through the positive real number θ defined by

$$\theta = \max\left\{\frac{h_K}{r_K},\ K\in\mathcal{M}\right\} \cup \left\{\frac{h_K}{d_\epsilon},\ K\in\mathcal{M},\ \epsilon\subset K\right\} \cup \left\{\frac{d_{\sigma,\epsilon}}{d_{\sigma',\epsilon}},\ \sigma,\sigma'\in\mathcal{E},\ \epsilon=\sigma|\sigma'\right\} \tag{3.46}$$

We define the space $H_{\mathcal{E}}(\Omega\times(0,T))$ of functions constant on every $D_\sigma\times(t^n,t^{n+1})$, $\sigma\in\mathcal{E}$, $n\in[\![0,N-1]\!]$. We denote by $H_{\mathcal{E},0}(\Omega\times(0,T))$ the subspace of functions null on every D_σ, $\sigma\in\mathcal{E}_{\mathrm{ext}}$. We naturally define its interpolation operator $\mathcal{P}_{\mathcal{E}}$

$$\mathcal{P}_{\mathcal{E}} : \left| \begin{array}{lcl} C((0,T);H_0^1(\Omega)) & \longrightarrow & H_{\mathcal{E}}(\Omega\times(0,T)) \\ \varphi & \mapsto & \mathcal{P}_{\mathcal{E}}\varphi(\boldsymbol{x}) = \displaystyle\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}} \varphi_\sigma^{n+1}\,\mathcal{X}_{D_\sigma}\,\mathcal{X}_{(t^n,t^{n+1}]}, \end{array}\right. \tag{3.47}$$

with $\varphi_\sigma^n = \dfrac{1}{|\sigma|}\displaystyle\int_\sigma \varphi(\boldsymbol{x},t^n)\mathrm{d}\gamma(\boldsymbol{x})$.

For a discrete function $\boldsymbol{v}\in H_{\mathcal{E},0}(\Omega\times(0,T))^d$, we define a discrete $\mathrm{L}^1((0,T);\mathrm{BV}(\Omega))$ norm by :

$$\|\boldsymbol{v}\|_{\mathcal{T},x,\mathrm{BV}} = \sum_{n=0}^{N}\delta t \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}_{\mathrm{int}}} |\epsilon|\|\boldsymbol{v}_{\sigma'}^n - \boldsymbol{v}_\sigma^n|,$$

and a discrete $\mathrm{L}^1(\Omega;\mathrm{BV}((0,T)))$ norm by :

$$\|\boldsymbol{v}\|_{\mathcal{T},t,\mathrm{BV}} = \sum_{\sigma\in\mathcal{E}} |D_\sigma| \sum_{n=0}^{N-1} |\boldsymbol{v}_\sigma^{n+1} - \boldsymbol{v}_\sigma^n|.$$

Définition 3.2 (*Interpolates on multi-dimensional meshes*)

Let Ω be an open bounded interval of $\mathbb{R}$, and let $\mathcal{M}$ be a mesh over Ω. Let $\varphi_{\mathcal{M}}\in L_{\mathcal{M}}(\Omega\times(0,T))$. The discrete time derivative of the discrete function $\varphi_{\mathcal{M}}$ is defined by :

$$\eth_t\varphi_{\mathcal{M}} = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} \frac{\varphi_K^{n+1}-\varphi_K^n}{\delta t}\,\mathcal{X}_K\,\mathcal{X}_{(t^n,t^{n+1}]},$$

and its discrete space gradient by :

$$\boldsymbol{\nabla}_{\mathcal{E}}\varphi_{\mathcal{M}} = \sum_{n=0}^{N-1}\sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}_{\mathrm{int}}} \frac{|\sigma|}{|D_\sigma|}\left(\varphi_L^{n+1}-\varphi_K^{n+1}\right)\boldsymbol{n}_{K,\sigma}\,\mathcal{X}_{D_\sigma}\,\mathcal{X}_{(t^n,t^{n+1}]}.$$

In case of a vector $\boldsymbol{\varphi}_{\mathcal{M}}\in L_{\mathcal{M}}(\Omega\times(0,T))^d$, we define the discrete gradient matrice

$$\underline{\underline{\boldsymbol{\nabla}_{\mathcal{E}}}}\boldsymbol{\varphi}_{\mathcal{M}} = \sum_{n=0}^{N-1}\sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}_{\mathrm{int}}} \frac{|\sigma|}{|D_\sigma|}\left(\boldsymbol{\varphi}_L^{n+1}-\boldsymbol{\varphi}_K^{n+1}\right)\otimes\boldsymbol{n}_{K,\sigma}\,\mathcal{X}_{D_\sigma}\,\mathcal{X}_{(t^n,t^{n+1}]},$$

where $\otimes$ design the tensorial product. Now consider $\varphi_{\mathcal{E}}\in H_{\mathcal{E},0}(\Omega\times(0,T))$ We also define

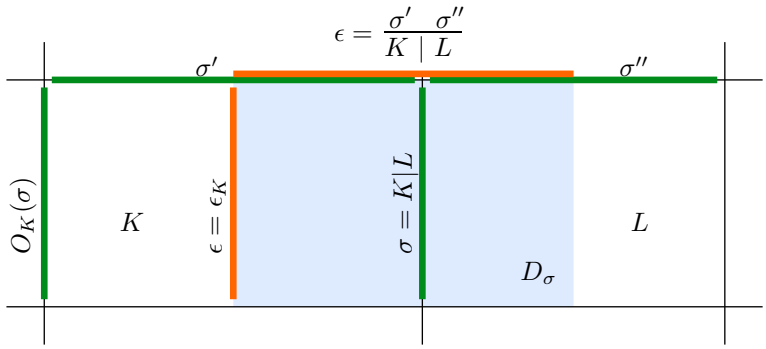


Figure 3.2 – Notations for the dual mesh.

the time derivative of this function by :

$$\eth_t \varphi_{\mathcal{E}} = \sum_{n=0}^{N-1} \sum_{\sigma \in \mathcal{E}} \frac{\varphi_\sigma^{n+1} - \varphi_\sigma^n}{\delta t} \mathcal{X}_{D_\sigma} \mathcal{X}_{(t^n, t^{n+1}]},$$

and, in case of a vector, the discrete divergence operator :

$$\mathrm{div}_{\mathcal{M}} \boldsymbol{\varphi}_{\mathcal{E}} = \sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}} \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} |\sigma| \, \boldsymbol{\varphi}_\sigma^{n+1} . \boldsymbol{n}_{K,\sigma}.$$

Finally, for $\varphi \in H_{\mathcal{E}}(\Omega \times (0,T))$ and $\psi \in L_{\mathcal{M}}(\Omega \times (0,T))$, we define a gradient operator $\boldsymbol{\nabla}^h$ based on both discretization by :

$$\boldsymbol{\nabla}^h(\varphi, \psi) = \sum_{n=0}^{N-1} \sum_{D_{K,\sigma} \in \boldsymbol{Q}} \frac{|\sigma|}{|D_{K,\sigma}|} \left(\varphi_\sigma^{n+1} - \psi_K^{n+1}\right) \boldsymbol{n}_{K,\sigma} \mathcal{X}_{D_{K,\sigma}} \mathcal{X}_{(t^n, t^{n+1}]},$$

with $\boldsymbol{Q}$ the set of half-diamond cells $D_{K,\sigma}$.

We now turn to the specific notations of the MAC discretization

MAC scheme

Many notations introduced for the RT case are still valid here. The regularity of the mesh is measured through

$$\theta = \max \left\{ \frac{h_K}{r_K}, \ K \in \mathcal{M} \right\} \cup \left\{ \frac{d_{\sigma,\epsilon}}{d_{\sigma',\epsilon}}, \ \sigma, \sigma' \in \mathcal{E}, \ \epsilon = \sigma | \sigma' \right\} \tag{3.48}$$

We will introduce the following notations for the dual fluxes $F_{\sigma,\epsilon}$ which will be useful later on. Let us consider a direction (i), $i \in [\![1, d]\!]$ and an edge $\sigma \in \mathcal{E}_{\mathrm{int}}^{(i)}$. If $\sigma \in \mathcal{E}(K)$, we denote by $O_K(\sigma)$ the opposite face to σ in K. For a dual face $\epsilon \in \tilde{\mathcal{E}}(D_\sigma)$, we distinguish two cases :

— The vector $\boldsymbol{e}^{(i)}$ is normal to ϵ and included in cell K. We write $\epsilon = \epsilon_K$.

— The vector $\boldsymbol{e}^{(i)}$ is tangent to ϵ, and ϵ is the union of the half of two primal faces. Let us denote by σ' and σ'' these two primal faces, and let us suppose that $\sigma = K|L$ with $\sigma' \in \mathcal{E}(K)$ and $\sigma'' \in \mathcal{E}(L)$. Then we write $\sigma = \dfrac{\sigma' \quad \sigma''}{K \mid L}$.

Let u be a discrete function defined on $\sigma \in \mathcal{E}^{(i)}$. For $\sigma \in \mathcal{E}$, let us define $\hat{u}_\sigma^{(i)}$ by :

$$\hat{u}_\sigma^{(i)} = u_\sigma \text{ if } \sigma \in \mathcal{E}^{(i)}, \ \hat{u}_\sigma^{(i)} = \frac{1}{\mathrm{card}(\mathcal{N}_\sigma)} \sum_{\sigma' \in \mathcal{N}_\sigma} u_{\sigma'} \text{ otherwise.} \tag{3.49}$$

where, for $\sigma \in \mathcal{E} \setminus \mathcal{E}^{(i)}$, $\mathcal{N}_\sigma = \{\sigma' \in \mathcal{E}^{(i)},\ \bar{D}_\sigma \cap \sigma' \neq \emptyset\}$. We also denote by u_K the quantity :

$$u_K = \frac{1}{2} \sum_{\sigma \in \mathcal{E}^{(i)}(K)} u_\sigma \tag{3.50}$$

The main difference lies in the definition of spaces based on the dual mesh, because of the fact that there is a dual mesh for each direction of space. Consequently we denote by $H^{(i)}_{\mathcal{E}}(\Omega \times (0,T))$ the space of constant functions on each $D_\sigma \times (t^n, t^{n+1})$, $\sigma \in \mathcal{E}^{(i)}$, $i \in [\![1,d]\!]$ and by $H^{(i)}_{\mathcal{E},0}(\Omega \times (0,T))$ the subspace of functions null on the boundary. We denote by $\mathcal{P}^{(i)}_{\mathcal{E}^{(i)}_{\mathcal{S}}}$ its natural interpolation operator (we use a similar interpolation as in (3.47)).

BV norms are now defined for each direction so we introduce, for $v \in H^{(i)}_{\mathcal{E},0}(\Omega \times (0,T))$, a discrete $\mathrm{L}^1(\Omega; \mathrm{BV}((0,T)))$ norm by :

$$\|v\|_{\mathcal{T},x,\mathrm{BV},(i)} = \sum_{n=0}^{N} \delta t \sum_{\epsilon = D_\sigma | D_{\sigma'} \in \tilde{\mathcal{E}}^{(i)}_{\mathrm{int}}} |\epsilon| |v^n_{\sigma'} - v^n_\sigma|,$$

and a discrete $\mathrm{L}^1(\Omega; \mathrm{BV}((0,T)))$ norm by :

$$\|v\|_{\mathcal{T},t,\mathrm{BV},(i)} = \sum_{\sigma \in \mathcal{E}^{(i)}} |D_\sigma| \sum_{n=0}^{N-1} |v^{n+1}_\sigma - v^n_\sigma|.$$

For the vector $\boldsymbol{v} = (v^{(1)}, \cdots v^{(d)}) \in H^{(1)}_{\mathcal{E},0}(\Omega \times (0,T)) \times \cdots \times H^{(d)}_{\mathcal{E},0}(\Omega \times (0,T))$:

$$\|\boldsymbol{v}\|_{\mathcal{T},t,\mathrm{BV}} = \max_{i \in [\![1,d]\!]} \|v^{(i)}\|_{\mathcal{T},x,\mathrm{BV},(i)} \qquad \|\boldsymbol{v}\|_{\mathcal{T},x,\mathrm{BV}} = \max_{i \in [\![1,d]\!]} \|v^{(i)}\|_{\mathcal{T},x,\mathrm{BV},(i)}.$$

We turn to the new definitions of discrete derivatives for functions defined on $H^{(i)}_{\mathcal{E},0}(\Omega \times (0,T))$.

Définition 3.3 (*Interpolates on multi-dimensional meshes*)

Let Ω be an open bounded interval of $\mathbb{R}$, and let $\mathcal{M}$ be a mesh over Ω. Let $\varphi_{\mathcal{M}} \in L_{\mathcal{M}}(\Omega \times (0,T))$. We define its space discrete gradient by :

$$\boldsymbol{\nabla}_{\mathcal{E}} \varphi_{\mathcal{M}} = \sum_{i=1}^{d} \eth_{x,i} \varphi_{\mathcal{M}} \boldsymbol{e}^{(i)}$$

with :

$$\eth_{x,i} \varphi_{\mathcal{M}} = \sum_{n=0}^{N-1} \sum_{\sigma = \overrightarrow{K|L} \in \mathcal{E}^{(i)}_{\mathrm{int}}} \frac{|\sigma|}{|D_\sigma|} \left(\varphi^{n+1}_L - \varphi^{n+1}_K\right) \mathcal{X}_{D_\sigma} \mathcal{X}_{(t^n, t^{n+1}]}.$$

For a vector $\boldsymbol{\varphi}_{\mathcal{M}} = (\varphi^{(1)}_{\mathcal{M}}, \cdots \varphi^{(d)}_{\mathcal{M}}) \in H^{(1)}_{\mathcal{E},0}(\Omega \times (0,T)) \times \cdots \times H^{(d)}_{\mathcal{E},0}(\Omega \times (0,T))$ we naturally define the discrete gradient matrix by :

$$\left(\underline{\underline{\boldsymbol{\nabla}_{\mathcal{E}}}} \boldsymbol{\varphi}_{\mathcal{M}}\right)_{(i,j)} = \eth_{x,j} \varphi^{(i)}_{\mathcal{M}}.$$

Finally the hybrid gradient is defined, for $\boldsymbol{\varphi} = (\varphi^{(1)}, \cdots \varphi^{(d)}) \in H^{(1)}_{\mathcal{E},0}(\Omega \times (0,T)) \times \cdots \times H^{(d)}_{\mathcal{E},0}(\Omega \times (0,T))$ and $\psi \in L_{\mathcal{M}}(\Omega \times (0,T))$, by :

$$\boldsymbol{\nabla}^h(\boldsymbol{\varphi},\ \psi) = \sum_{i=1}^{d} \partial^h_{x,i}(\varphi^{(i)},\ \psi) \boldsymbol{e}^{(i)},$$

with

$$\partial^h_{x,i}(\varphi^{(i)},\ \psi) = \sum_{n=0}^{N-1} \sum_{D_{K,\sigma}\in \boldsymbol{Q}^{(i)}} \frac{|\sigma|}{|D_{K,\sigma}|} \left((\varphi^{(i)})^{n+1}_\sigma - \psi^{n+1}_K\right) \boldsymbol{n}_{K,\sigma}\, \mathcal{X}_{D_{K,\sigma}}\, \mathcal{X}_{(t^n,t^{n+1}]},$$

with $\boldsymbol{Q}^{(i)}$ the set of half cells $D_{K,\sigma}$ for the dual mesh associated to the direction i.

Definition of global solution

Let a sequence of discretizations $(\mathcal{M}^{(m)},\ \delta t^{(m)})_{m\in\mathbb{N}}$ be given. Let us denote by $h^{(m)}$ the size of the mesh. Let $\rho^{(m)}$, $p^{(m)}$, $e^{(m)}$ and $\boldsymbol{u}^{(m)}$ be the solution given by the scheme (3.5) with the mesh $\mathcal{M}^{(m)}$ and the time step $\delta t^{(m)}$. To the discrete unknowns, we associate piecewise constant functions on time intervals and on primal or dual meshes, so the density $\rho^{(m)}$, the pressure $p^{(m)}$, the internal energy $e^{(m)}$ and the velocity $u^{(m)}$ are defined almost everywhere on $\Omega\times(0,T)$ by :

$$\begin{aligned}
\rho^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (\rho^{(m)})^n_K\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t),\\
p^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (p^{(m)})^n_K\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t),\\
e^{(m)}(\boldsymbol{x},t) &= \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (e^{(m)})^n_K\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t),\\
\boldsymbol{u}^{(m)}(\boldsymbol{x},t) &= \begin{cases} \displaystyle\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}} (\boldsymbol{u}^{(m)})^n_\sigma\, \mathcal{X}_{D_\sigma}(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t), & \text{RT-scheme}\\ \displaystyle\sum_{i=1}^{d}\sum_{n=0}^{N-1}\sum_{\sigma\in\mathcal{E}^{(i)}} (u^{(m)})^n_{\sigma,i}\, \mathcal{X}_{D_\sigma}(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t). & \text{MAC-scheme}\end{cases}
\end{aligned} \tag{3.51}$$

Likewise one can define $\eta^{(m)} \in L_{\mathcal{M}^{(m)}}(\Omega\times(0,T))$ by

$$\eta^{(m)}(\boldsymbol{x},t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (\eta^{(m)})^n_K\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{[n,n+1)}(t), \tag{3.52}$$

with the entropy $(\eta^{(m)})^n_K$ defined in (3.33).

Assumed estimates

In order to be able to obtain a consistency result (see theorem hereafter), we need to assume some estimates on the discrete solution. Under the CFL conditions 3.32, thanks to the lemma 3.3 a sequence of discrete solutions $(\rho^{(m)}, p^{(m)}, e^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ satisfies $\rho^{(m)} > 0$, $p^{(m)} > 0$ and $e^{(m)} > 0$, $\forall m \in \mathbb{N}$. Let us suppose that it is uniformly bounded in $\mathrm{L}^\infty(\Omega\times(0,T))^4$, *i.e.*, for $m \in \mathbb{N}$ and $0 \le n \le N^{(m)}$:

$$0 < (\rho^{(m)})^n_K \le C,\ 0 < (p^{(m)})^n_K \le C,\ 0 < (e^{(m)})^n_K \le C, \quad \forall K \in \mathcal{M}^{(m)}, \tag{3.53}$$

and

$$|(\boldsymbol{u}^{(m)})^n_\sigma| \le C, \quad \forall \sigma \in \mathcal{E}^{(m)}, \tag{3.54}$$

where C is a positive real number. We also suppose that $\dfrac{1}{\rho^{(m)}}$ and $\dfrac{1}{e^{(m)}}$ are in $\mathrm{L}^\infty(\Omega\times(0,T))$.

Note that, by definition of the initial conditions of the scheme, these inequalities imply that the functions ρ_0, e_0, u_0 belong to $\mathrm{L}^\infty(\Omega)$ and that

$$\rho_0(\boldsymbol{x}) > \rho_{\min} > 0 \qquad e_0 > e_{\min} > 0.$$

We also have to assume that a sequence of discrete solutions satisfies the following BV-stability assumption :

$$\lim_{m\to\infty} \left(h^{(m)} + \delta t^{(m)}\right)\left[\|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|p^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|e^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},t,\mathrm{BV}}\right] = 0. \tag{3.55}$$

Note that this is a much weaker assumption than the uniform bound on the discrete BV-norms.

We assume an additional hypothesis which is a strengthened CFL condition, namely :

$$\lim_{m\to+\infty} \frac{\delta t^{(m)}}{\min\limits_{K\in\mathcal{M}^{(m)}} h_K}\left(\|\rho^{(m)}\|_{\mathcal{T},t,\mathrm{BV}} + \|e^{(m)}\|_{\mathcal{T},t,\mathrm{BV}}\right) = 0 \tag{3.56}$$

3.5.2 Preliminary results

Before stating the main result of this paper, we give the two following lemmas.

Lemma 3.7

(Weak convergence of the discrete gradient).
Let $(\mathcal{M}^{(m)}, \delta t^{(m)})_{m\in\mathbb{N}}$ be a sequence of discretizations such that both the time step $\delta t^{(m)}$ and the size $h^{(m)}$ of the mesh $\mathcal{M}^{(m)}$ tend to zero as $m \to \infty$ and $\theta^{(m)} \le \theta_0$ for all $m \in \mathbb{N}$. Let us denote by $L_{\mathcal{M}^{(m)}}(\Omega)$ the space of piecewise constant functions on the primal mesh. For $m \in \mathbb{N}$, let $q^{(m)} \in L_{\mathcal{M}^{(m)}}(\Omega)$ and assume that there exists C in $\mathbb{R}_+$ such that, for all $m \in \mathbb{N}$, $\|\nabla_{\mathcal{E}^{(m)}} q^{(m)}\|_{\mathrm{L}^p(\Omega)^d} \le C$ for some p in $[1,\infty]$. Assume also that there exists $\bar{q}$ in $\mathrm{W}^{1,p}(\Omega)$ such that $q^{(m)}$ converges to $\bar{q}$ in the distribution sense as m tends to $+\infty$, i.e. *:*

$$\forall \varphi \in \mathrm{C}_c^\infty(\Omega), \quad \lim_{m\to+\infty} \int_\Omega q^{(m)}(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_\Omega \bar{q}(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Then $\nabla_{\mathcal{E}^{(m)}} q^{(m)}$ converges to $\nabla\bar{q}$ in the distribution sense :

$$\forall \boldsymbol{\varphi} \in \mathrm{C}_c^\infty(\Omega)^d, \quad \lim_{m\to+\infty} \int_\Omega \nabla_{\mathcal{E}^{(m)}} q^{(m)}(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_\Omega \nabla\bar{q}(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

In addition, for $p \in (1,\infty)$ (resp. $q = +\infty$), if $q^{(m)}$ weakly (resp. weakly-$\star$) converges to $\bar{q}$ in $\mathrm{L}^p(\Omega)$, then $\nabla_{\mathcal{E}^{(m)}} q^{(m)}$ also converges to $\nabla\bar{q}$ weakly (resp. weakly-$\star$) in $\mathrm{L}^p(\Omega)^d$.

Proof : Let $\boldsymbol{\varphi} \in \mathcal{C}_c^\infty(\Omega)^d$. For a given discretization $\mathcal{D} = (\mathcal{M},\mathcal{E})$, for $\sigma \in \mathcal{E}$, let $\boldsymbol{\varphi}_\sigma = |\sigma|^{-1}\int_\sigma \boldsymbol{\varphi}$, and let $P_{\mathcal{E}}\boldsymbol{\varphi}$ be the function defined by $P_{\mathcal{E}}\boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{\varphi}_\sigma$ if $\boldsymbol{x} \in D_\sigma$. With the assumptions of the lemma, an easy calculation shows that $\|P_{\mathcal{E}^{(m)}}\boldsymbol{\varphi} - \boldsymbol{\varphi}\|_{\mathrm{L}^{p'}(\Omega)^d} \le |\Omega|^{\frac{1}{p'}}\|\nabla\boldsymbol{\varphi}\|_{\mathrm{L}^\infty(\Omega)^{d\times d}} h^{(m)}$ where $\frac{1}{p} + \frac{1}{p'} = 1$. We may write

$$\int_\Omega \nabla_{\mathcal{E}^{(m)}} q^{(m)}(\boldsymbol{x}).\boldsymbol{\varphi}(\boldsymbol{x}) = \int_\Omega \nabla_{\mathcal{E}^{(m)}} q^{(m)}(\boldsymbol{x}).P_{\mathcal{E}^{(m)}}\boldsymbol{\varphi}(\boldsymbol{x}) + R,$$

with $|R| \le \|\nabla_{\mathcal{E}^{(m)}} q^{(m)}\|_{\mathrm{L}^p(\Omega)^d}\|P_{\mathcal{E}^{(m)}}\boldsymbol{\varphi} - \boldsymbol{\varphi}\|_{\mathrm{L}^{p'}(\Omega)^d} \le C|\Omega|^{\frac{1}{p'}}\|\nabla\boldsymbol{\varphi}\|_{\mathrm{L}^\infty(\Omega)^{d\times d}} h^{(m)} \to 0$ as $m \to +\infty$. By construction,

$$\int_\Omega \nabla_{\mathcal{E}^{(m)}} q^{(m)}(\boldsymbol{x}).P_{\mathcal{E}^{(m)}}\boldsymbol{\varphi}(\boldsymbol{x}) = -\int_\Omega q^{(m)}(\boldsymbol{x})\mathrm{div}_{\mathcal{M}^{(m)}} P_{\mathcal{E}^{(m)}}\boldsymbol{\varphi}(\boldsymbol{x}).$$

A quick calculation shows that :

$$\int_\Omega q^{(m)}(\boldsymbol{x})\mathrm{div}_{\mathcal{M}} P_{\mathcal{E}^{(m)}}\boldsymbol{\varphi}(\boldsymbol{x}) = \int_\Omega q^{(m)}(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}).$$

Passing to the limit, we get :

$$\int_\Omega \nabla_{\mathcal{E}^{(m)}} q^{(m)}(x).\varphi(x) \to -\int_\Omega \overline{q}(x)\mathrm{div}\varphi(x),$$

so that $\nabla_{\mathcal{E}^{(m)}} q^{(m)}$ tends to $\nabla\overline{q}$ in the distributional sense. The weak or weak-$\star$ convergence follows by density. ■

Remark 3.2

This lemma is true for both RT and MAC discretizations. The result is stronger in the MAC case as the convergence of the discrete gradient is strong. However, only the weak convergence is needed for the consistency result we are seeking.

This lemma only concerns the RT discretization.

Lemma 3.8

(Weak convergence of the hybrid gradient in the RT case).
Let $(\mathcal{M}^{(m)}, \delta t^{(m)})_{m\in\mathbb{N}}$ be a sequence of discretizations such that both the time step $\delta t^{(m)}$ and the size $h^{(m)}$ of the mesh $\mathcal{M}^{(m)}$ tend to zero as $m \to \infty$ and $\theta^{(m)} \le \theta_0$ for all $m \in \mathbb{N}$. Let us denote by $H_{\mathcal{E}^{(m)}}(\Omega)$ the space of functions constant in each dual cell and $L_{\mathcal{M}^{(m)}}(\Omega)$ the space of functions constant in each primal cell. For $m \in \mathbb{N}$, let $\hat{q}^{(m)} \in H_{\mathcal{E}^{(m)}}(\Omega)$, $q^{(m)} \in L_{\mathcal{M}^{(m)}}(\Omega)$ and assume that there exists $C \in \mathbb{R}$ such that, for all $m \in \mathbb{N}$,

$$\sum_{K\in\mathcal{M}^{(m)}} \sum_{\sigma\in\mathcal{E}^{(m)}(K)} \frac{|\sigma|^p}{|D_{K,\sigma}|^{p-1}} |q_K^{(m)} - \hat{q}_\sigma^{(m)}|^p \le C, \tag{3.57}$$

for some $p \in [1, +\infty]$. Assume also that there exists $\bar{q}$ in $\mathrm{W}^{1,p}(\Omega)$ such that $q^{(m)}$ converges to $\bar{q}$ in the distribution sense as m tends to $+\infty$. Then $\nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})$ converges to $\nabla\bar{q}$ in the distribution sense. In addition, for $p \in (1, \infty)$ (resp. $p = +\infty$), if $q^{(m)}$ weakly (resp. weakly-$\star$) converges to $\bar{q}$ in $\mathrm{L}^p(\Omega)$, then $\nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})$ also converges to $\nabla\bar{q}$ weakly (resp. weakly-$\star$) in $\mathrm{L}^p(\Omega)^d$.

Proof : Let $\varphi \in C_c^\infty(\Omega)^d$. For a given discretization $\mathcal{D} = (\mathcal{M}, \mathcal{E})$, for $\sigma \in \mathcal{E}$, let $\varphi_\sigma = \varphi(x_\sigma)$, $\varphi_K = \varphi(x_K)$, and let $P_\mathcal{E}\varphi$ be the function defined by $P_\mathcal{E}\varphi(x) = \varphi_\sigma$ if $x \in D_\sigma$ and $P_\mathcal{M}\varphi$ by $P_\mathcal{M}\varphi(x) = \varphi_K$ if $x \in K$. We have :

$$\int_\Omega \nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})(x).\varphi(x) = \int_\Omega \nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})(x).P_{\mathcal{E}^{(m)}}\varphi(x) + R,$$

with $|R| \le \|\nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})\|_{\mathrm{L}^p(\Omega)^d} \|P_{\mathcal{E}^{(m)}}\varphi - \varphi\|_{\mathrm{L}^{p'}(\Omega)^d} \le C|\Omega|^{\frac{1}{p'}} \|\nabla\varphi\|_{\mathrm{L}^\infty(\Omega)^{d\times d}} h^{(m)} \to 0$ as $m \to +\infty$. Consequently, $|R| \to 0$ as $h^{(m)} \to 0$, with $\frac{1}{p} + \frac{1}{p'} = 1$. After a quick computation we have :

$$\int_\Omega \nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})(x).P_{\mathcal{E}^{(m)}}\varphi(x) = \int_\Omega \nabla_{\mathcal{E}^{(m)}} q^{(m)}(x).P_{\mathcal{E}^{(m)}}\varphi(x).$$

As a result we get that :

$$\int_\Omega \nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})(x).P_{\mathcal{E}^{(m)}}\varphi(x) = -\int_\Omega q^{(m)}(x)\mathrm{div}\varphi(x).$$

$q^{(m)}$ converges towards $\overline{q}$ in the distribution sense and finally,

$$\int_\Omega \nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})(x).\varphi(x) \to -\int_\Omega \overline{q}(x)\mathrm{div}\varphi(x), \quad \text{as } m \to \infty.$$

$\nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})$ tends to $\nabla\overline{q}$ in the distributional sense. The weak or weak-$\star$ convergence of $\nabla^{h^{(m)}}(\hat{q}^{(m)}, q^{(m)})$ follows by density. ■

Remark 3.3 (*Weak convergence of the hybrid gradient in the MAC case*)

The hybrid gradient defined in the MAC case is different, because there is one dual mesh per direction of space. Therefore the previous lemma cannot be applied directly to the hybrid gradient in the MAC case. However, with few adjustments, we can proove the weak convergence of each component of the hybrid gradient $\partial^h_{x,i}$, $i \in [\![1,d]\!]$. We just point out the minor changes to perform in the previous lemma without writing it again as it would be redundant. We consider this time a sequence of discrete functions in $\hat{q}^{(m)} \in H_{\mathcal{E}^{(m)}}(\Omega)^{(i)}$. The condition (3.57) becomes :

$$\sum_{D_{K,\sigma}\in\mathcal{Q}^{(i)}} \frac{|\sigma|^p}{|D_{K,\sigma}|^{p-1}} |q_K^{(m)} - \hat{q}_\sigma^{(m)}|^p \leq C. \tag{3.58}$$

Everything else is identical. The weak convergence of each component leads to the weak convergence of the hybrid gradient.

3.5.3 Consistency Theorem

The main result of this paper is given hereafter

Theorem 3.9 (*Consistency of the multi-dimensional decoupled scheme*)

Let Ω be an open bounded interval of $\mathbb{R}$. We suppose that the initial data satisfies $\rho_0 \in \mathrm{L}^\infty(\Omega)$, $p_0 \in \mathrm{BV}(\Omega)$, $e_0 \in \mathrm{L}^\infty(\Omega)$ and $\boldsymbol{u}_0 \in \mathrm{L}^\infty(\Omega)^d$. Let $(\mathcal{M}^{(m)}, \delta t^{(m)})_{m\in\mathbb{N}}$ be a sequence of discretizations such that both the time step $\delta t^{(m)}$ and the size $h^{(m)}$ of the mesh $\mathcal{M}^{(m)}$ tend to zero as $m \to \infty$, and let $(\rho^{(m)}, p^{(m)}, e^{(m)}, \boldsymbol{u}^{(m)})_{m\in\mathbb{N}}$ be the corresponding sequence of solutions. We suppose that this sequence satisfies the estimates (3.53)–(3.55) and converges in $\mathrm{L}^r(\Omega\times(0,T))^3 \times \mathrm{L}^r(\Omega\times(0,T))^d$, for $1 \leq r < \infty$, to $(\bar{\rho}, \bar{p}, \bar{e}, \bar{\boldsymbol{u}}) \in \mathrm{L}^\infty(\Omega\times(0,T))^3 \times \mathrm{L}^\infty(\Omega\times(0,T))^d$. Furthermore suppose the additional limitation (3.34) on the MUSCL interpolation and suppose that the cfl condition (3.56) is satisfied.

Then the limit $(\bar{\rho}, \bar{p}, \bar{e}, \bar{\boldsymbol{u}})$ satisfies the system (3.44)-(3.52).

RT Case

Proof : It is clear that with the assumed convergence for the sequence of solutions, the limit satisfies the equation of state. Let $\varphi \in C_C^\infty(\Omega\times[0,T))$, $m \in \mathbb{N}$, $\mathcal{M}^{(m)}$ and $\delta t^{(m)}$ be given. Dropping the superscript $^{(m)}$, thanks to the regularity of φ and the lemma (3.7), $\eth_t\mathcal{P}_\mathcal{M}\varphi$ and $\mathcal{P}_\mathcal{M}\varphi$ converges in $\mathrm{L}^r(\Omega\times(0,T))$, for $r \geq 1$, to $\partial_t\varphi$ and φ respectively, and $\boldsymbol{\nabla}_\mathcal{E}\mathcal{P}_\mathcal{M}\varphi$ converges weakly in $\mathrm{L}^r(\Omega\times(0,T))$ to $\boldsymbol{\nabla}\varphi$. Besides, $\mathcal{P}_\mathcal{M}\varphi(.,0)$ converges to $\varphi(.,0)$ in $\mathrm{L}^r(\Omega)$.

Similarly, $\mathcal{P}_\mathcal{E}\varphi$, $\eth_t\mathcal{P}_\mathcal{E}\varphi$ converge in $\mathrm{L}^r(\Omega\times(0,T))$, for $r \geq 1$ towards φ and $\partial_t\varphi$, while the hybrid gradient $\boldsymbol{\nabla}^h(\mathcal{P}_\mathcal{M}\varphi, \mathcal{P}_\mathcal{E}\varphi)$ converges weakly in $\mathrm{L}^r(\Omega\times(0,T))$ towards $\boldsymbol{\nabla}\varphi$. Besides, $\mathcal{P}_\mathcal{E}\varphi(.,0)$ converges to $\varphi(.,0)$ in $\mathrm{L}^r(\Omega)$.

Let $\boldsymbol{\varphi} \in C_C^\infty(\Omega\times[0,T))^d$. Let us denote $\mathcal{P}_\mathcal{E}\boldsymbol{\varphi}$ its interpolate on the dual mesh. We introduce an interpolation on the primal mesh $\boldsymbol{\varphi}_\mathcal{M}$ derived from the interpolation on the dual mesh thanks to the relation $\boldsymbol{\varphi}_K = \sum_{\sigma\in\mathcal{E}(K)} \frac{|D_{K,\sigma}|}{|K|}\boldsymbol{\varphi}_\sigma$, $\forall K \in \mathcal{M}$. Thanks to the regularity of $\boldsymbol{\varphi}$, $\boldsymbol{\varphi}_\mathcal{M}$ converges in $\mathrm{L}^r(\Omega\times(0,T))^d$, for $r \geq 1$ to $\boldsymbol{\varphi}$ and its discrete gradient $\underline{\underline{\boldsymbol{\nabla}_\mathcal{E}}}\boldsymbol{\varphi}_\mathcal{M}$ converges weakly in $\mathrm{L}^r(\Omega\times(0,T))^{d\times d}$ towards $\underline{\underline{\boldsymbol{\nabla}}}\boldsymbol{\varphi}$.

Mass balance equation – Since the support of φ is compact in $\Omega\times[0,T)$, for m large enough, the interpolates of φ vanish on the boundary cells and at the last time step(s) ; hereafter, we assume that we are in this case.

Let us multiply the first equation (3.5a) of the scheme by $\delta t\, \varphi_K^{n+1}$, and sum the result for $0 \le n \le N-1$ and $K \in \mathcal{M}$, to obtain $T_1^{(m)} + T_2^{(m)} = 0$ with

$$T_1^{(m)} = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{M}} |K|(\rho_K^{n+1} - \rho_K^n)\, \varphi_K^{n+1}, \qquad T_2^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n\, \varphi_K^{n+1}.$$

Reordering the sums in $T_1^{(m)}$ yields :

$$T_1^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} |K|\, \rho_K^n\, \frac{\varphi_K^{n+1} - \varphi_K^n}{\delta t} - \sum_{K\in\mathcal{M}} |K|\, \rho_K^0\, \varphi_K^0,$$

so that :

$$T_1^{(m)} = -\int_0^T \int_\Omega \rho^{(m)} \eth_t \mathcal{P}_\mathcal{M}\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega (\rho^{(m)})^0(x)\, \mathcal{P}_\mathcal{M}\varphi(x,0)\,\mathrm{d}x.$$

The boundedness of ρ_0 and the definition (3.23) of the initial conditions for the scheme ensures that the sequence $((\rho^{(m)})^0)_{m\in\mathbb{N}}$ converges to ρ_0 in $\mathrm{L}^r(\Omega)$ for $r \ge 1$. Since, by assumption, the sequence of discrete solutions and of the interpolate time derivatives converge in $\mathrm{L}^r\big(\Omega \times (0,T)\big)$ for $r \ge 1$, we thus obtain :

$$\lim_{m\to+\infty} T_1^{(m)} = -\int_0^T \int_\Omega \bar\rho\, \partial_t\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \rho_0(x)\, \varphi(x,0)\,\mathrm{d}x.$$

Using the expression of the mass flux $F_{K,\sigma}$ and reordering the sum in $T_2^{(m)}$ we get :

$$T_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} |D_\sigma| \rho_\sigma^n \boldsymbol{u}_\sigma^n.\, \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1})\boldsymbol{n}_{K,\sigma}.$$

We decompose the sum in two terms, $T_2^{(m)} = \mathcal{T}_2^{(m)} + \mathcal{R}_2^{(m)}$ with

$$\mathcal{T}_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \big(|D_{K,\sigma}|\rho_K^n + |D_{L,\sigma}|\rho_L^n\big)\boldsymbol{u}_\sigma^n.\, \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1})\boldsymbol{n}_{K,\sigma},$$

$$\mathcal{R}_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \big(|D_\sigma|\rho_\sigma^n - |D_{K,\sigma}|\rho_K^n - |D_{L,\sigma}|\rho_L^n\big)\, \boldsymbol{u}_\sigma^n.\, \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1})\boldsymbol{n}_{K,\sigma},$$

We have, for the term $\mathcal{T}_2^{(m)}$:

$$\mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \rho^{(m)} \boldsymbol{u}^{(m)}.\boldsymbol{\nabla}_{\mathcal{E}^{(m)}} \mathcal{P}_{\mathcal{M}^{(m)}}\varphi$$

and therefore

$$\lim_{m\to+\infty} \mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \bar\rho\, \overline{\boldsymbol{u}}.\boldsymbol{\nabla}\varphi.$$

The remainder term $\mathcal{R}_2^{(m)}$ can be expressed thanks to the MUSCL property (3.10). There exists $\alpha_\sigma^n \in [0,1]$ such that :

$$\mathcal{R}_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \Big(|D_{L,\sigma}|\big(\rho_K^n - \rho_\mathrm{L}^n\big) - (1-\alpha_\sigma^n)\,|D_{K,\sigma}|\big(\rho_K^n - \rho_\mathrm{L}^n\big)\Big)$$
$$\boldsymbol{u}_\sigma^n.\frac{|\sigma|}{|D_\sigma|}(\varphi_\mathrm{L}^{n+1} - \varphi_K^{n+1})\boldsymbol{n}_{K,\sigma}$$

We can consequently bound the remainder term as follows :

$$|\mathcal{R}_2^{(m)}| \leq C\|\nabla\varphi\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d} \sum_{n=0}^{N-1} \delta t \sum_{\sigma=K|L\in\mathcal{E}} |\rho_K^n - \rho_L^n|\, |D_\sigma|\, |\boldsymbol{u}_\sigma^n|$$
$$\leq C\|\nabla\varphi\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d} \|\boldsymbol{u}^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d}\, \|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}\, h^{(m)},$$

which tends to zero when m tends to ∞, by the assumed stability of the solution.

Momentum balance equation – Let us multiply Equation (3.5d) by $\delta t\, \boldsymbol{\varphi}_\sigma^{n+1}$, and sum the result for $0 \leq n \leq N-1$ and $\sigma \in \mathcal{E}_{\mathrm{int}}$. We obtain $T_1^{(m)} + T_2^{(m)} + T_3^{(m)} + T_{visco}^{(m)} = 0$ with

$$T_1^{(m)} = \sum_{n=0}^{N-1} \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma|\, (\rho_{D_\sigma}^{n+1} \boldsymbol{u}_\sigma^{n+1} - \rho_{D_\sigma}^n \boldsymbol{u}_\sigma^n).\, \boldsymbol{\varphi}_\sigma^{n+1},$$
$$T_2^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \sum_{\epsilon\in\tilde{\mathcal{E}}(\sigma)} F_{\sigma,\epsilon}^n \boldsymbol{u}_{\epsilon,\mathrm{cen}}^n . \boldsymbol{\varphi}_\sigma^{n+1},$$
$$T_3^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma| (\nabla_{\mathcal{E}} p^{(m)})_\sigma . \boldsymbol{\varphi}_\sigma^{n+1},$$
$$T_{visco}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}} \sum_{\epsilon=\sigma|\sigma'\in\tilde{\mathcal{E}}(\sigma)} \left(\mu_\epsilon^n + \frac{|F_{\sigma,\epsilon}^n|}{2}\right) (\boldsymbol{u}_\sigma^n - \boldsymbol{u}_{\sigma'}^n).\boldsymbol{\varphi}_\sigma^{n+1}.$$

with $\boldsymbol{u}_{\epsilon,\mathrm{cen}}^n$ the centered interpolation of the velocity at the dual faces (we add the numerical diffusion from the upwind interpolation in the viscosity term). Reordering the sums, we get for $T_1^{(m)}$:

$$T_1^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma|\, \rho_{D_\sigma}^n \boldsymbol{u}_\sigma^n .\, \frac{\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_\sigma^n}{\delta t} - \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma|\, \rho_{D_\sigma}^0\, \boldsymbol{u}_\sigma^0 .\, \boldsymbol{\varphi}_\sigma^0.$$

Thanks to the definition of the quantity ρ_{D_σ} (namely the fact that $|D_\sigma|\, \rho_{D_\sigma}^n = (|D_{K,\sigma}|\, \rho_K^n + |D_{L,\sigma}|\, \rho_L^n)$, we have :

$$T_1^{(m)} = -\int_0^T \int_\Omega \rho^{(m)}\, \boldsymbol{u}^{(m)} .\, \eth_t \mathcal{P}_{\mathcal{E}} \boldsymbol{\varphi}\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega (\rho^{(m)})^0(x)\, (\boldsymbol{u}^{(m)})^0(x) .\, \mathcal{P}_{\mathcal{E}} \boldsymbol{\varphi}(x,0)\, \mathrm{d}\boldsymbol{x}.$$

By the same arguments as for the mass balance equation, we therefore obtain :

$$\lim_{m\to+\infty} T_1^{(m)} = -\int_0^T \int_\Omega \bar{\rho}\, \overline{\boldsymbol{u}} .\partial_t \boldsymbol{\varphi}\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(x)\, \boldsymbol{u}_0(x) .\boldsymbol{\varphi}(x,0)\, \mathrm{d}\boldsymbol{x}.$$

We now turn to the viscosity term $T_{visco}^{(m)}$. Reordering the sum we get that :

$$T_{visco}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\epsilon\in\tilde{\mathcal{E}}_{\mathrm{int}}} \left(\mu_\epsilon^n + \frac{|F_{\sigma,\epsilon}^n|}{2}\right) (\boldsymbol{u}_\sigma^n - \boldsymbol{u}_{\sigma'}^n).(\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_{\sigma'}^{n+1}).$$

An important property of the dual flux is that they are bounded with respect to the primal flux. Consequently,

$$|F_{\sigma,\epsilon}^n| \leq C\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \|\boldsymbol{u}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d} h_K^{d-1}.$$

The viscosity has the same behaviour by construction :

$$\mu_\epsilon^n \leq C h_K^{d-1}.$$

We then write

$$T_{visco}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\epsilon\in\tilde{\mathcal{E}}_{\mathrm{int}}} d_\epsilon \frac{\mu_\epsilon^n + \frac{|F_{\sigma,\epsilon}^n|}{2}}{|\epsilon|} |\epsilon| (\boldsymbol{u}_\sigma^n - \boldsymbol{u}_{\sigma'}^n).\frac{\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_{\sigma'}^{n+1}}{d_\epsilon},$$

and so we can bound the viscosity term,

$$|T_{visco}^{(m)}| \leq h^{(m)} C\|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}$$

, with C only depending on $\boldsymbol{\varphi}$, and the uniform estimates of the discrete solutions. Therefore this term tends to zero as m tends to $+\infty$.

Concerning $T_3^{(m)}$, thanks to the duality relation (3.21) we get :

$$T_3^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} |K| p_K^{n+1} (\mathrm{div}\boldsymbol{\varphi})_K^{n+1}.$$

Reordering the sums, we obtain $T_3^{(m)} = \mathcal{T}_3^{(m)} + \mathcal{R}_3^{(m)}$ with :

$$\mathcal{T}_3^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} |K| p_K^{n} (\mathrm{div}\boldsymbol{\varphi})_K^{n+1},$$
$$\mathcal{R}_3^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} |K| (p_K^{n+1} - p_K^{n}) (\mathrm{div}\boldsymbol{\varphi})_K^{n+1}.$$

The remainder term reads :

$$\mathcal{R}_3^{(m)} = \sum_{n=1}^{N} \delta t \sum_{K\in\mathcal{M}} |K| p_K^{n} \left[(\mathrm{div}\boldsymbol{\varphi})_K^{n+1} - (\mathrm{div}\boldsymbol{\varphi})_K^{n} \right] + \delta t \sum_{K\in\mathcal{M}} p_K^0 (\mathrm{div}\boldsymbol{\varphi})_K^0.$$

Thanks to the regularity of $\boldsymbol{\varphi}$ we can bound this term as follows :

$$|\mathcal{R}_3^{(m)}| \leq C_{\boldsymbol{\varphi}} \left(\delta t^{(m)} + h^{(m)}\right) \|p\|_{\mathrm{L}^\infty(\Omega\times(0,T))},$$

where the real number $C_{\boldsymbol{\varphi}}$ only depends on $\boldsymbol{\varphi}$. As a result this term tends to zero when m tends to ∞ and, since

$$\mathcal{T}_3^{(m)} = -\int_0^T \int_\Omega p^{(m)} \,\mathrm{div}_{\mathcal{M}} \mathcal{P}_{\mathcal{E}} \boldsymbol{\varphi},$$

we obtain that :

$$\lim_{m\to+\infty} T_3^{(m)} = -\int_0^T \int_\Omega \bar{p} \,\mathrm{div}\boldsymbol{\varphi},$$

because $p^{(m)}$ converges strongly to $\bar{p}$ and $\mathrm{div}_{\mathcal{M}}\mathcal{P}_{\mathcal{E}}\boldsymbol{\varphi}$ weakly converges to $\mathrm{div}\boldsymbol{\varphi}$ (this is a direct consequence of the lemma (3.7) and the fact that $\mathcal{P}_{\mathcal{E}}\boldsymbol{\varphi}$ converges strongly to $\boldsymbol{\varphi}$) .

Finally we need to analyze the convection term $T_2^{(m)}$. We need to write this term on the primal mesh because we do not have an easy access to the dual fluxes $F_{\sigma,\epsilon}^n$. We set $\boldsymbol{\varphi}_K^{n+1} = \sum_{\sigma\in\mathcal{E}(K)} \underbrace{\frac{|D_{K,\sigma}|}{|K|}}_{\xi_K^\sigma} \boldsymbol{\varphi}_\sigma$.

We have $T_2^{(m)} = \mathcal{T}_2^{(m)} + R^{(m)}$ with :

$$\mathcal{T}_2^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n \boldsymbol{u}_\sigma^n . \boldsymbol{\varphi}_K^{n+1},$$
$$R^{(m)} = \underbrace{\sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \sum_{\epsilon\in\tilde{\mathcal{E}}(\sigma)} F_{\sigma,\epsilon}^n \boldsymbol{u}_{\epsilon,\mathrm{cen}}^n . \boldsymbol{\varphi}_\sigma^{n+1}}_{Q_{\mathcal{E}}^{(m)}} - \underbrace{\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n \boldsymbol{u}_\sigma^n . \boldsymbol{\varphi}_K^{n+1}}_{Q_{\mathcal{M}}^{(m)}},$$

We reconstruct $Q_{\mathcal{E}}^{(m)}$ by summing over the elements of the mesh and we use the conservativity of the primal fluxes to get :

$$Q_{\mathcal{E}}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} \boldsymbol{\varphi}_\sigma^{n+1} . \Big[F_{K,\sigma}^n \boldsymbol{u}_\sigma^n + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}_S^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset,\\ \epsilon=D_\sigma|D_{\sigma'}}} F_{\sigma,\epsilon}^n \boldsymbol{u}_{\epsilon,\mathrm{cen}}^n \Big]$$

Let us write $Q_{\mathcal{E}}^{(m)} = Q_{\mathcal{E},1}^{(m)} + Q_{\mathcal{E},2}^{(m)}$ with :

$$Q_{\mathcal{E},1}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} \boldsymbol{\varphi}_K^{n+1}.\Big[F_{K,\sigma}^n \boldsymbol{u}_\sigma^n + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset,\\ \epsilon=D_\sigma|D_{\sigma'}}} F_{\sigma,\epsilon}^n \boldsymbol{u}_{\epsilon,\mathrm{cen}}^n\Big],$$

$$Q_{\mathcal{E},2}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} (\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_K^{n+1}).\Big[F_{K,\sigma}^n \boldsymbol{u}_\sigma^n + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset,\\ \epsilon=D_\sigma|D_{\sigma'}}} F_{\sigma,\epsilon}^n \boldsymbol{u}_{\epsilon,\mathrm{cen}}^n\Big].$$

Using the conservativity of the dual fluxes we get that $Q_{\mathcal{E},1}^{(m)} = Q_{\mathcal{M}}^{(m)}$. Consequently, $Q_{\mathcal{E},2}^{(m)} = R^{(m)}$. By construction of the dual fluxes we recall that we have :

$$\forall K \in \mathcal{M},\ \forall \sigma \in \mathcal{E}(K), \quad F_{K,\sigma}^n + \sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma),\ \epsilon\subset K} F_{\sigma,\epsilon}^n = \xi_K^\sigma \sum_{\sigma'\in\mathcal{E}(K)} F_{K,\sigma'}^n,$$

so we can write $R^{(m)} = R_1^{(m)} + R_2^{(m)}$, with :

$$R_1^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} (\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_K^{n+1}). \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset,\\ \epsilon=D_\sigma|D_{\sigma'}}} F_{\sigma,\epsilon}^n \frac{\boldsymbol{u}_{\sigma'}^n - \boldsymbol{u}_\sigma^n}{2}$$

$$R_2^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} (\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_K^{n+1}).\boldsymbol{u}_\sigma^n \xi_K^\sigma \left[\sum_{\sigma'\in\mathcal{E}(K)} F_{K,\sigma'}\right]$$

By construction of the dual fluxes we have $|F_{\sigma,\epsilon}^n| \leq C\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\|\boldsymbol{u}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d} h_K^{d-1}$. As a consequence, since $\boldsymbol{\varphi}_K^{n+1}$ is a convex combination of the $(\boldsymbol{\varphi}_\sigma^{n+1})_{\sigma\in\mathcal{E}(K)}$, we have for any $K \in \mathcal{M}$:

$$\Big|\sum_{\sigma\in\mathcal{E}(K)} (\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_K^{n+1}). \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset,\\ \epsilon=D_\sigma|D_{\sigma'}}} F_{\sigma,\epsilon}^n \frac{\boldsymbol{u}_{\sigma'}^n - \boldsymbol{u}_\sigma^n}{2}\Big| \leq$$
$$C\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\|\boldsymbol{u}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d} h^{(m)} \sum_{\sigma,\sigma',\sigma'',\sigma'''\in\mathcal{E}(K)} h_K^{d-2}|\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_{\sigma'}^{n+1}|\|\boldsymbol{u}_{\sigma''}^n - \boldsymbol{u}_{\sigma'''}^n|.$$

Hence we can deduce that there exists $C' \in \mathbb{R}^+$ only depending on $\boldsymbol{\varphi}$ and θ_0 such that :

$$|R_1^{(m)}| \leq C'\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\|\boldsymbol{u}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d}\|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} h^{(m)},$$

so $R_1^{(m)}$ tends to zero as m tends to infinity.

Let us now focus on $R_2^{(m)}$. By definition of $\boldsymbol{\varphi}_K^{n+1}$, we have $\sum_{\sigma\in\mathcal{E}(K)} \xi_K^\sigma(\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_K^{n+1}) = 0$, so we can write :

$$R_2^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} (\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_K^{n+1}).(\boldsymbol{u}_\sigma^n - \boldsymbol{u}_K^n)\xi_K^\sigma \left[\sum_{\sigma'\in\mathcal{E}(K)} F_{K,\sigma'}\right],$$

and then we get :

$$|R_2^{(m)}| \leq C\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\|\boldsymbol{u}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d} h^{(m)} \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} h_K^{d-2}|\boldsymbol{\varphi}_\sigma^{n+1} - \boldsymbol{\varphi}_K^{n+1}||\boldsymbol{u}_\sigma^n - \boldsymbol{u}_K^n|$$
$$\leq C'\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\|\boldsymbol{u}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d}\|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} h^{(m)}$$

so $R_2^{(m)}$ tends to zero as m tends to infinity. Finally we have

$$\begin{aligned}
\mathcal{T}_2^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n \boldsymbol{u}_\sigma^n.\boldsymbol{\varphi}_K^{n+1} \\
&= \sum_{n=0}^{N-1} \delta t \sum_{\substack{\sigma\in\mathcal{E}\\ \sigma=K|L}} |\sigma|(\rho_\sigma^n \boldsymbol{u}_\sigma^n.\boldsymbol{n}_{K,\sigma})(\boldsymbol{u}_\sigma^n.[\boldsymbol{\varphi}_K^{n+1} - \boldsymbol{\varphi}_L^{n+1}]) \\
&= -\sum_{n=0}^{N-1} \delta t \sum_{\substack{\sigma\in\mathcal{E}\\ \sigma=K|L}} (|D_\sigma|\rho_\sigma^n \boldsymbol{u}_\sigma^n \otimes \boldsymbol{u}_\sigma^n) : \underline{\underline{\nabla_\mathcal{E}}}\boldsymbol{\varphi}_\sigma^{n+1}.
\end{aligned}$$

We use the same process as for the mass balance equation. We can split up $\mathcal{T}_2^{(m)} = \mathcal{T}_{2,1}^{(m)} + \mathcal{T}_{2,2}^{(m)}$ with

$$\mathcal{T}_{2,1}^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\substack{\sigma\in\mathcal{E}\\ \sigma=K|L}} \left[\left(|D_{K,\sigma}|\rho_K^n + |D_{L,\sigma}|\rho_L^n\right)\boldsymbol{u}_\sigma^n \otimes \boldsymbol{u}_\sigma^n\right] : \underline{\underline{\nabla_\mathcal{E}}}\boldsymbol{\varphi}_\sigma^{n+1},$$

and $\mathcal{T}_{2,2}^{(m)}$ vanishing when m tends to infinity. Consequently we have

$$\mathcal{T}_{2,1}^{(m)} = -\int_0^T \int_\Omega (\rho^{(m)}\boldsymbol{u}^{(m)} \otimes \boldsymbol{u}^{(m)}) : \underline{\underline{\nabla_\mathcal{E}}}\boldsymbol{\varphi}_\mathcal{M},$$

and passing to the limit, we get :

$$\lim_{m\to+\infty} \mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega (\bar{\rho}\bar{\boldsymbol{u}} \otimes \bar{\boldsymbol{u}}) : \underline{\underline{\nabla}}\boldsymbol{\varphi},$$

Total energy balance equation – On one hand, let us multiply the discrete internal energy balance equation (3.5b) by $\delta t\, \varphi_K^{n+1}$, and sum the result for $0 \le n \le N-1$ and $K \in \mathcal{M}$. On the other hand, let us multiply the discrete kinetic energy balance (3.25) by $\delta t\, \varphi_\sigma^{n+1}$, and sum the result for $0 \le n \le N-1$ and $\sigma \in \mathcal{E}_{\text{int}}$. Finally, adding the two obtained relations, we get :

$$T_1^{(m)} + T_2^{(m)} + T_3^{(m)} + \tilde{T}_1^{(m)} + \tilde{T}_2^{(m)} + \tilde{T}_3^{(m)} = S^{(m)} - \tilde{R}^{(m)}, \tag{3.59}$$

where :

$$\begin{aligned}
T_1^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \frac{|K|}{\delta t}\left[\rho_K^{n+1} e_K^{n+1} - \rho_K^n e_K^n\right]\varphi_K^{n+1}, \\
T_2^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n e_\sigma^n \varphi_K^{n+1}, \\
T_3^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} |K| p_K^n (\operatorname{div}(\boldsymbol{u}))_K^n \varphi_K^{n+1}, \\
\tilde{T}_1^{(m)} &= \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\text{int}}} \frac{|D_\sigma|}{\delta t}\left[\rho_{D_\sigma}^{n+1}|\boldsymbol{u}_\sigma^{n+1}|^2 - \rho_{D_\sigma}^n|\boldsymbol{u}_\sigma^n|^2\right]\varphi_\sigma^{n+1}, \\
\tilde{T}_2^{(m)} &= \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\text{int}}} \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n \frac{|\boldsymbol{u}_\sigma^n|^2 + |\boldsymbol{u}_{\sigma'}^n|^2}{2}\varphi_\sigma^{n+1} \\
\tilde{T}_{2,R}^{(m)} &= \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\text{int}}} \varphi_\sigma^{n+1} \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n \left(\boldsymbol{u}_{\sigma,i}^n \cdot \boldsymbol{u}_{\sigma',i}^n - \frac{|\boldsymbol{u}_\sigma^n|^2 + |\boldsymbol{u}_{\sigma'}^n|^2}{2}\right) + \mu_\epsilon^n(\boldsymbol{u}_\sigma^n - \boldsymbol{u}_{\sigma'}^n)\cdot(\boldsymbol{u}_{\sigma'}^n + \boldsymbol{u}_\sigma^n) \\
\tilde{T}_3^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\text{int}}} |D_\sigma|(\nabla p)_\sigma^{n+1}.\boldsymbol{u}_\sigma^{n+1}\,\varphi_\sigma^{n+1},
\end{aligned}$$

$$S^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} S_K^n \, \varphi_K^{n+1}, \qquad \tilde{R}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{i=1}^{d} \sum_{\sigma\in\mathcal{E}_{\text{int}}} R_{\sigma,i}^{n+1} \, \varphi_\sigma^{n+1},$$

and the quantities S_K^n and $R_{\sigma,i}^{n+1}$ are given by Equation (3.31) and Equation (3.26) respectively.

Reordering the sums in $T_1^{(m)}$ yields :

$$T_1^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} |K| \, \rho_K^n \, e_K^n \, \frac{\varphi_K^{n+1} - \varphi_K^n}{\delta t} - \sum_{K\in\mathcal{M}} |K| \, \rho_K^0 \, e_K^0 \, \varphi_K^0,$$

so that :

$$T_1^{(m)} = -\int_0^T \int_\Omega \rho^{(m)} \, e^{(m)} \check{\partial}_t \mathcal{P}_\mathcal{M}\varphi \, \mathrm{d}x \, \mathrm{d}t - \int_\Omega (\rho^{(m)})^0(x) \, (e^{(m)})^0(x) \, \mathcal{P}_\mathcal{M}\varphi(x,0) \, \mathrm{d}x.$$

The boundedness of ρ_0, e_0 and the definition (3.23) of the initial conditions for the scheme ensures that the sequences $((\rho^{(m)})^0)_{m\in\mathbb{N}}$ and $((e^{(m)})^0)_{m\in\mathbb{N}}$ converge to ρ_0 and e_0 respectively in $\mathrm{L}^r(\Omega)$ for $r \geq 1$. Since, by assumption, the sequence of discrete solutions and of the interpolate time derivatives converge in $\mathrm{L}^r\big(\Omega \times (0,T)\big)$ for $r \geq 1$, we thus obtain :

$$\lim_{m\to+\infty} T_1^{(m)} = -\int_0^T \int_\Omega \bar{\rho} \, \bar{e} \, \partial_t \varphi \, \mathrm{d}x \, \mathrm{d}t - \int_\Omega \rho_0(x) \, e_0(x) \, \varphi(x,0) \, \mathrm{d}x.$$

Using the expression of the mass flux $F_{K,\sigma}$ and reordering the sum in $T_2^{(m)}$ we get :

$$-\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} |D_\sigma| \rho_\sigma^n e_\sigma^n \boldsymbol{u}_\sigma^n . \, \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1}) \boldsymbol{n}_{K,\sigma}.$$

We decompose the sum in two terms, $T_2^{(m)} = \mathcal{T}_2^{(m)} + \mathcal{R}_2^{(m)}$ with

$$\mathcal{T}_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \Big(|D_{K,\sigma}| \rho_K^n e_K^n + |D_{L,\sigma}| \rho_L^n e_L^n\Big) \boldsymbol{u}_\sigma^n . \, \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1}) \boldsymbol{n}_{K,\sigma},$$

$$\mathcal{R}_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \Big(|D_\sigma| \rho_\sigma^n e_\sigma^n - |D_{K,\sigma}| \rho_K^n e_K^n - |D_{L,\sigma}| \rho_L^n e_L^n\Big) \boldsymbol{u}_\sigma^n . \, \frac{|\sigma|}{|D_\sigma|} \, (\varphi_L^{n+1} - \varphi_K^{n+1}) \boldsymbol{n}_{K,\sigma}.$$

We have, for the term $\mathcal{T}_2^{(m)}$:

$$\mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \rho^{(m)} e^{(m)} \boldsymbol{u}^{(m)} . \nabla_\mathcal{E} \mathcal{P}_\mathcal{M}\varphi$$

and therefore

$$\lim_{m\to+\infty} \mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \bar{\rho} \bar{e} \, \overline{\boldsymbol{u}} . \nabla\varphi.$$

The remainder term $\mathcal{R}_2^{(m)}$ can be expressed thanks to the MUSCL property (3.12). There exists $\alpha_\sigma^n \in [0,1]$ such that :

$$\mathcal{R}_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \Big(\alpha_\sigma^n |D_{K,\sigma}| \Big(\rho_K^n e_K^n - \rho_L^n e_L^n\Big) - (1-\alpha_\sigma^n) \, |D_{L,\sigma}| \, \Big(\rho_K^n e_K^n - \rho_L^n e_L^n\Big)\Big) \boldsymbol{u}_\sigma^n . \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1}) \boldsymbol{n}_{K,\sigma}.$$

We thus get :

$$|\mathcal{R}_2^{(m)}| \leq C_\varphi \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} |D_\sigma||\rho_K^n e_K^n - \rho_L^n e_L^n||u_\sigma^n|,$$

with C_φ only depending on φ. Applying the identity $2\,(ab-cd) = (a-c)\,(b+d) + (a+c)\,(b-d)$, which holds for any a, b, c, d real, to the quantity $\rho_K^n e_K^n - \rho_L^n e_L^n$, we obtain :

$$|\mathcal{R}_2^{(m)}| \leq C_\varphi\, h^{(m)}\, \|u^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \Big[\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\, \|e^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|e^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\, \|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}\Big],$$

and thus $|\mathcal{R}_2^{(m)}|$ tends to zero when m tends to $+\infty$. For the term $\tilde{T}_1^{(m)}$, the definition (3.16) of ρ_{D_σ} and a reordering in the summation yield :

$$\tilde{T}_1^{(m)} = -\frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{\sigma=K|L\in\mathcal{E}} \Big[|D_{K,\sigma}|\,\rho_K^n + |D_{L,\sigma}|\,\rho_L^n\Big]\,|\boldsymbol{u}_\sigma^n|^2\, \frac{\varphi_\sigma^{n+1} - \varphi_\sigma^n}{\delta t} - \frac{1}{2}\sum_{\sigma=K|L\in\mathcal{E}} \Big[|D_{K,\sigma}|\,\rho_K^0 + |D_{L,\sigma}|\,\rho_L^0\Big]\,|\boldsymbol{u}_\sigma^0|^2\,\varphi_\sigma^0,$$

so that, by similar arguments as for the term $T_1^{(m)}$, we get :

$$\lim_{m\to+\infty} \tilde{T}_1^{(m)} = -\int_0^T \int_\Omega \frac{1}{2}\,\bar{\rho}\,\bar{\boldsymbol{u}}^2\,\partial_t\varphi\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \frac{1}{2}\,\rho_0(x)\,u_0(x)^2\,\varphi(x,0)\,\mathrm{d}x.$$

In order to analyze the term $\tilde{T}_{2,R}^{(m)}$ we notice that :

$$F_{\sigma,\epsilon}^n \boldsymbol{u}_\sigma^n \cdot \boldsymbol{u}_{\sigma'}^n - \frac{1}{2}F_{\sigma,\epsilon}^n(|\boldsymbol{u}_\sigma^n|^2 + |\boldsymbol{u}_{\sigma'}^n|^2) = -\frac{1}{2}|F_{\sigma,\epsilon}^n|(\boldsymbol{u}_\sigma^n - \boldsymbol{u}_{\sigma'}^n)^2$$

so thanks to the boundedness of the dual fluxes and the numerical diffusion, we have :

$$|\tilde{T}_{2,R}^{(m)}| \leq C_\varphi h^{(m)} \|\boldsymbol{u}^{(m)}\|^2_{\mathrm{L}^\infty(\Omega\times(0,T))^d} \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}},$$

which tends to zero for vanishing space and time steps. Concerning the term $\tilde{T}_2^{(m)}$, the idea is the same as for the convection term in the momentum balance equation, *i.e.* defining an equivalent of $\tilde{T}_2^{(m)}$ on the primal mesh. By denoting $|\boldsymbol{u}_{\epsilon,\mathrm{cen}}|^2 = \frac{1}{2}(|u_\sigma^n|^2 + |u_{\sigma'}^n|^2)$, we set $\varphi_K^{n+1} = \sum_{\sigma\in\mathcal{E}(K)} \underbrace{\frac{|D_{K,\sigma}|}{|K|}}_{\xi_K^\sigma} \varphi_\sigma^{n+1}$.

We have $\tilde{T}_2^{(m)} = \tilde{\mathcal{T}}_2^{(m)} + R^{(m)}$ with :

$$\tilde{\mathcal{T}}_2^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n |\boldsymbol{u}_\sigma^n|^2 \varphi_K^{n+1},$$

$$R^{(m)} = \underbrace{\frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \sum_{\epsilon\in\tilde{\mathcal{E}}(\sigma)} F_{\sigma,\epsilon}^n |\boldsymbol{u}_{\epsilon,\mathrm{cen}}^n|^2 \varphi_\sigma^{n+1}}_{Q_{\mathcal{E}}^{(m)}} - \underbrace{\frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n |\boldsymbol{u}_\sigma^n|^2 \varphi_K^{n+1}}_{Q_{\mathcal{M}}^{(m)}},$$

The computations are the same as for the momentum balance equation. One can directly deduce that $R^{(m)} = R_1^{(m)} + R_2^{(m)}$, with :

$$R_1^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} (\varphi_\sigma^{n+1} - \varphi_K^{n+1}) \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}_S^{(i)},\ \epsilon\cap\bar{K}\neq\emptyset,\\ \epsilon=D_\sigma|D_{\sigma'}}} F_{\sigma,\epsilon}^n \frac{|\boldsymbol{u}_{\sigma'}^n|^2 - |\boldsymbol{u}_\sigma^n|^2}{2}$$

$$R_2^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} (\varphi_\sigma^{n+1} - \varphi_K^{n+1}) |\boldsymbol{u}_\sigma^n|^2 \xi_K^\sigma \left[\sum_{\sigma'\in\mathcal{E}(K)} F_{K,\sigma'}\right]$$

We have $\dfrac{|\boldsymbol{u}^n_{\sigma'}|^2 - |\boldsymbol{u}^n_{\sigma}|^2}{2} = \dfrac{1}{2}(\boldsymbol{u}^n_{\sigma} + \boldsymbol{u}^n_{\sigma'}).(\boldsymbol{u}^n_{\sigma} - \boldsymbol{u}^n_{\sigma'})$ so we can bound $R_1^{(m)}$ as follows (adopting the same notation as in the momentum balance equation) :

$$|R_1^{(m)}| \leq C' \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \|\boldsymbol{u}\|^2_{\mathrm{L}^\infty(\Omega\times(0,T))^d} \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} h^{(m)},$$

and then conclude that $R_1^{(m)}$ tends to zero as m tends to infinity. For the term $R_2^{(m)}$, we set $|\boldsymbol{u}^n_K|^2 = \sum_{\sigma\in\mathcal{E}(K)} \xi^\sigma_K |\boldsymbol{u}^n_\sigma|^2$ so we have :

$$R_2^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} (\varphi^{n+1}_\sigma - \varphi^{n+1}_K)(|\boldsymbol{u}^n_\sigma|^2 - |\boldsymbol{u}^n_K|^2)\xi^\sigma_K \left[\sum_{\sigma'\in\mathcal{E}(K)} F_{K,\sigma'}\right].$$

As a result :

$$|R_2^{(m)}| \leq C' \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \|\boldsymbol{u}\|^2_{\mathrm{L}^\infty(\Omega\times(0,T))^d} \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} h^{(m)}$$

and tends to zero as m tends to infinity.

$$\begin{aligned}
\mathcal{T}_2^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F^n_{K,\sigma} \frac{1}{2}|\boldsymbol{u}^n_\sigma|^2 \varphi^{n+1}_K \\
&= \sum_{n=0}^{N-1} \delta t \sum_{\substack{\sigma\in\mathcal{E}\\ \sigma=K|L}} |\sigma|\rho^n_\sigma \frac{1}{2}|\boldsymbol{u}^n_\sigma|^2 \boldsymbol{u}^n_\sigma.\boldsymbol{n}_{K,\sigma}[\varphi^{n+1}_K - \varphi^{n+1}_L] \\
&= -\sum_{n=0}^{N-1} \delta t \sum_{\substack{\sigma\in\mathcal{E}\\ \sigma=K|L}} (|D_\sigma|\rho^n_\sigma \frac{1}{2}|\boldsymbol{u}^n_\sigma|^2 \boldsymbol{u}^n_\sigma).\boldsymbol{\nabla}_{\mathcal{E}}\varphi^{n+1}_\sigma.
\end{aligned}$$

We use the same process as for the mass balance equation. We can split up $\mathcal{T}_2^{(m)} = \mathcal{T}_{2,1}^{(m)} + \mathcal{T}_{2,2}^{(m)}$ with

$$\mathcal{T}_{2,1}^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\substack{\sigma\in\mathcal{E}\\ \sigma=K|L}} \left[\left(|D_{K,\sigma}|\rho^n_K + |D_{L,\sigma}|\rho^n_L\right)\frac{1}{2}|\boldsymbol{u}^n_\sigma|^2 \boldsymbol{u}^n_\sigma\right].\boldsymbol{\nabla}_{\mathcal{E}}\varphi^{n+1}_\sigma,$$

and $\mathcal{T}_{2,2}^{(m)}$ vanishing when m tends to infinity. Consequently we have

$$\mathcal{T}_{2,1}^{(m)} = -\int_0^T \int_\Omega (\rho^{(m)} \frac{1}{2}|\boldsymbol{u}^{(m)}|^2 \boldsymbol{u}^{(m)}).\boldsymbol{\nabla}_{\mathcal{E}}\mathcal{P}_{\mathcal{M}}\varphi,$$

and passing to the limit, we get :

$$\lim_{m\to+\infty} \mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega (\bar{\rho}\frac{1}{2}|\bar{\boldsymbol{u}}|^2 \bar{\boldsymbol{u}}).\boldsymbol{\nabla}\varphi,$$

The terms $T_3^{(m)}$ and $\tilde{T}_3^{(m)}$ have to be analyzed together.

$$\tilde{T}_3^{(m)} = \sum_{n=1}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma|(\boldsymbol{\nabla}p)^n_\sigma.\boldsymbol{u}^n_\sigma\, \varphi^n_\sigma = \tilde{\mathcal{T}}_3^{(m)} + \tilde{\mathcal{R}}_3^{(m)},$$

with :

$$\begin{aligned}
\tilde{\mathcal{T}}_3^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma|(\boldsymbol{\nabla}p)^n_\sigma.\boldsymbol{u}^n_\sigma\, \varphi^{n+1}_\sigma, \\
\tilde{\mathcal{R}}_3^{(m)} &= -\delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma|(\boldsymbol{\nabla}p)^0_\sigma\, \boldsymbol{u}^0_\sigma\, \varphi^0_\sigma + \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma|(\boldsymbol{\nabla}p)^n_\sigma\, \boldsymbol{u}^n_\sigma\, (\varphi^n_\sigma - \varphi^{n+1}_\sigma).
\end{aligned}$$

We have, thanks to the regularity of φ :

$$|\tilde{\mathcal{R}}_3^{(m)}| \leq C_\varphi\, \delta t^{(m)}\Big[\|(u^{(m)})^0\|_{\mathrm{L}^\infty(\Omega)}\, \|(p^{(m)})^0\|_{\mathrm{BV}(\Omega)} + \|u^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\, \|p^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}\Big].$$

Therefore, invoking the regularity of the initial conditions, this term tends to zero when m tends to $+\infty$. By reordering the sum in $T_3^{(m)}$, we get that :

$$\tilde{\mathcal{T}}_3^{(m)} + T_3^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}_{\text{int}}} \left[p_{\mathrm{L}}^n(\varphi_\sigma^{n+1} - \varphi_L^{n+1}) + p_K^n(\varphi_K^{n+1} - \varphi_\sigma^{n+1})\right] |\sigma| \boldsymbol{u}_\sigma^n.\boldsymbol{n}_{K,\sigma}.$$

We can rewrite the sum as :

$$\tilde{\mathcal{T}}_3^{(m)} + T_3^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{D_{K,\sigma}\in\boldsymbol{Q}} |D_{K,\sigma}| p_K^n \boldsymbol{u}_\sigma^n.(\varphi_\sigma^{n+1} - \varphi_L^{n+1})\boldsymbol{n}_{K,\sigma},$$

$$\tilde{\mathcal{T}}_3^{(m)} + T_3^{(m)} = -\int_0^T \int_\Omega p^{(m)} \boldsymbol{u}^{(m)}.\nabla^h(\mathcal{P}_{\mathcal{M}^{(m)}}\varphi, \mathcal{P}_{\mathcal{E}^{(m)}}\varphi),$$

so we can conclude that

$$\lim_{m\to+\infty} \mathcal{T}_3^{(m)} + T_3^{(m)} = -\int_0^T \int_\Omega \bar{p}\bar{\boldsymbol{u}}.\nabla\varphi.$$

Finally, it now remains to check that $\lim_{m\to+\infty}(S^{(m)} - \tilde{R}^{(m)}) = 0$. Let us write this quantity as $S^{(m)} - \tilde{R}^{(m)} = \mathcal{R}_1^{(m)} + \mathcal{R}_2^{(m)}$ where, using $S_K^0 = 0, \forall K \in \mathcal{M}$:

$$\mathcal{R}_1^{(m)} = \sum_{n=0}^{N-1} \delta t \Big[\sum_{K\in\mathcal{M}} S_K^{n+1}\varphi_K^{n+1} - \sum_{\sigma\in\mathcal{E}} R_\sigma^{n+1}\varphi_\sigma^{n+1}\Big],$$
$$\mathcal{R}_2^{(m)} = \sum_{n=1}^{N-1} \delta t \sum_{K\in\mathcal{M}} S_K^n\,(\varphi_K^{n+1} - \varphi_K^n).$$

First, we prove that $\lim_{m\to+\infty} \mathcal{R}_1^{(m)} = 0$. Gathering and reordering the sums, we obtain $\mathcal{R}_1^{(m)} = \mathcal{R}_{1,1}^{(m)} + \mathcal{R}_{1,2}^{(m)} + \mathcal{R}_{1,3}^{(m)}$ with

$$\begin{aligned}
\mathcal{R}_{1,1}^{(m)} &= \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{\sigma=K|L\in\mathcal{E}} \Big[\frac{|D_{K,\sigma}|}{\delta t}\rho_K^{n+1}|\boldsymbol{u}_\sigma^{n+1} - \boldsymbol{u}_\sigma^n|^2(\varphi_K^{n+1} - \varphi_\sigma^{n+1}) \\
&\qquad + \frac{|D_{L,\sigma}|}{\delta t}\rho_L^{n+1}|\boldsymbol{u}_\sigma^{n+1} - \boldsymbol{u}_\sigma^n|^2(\varphi_L^{n+1} - \varphi_\sigma^{n+1})\Big], \\
\mathcal{R}_{1,2}^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}_{\text{int}}} \frac{1}{2}\mu_\epsilon^n(\boldsymbol{u}_{\sigma'}^n - \boldsymbol{u}_\sigma^n)^2\left(\varphi_K^{n+1} - \varphi_\sigma^{n+1} + \varphi_K^{n+1} - \varphi_{\sigma'}^{n+1}\right) \\
\mathcal{R}_{1,3}^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}_{\text{int}}} (\mu_\epsilon^n - \frac{F_{\sigma,\epsilon}^n}{2})(\boldsymbol{u}_\sigma^{n+1} - \boldsymbol{u}_\sigma^n)\,(\boldsymbol{u}_\sigma^n - \boldsymbol{u}_{\sigma'}^n)(\varphi_K^{n+1} - \varphi_\sigma^{n+1}) \\
&\qquad + (\mu_\epsilon^n - \frac{F_{\sigma',\epsilon}^n}{2})(\boldsymbol{u}_{\sigma'}^{n+1} - \boldsymbol{u}_{\sigma'}^n)\,(\boldsymbol{u}_{\sigma'}^n - \boldsymbol{u}_\sigma^n)(\varphi_K^{n+1} - \varphi_{\sigma'}^{n+1})
\end{aligned}$$

We thus obtain :

$$|\mathcal{R}_{1,1}^{(m)}| \le h^{(m)}\, C_\varphi\, \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\, \|\boldsymbol{u}^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d}\, \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},t,\mathrm{BV}},$$

and

$$|\mathcal{R}_{1,2}^{(m)}| + |\mathcal{R}_{1,3}^{(m)}| \le h^{(m)}\, C_\varphi\, \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\, \|\boldsymbol{u}^{(m)}\|^2_{\mathrm{L}^\infty(\Omega\times(0,T))^d}\, \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}},$$

so these two terms tend to zero. The fact that $|\mathcal{R}_2^{(m)}|$ behaves as $\delta t^{(m)}$ may be proven by similar arguments. ■

Entropy inequality –

First of all we need to prove that $\eta^{(m)}$ converges strongly towards $\bar{\eta} = \ln(\bar{\rho}) + \dfrac{1}{1-\gamma}\ln(\bar{e})$. Thanks to the estimates on $\dfrac{1}{\rho^{(m)}}$ and $\dfrac{1}{e^{(m)}}$ the function $x \mapsto \ln(x)$ is Lipschitz continuous. Therefore there exists $C > 0$ such that $|\ln(e^{(m)}) - \ln(\bar{e})| \leq C|e^{(m)} - \bar{e}|$ and $|\ln(\rho^{(m)}) - \ln(\bar{\rho})| \leq C|\rho^{(m)} - \bar{\rho}|$. As a result

$$\|\eta^{(m)} - \eta\|_{\mathrm{L}^r(\Omega\times(0,T))} \leq C' \left[\|e^{(m)} - e\|_{\mathrm{L}^r(\Omega\times(0,T))} + \|\rho^{(m)} - \rho\|_{\mathrm{L}^r(\Omega\times(0,T))}\right]$$

which proves the strong convergence in $\mathrm{L}^r(\Omega \times (0,T))$. Now we turn to the entropy inequality. First we need to impose a stronger cfl condition, namely :

let us multiply the discrete entropy inequality (3.38) by $\delta t\, \varphi_K^{n+1} \geq 0$, and sum the result for $0 \leq n \leq N-1$ and $K \in \mathcal{M}$. We get that $T_1^{(m)} + T_2^{(m)} + R^{(m)} + T_{\mathrm{conv}}^{(m)} \leq 0$, with

$$\begin{aligned}
T_1^{(m)} &= \sum_{n=0}^{N-1} \sum_{K\in\mathcal{M}} |K|(\rho_K^{n+1}\eta_K^{n+1} - \rho_K^n\eta_K^n)\,\varphi_K^{n+1},\\
T_2^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n \eta_\sigma^n\, \varphi_K^{n+1},\\
R^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} R_K^n\, \varphi_K^{n+1},\\
T_{\mathrm{conv}}^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} T_{\mathrm{conv},K}^n\, \varphi_K^{n+1}.
\end{aligned}$$

Reordering the sums in $T_1^{(m)}$ yields :

$$T_1^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} |K|\, \rho_K^n\, \eta_K^n\, \frac{\varphi_K^{n+1} - \varphi_K^n}{\delta t} - \sum_{K\in\mathcal{M}} |K|\, \rho_K^0\, \eta_K^0\, \varphi_K^0,$$

so that :

$$T_1^{(m)} = -\int_0^T \int_\Omega \rho^{(m)}\, \eta^{(m)} \eth_t \mathcal{P}_\mathcal{M}\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega (\rho^{(m)})^0(x)\, (\eta^{(m)})^0(x)\, \mathcal{P}_\mathcal{M}\varphi(x,0)\,\mathrm{d}x.$$

The boundedness of ρ_0, η_0 and the definition (3.23) of the initial conditions for the scheme ensures that the sequences $((\rho^{(m)})^0)_{m\in\mathbb{N}}$ and $((\eta^{(m)})^0)_{m\in\mathbb{N}}$ converge to ρ_0 and η_0 respectively in $\mathrm{L}^r(\Omega)$ for $r \geq 1$. Since, by assumption, the sequence of discrete solutions and of the interpolate time derivatives converge in $\mathrm{L}^r\big(\Omega \times (0,T)\big)$ for $r \geq 1$, we thus obtain :

$$\lim_{m\to+\infty} T_1^{(m)} = -\int_0^T \int_\Omega \bar{\rho}\, \bar{\eta}\, \partial_t\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \rho_0(x)\, \eta_0(x)\, \varphi(x,0)\,\mathrm{d}x.$$

Similarly to the convective part of the internal energy balance, reordering the sum in $T_2^{(m)}$ leads to :

$$-\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} |D_\sigma|\rho_\sigma^n\eta_\sigma^n \boldsymbol{u}_\sigma^n.\, \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1})\boldsymbol{n}_{K,\sigma}.$$

We decompose the sum in two terms, $T_2^{(m)} = \mathcal{T}_2^{(m)} + \mathcal{R}_2^{(m)}$ with

$$\mathcal{T}_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \left(|D_{K,\sigma}|\rho_K^n\eta_K^n + |D_{L,\sigma}|\rho_L^n\eta_L^n\right)\boldsymbol{u}_\sigma^n.\ \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1})\boldsymbol{n}_{K,\sigma},$$

$$\mathcal{R}_2^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \left(|D_\sigma|\rho_\sigma^n\eta_\sigma^n - |D_{K,\sigma}|\rho_K^n\eta_K^n - |D_{L,\sigma}|\rho_L^n\eta_L^n\right)\boldsymbol{u}_\sigma^n.\ \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1})\boldsymbol{n}_{K,\sigma}.$$

We have, for the term $\mathcal{T}_2^{(m)}$:

$$\mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \rho^{(m)}\eta^{(m)}\boldsymbol{u}^{(m)}.\boldsymbol{\nabla}_\mathcal{E}\mathcal{P}_\mathcal{M}\varphi$$

and therefore

$$\lim_{m\to+\infty} \mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \bar{\rho}\bar{\eta}\,\bar{\boldsymbol{u}}\,.\boldsymbol{\nabla}\varphi.$$

Concerning the remainder term $\mathcal{R}_2^{(m)}$, one notice we can bound $|\eta_\sigma - \eta_K|$ as follows :

$$|\eta_\sigma - \eta_K| \leq |\ln(\rho_\sigma) - \ln(\rho_K)| + \frac{1}{\gamma-1}|\ln(e_\sigma) - \ln(e_K)|.$$

The lipschitz continuity of ln and the muscl interpolations of ρ and e induce :

$$|\eta_\sigma - \eta_K| \leq |\eta_L - \eta_K| \leq C\left(|\rho_L - \rho_K| + |e_L - e_K|\right).$$

with C only depending on γ and the uniform bounds on $1/e^{(m)}$ and $1/\rho^{(m)}$, which means the BV norm of η is controlled by BV norms of e and ρ :

$$\|\eta^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} \leq C\left(\|e^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}\right)$$

We apply the identity $2\,(ab - cd) = (a-c)\,(b+d) + (a+c)\,(b-d)$, which holds for any a, b, c, d real, to the quantities $\rho_\sigma^n\eta_\sigma^n - \rho_L^n\eta_L^n$ and $\rho_\sigma^n\eta_\sigma^n - \rho_K^n\eta_K^n$. Thanks to the uniform boundnesss of $\rho^{(m)}$, $e^{(m)}$, $\dfrac{1}{\rho^{(m)}}$, $\dfrac{1}{e^{(m)}}$, $\eta^{(m)}$ is uniformly bounded in $\mathrm{L}^\infty(\Omega\times(0,T))$.

$$|\mathcal{R}_2^{(m)}| \leq C_\varphi\ h^{(m)}\ \|u^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\left[\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\ \|\eta^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|\eta^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\ \|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}}\right],$$

and thus $|\mathcal{R}_2^{(m)}|$ tends to zero when m tends to $+\infty$.

We now turn to the remainder term $R^{(m)} = R_1^{(m)} + R_2^{(m)}$ with :

$$R_1^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n\left(\rho_K^{n+1} - \rho_K^n\right)\phi''(\rho_K^{(1)})\varphi_K^{n+1},$$

$$R_2^{(m)} = \sum_{\sigma\in\mathcal{E}(K)} |\sigma|p_K^n u_{K,\sigma}^n\left(e_K^{n+1} - e_K^n\right)\psi''(e_K^{(1)})\varphi_K^{n+1}.$$

We focus on the first term, the results concerning the second term following following immediately. The uniform boundness of $\frac{1}{\rho^{(m)}}$ leads to a uniform boundedness of $\phi''(\rho_K^{(1)})$ in $\mathrm{L}^\infty(\Omega\times(0,T))$. Therefore we consequently get that

$$|R_1^{(m)}| \leq C\frac{\delta t^{(m)}}{\min\limits_{K\in\mathcal{M}^{(m)}} h_K}\|\rho^{(m)}\|_{\mathcal{T},t,\mathrm{BV}}$$

which tends to zero thanks to the CFL condition (3.56).

For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$, let us denote by $RC^n(\phi,\rho)_{K,\sigma}$ and $RC^n(\psi,e)_{K,\sigma}$ the quantities

$$RC^n(\phi,\rho)_{K,\sigma} = (\rho^n_\sigma - \rho^n_K)\phi'(\rho^n_K) - \Big(\phi(\rho^n_\sigma) - \phi(\rho^n_K)\Big)$$
$$RC^n(\psi,e)_{K,\sigma} = (e^n_\sigma - e^n_K)\psi'(e^n_K) - \Big(\psi(e^n_\sigma) - \psi(e^n_K)\Big)$$

We reorder the sum in the term $T^{(m)}_{\text{conv}}$ to get :

$$T^{(m)}_{\text{conv}} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} |\sigma| u^n_{K,\sigma} \left\{R^n_{\sigma,\rho} + \rho^n_\sigma R^n_{\sigma,e}\right\}\left(\varphi^{n+1}_K - \varphi^{n+1}_L\right).$$

Next we can see that :

$$R^n_{\sigma,\rho} = (\rho^n_\sigma - \bar{\rho}^n_\sigma)(\frac{\phi'(\rho^n_K) + \phi'(\rho^n_L)}{2}) + (\bar{\rho}^n_\sigma - \rho^n_K)\phi'(\rho^n_K) + \phi(\rho^n_K) - \phi(\bar{\rho}^n_\sigma) + \phi(\bar{\rho}^n_\sigma) - \phi(\rho^n_\sigma),$$

and we notice that

$$|\phi(\bar{\rho}^n_\sigma) - \phi(\rho^n_\sigma)| \le C|\rho^n_L - \rho^n_K|$$

because ϕ is Lipschitz (thanks to the uniform bound on $1/\rho^{(m)}$). Besides, $\frac{\phi'(\rho^n_K)+\phi'(\rho^n_L)}{2}$ is bounded in L^∞ so we also have

$$(\rho^n_\sigma - \bar{\rho}^n_\sigma)(\frac{\phi'(\rho^n_K) + \phi'(\rho^n_L)}{2}) \le C|\rho^n_L - \rho^n_K|.$$

Using Taylor expansions, we get that

$$(\bar{\rho}^n_\sigma - \rho^n_K)\phi'(\rho^n_K) + \phi(\rho^n_K) - \phi(\bar{\rho}^n_\sigma) = -\frac{1}{2}(\bar{\rho}_\sigma - \rho^n_K)^2\phi''(\bar{\rho}^n_{\sigma,K}),$$

with $\bar{\rho}^n_{\sigma,K} \in |[\rho^n_K, \bar{\rho}^n_\sigma]|$, so we get :

$$|(\bar{\rho}^n_\sigma - \rho^n_K)\phi'(\rho^n_K) + \phi(\rho^n_K) - \phi(\bar{\rho}^n_\sigma)| \le C\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\|\frac{1}{\rho^{(m)}}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}|\rho^n_L - \rho^n_K|.$$

Applying the same reasoning with $R^n_{\sigma,e}$, we finally obtain :

$$|R^n_{\sigma,\rho} + \rho^n_\sigma R^n_{\sigma,e}| \le C(|\rho^n_L - \rho^n_K| + |e^n_L - e^n_K|),$$

which means that

$$T^{(m)}_{\text{conv}} \le C(\|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|e^{(m)}\|_{\mathcal{T},x,\mathrm{BV}})h^{(m)},$$

with C only depending on φ, uniform bounds on solutions norm estimates, and the regularity of the mesh. Consequently, it tends to zero as m tends to infinity and the solution satisfies a weak entropy inequality at the limit.

MAC case

Proof : It is clear that with the assumed convergence for the sequence of solutions, the limit satisfies the equation of state. Let $\varphi \in C^\infty_C(\Omega\times[0,T))$, $m \in \mathbb{N}$, $\mathcal{M}^{(m)}$ and $\delta t^{(m)}$ be given. Dropping for short the superscript $^{(m)}$, let $\mathcal{P}_\mathcal{M}\varphi$ the interpolate of φ on the primal mesh, $\eth_t\mathcal{P}_\mathcal{M}\varphi$ its time discrete derivative and $\nabla_\mathcal{E}\mathcal{P}_\mathcal{M}\varphi$ its discrete gradient. Thanks to the regularity of φ and the lemma (3.7), $\eth_t\mathcal{P}_\mathcal{M}\varphi$ and $\mathcal{P}_\mathcal{M}\varphi$ converges in $\mathrm{L}^r(\Omega\times(0,T))$, for $r \ge 1$, to $\partial_t\varphi$ and φ respectively, and $\nabla_\mathcal{E}\mathcal{P}_\mathcal{M}\varphi$ converges weakly in $\mathrm{L}^r(\Omega\times(0,T))$ to $\nabla\varphi$. Besides, $\mathcal{P}_\mathcal{M}\varphi(.,0)$ converges to $\varphi(.,0)$ in $\mathrm{L}^r(\Omega)$.

We recall that $\mathcal{P}^{(i)}_{\mathcal{E}^{(i)}_S}\varphi$, $\eth_t\mathcal{P}^{(i)}_{\mathcal{E}^{(i)}_S}\varphi$ and $\eth^h_{x,i}(\mathcal{P}^{(i)}_{\mathcal{E}^{(i)}_S}\varphi,\ \mathcal{P}_\mathcal{M}\varphi)$ stand for the interpolate of φ on $H^{(i)}_{\mathcal{E},0}(\Omega\times(0,T))$, its discrete time derivative and its ith discrete hybrid gradient component respectively.

The first two converges in $\mathrm{L}^r(\Omega\times(0,T))$, for $r\geq 1$ towards φ and $\partial_t\varphi$, while the hybrid gradient converges weakly in $\mathrm{L}^r(\Omega\times(0,T))$ towards $\partial_{x,i}\varphi$. Besides, $\mathcal{P}^{(i)}_{\mathcal{E}^{(i)}_S}\varphi(.,0)$ converges to $\varphi(.,0)$ in $\mathrm{L}^r(\Omega)$. Futhermore we denote by $\underline{\mathcal{P}}_{\mathcal{E}}\varphi=\sum_{i=1}^d\mathcal{P}^{(i)}_{\mathcal{E}^{(i)}_S}\varphi\boldsymbol{e}^{(i)}$.

Let $\boldsymbol{\varphi}\in C^\infty_C(\Omega\times[0,T))^d$. Let us denote by $\mathcal{P}_{\mathcal{E}}\boldsymbol{\varphi}=\displaystyle\sum_{i=1}^d\mathcal{P}^{(i)}_{\mathcal{E}^{(i)}_S}\varphi^{(i)}\boldsymbol{e}^{(i)}$ its interpolate on the dual meshes (one for each component of the vector). As in the proof of lemma (3.8), we introduce an interpolation on the primal mesh $\boldsymbol{\varphi}_{\mathcal{M}}$ derived from the interpolation on the dual mesh thanks to the relation $\varphi_{K,i}=\dfrac{1}{2}\sum_{\sigma\in\mathcal{E}^{(i)}(K)}\varphi_{\sigma,i}$, $\forall i\in[\![1,d]\!]$ and $\forall K\in\mathcal{M}$. Thanks to the regularity of $\boldsymbol{\varphi}$, $\mathcal{P}_{\mathcal{E}}\boldsymbol{\varphi}$ converges in $\mathrm{L}^r(\Omega\times(0,T))^d$, for $r\geq 1$ to $\boldsymbol{\varphi}$ and its discrete gradient $\underline{\underline{\nabla}}_{\mathcal{E}}\mathcal{P}_{\mathcal{E}}\boldsymbol{\varphi}$ converges weakly in $\mathrm{L}^r(\Omega\times(0,T))^{d\times d}$ towards $\underline{\underline{\nabla}}\boldsymbol{\varphi}$. Since the support of φ is compact in $\overline{\Omega}\times[0,T)$, for m large enough, the interpolates of φ vanish on the boundary cells and at the last time step(s) ; hereafter, we assume that we are in this case.

Let us multiply the first equation (3.5a) of the scheme by $\delta t\,\varphi_K^{n+1}$, and sum the result for $0\leq n\leq N-1$ and $K\in\mathcal{M}$, to obtain $T_1^{(m)}+T_2^{(m)}=0$ with

$$T_1^{(m)}=\sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}}|K|(\rho_K^{n+1}-\rho_K^n)\,\varphi_K^{n+1},\qquad T_2^{(m)}=\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}(K)}F^n_{K,\sigma}\,\varphi_K^{n+1}.$$

Reordering the sums in $T_1^{(m)}$ yields :

$$T_1^{(m)}=-\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}}|K|\,\rho_K^n\,\frac{\varphi_K^{n+1}-\varphi_K^n}{\delta t}-\sum_{K\in\mathcal{M}}|K|\,\rho_K^0\,\varphi_K^0,$$

so that :

$$T_1^{(m)}=-\int_0^T\int_\Omega\rho^{(m)}\eth_t\mathcal{P}_{\mathcal{M}}\varphi\,\mathrm{d}x\,\mathrm{d}t-\int_\Omega(\rho^{(m)})^0(x)\,\mathcal{P}_{\mathcal{M}}\varphi(x,0)\,\mathrm{d}x.$$

The boundedness of ρ_0 and the definition (3.23) of the initial conditions for the scheme ensures that the sequence $((\rho^{(m)})^0)_{m\in\mathbb{N}}$ converges to ρ_0 in $\mathrm{L}^r(\Omega)$ for $r\geq 1$. Since, by assumption, the sequence of discrete solutions and of the interpolate time derivatives converge in $\mathrm{L}^r\big(\Omega\times(0,T)\big)$ for $r\geq 1$, we thus obtain :

$$\lim_{m\to+\infty}T_1^{(m)}=-\int_0^T\int_\Omega\bar{\rho}\,\partial_t\varphi\,\mathrm{d}x\,\mathrm{d}t-\int_\Omega\rho_0(x)\,\varphi(x,0)\,\mathrm{d}x.$$

Using the expression of the mass flux $F_{K,\sigma}$ and reordering the sum in $T_2^{(m)}$ we get :

$$-\sum_{i=1}^d\sum_{n=0}^{N-1}\delta t\sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}^{(i)}_{\mathrm{int}}}|D_\sigma|\rho^n_\sigma u^n_{\sigma,i}\,\frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1}-\varphi_K^{n+1}).$$

We decompose the sum in two terms, $T_2^{(m)}=\mathcal{T}_2^{(m)}+\mathcal{R}_2^{(m)}$ with

$$\mathcal{T}_2^{(m)}=-\sum_{i=1}^d\sum_{n=0}^{N-1}\delta t\sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}}\big(|D_{K,\sigma}|\rho_K^n+|D_{L,\sigma}|\rho_L^n\big)u^n_{\sigma,i}\,\frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1}-\varphi_K^{n+1}),$$

$$\mathcal{R}_2^{(m)}=-\sum_{i=1}^d\sum_{n=0}^{N-1}\delta t\sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}}\big(|D_\sigma|\rho^n_\sigma-|D_{K,\sigma}|\rho_K^n-|D_{L,\sigma}|\rho_L^n\big)\,u^n_{\sigma,i}\,\frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1}-\varphi_K^{n+1}),$$

We have, for the term $\mathcal{T}_2^{(m)}$:

$$\mathcal{T}_2^{(m)}=-\sum_{i=1}^d\int_0^T\int_\Omega\rho^{(m)}u_i^{(m)}\eth_{x,i}\mathcal{P}_{\mathcal{M}}\varphi$$

$$\mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \rho^{(m)} \boldsymbol{u}^{(m)} \nabla \mathcal{P}_\mathcal{M} \varphi$$

and therefore

$$\lim_{m\to+\infty} \mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \bar{\rho}\, \overline{\boldsymbol{u}} . \nabla \varphi.$$

The remainder term $\mathcal{R}_2^{(m)}$ can be expressed thanks to the MUSCL property (3.10). There exists $\alpha_\sigma^n \in [0,1]$ such that :

$$\mathcal{R}_2^{(m)} = -\sum_{i=1}^d \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \left(\alpha_\sigma^n |D_{K,\sigma}| \left(\rho_K^n - \rho_L^n\right) - (1-\alpha_\sigma^n)\, |D_{L,\sigma}| \left(\rho_K^n - \rho_L^n\right)\right) u_{\sigma,i}^n \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1})$$

We can consequently bound the remainder term as follows :

$$\begin{aligned} |\mathcal{R}_2^{(m)}| &\le C \|\nabla\varphi\|_{L^\infty(\Omega\times(0,T))^d} \sum_{i=1}^d \sum_{n=0}^{N-1} \delta t \sum_{\sigma=K|L\in\mathcal{E}} |\rho_K^n - \rho_L^n|\, |D_\sigma|\, |u_{\sigma,i}^n| \\ &\le C \|\nabla\varphi\|_{L^\infty(\Omega\times(0,T))^d} \|\boldsymbol{u}^{(m)}\|_{L^\infty(\Omega\times(0,T))^d} \sum_{i=1}^d \|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)}\, h^{(m)}, \end{aligned}$$

and tends to zero when m tends to ∞, by the assumed stability of the solution.

Momentum balance equation – Let us multiply Equation (3.5d) by $\delta t\, \boldsymbol{\varphi}_\sigma^{n+1}$, and sum the result for $0 \le n \le N-1$ and $\sigma \in \mathcal{E}_{\mathrm{int}}$. We obtain $\sum_{i=1}^d T_{1,i}^{(m)} + T_{2,i}^{(m)} + T_{3,i}^{(m)} + T_{visco,i}^{(m)} = 0$ with

$$\begin{aligned} T_{1,i}^{(m)} &= \sum_{n=0}^{N-1} \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} |D_\sigma|\, (\rho_{D_\sigma}^{n+1} u_{\sigma,i}^{n+1} - \rho_{D_\sigma}^n u_{\sigma,i}^n)\, \varphi_{\sigma,i}^{n+1}, \\ T_{2,i}^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} \sum_{\epsilon\in\tilde{\mathcal{E}}(\sigma)} F_{\sigma,\epsilon}^n u_{\epsilon,i}^n \varphi_{\sigma,i}^{n+1}, \\ T_{3,i}^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} |D_\sigma| (\eth_{x,i} p^{(m)})_\sigma^{n+1} \varphi_{\sigma,i}^{n+1}, \\ T_{visco,i}^{(m)} &= \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}^{(i)}} \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} \mu_\epsilon^n (u_{\sigma,i}^n - u_{\sigma',i}^n) \varphi_{\sigma,i}^{n+1}. \end{aligned}$$

Reordering the sums, we get for $T_{1,i}^{(m)}$:

$$T_{1,i}^{(m)} = -\sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} |D_\sigma|\, \rho_{D_\sigma}^n u_{\sigma,i}^n \frac{\varphi_{\sigma,i}^{n+1} - \varphi_{\sigma,i}^n}{\delta t} - \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} |D_\sigma|\, \rho_{D_\sigma}^0\, u_{\sigma,i}^0\, \varphi_{\sigma,i}^0.$$

Thanks to the definition of the quantity ρ_{D_σ} (namely the fact that $|D_\sigma|\, \rho_{D_\sigma}^n = (|K|\, \rho_K^n + |L|\, \rho_L^n)/2$), we have :

$$\sum_{i=1}^d T_{1,i}^{(m)} = -\int_0^T \int_\Omega \rho^{(m)}\, \boldsymbol{u}^{(m)} . \eth_t \mathcal{P}_\mathcal{E} \boldsymbol{\varphi}\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega (\rho^{(m)})^0(x)\, (\boldsymbol{u}^{(m)})^0(x) . \mathcal{P}_\mathcal{E} \boldsymbol{\varphi}(x,0)\, \mathrm{d}\boldsymbol{x}.$$

By the same arguments as for the mass balance equation, we therefore obtain :

$$\lim_{m\to+\infty} \sum_{i=1}^d T_{1,i}^{(m)} = -\int_0^T \int_\Omega \bar{\rho}\, \overline{\boldsymbol{u}} . \partial_t \boldsymbol{\varphi}\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t - \int_\Omega \rho_0(x)\, \boldsymbol{u}_0(x) . \boldsymbol{\varphi}(x,0)\, \mathrm{d}\boldsymbol{x}.$$

We now turn to the viscosity term $T^{(m)}_{visco,i}$. Reordering the sum we get that :

$$T^{(m)}_{visco,i} = \sum_{\epsilon\in\tilde{\mathcal{E}}^{(i)}_{\text{int}}} \mu^n_\epsilon (u^n_{\sigma,i} - u^n_{\sigma',i}).(\varphi^{n+1}_{\sigma,i} - \varphi^{n+1}_{\sigma',i}).$$

An important property of the dual flux is that they are bounded with respect to the primal flux. Consequently,

$$|F^n_{\sigma,\epsilon}| \leq C\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\|\boldsymbol{u}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d} h_K^{d-1}.$$

The viscosity has the same behaviour by construction :

$$\nu^n_\epsilon \leq C h_K^{d-1}.$$

We then write

$$T^{(m)}_{visco,i} = \sum_{\epsilon\in\tilde{\mathcal{E}}_{\text{int}}} d_\epsilon \frac{\nu^n_\epsilon + \frac{|F^n_{\sigma,\epsilon}|}{2}}{|\sigma|}|\sigma|(u^n_{\sigma,i} - u^n_{\sigma',i})\frac{\varphi^{n+1}_{\sigma,i} - \varphi^{n+1}_{\sigma',i}}{d_\epsilon},$$

and so we can bound the viscosity term,

$$|T^{(m)}_{visco,i}| \leq h^{(m)} C_\varphi \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)},$$

which tends to zero as m tends to $+\infty$.

We now turn to the convection part of the momentum balance equation. Let $\psi = (\psi_\sigma)_{\sigma\in\mathcal{E}^{(i)}_{\text{int}}}$ be a discrete scalar function. Within the next paragraph we will omit the time superscript and simply denote by φ_{edge} the term $\varphi_{\sigma,i}$, for readability. We denote by $\overline{C}_i(\psi,\varphi)$ the quantity :

$$\overline{C}(\psi,\varphi) = \sum_{\sigma\in\mathcal{E}^{(i)}} \varphi_\sigma \left[\sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}\frac{\psi_\sigma + \psi_{\sigma'}}{2}\right],$$

and $C_i(\psi,\varphi)$ by :

$$C_i(\psi,\varphi) = \sum_{K\in\mathcal{M}} \varphi_K \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\hat{\psi}_\sigma,$$

adopting the notations (3.49) and (3.50). Let $R_i(\psi,\varphi) = \overline{C}(\psi,\varphi) - C_i(\psi,\varphi)$. We first decompose the sum in $\overline{C}(\psi,\varphi)$ as a sum over the primal cells, to obtain :

$$\overline{C}(\psi,\varphi) = \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}^{(i)}(K)} \varphi_\sigma \left[F_{\sigma,\epsilon_K}\psi_{\epsilon_K} + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)\\ \epsilon\neq\epsilon_K,\ (\epsilon\cap\overline{K})\subset\sigma'}} \frac{F_{K,\sigma'}}{2}\psi_\epsilon\right],$$

with ψ_ϵ the centered interpolation. By conservativity, adding $F_{K,\sigma}\psi_\sigma$ in the internal sum does not change the value of the sum, so

$$\overline{C}(\psi,\varphi) = \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}^{(i)}(K)} \varphi_\sigma \left[F_{K,\sigma}\psi_\sigma + F_{\sigma,\epsilon_K}\psi_{\epsilon_K} + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)\\ \epsilon\neq\epsilon_K,\ (\epsilon\cap\overline{K})\subset\sigma'}} \frac{F_{K,\sigma'}}{2}\psi_\epsilon\right].$$

We now remark that, by conservativity :

$$\sum_{\sigma\in\mathcal{E}^{(i)}(K)} \varphi_K F_{\sigma,\epsilon_K}\psi_{\epsilon_K} = 0,$$

and that :

$$\sum_{\sigma\in\mathcal{E}^{(i)}(K)} \varphi_\sigma \left[F_{K,\sigma}\psi_\sigma + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)\\ \epsilon\neq\epsilon_K,\ (\epsilon\cap\overline{K})\subset\sigma'}} \frac{F_{K,\sigma'}}{2}\psi_\epsilon\right] = \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\hat{\psi}_\sigma,$$

so the $C_i(\psi,\varphi)$ may be written as :

$$C_i(\psi,\varphi) = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}^{(i)}(K)} \varphi_K \left[F_{K,\sigma}\psi_\sigma + F_{\sigma,\epsilon_K}\psi_{\epsilon_K} + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)\\ \epsilon\neq\epsilon_K,\ (\epsilon\cap\overline{K})\subset\sigma'}} \frac{F_{K,\sigma'}}{2}\psi_\epsilon \right].$$

We thus get :

$$R_i(\psi,\varphi) = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}^{(i)}(K)} (\varphi_\sigma - \varphi_K) \left[F_{K,\sigma}\psi_\sigma + F_{\sigma,\epsilon_K}\psi_{\epsilon_K} + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)\\ \epsilon\neq\epsilon_K,\ (\epsilon\cap\overline{K})\subset\sigma'}} \frac{F_{K,\sigma'}}{2}\psi_\epsilon \right].$$

We remark that a discrete mass balance is satisfied over the half-diamond cells, in the sense that :

$$F_{K,\sigma} + F_{\sigma,\epsilon_K} + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)\\ \epsilon\neq\epsilon_K,\ (\epsilon\cap\overline{K})\subset\sigma'}} \frac{F_{K,\sigma'}}{2} = \frac{1}{2}\sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma},$$

and hence :

$$\sum_{\sigma\in\mathcal{E}^{(i)}(K)} (\varphi_\sigma - \varphi_K) \left[F_{K,\sigma} + F_{\sigma,\epsilon_K} + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)\\ \epsilon\neq\epsilon_K,\ (\epsilon\cap\overline{K})\subset\sigma'}} \frac{F_{K,\sigma'}}{2} \right] =$$
$$\left[\sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma} \right] \left[\left(\sum_{\sigma\in\mathcal{E}^{(i)}(K)} \frac{\varphi_\sigma}{2} \right) - \varphi_K \right] = 0.$$

We may thus write :

$$R_i(\psi,\varphi) = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}^{(i)}(K)} (\varphi_\sigma - \varphi_K) \Big[F_{K,\sigma}(\psi_\sigma - \psi_K) + \sum_{\substack{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)\\ \epsilon\neq\epsilon_K,\ (\epsilon\cap\overline{K})\subset\sigma'}} \frac{F_{K,\sigma'}}{2}(\psi_\epsilon - \psi_K) \Big]. \tag{3.60}$$

We are now able to rewrite $T_{2,i}^{(m)}$. We have :

$$T_{2,i}^{(m)} = \sum_{n=0}^{N-1} \delta t\ C_i(\boldsymbol{u}^n, \boldsymbol{\varphi}^{n+1}) + \sum_{n=0}^{N-1} \delta t\ R_i(\boldsymbol{u}^n, \boldsymbol{\varphi}^{n+1}).$$

Looking at expression (3.60) it appears that :

$$\sum_{n=0}^{N-1} \delta t\ R_i(\boldsymbol{u}^n, \boldsymbol{\varphi}^{n+1}) \leq h^{(m)} C_\varphi \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)} \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \|\boldsymbol{u}^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d},$$

so the second term of $T_{2,i}^{(m)}$ tends to zero as $m\to\infty$. On the other hand we have :

$$\mathcal{T}_{2,i}^{(m)} = \sum_{n=0}^{N-1} \delta t\ C_i(\boldsymbol{u}^n, \boldsymbol{\varphi}^{n+1}) = -\sum_{j=1}^{d}\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}_{\mathrm{int}}^{(j)}} \rho_\sigma^n u_{\sigma,j}^n (\hat{u}^{(i)})_\sigma^n (\varphi_{L,i}^{n+1} - \varphi_{K,i}^{n+1}).$$

We set $\mathcal{T}_{2,i}^{(m)} = \mathcal{T}_{2,1,i}^{(m)} + \mathcal{R}_{2,1,i}^{(m)}$, with :

$$\mathcal{T}_{2,1,i}^{(m)} = -\sum_{j=1}^{d}\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \left(|D_{K,\sigma}|\rho_K^n + |D_{L,\sigma}|\rho_L^n \right) u_{\sigma,j}^n\ (\hat{u}^{(i)})_\sigma^n \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1}),$$

$$\mathcal{R}_{2,1,i}^{(m)} = -\sum_{j=1}^{d}\sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \left(|D_\sigma|\rho_\sigma^n - |D_{K,\sigma}|\rho_K^n - |D_{L,\sigma}|\rho_L^n \right) u_{\sigma,j}^n\ (\hat{u}^{(i)})_\sigma^n \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1}).$$

Bounding $\mathcal{R}^{(m)}_{2,1,i}$ is straightforward :

$$|\mathcal{R}^{(m)}_{2,1,i}| \leq C_\varphi \|\boldsymbol{u}^{(m)}\|^2_{\mathrm{L}^\infty(\Omega\times(0,T))^d} \|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)} h^{(m)},$$

so it tends to zero as $m \to \infty$. For $\mathcal{T}^{(m)}_{2,1,i}$ we have :

$$\mathcal{T}^{(m)}_{2,1,i} = -\sum_{j=1}^{d} \int_0^T \int_\Omega \rho^{(m)} u^{(m)}_j (\hat{u}^{(i)})^{(m)} \eth_{x,j} \varphi_{\mathcal{M},i}$$

$$\mathcal{T}^{(m)}_{2,1,i} = -\int_0^T \int_\Omega \rho^{(m)} \boldsymbol{u}^{(m)} (\hat{u}^{(i)})^{(m)} . \nabla_{\mathcal{E}} \varphi_{\mathcal{M},i}$$

Summing over i leads to :

$$\sum_{i=1}^{d} \mathcal{T}^{(m)}_{2,1,i} = -\int_0^T \int_\Omega (\rho^{(m)} \boldsymbol{u}^{(m)} \otimes \hat{\boldsymbol{u}}^{(m)}) . \underline{\underline{\nabla_{\mathcal{E}}}} \varphi_{\mathcal{M}},$$

with $\hat{\boldsymbol{u}}^{(m)} = \sum_{i=1}^{d} (\hat{u}^{(i)})^{(m)} \boldsymbol{e}^{(i)}$. We need to prove that $(\hat{u}^{(i)})^{(m)}$ strongly converges to $\bar{\boldsymbol{u}}_i$ as $m \to \infty$. Let $R = \int_0^T \int_\Omega |u^{(m)}_i - (\hat{u}^{(i)})^{(m)}|$. Let us denote, for $\sigma \in \mathcal{E}^{(j)}$ and $\sigma' \in \mathcal{E}^{(i)}$ such that $D_\sigma \cap D_{\sigma'} \neq \emptyset$, $D_{\sigma,\sigma'} = D_\sigma \cap D_{\sigma'}$. We have :

$$R = \sum_{n=0}^{N-1} \delta t \sum_{D_{\sigma,\sigma'}} |D_{\sigma,\sigma'}| |u^n_{\sigma',i} - \frac{1}{\mathrm{card}(\mathcal{N}_\sigma)} \sum_{\sigma'' \in \mathcal{N}_\sigma} u^n_{\sigma'',i}|,$$

$$R = \sum_{n=0}^{N-1} \delta t \sum_{D_{\sigma,\sigma'}} \frac{|D_{\sigma,\sigma'}|}{\mathcal{N}_\sigma} |\sum_{\sigma'' \in \mathcal{N}_\sigma} (u^n_{\sigma',i} - u^n_{\sigma'',i})|,$$

$$R \leq \sum_{n=0}^{N-1} \delta t \sum_{D_{\sigma,\sigma'}} \frac{|D_{\sigma,\sigma'}|}{\mathcal{N}_\sigma} \sum_{\sigma'' \in \mathcal{N}_\sigma} |u^n_{\sigma',i} - u^n_{\sigma'',i}|.$$

By means of triangular inequalities one can see that we get that :

$$R \leq C \sum_{n=0}^{N-1} \delta t \sum_{\sigma \in \mathcal{E}^{(i)}} |D_\sigma| \sum_{\epsilon = D_\sigma | D_{\sigma'} \in \tilde{\mathcal{E}}(D_\sigma)} |u^n_{\sigma,i} - u^n_{\sigma',i}|.$$

Reordering the sum leads to :

$$R \leq C h^{(m)} \sum_{n=0}^{N-1} \delta t \sum_{\epsilon = D_\sigma | D_{\sigma'} \in \tilde{\mathcal{E}}^{(i)}_{\mathrm{int}}} |\epsilon| |u^n_{\sigma,i} - u^n_{\sigma',i}|,$$

so we have :

$$R \leq C h^{(m)} \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)} \underset{m \to +\infty}{\longrightarrow} 0.$$

Consequently, as $\boldsymbol{u}^{(m)}$ converges strongly to $\bar{\boldsymbol{u}}$, so does $\hat{\boldsymbol{u}}^{(m)}$. We are now in position to conclude as :

$$\sum_{i=1}^{d} \mathcal{T}^{(m)}_{2,1,i} \underset{m \to +\infty}{\longrightarrow} -\int_0^T \int_\Omega (\bar{\rho} \bar{\boldsymbol{u}} \otimes \bar{\boldsymbol{u}}) . \underline{\underline{\nabla}} \boldsymbol{\varphi},$$

Total energy balance equation – On one hand, let us multiply the discrete internal energy balance equation (3.5b) by $\delta t\, \varphi^{n+1}_K$, and sum the result for $0 \leq n \leq N-1$ and $K \in \mathcal{M}$. On the other hand, let us multiply the discrete kinetic energy balance (3.25) by $\delta t\, \varphi^{n+1}_{\sigma,i}$, and sum the result for $0 \leq n \leq N-1$, $\sigma \in \mathcal{E}^{(i)}_{\mathrm{int}}$ and $i \in [1,d]$. Finally, adding the two obtained relations, we get :

$$T^{(m)}_1 + T^{(m)}_2 + T^{(m)}_3 + \sum_{i=1}^{d} \tilde{T}^{(m)}_{1,i} + \tilde{T}^{(m)}_{2,i} + \tilde{T}^{(m)}_{2,R,i} + \tilde{T}^{(m)}_{3,i} + R = S^{(m)} - \sum_{i=1}^{d} \tilde{R}^{(m)}_i, \tag{3.61}$$

where :

$$T_1^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \frac{|K|}{\delta t} \left[\rho_K^{n+1} e_K^{n+1} - \rho_K^n e_K^n\right] \varphi_K^{n+1},$$

$$T_2^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n e_\sigma^n \varphi_K^{n+1},$$

$$T_3^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} |K| p_K^n (\mathrm{div}(\boldsymbol{u}))_K^n \varphi_K^{n+1},$$

$$\tilde{T}_{1,i}^{(m)} = \frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} \frac{|D_\sigma|}{\delta t} \left[\rho_{D_\sigma}^{n+1} |u_{\sigma,i}^{n+1}|^2 - \rho_{D_\sigma}^n |u_{\sigma,i}^n|^2\right] \varphi_{\sigma,i}^{n+1},$$

$$\tilde{T}_{2,i}^{(m)} = \frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} \varphi_{\sigma,i}^{n+1} \sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n \frac{(u_{\sigma,i}^n)^2 + (u_{\sigma',i}^n)^2}{2}$$

$$\tilde{T}_{2,R,i}^{(m)} = \frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} \varphi_{\sigma,i}^{n+1} \sum_{\epsilon\in\tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\epsilon}^n \left(u_{\sigma,i}^n u_{\sigma',i}^n - \frac{(u_{\sigma,i}^n)^2 + (u_{\sigma',i}^n)^2}{2}\right) + \mu_\epsilon^n (u_{\sigma,i}^n - u_{\sigma',i}^n)(u_{\sigma',i}^n + u_{\sigma,i}^n)$$

$$\tilde{T}_{3,i}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}_{\mathrm{int}}^{(i)}} |\sigma| (p_{\mathrm{L}}^{n+1} - p_K^{n+1}).u_{\sigma,i}^{n+1} \varphi_{\sigma,i}^{n+1},$$

$$S^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}} S_K^n \varphi_K^{n+1}, \qquad \tilde{R}_i^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}^{(i)}} R_{\sigma,i}^{n+1} \varphi_{\sigma,i}^{n+1},$$

and the quantities S_K^n and $R_{\sigma,i}^{n+1}$ are given by Equation (3.31) and Equation (3.26) respectively.

The consistency of the term $T_1^{(m)}$ is similar to the one written in the RT case. Using the expression of the mass flux $F_{K,\sigma}$ and reordering the sum in $T_2^{(m)}$ we get :

$$T_2^{(m)} = -\sum_{i=1}^{d} \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}^{(i)}} |D_\sigma| \rho_\sigma^n e_\sigma^n u_{\sigma,i}^n \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1}).$$

We decompose the sum in two terms, $T_2^{(m)} = \mathcal{T}_2^{(m)} + \mathcal{R}_2^{(m)}$ with

$$\mathcal{T}_2^{(m)} = -\sum_{i=1}^{d} \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}^{(i)}} \left(|D_{K,\sigma}| \rho_K^n e_K^n + |D_{L,\sigma}| \rho_L^n e_{\mathrm{L}}^n\right) u_{\sigma,i}^n \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1}),$$

$$\mathcal{R}_2^{(m)} = -\sum_{i=1}^{d} \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}^{(i)}} \left(|D_\sigma| \rho_\sigma^n - |D_{K,\sigma}| \rho_K^n e_K^n - |D_{L,\sigma}| \rho_L^n e_{\mathrm{L}}^n\right) u_{\sigma,i}^n \frac{|\sigma|}{|D_\sigma|} (\varphi_L^{n+1} - \varphi_K^{n+1}).$$

We have, for the term $\mathcal{T}_2^{(m)}$:

$$\mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \rho^{(m)} e^{(m)} \boldsymbol{u}^{(m)}.\nabla_{\mathcal{E}} \mathcal{P}_{\mathcal{M}} \varphi$$

and therefore

$$\lim_{m\to+\infty} \mathcal{T}_2^{(m)} = -\int_0^T \int_\Omega \bar{\rho}\bar{e}\,\overline{\boldsymbol{u}}.\nabla\varphi.$$

The remainder term $\mathcal{R}_2^{(m)}$ can be expressed thanks to the MUSCL property (3.12). There exists $\alpha_\sigma^n \in [0,1]$ such that :

$$\mathcal{R}_2^{(m)} = -\sum_{i=1}^{d}\sum_{n=0}^{N-1}\delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}^{(i)}} \Big(\alpha_\sigma^n |D_{K,\sigma}|\left(\rho_K^n e_K^n - \rho_L^n e_L^n\right) - (1-\alpha_\sigma^n)\,|D_{L,\sigma}| \left(\rho_K^n e_K^n - \rho_L^n e_L^n\right)\Big) u_{\sigma,i}^n \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1}).$$

We thus get :

$$|\mathcal{R}_2^{(m)}| \le C_\varphi \sum_{i=1}^{d}\sum_{n=0}^{N-1}\delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}^{(i)}} |D_\sigma||\rho_K^n e_K^n - \rho_L^n e_L^n||u_{\sigma,i}^n|,$$

with C_φ only depending on φ. Applying the identity $2\,(ab-cd) = (a-c)\,(b+d) + (a+c)\,(b-d)$, which holds for any a,b,c,d real, to the quantity $\rho_K^n e_K^n - \rho_L^n e_L^n$, we obtain :

$$|\mathcal{R}_2^{(m)}| \le C_\varphi\, h^{(m)}\, \|u^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \Big[\|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \sum_{i=1}^{d} \|e^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)} + \|e^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \sum_{i=1}^{d} \|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)}\Big],$$

and thus $|\mathcal{R}_2^{(m)}|$ tends to zero when m tends to $+\infty$.

We now turn to $\tilde{T}_{1,i}^{(m)}$. The definition (3.16) of ρ_{D_σ} and a reordering in the summation yields :

$$\tilde{T}_{1,i}^{(m)} = -\frac{1}{2}\sum_{n=0}^{N-1}\delta t \sum_{\sigma=K|L\in\mathcal{E}^{(i)}} \left[|D_{K,\sigma}|\,\rho_K^n + |D_{L,\sigma}|\,\rho_L^n\right]|u_{\sigma,i}^n|^2\, \frac{\varphi_{\sigma,i}^{n+1} - \varphi_{\sigma,i}^n}{\delta t} - \frac{1}{2}\sum_{\sigma=K|L\in\mathcal{E}^{(i)}} \left[|D_{K,\sigma}|\,\rho_K^0 + |D_{L,\sigma}|\,\rho_L^0\right]|\boldsymbol{u}_{\sigma,i}^0|^2\,\varphi_{\sigma,i}^0,$$

so that, by similar arguments as for the term $T_1^{(m)}$, we get :

$$\lim_{m\to+\infty} \tilde{T}_{1,i}^{(m)} = -\int_0^T\int_\Omega \frac{1}{2}\,\bar{\rho}\,\bar{\boldsymbol{u}}_i^2\,\partial_t\varphi_i\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \frac{1}{2}\,\rho_0(x)\,u_{0,i}(x)^2\,\varphi(x,0)_i\,\mathrm{d}x.$$

The convergence to zero of the term $\tilde{T}_{2,R,i}$ is similar to the RT case. To lighten the expression we denote by $\chi_\sigma^n = \frac{1}{2}|\boldsymbol{u}_\sigma^n|^2$, $\chi_{\sigma,i}^n = \frac{1}{2}(u_{\sigma,i}^n)^2$ and we run through the same computations as in the momentum balance equation, with, $\psi_\sigma = \chi_\sigma^n$, to get :

$$\mathcal{T}_{2,i}^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1}\delta t \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^n \hat{\chi}_{\sigma,i}^n \varphi_{K,i}^{n+1} + \sum_{n=0}^{N-1}\delta t R_i(\chi^n, \boldsymbol{\varphi}^{n+1})$$

Thanks to (3.60), we get that :

$$\sum_{n=0}^{N-1}\delta t\, R_i(\chi^n, \boldsymbol{\varphi}^{n+1}) \le h^{(m)} C_\varphi \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)} \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \|\boldsymbol{u}^{(m)}\|^2_{\mathrm{L}^\infty(\Omega\times(0,T))^d},$$

so it tends to zero as $m\to\infty$. The strong convergence of $\hat{\chi}^{(m)}$ is straightforward (see the proof of convergence of $\hat{\boldsymbol{u}}^{(m)}$). We set $\mathcal{T}_{2,i}^{(m)} = \mathcal{T}_{2,1,i}^{(m)} + \mathcal{R}_{2,1,i}^{(m)}$, with :

$$\mathcal{T}_{2,1,i}^{(m)} = -\sum_{j=1}^{d}\sum_{n=0}^{N-1}\delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \left(|D_{K,\sigma}|\rho_K^n + |D_{L,\sigma}|\rho_L^n\right) u_{\sigma,j}^n\, (\hat{\chi}^{(i)})_\sigma^n \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1}),$$

$$\mathcal{R}_{2,1,i}^{(m)} = -\sum_{j=1}^{d}\sum_{n=0}^{N-1}\delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}} \left(|D_\sigma|\rho_\sigma^n - |D_{K,\sigma}|\rho_K^n - |D_{L,\sigma}|\rho_L^n\right) u_{\sigma,j}^n\, (\hat{\chi}^{(i)})_\sigma^n \frac{|\sigma|}{|D_\sigma|}(\varphi_L^{n+1} - \varphi_K^{n+1}).$$

It is straightforward that :

$$\lim_{m\to+\infty} \mathcal{R}^{(m)}_{2,1,i} = 0,$$

and

$$\mathcal{T}^{(m)}_{2,1,i} = -\sum_{j=1}^{d} \int_0^T \int_\Omega \rho^{(m)} \boldsymbol{u}^{(m)}_j (\hat{\chi}^{(i)})^{(m)} \eth_{x,j} \varphi_{\mathcal{M},i},$$

Passing to the limit we finally get :

$$\lim_{m\to+\infty} \mathcal{T}^{(m)}_{2,1,i} = -\int_0^T \int_\Omega \bar{\rho}\bar{\boldsymbol{u}} \left(\frac{1}{2}|\bar{\boldsymbol{u}}_i|^2\right).\nabla\varphi_i$$

The terms $\tilde{T}^{(m)}_{3,i}$ and $T^{(m)}_3$ are analyzed together.

$$\tilde{T}^{(m)}_{3,i} = \sum_{n=1}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}^{(i)}_{\text{int}}} |D_\sigma| (\eth_{x,i} p^{(m)})^n_\sigma u^n_{\sigma,i} \varphi^n_{\sigma,i} = \tilde{\mathcal{T}}^{(m)}_{3,i} + \tilde{\mathcal{R}}^{(m)}_{3,i},$$

with :

$$\tilde{\mathcal{T}}^{(m)}_{3,i} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}^{(i)}_{\text{int}}} |D_\sigma| (\eth_{x,i} p^{(m)})^n_\sigma u^n_{\sigma,i} \varphi^{n+1}_{\sigma,i},$$

$$\tilde{\mathcal{R}}^{(m)}_{3,i} = -\delta t \sum_{\sigma\in\mathcal{E}^{(i)}_{\text{int}}} |D_\sigma| (\eth_{x,i} p^{(m)})^0_\sigma u^0_{\sigma,i} \varphi^0_{\sigma,i}$$

$$+ \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}^{(i)}_{\text{int}}} |D_\sigma| (\eth_{x,i} p^{(m)})^n_\sigma u^n_{\sigma,i} \, (\varphi^n_{\sigma,i} - \varphi^{n+1}_{\sigma,i}).$$

We have, thanks to the regularity of φ :

$$|\tilde{\mathcal{R}}^{(m)}_{3,i}| \quad \leq \quad C_\varphi \quad \delta t^{(m)} \Big[\|(u^{(m)})^0\|_{\mathrm{L}^\infty(\Omega)} \quad \|(p^{(m)})^0\|_{\mathrm{BV}(\Omega)} \; + \; \|u^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))} \quad \|p^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)} \Big].$$

Therefore, invoking the regularity of the initial conditions, this term tends to zero when m tends to $+\infty$. A simple computation yields :

$$T^{(m)}_3 \; + \; \sum_{i=1}^{d} \tilde{\mathcal{T}}^{(m)}_{3,i} \quad = \quad \sum_{i=1}^{d} \sum_{n=0}^{N-1} \delta t \sum_{\sigma=\overrightarrow{K|L}\in\mathcal{E}^{(i)}_{\text{int}}} |\sigma| \Big[p^n_{\mathrm{L}} (\varphi^{n+1}_{\sigma,i} \; - \; \varphi^{n+1}_{\mathrm{L}}) \; + \; p^n_K (\varphi^{n+1}_K \; - \; \varphi^{n+1}_{\sigma,i}) \Big] \boldsymbol{u}^n_{\sigma,i}.$$

The sum is equal to :

$$T^{(m)}_3 + \sum_{i=1}^{d} \tilde{\mathcal{T}}^{(m)}_{3,i} = -\sum_{i=1}^{d} \sum_{n=0}^{N-1} \delta t \sum_{D_{K,\sigma}\in\boldsymbol{Q}^{(i)}} |D_{K,\sigma}| p^n_K u^n_{\sigma,i} \eth_{x,i} \varphi^{(m)}_{\mathcal{M},\mathcal{E}},$$

$$= -\int_0^T \int_\Omega p^{(m)} \boldsymbol{u}^{(m)}.\nabla^h (\underline{\mathcal{P}}_{\mathcal{E}^{(m)}} \varphi, \mathcal{P}_{\mathcal{M}^{(m)}} \varphi),$$

so we can deduce that :

$$\lim_{m\to+\infty} \mathcal{T}^{(m)}_3 + T^{(m)}_3 = -\int_0^T \int_\Omega \bar{p}\bar{\boldsymbol{u}}.\nabla\varphi.$$

We finally focus on the remainder terms $S^{(m)}_i - \tilde{R}^{(m)}_i$. Let us write this quantity as $S^{(m)}_i - \tilde{R}^{(m)}_i = \mathcal{R}^{(m)}_1 + \mathcal{R}^{(m)}_2$ where, using $S^0_{K,i} = 0, \forall K \in \mathcal{M}$:

$$\mathcal{R}^{(m)}_1 = \sum_{n=0}^{N-1} \delta t \Big[\sum_{K\in\mathcal{M}} S^{n+1}_{K,i} \varphi^{n+1}_K - \sum_{\sigma\in\mathcal{E}^{(i)}} R^{n+1}_{\sigma,i} \varphi^{n+1}_{\sigma,i} \Big],$$

$$\mathcal{R}^{(m)}_2 = \sum_{n=1}^{N-1} \delta t \sum_{K\in\mathcal{M}} S^n_{K,i} \, (\varphi^{n+1}_K - \varphi^n_K).$$

First, we prove that $\lim_{m\to+\infty} \mathcal{R}_1^{(m)} = 0$. Gathering and reordering the sums, we obtain $\mathcal{R}_1^{(m)} = \mathcal{R}_{1,1}^{(m)} + \mathcal{R}_{1,2}^{(m)} + \mathcal{R}_{1,3}^{(m)}$ with

$$\mathcal{R}_{1,1}^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1} \delta t \sum_{\sigma=K|L\in\mathcal{E}} \Big[\frac{|D_{K,\sigma}|}{\delta t}\rho_K^{n+1}|u_{\sigma,i}^{n+1} - u_{\sigma,i}^n|^2(\varphi_K^{n+1} - \varphi_{\sigma,i}^{n+1}) + \frac{|D_{L,\sigma}|}{\delta t}\rho_L^{n+1}|u_{\sigma,i}^{n+1} - u_{\sigma,i}^n|^2(\varphi_L^{n+1} - \varphi_{\sigma,i}^{n+1})\Big],$$

$$\mathcal{R}_{1,2}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}_{\text{int}}} \frac{1}{2}\mu_\epsilon^n(u_{\sigma'}^n - u_{\sigma,i}^n)^2\left(\varphi_K^{n+1} - \varphi_{\sigma,i}^{n+1} + \varphi_K^{n+1} - \varphi_{\sigma',i}^{n+1}\right)$$

$$\mathcal{R}_{1,3}^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\epsilon=D_\sigma|D_{\sigma'}\in\tilde{\mathcal{E}}_{\text{int}}} (\mu_\epsilon^n - \frac{F_{\sigma,\epsilon}^n}{2})(u_{\sigma,i}^{n+1} - u_{\sigma,i}^n)\,(u_{\sigma,i}^n - u_{\sigma',i}^n)(\varphi_K^{n+1} - \varphi_{\sigma,i}^{n+1}) + (\mu_\epsilon^n - \frac{F_{\sigma',\epsilon}^n}{2})(u_{\sigma',i}^{n+1} - u_{\sigma',i}^n)\,(u_{\sigma',i}^n - u_{\sigma,i}^n)(\varphi_K^{n+1} - \varphi_{\sigma',i}^{n+1})$$

We thus obtain :

$$|\mathcal{R}_{1,1}^{(m)}| \le h^{(m)}\, C_\varphi\, \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\, \|\boldsymbol{u}^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d}\, \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},t,\mathrm{BV},(i)},$$

and

$$|\mathcal{R}_{1,2}^{(m)}| + |\mathcal{R}_{1,3}^{(m)}| \le h^{(m)}\, C_\varphi\, \|\rho^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))}\, \|\boldsymbol{u}^{(m)}\|^2_{\mathrm{L}^\infty(\Omega\times(0,T))^d}\, \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)},$$

so these two terms tend to zero. With straightforward computations we can also bound the term $\mathcal{R}_{1,visco}^{(m)}$:

$$|\mathcal{R}_{1,visco}^{(m)}| \le C_\varphi\, h^{(m)}\, \|\boldsymbol{u}^{(m)}\|_{\mathrm{L}^\infty(\Omega\times(0,T))^d}\, \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},x,\mathrm{BV},(i)}\, \|\mu^{(m)}\|_\infty$$

Then $\lim_{m\to+\infty} \mathcal{R}_{1,visco}^{(m)} = 0$ The fact that $|\mathcal{R}_2^{(m)}|$ behaves as $\delta t^{(m)}$ may be proven by similar arguments.

Entropy inequality – The proof of the consistency of the MAC scheme with en entropy inequality is very similar to the one in the RT case. The only difference is that you have to decompose your terms according to each direction of space (see the mass balance equation consistency with the MAC scheme). ■

3.5.4 Pressure correction scheme

Some adjustments need to be made in order to perform the same proof of consistency as the decoupled case. First of all the temporal indices are changed. Concerning the definition of the discrete solutions of the scheme we have now

$$\rho^{(m)}(\boldsymbol{x}, t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} (\rho^{(m)})_K^{n+1}\, \mathcal{X}_K(\boldsymbol{x})\, \mathcal{X}_{(n,n+1]}(t),$$

and the same changes concerning the others variables. The discrete interpolator are changed as a consequence.

$$\mathcal{P}_\mathcal{M} : \left|\begin{array}{lcl} C((0,T); H_0^1(\Omega)) & \longrightarrow & L_\mathcal{M}(\Omega\times(0,T)) \\ \varphi & \mapsto & \mathcal{P}_\mathcal{M}\varphi(\boldsymbol{x}) = \displaystyle\sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}} \varphi_K^n\, \mathcal{X}_K\, \mathcal{X}_{[t^n,t^{n+1})}, \end{array}\right.$$

This modification also impacts the other interpolators and the discrete derivatives.

Concerning the estimates on the discrete solutions, we consider the following hypothesis

$$\lim_{m\to\infty}\left(h^{(m)} + \delta t^{(m)}\right)\Big[\|\rho^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|p^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|e^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|\tilde{\boldsymbol{u}}^{(m)}\|_{\mathcal{T},x,\mathrm{BV}} + \|\boldsymbol{u}^{(m)}\|_{\mathcal{T},t,\mathrm{BV}}\Big] = 0.$$

We furthermore suppose a uniform bound on $\tilde{\boldsymbol{u}}^{(m)}$. Note that we do not need any control on time discrete derivative of $\tilde{\boldsymbol{u}}^{(m)}$ or space discrete derivative of $\boldsymbol{u}^{(m)}$.

We now turn to the theorem for the pressure correction scheme.

Theorem 3.10 (*Consistency of the multi-dimensional pressure correction scheme*)

Let Ω be an open bounded interval of $\mathbb{R}$. We suppose that the initial data satisfies $\rho_0 \in \mathrm{L}^\infty(\Omega)$, $p_0 \in \mathrm{BV}(\Omega)$, $e_0 \in \mathrm{L}^\infty(\Omega)$ and $\boldsymbol{u}_0 \in \mathrm{L}^\infty(\Omega)^d$. Let $(\mathcal{M}^{(m)}, \delta t^{(m)})_{m\in\mathbb{N}}$ be a sequence of discretizations such that both the time step $\delta t^{(m)}$ and the size $h^{(m)}$ of the mesh $\mathcal{M}^{(m)}$ tend to zero as $m \to \infty$, and let $(\rho^{(m)}, p^{(m)}, e^{(m)}, \boldsymbol{u}^{(m)}, \tilde{\boldsymbol{u}}^{(m)})_{m\in\mathbb{N}}$ be the corresponding sequence of solutions. We suppose that this sequence satisfies the estimates (3.53)–(3.55) and converges in $\mathrm{L}^r(\Omega\times(0,T))^3\times(\mathrm{L}^r(\Omega\times(0,T))^d)^2$, for $1 \le r < \infty$, to $(\bar{\rho}, \bar{p}, \bar{e}, \bar{\boldsymbol{u}}, \bar{\tilde{\boldsymbol{u}}}) \in \mathrm{L}^\infty(\Omega\times(0,T))^3 \times (\mathrm{L}^\infty(\Omega\times(0,T))^d)^2$. Futhermore suppose that the CFL condition (3.56) is satisfied.

Then $\bar{\tilde{\boldsymbol{u}}} = \bar{\boldsymbol{u}}$ and the limit $(\bar{\rho}, \bar{p}, \bar{e}, \bar{\boldsymbol{u}})$ satisfies the system (3.44)-(3.52).

Proof : Most of the proof is very similar to the decoupled case. We only focus on the major differences for the sake of clarity. We first need to check that $\bar{\tilde{\boldsymbol{u}}} = \bar{\boldsymbol{u}}$. Let us multiply the correction equation by $\delta t^2(u^{n+1}_{\sigma,i} - \tilde{u}^{n+1}_{\sigma,i})$ and sum it over i, n and $\sigma \in \mathcal{E}_{\mathrm{int}}$ (we consider RT discretization for simplicity). We get $T^{(m)} + R^{(m)} = 0$, with

$$T^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |D_\sigma||\boldsymbol{u}^{n+1}_\sigma - \tilde{\boldsymbol{u}}^{n+1}_\sigma|^2$$

$$R^{(m)} = \sum_{n=0}^{N-1} \delta t^2 \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} |\sigma| \left[\frac{1}{\rho^n_{D_\sigma}}(p^{n+1}_L - p^{n+1}_K) - \frac{1}{\sqrt{\rho^n_{D_\sigma}\rho^{n-1}_{D_\sigma}}}(p^n_L - p^n_K) \right] (u^{n+1}_{K,\sigma} - \tilde{u}^{n+1}_{K,\sigma}) = 0.$$

Obviously, we have $T^{(m)} = \|\boldsymbol{u}^{(m)} - \tilde{\boldsymbol{u}}^{(m)}\|_{\mathrm{L}^2(\Omega\times(0,T))}$. On the other hand the uniform bound on $\dfrac{1}{\rho^{(m)}}$ and $\boldsymbol{u}^{(m)}$ leads to

$$|R^{(m)}| \le C\delta t \|p^{(m)}\|_{\mathcal{T},\mathcal{X},\mathrm{BV}},$$

with C only depending on ρ_0 and $\boldsymbol{u}_0$. Passing to the limit induces the desired result. The consistency of the scheme with the momentum balance equation (3.44b) is then obtained by summing the prediction equation (3.6b) together with the correction equation (3.6c), taking the scalar product with $\delta t\boldsymbol{\varphi}^n_\sigma$, summing on the internal edges, and following the same steps as in the decoupled case. Finally, when seeking the consistency of the total energy balance equation (3.1c), we have an additional term, namely

$$P^{(m)} = \sum_{n=0}^{N-1} \delta t \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} P^{n+1}_\sigma \varphi^n_\sigma,$$

with P^{n+1}_σ defined in (3.28). Lemma 3.20 in [36] insures that this remaining term tends to zero as m tends to $+\infty$. ■

Chapitre 4

A class of finite volume schemes for the G-equation

4.1 Introduction

The problem addressed in this paper is the so called G-equation, which reads :

$$\partial_t(\rho G) + \operatorname{div}(\rho \boldsymbol{u} G) + \rho u_f |\nabla G| = 0, \tag{4.1}$$

where ρ is the density of the fluid, G stands for the front indicator, $\boldsymbol{u}$ is a convective velocity and u_f is a front propagation speed. This equation, used to model the propagation of fronts in fluids, is a particular Hamilton-Jacobi equation when coupled in a system with the mass balance equation, namely

$$\partial_t \rho + \operatorname{div}(\rho \boldsymbol{u}) = 0.$$

Indeed, the convective part of the equation is a transport operator and we get :

$$\partial_t G + \boldsymbol{u} \cdot \nabla G + u_f |\nabla G| = 0, \tag{4.2}$$

provided that the density never vanish. The unknown is the indicator G, and we consider the other quantities as known data.

The discrete and continuous theories of Hamilton-Jacobi equations were vastly developed by J.-L. Lions [21, 46]. A converging finite difference scheme was developed in [20]. From this point high order extensions to this scheme were given by S.Osher and James A. Sethian in [51], and a simple finite volume scheme was derived in [41], inspired from a unstructured finite difference scheme based on triangular meshes developed by R. Abrall in [1]. The convergence theory of numerical approximations of Hamilton Jacobi equations, was first developed for finite difference scheme in [20] and a generalized formulation was given in [6, 58]. Since then, various schemes were developed for Hamilton-Jacobi equations ; high-order finite difference schemes in [13, 56, 52] and schemes for unstructured meshes [9, 57, 65, 5].

The aim in this paper is to propose and study a new finite volume scheme to solve the equation (4.2) on Cartesian and non Cartesian grids. Unlike the scheme proposed in [41], it can be applied for various space discretizations and the discrete spatial operator computation is straightforward. The general framework is based on the theory of viscosity solutions of Hamilton-Jacobi equations (see [20]). Consider the following Cauchy problem :

$$\begin{cases} \partial_t G + H(\nabla G) = 0, \\ G(0, \boldsymbol{x}) = G_0(\boldsymbol{x}), \end{cases} \tag{4.3}$$

defined on $[0, T] \times \mathbb{R}^d$, with $H \in C(\mathbb{R}^d)$ and $G_0 \in \mathrm{BUC}(\mathbb{R}^d)$ ($\mathrm{BUC}(\Omega)$ stands for the set of bounded uniformly continuous functions on Ω). There is exactly one function $G \in \mathrm{BUC}([0, T] \times \mathbb{R}^d)$ such

that $G(0,\boldsymbol{x}) = G_0(\boldsymbol{x})$, and for every $\phi \in C^1((0,\infty)\times\mathbb{R}^d$:

$$\begin{cases} \forall \phi \in C^1(\mathbb{R}^d \times (0,\infty)), \text{ if } (x_0;t_0) \text{ is a local maximum of } G-\phi \text{ on } \mathbb{R}^d\times(0,T], \text{ then,} \\ \partial_t\phi(x_0,t_0) + H(\nabla\phi(x_0,t_0)) \leq 0 \end{cases} \tag{4.4}$$

and

$$\begin{cases} \forall \phi \in C^1(\mathbb{R}^d \times (0,\infty)), \text{ if } (x_0;t_0) \text{ is a local minimum of } G-\phi \text{ on } \mathbb{R}^d\times(0,T], \text{ then,} \\ \partial_t\phi(x_0,t_0) + H(\nabla\phi(x_0,t_0)) \geq 0. \end{cases} \tag{4.5}$$

Equation (4.2) is a particular Hamilton-Jacobi equation, with $H(\boldsymbol{x}) = \boldsymbol{u}\cdot\boldsymbol{x} + u_f|\boldsymbol{x}|$. For the sake of clarity, we suppose that $\boldsymbol{u} = 0$ and $u_f = 1$, so the problem considered here is the unsteady eikonal equation,

$$\partial_t G + |\nabla G| = 0, \tag{4.6a}$$

$$G(0,\boldsymbol{x}) = G_0(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{4.6b}$$

$G_0 \in \mathrm{BUC}(\mathbb{R}^d)$. The scheme proposed here is consistent and monotone on Cartesian grids. The L^∞ convergence is then prooved thanks to the theory developped in [6]. Numerical results are given to highlight this convergence results as well as the numerical convergence of the scheme on unstructured discretizations.

The paper is organized as follows. We start by the description of the spatial discretization and the corresponding notations that are used throughout the paper. We present the scheme and its properties in the second part. We finish with some convergence and numerical results.

4.2 Spatial discretization

In this section, we focus on the discretization of a multi-dimensional domain (*i.e.* $d = 2$ or $d = 3$) ; the extension to the one-dimensional case is straightforward.

Let $\mathcal{M}$ be a mesh of the domain Ω (which is an open bounded connected subset of $\mathbb{R}^d$ or $\mathbb{R}^d$ itself), supposed to be regular in the usual sense of the finite element literature (*e.g.* [19]). The cells of the mesh are assumed to be :

- for a general domain Ω, either non-degenerate quadrilaterals ($d = 2$) or hexahedra ($d = 3$) or simplices, both types of cells being possibly combined in a same mesh,
- for a domain whose boundaries are hyperplanes normal to a coordinate axis, rectangles ($d = 2$) or rectangular parallelepipeds ($d = 3$) (the faces of which, of course, are then also necessarily normal to a coordinate axis).

By $\mathcal{E}$ and $\mathcal{E}(K)$ we denote the set of all $(d-1)$-faces σ of the mesh and of the element $K \in \mathcal{M}$ respectively. The set of faces included in the boundary of Ω is denoted by $\mathcal{E}_{\text{ext}}$ and the set of internal faces (*i.e.* $\mathcal{E} \setminus \mathcal{E}_{\text{ext}}$) is denoted by $\mathcal{E}_{\text{int}}$; a face $\sigma \in \mathcal{E}_{\text{int}}$ separating the cells K and L is denoted by $\sigma = K|L$. The outward normal vector to a face σ of K is denoted by $\boldsymbol{n}_{K,\sigma}$. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, we denote by $|K|$ the measure of K and by $|\sigma|$ the $(d-1)$-measure of the face σ. The mass center of a face is denoted by $\boldsymbol{x}_\sigma$. We denote by $\mathcal{P} = (\boldsymbol{x}_K)_{K\in\mathcal{M}}$ a set of points of Ω such that, $\forall K \in \mathcal{M}$, $\boldsymbol{x}_K \in K$. Most of the time $\mathcal{P}$ consists of mass center of the cells $K \in \mathcal{M}$. For all $\sigma \in \mathcal{E}_{\text{int}}$, we suppose that there is a set of neighbouring cells V_σ such that :

$$\boldsymbol{x}_\sigma = \sum_{K\in V_\sigma} \alpha_{K,\sigma}\boldsymbol{x}_K \qquad \sum_{K\in V_\sigma} \alpha_{K,\sigma} = 1.$$

For unstructured meshes, D_σ refers to the diamond cell whose vertices are $\boldsymbol{x}_L$, $\boldsymbol{x}_K$ and the vertices of σ. If $\sigma \in \mathcal{E}(K)$ lies on the boundary, D_σ is the cone of basis σ and of vertex $\boldsymbol{x}_K$. For Cartesian grids and $\sigma = K|L \in \mathcal{E}_{\text{int}}$, D_σ is composed of two rectangles of basis σ and of measure $|K|/2$ and $|L|/2$. For $\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\text{ext}}$, D_σ is the rectangle of basis σ and of measure $|K|/2$. Finally we denote by d_σ the measure of $\overrightarrow{\boldsymbol{x}_K\boldsymbol{x}_L}$.

We now give the definition of an admissible mesh.

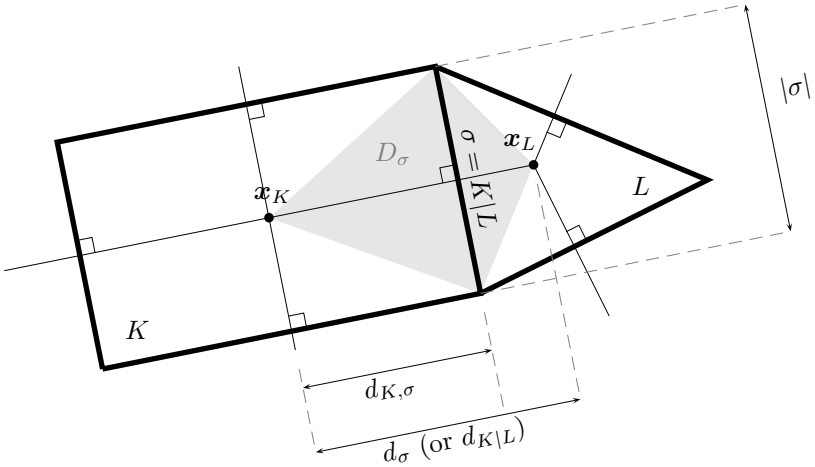


Figure 4.1 – Admissible mesh

Definition 4.1 (*Admissible mesh*)

The set $(\mathcal{M}, \mathcal{E}, \mathcal{P})$ is said to be admissible if and only if :
— For all $\sigma = K|L \in \mathcal{E}_{\text{int}}$, $\overrightarrow{x_K x_L}$ is perpendicular to the face σ.

The unknown discrete function G is piecewise constant on the cells K. We denote by $H_\mathcal{M}$ the space of such piecewise constant functions.

$$G_\mathcal{M} \in H_\mathcal{M} \Longleftrightarrow G_\mathcal{M} = \sum_{K \in \mathcal{M}} G_K \mathcal{X}_K,$$

where $\mathcal{X}_O$ stands for the characteristic function of the set O.

4.3 The scheme

The problem (4.6) is posed over $\mathbb{R}^d \times (0, T)$, where $(0, T)$ is a finite time interval. Concerning the initial data, we have $G_0 \in \text{BUC}(\mathbb{R}^d)$. According to the known results at the continuous level, the problem has a unique viscosity solution in $\text{BUC}([0, T] \times \mathbb{R}^d)$, that we denote $\bar{G}$. In order to be able to perform computations, the domain can be reduced to an open bounded connected subset Ω of $\mathbb{R}^d$. In order to simulate free boundaries, Neumann homogeneous boundary conditions are sufficient. We propose three versions of the scheme depending on the regularity of the mesh. The finite volume scheme is written on an alternative form of Equation (4.6a) :

$$\partial_t G + \left(\frac{\nabla G}{|\nabla G|}\right) \cdot \nabla G = 0, \tag{4.7}$$

and makes use of the classical identity :

$$\boldsymbol{u}.\nabla\phi = \text{div}(\phi \boldsymbol{u}) - \phi \text{div}(\boldsymbol{u}). \tag{4.8}$$

Let us consider a partition $0 = t_0 < t_1 < \ldots < t_N = T$ of the time interval $(0, T)$, which we suppose uniform for the sake of simplicity, and let $\delta t = t_{n+1} - t_n$ for $n = 0, 1, \ldots, N-1$ be the (constant) time step. We consider an explicit-in-time scheme, which reads, for $0 \leq n \leq N-1$ and $K \in \mathcal{M}$:

$$\eth_t G^n + F_\mathcal{M}(G^n) = 0, \tag{4.9}$$

with,

$$\eth_t G^n = \sum_{K\in\mathcal{M}} \frac{G_K^{n+1} - G_K^n}{\delta t} \mathcal{X}_K, \tag{4.10}$$

and

$$F_{\mathcal{M}}(G^n) = \operatorname{div}\left(\frac{\nabla_{\mathcal{E}} G^n}{|\nabla_{\mathcal{E}} G^n|} G^n\right)_K - G_K^n \operatorname{div}\left(\frac{\nabla_{\mathcal{E}} G^n}{|\nabla_{\mathcal{E}} G^n|}\right)_K. \tag{4.11}$$

The discrete divergence operator is given by :

$$\text{for } K \in \mathcal{M}, \qquad (\operatorname{div}\boldsymbol{u})_K = \frac{1}{|K|} \sum_{\sigma=K|L\in\mathcal{E}(K)} \kappa_{K,\sigma}^{\mathcal{M}} |\sigma|\, \boldsymbol{u}_\sigma.\boldsymbol{n}_{K,\sigma}, \tag{4.12}$$

where $\kappa_{K,\sigma}^{\mathcal{M}}$ is a coefficient equal to 1 for unstructured meshes, and equal to $\kappa_{K,\sigma}^{\mathcal{M}} = \dfrac{|K|}{|D_\sigma|}$ on Cartesian grids. Likewise

$$\text{for } K \in \mathcal{M}, \qquad (\operatorname{div} G\boldsymbol{u})_K = \frac{1}{|K|} \sum_{\sigma=K|L\in\mathcal{E}(K)} \kappa_{K,\sigma}^{\mathcal{M}} |\sigma|\, G_\sigma \boldsymbol{u}_\sigma.\boldsymbol{n}_{K,\sigma}, \tag{4.13}$$

where G_σ denotes an interpolation of G on the edge σ that is :

$$\text{for } \sigma = K|L \in \mathcal{E}_{\text{int}}, \quad G_\sigma = \left| \begin{array}{ll} G_K & \text{if } \boldsymbol{u}_\sigma.\boldsymbol{n}_{K,\sigma} \geq 0, \\ G_L & \text{otherwise.} \end{array} \right.$$

For a face $\sigma \in \mathcal{E}_{\text{ext}}$ one simply take $G_\sigma = G_K$ so that

$$\nabla G \cdot \boldsymbol{n}_{K,\sigma} = \frac{|\sigma|}{|K|} (G_\sigma - G_K) = 0.$$

The expression of the discrete spatial operator (4.11) becomes

$$F_{\mathcal{M}}(G_{\mathcal{M}}^n) = \sum_{K\in\mathcal{M}} \left[\sum_{\sigma=K|L\in\mathcal{E}(K)} \kappa_{K,\sigma}^{\mathcal{M}} \frac{|\sigma|}{|K|} \frac{(\nabla_{\mathcal{E}} G^n)_\sigma}{|(\nabla_{\mathcal{E}} G^n)_\sigma|} \cdot \boldsymbol{n}_{K,\sigma} (G_\sigma^n - G_K^n) \right] \mathcal{X}_K, \tag{4.14}$$

where $\nabla_{\mathcal{E}}$ refers to a discrete gradient operator which is piecewise constant on every $D_\sigma, \sigma \in \mathcal{E}$.

4.3.1 Unstructured meshes

For $\sigma = K|L \in \mathcal{E}_{\text{int}}$, we take :

$$(\nabla_{\mathcal{E}} G)_\sigma = \sum_{\sigma\in\partial(K\cup L)} \frac{|\sigma|}{|K\cup L|} \tilde{G}_\sigma \boldsymbol{n}_{K\cup L,\sigma}, \tag{4.15}$$

with $\tilde{G}_\sigma$ defined as follows. Thanks to the definition of the discretization, there exists a set of neighbouring cells V_σ such that :

$$\exists (\alpha_{K,\sigma})_{K\in V_\sigma}, \quad \boldsymbol{x}_\sigma = \sum_{K\in V_\sigma} \alpha_{K,\sigma} \boldsymbol{x}_K \quad \text{and} \quad \sum_{K\in V_\sigma} \alpha_{K,\sigma} = 1.$$

We take the same combination for $\tilde{G}_\sigma$:

$$\tilde{G}_\sigma = \sum_{K\in V_\sigma} \alpha_{K,\sigma} G_K \tag{4.16}$$

4.3.2 Admissible mesh

The discrete gradient operator $\nabla_{\mathcal{E}}$, is splited over two components, one normal to the face and an other colinear to it :

$$\text{For } \sigma \in \mathcal{E}_{\text{int}}, \qquad (\nabla_{\mathcal{E}} G)_\sigma = \frac{G_L - G_K}{d_\sigma} \boldsymbol{n}_{K,\sigma} + \nabla_{//\sigma} G. \tag{4.17}$$

For the sake of readability we denote, for $\sigma = K|L \in \mathcal{E}_{\text{int}}$, by $\nabla_{K\cup L}$ the gradient defined on the unstructured mesh (4.15). Let us consider an orthonormal basis of σ, $\boldsymbol{n}_{//\sigma}$ in the 2D case and $(\boldsymbol{n}^1_{//\sigma}, \boldsymbol{n}^2_{//\sigma})$ in 3D. We take :

$$\begin{aligned} \nabla_{//\sigma} G &= \left(\nabla_{K\cup L} G \cdot \boldsymbol{n}_{//\sigma}\right) \boldsymbol{n}_{//\sigma} && \text{(2D)} \\ \nabla_{//\sigma} G &= \left(\nabla_{K\cup L} G \cdot \boldsymbol{n}^1_{//\sigma}\right) \boldsymbol{n}^1_{//\sigma} + \left(\nabla_{K\cup L} G \cdot \boldsymbol{n}^2_{//\sigma}\right) \boldsymbol{n}^2_{//\sigma} && \text{(3D)} \end{aligned} \tag{4.18}$$

4.3.3 Cartesian meshes

When the scheme is based on Cartesian grids, we have for $\sigma = \overrightarrow{K|L}$ (which means the flow goes from K to L) :

$$\text{For } \sigma \in \mathcal{E}_{\text{int}}, \qquad (\nabla_{\mathcal{E}} G)_\sigma = \left[\frac{G_L - G_K}{d_\sigma} \boldsymbol{n}_{K,\sigma} + \nabla_{//\sigma} G\right], \tag{4.19}$$

where $\nabla^C_{//\sigma}$ is defined by :

$$(\nabla G)^C_{//\sigma} = \sum_{i=1,\ \boldsymbol{e}^{(i)}\cdot \boldsymbol{n}_{K,\sigma}=0}^{d} \frac{(G_{K_i^+} - G_K)^+}{d_{\sigma_i^+}} - \left(1 - \text{sgn}(G_{K_i^+} - G_K)^+\right) \frac{(G_K - G_{K_i^-})^-}{d_{\sigma_i^-}} \boldsymbol{e}^{(i)}, \tag{4.20}$$

with $\sigma = \overrightarrow{K|L}$. For a cell $K \in \mathcal{M}$, σ_i^+ and σ_i^- stand for the two faces of K normal to $\boldsymbol{e}^{(i)}$. Superscripts $-$ and $+$ refer to the up and down faces of K respectively. We set $\sigma_i^+ = K|K_i^+$ and $\sigma_i^- = K|K_i^-$. We illustrate these notations in the following figure. We recall that $a^+ = \max(a, 0)$ and $a^- = \max(-a, 0)$, for $a \in \mathbb{R}$.

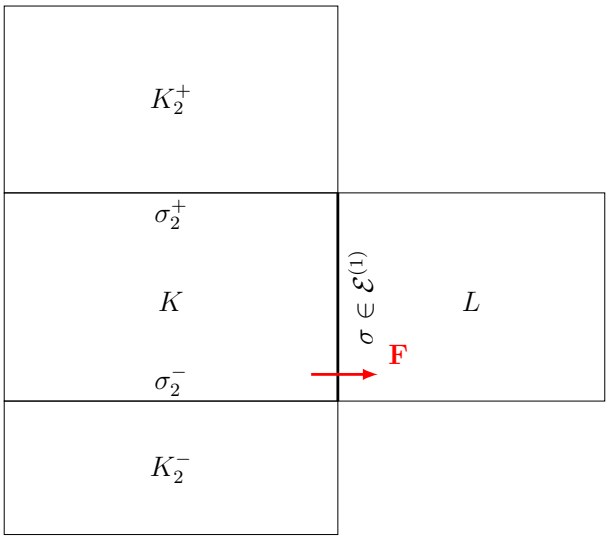


Figure 4.2 – Notations for the alternative gradient definition on Cartesian grids with $\mathbf{F} = (G_L - G_K)\boldsymbol{n}_{K,\sigma}$.

4.3.4 High order extension

It is possible to replace the Upwind interpolation by a higher order interpolation based on a MUSCL reconstruction. Adopting the same notations as in (4.13), its important property, based on [53] is stated below. For any $K \in \mathcal{M}$, and for any $\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\text{int}}$, there exists $\alpha_{K,\sigma} \in [0,1]$ such that :

$$G_\sigma - G_K = \begin{cases} \beta_{K,\sigma}(G_K - G_{M_\sigma^K}) & \text{if } \dfrac{\nabla_{\mathcal{E}} G_\sigma^n}{|\nabla_{\mathcal{E}} G_\sigma^n|} \cdot \boldsymbol{n}_{K,\sigma} \geq 0, \\ \beta_{K,\sigma}(G_{M_\sigma^K} - G_K) & \text{otherwise.} \end{cases} \tag{4.21}$$

The procedure is the following :

— We define a tentative value $\tilde{G}_\sigma$ based on the interpolation (4.16).

— The next step is to create a limitation procedure for ρ_σ and e_σ. Let $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = \overrightarrow{K|L}$ and V_K a set of neighbouring cells to K. We make the two following assumptions :

$$\begin{aligned} &(H1) \quad G_\sigma \in |[G_K, G_K + \tfrac{\zeta^+}{2}(G_L - G_K)]| \\ &(H2) \quad \text{there exists } M \in V_K \text{ such that } G_\sigma \in |[G_K, G_K + \tfrac{\zeta^-}{2}\tfrac{d_\sigma}{d_{K|M}}(G_K - G_M)]|, \end{aligned} \tag{4.22}$$

where, for $a, b \in \mathbb{R}$, we denote by $|[a,b]|$ the interval $\{\alpha a + (1-\alpha)\, b,\ \alpha \in [0,1]\}$ and $\overrightarrow{K|L}$ means that the flow is going from K to L ($\dfrac{\nabla_{\mathcal{E}} G_\sigma^n}{|\nabla_{\mathcal{E}} G_\sigma^n|}.\boldsymbol{n}_{K,\sigma} \geq 0$). The parameters ζ^+ and ζ^- lie in $[0,2]$.

— We compute G_σ as the nearest point to $\tilde{G}_\sigma$ in the limitation interval.

Whenever it is possible (*i.e.* with a mesh obtained by Q_1 mappings from the $(0,1)^d$ reference element), V_K may be chosen as the opposite cells to σ in K. Otherwise V_K is defined as the set of "upstream cells" to K. Note that, for a structured mesh, the first choice allows to recover the usual minmod limiter.

Remark 4.1 (*Cartesian grids*)

We impose $\zeta^+ = \zeta^- = 1$ for the Cartesian version of the scheme. This particular choice of parameters is the only one possible if we want to get consistency properties for the discrete spatial operator of the scheme.

4.4 Properties of the scheme

4.4.1 Stability

Thanks to the definition of the discrete convective operator, we have the following property :

Lemma 4.1 (*Maximum principle*)

Let $G_{\mathcal{M}}^n \in H_{\mathcal{M}}$, $n \in [0,N]$, be the solution of the scheme (4.9). For all $K \in \mathcal{M}$ and $n \in [0, N-1]$, we have :

$$\min_{L \in \mathcal{M}} G_L^n \leq G_K^{n+1} \leq \max_{L \in \mathcal{M}} G_L^n,$$

under the CFL condition :

$$\delta t \leq \min_{K \in \mathcal{M}} \frac{|K|}{\displaystyle\sum_{\sigma \in \mathcal{E}(K)} |\sigma|} \tag{4.23}$$

Proof : We have, for $K \in \mathcal{M}$ and $n \in [0, N-1]$:

$$G_K^{n+1} = \left(1 - \delta t \sum_{\sigma \in \mathcal{E}(K)} \frac{|\sigma|}{|K|} \left(\frac{\nabla_{\mathcal{E}} G_\sigma^n}{|\nabla_{\mathcal{E}} G_\sigma^n|} \cdot \boldsymbol{n}_{K,\sigma}\right)^-\right) G_K^n + \delta t \sum_{\sigma = K|L \in \mathcal{E}(K)} \frac{|\sigma|}{|K|} \left(\frac{\nabla_{\mathcal{E}} G_\sigma^n}{|\nabla_{\mathcal{E}} G_\sigma^n|} \cdot \boldsymbol{n}_{K,\sigma}\right)^- G_L^n.$$

Consequently, G_K^{n+1} is a convex combination of its neighbours at time n if (4.23) is verified, which completes the proof. ■

Remark 4.2 (*Cartesian grids*)

The property remains the same with the scheme on Cartesian grids, only the CFL is modified. One must replace $|K|$ by $d|D_\sigma|$ in (4.23).

Remark 4.3 (*MUSCL interpolation*)

Concerning the MUSCL interpolation, we use the property (4.21) and use it in the scheme to get :

$$G_K^{n+1} = \left(1 - \delta t \sum_{\sigma\in\mathcal{E}(K)} \frac{|\sigma|}{|K|}\beta_{K,\sigma}\left|\frac{\nabla_{\mathcal{E}} G_\sigma^n}{|\nabla_{\mathcal{E}} G_\sigma^n|}\cdot \boldsymbol{n}_{K,\sigma}\right|\right) G_K^n + \delta t \sum_{\sigma\in\mathcal{E}(K)} \frac{|\sigma|}{|K|}\beta_{K,\sigma}\left|\frac{\nabla_{\mathcal{E}} G_\sigma^n}{|\nabla_{\mathcal{E}} G_\sigma^n|}\cdot \boldsymbol{n}_{K,\sigma}\right| G_{M_K^\sigma}^n.$$

The maximum principle is still satisfied with the same CFL condition.

4.4.2 Consistency

We need to define interpolates of test functions on the mesh. Let $\phi \in C_c^\infty(\Omega)$. We set :

$$\phi_{\mathcal{M}} = \sum_{K\in\mathcal{M}} \phi_K \mathcal{X}_K, \qquad \phi_K = \phi(\boldsymbol{x}_K). \tag{4.24}$$

We give the definition of the consistency property.

Definition 4.2 (*Consistency*)

Let $F(G)$ be an operator approximated by $F_{\mathcal{M}}(G_{\mathcal{M}})$. Let $h_{\mathcal{M}} = \max_{K\in\mathcal{M}} \operatorname{diam}(K)$. Let $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}\right\}$ be a sequence of discretizations such that the size $h_{\mathcal{M}}^{(m)}$ tends to zero as $m \to \infty$. The discrete spatial operator $H_{\mathcal{M}}$ is said to be strongly consistent with H if for every $\phi \in C_c^\infty(\Omega)$:

$$\lim_{m\to\infty} \|F_{\mathcal{M}^{(m)}}(\phi_{\mathcal{M}^{(m)}}) - F(\phi)\|_{L^\infty(\Omega)} = 0.$$

Lemma 4.2 (*Strong consistency of the discrete gradient*)

For $\phi \in H_{\mathcal{M}}$, the discrete gradient operator $\nabla_{\mathcal{E}}\phi$ defined by :

$$\nabla_{\mathcal{E}}\phi = \sum_{\sigma\in\mathcal{E}} (\nabla_{\mathcal{E}}\phi)_\sigma \mathcal{X}_{D_\sigma},$$

is strongly consistent.

Proof : We focus on proving this lemma for admissible meshes as the strong consistency of $\nabla_{K\cup L} G$ (gradient for unstructured meshes) is required. Let $\phi \in C_c^\infty(\Omega)$ and $\phi_{\mathcal{M}} \in H_{\mathcal{M}}$ its natural interpolation on the mesh. We recall that :

$$\phi_{\mathcal{M}} = \sum_{K\in\mathcal{M}} \underbrace{\phi(\boldsymbol{x}_K)}_{\phi_K} \mathcal{X}_K.$$

For any internal edge $\sigma = K|L$, the gradient has two components. We first analyze the component normal to the face :

$$(\nabla_{\mathcal{E}}\phi_{\mathcal{M}})_{\perp\sigma} = \frac{1}{d_\sigma}(\phi_L - \phi_K)\boldsymbol{n}_{K,\sigma}.$$

We then do a Taylor expansion to the first order.

$$(\nabla_{\mathcal{E}}\phi_{\mathcal{M}})_{\perp\sigma} = \Big(\nabla\phi(\boldsymbol{x}_\sigma)\cdot \boldsymbol{n}_{K,\sigma}\Big)\boldsymbol{n}_{K,\sigma} + O(h_{\mathcal{M}}).$$

We focus on the other component. For the sake of simplicity we limit ourselves to the 2D case. We consider the following gradient operator :

$$\nabla'_{K\cup L}\phi = \sum_{\sigma\in\partial K\cup L}\frac{|\sigma|}{|K\cup L|}\left\{\frac{1}{\sigma}\int_\sigma \phi\right\}\boldsymbol{n}_{K\cup L,\sigma}.$$

Using the divergence theorem, we get that :

$$\nabla'_{K\cup L}\phi = \frac{1}{|K\cup L|}\int_{K\cup L}\nabla\phi = \nabla\phi(\boldsymbol{x}_\sigma) + O(h_{\mathcal{M}}),$$

thanks to the regularity of ϕ. Furthermore, using (4.15) and (4.16) we get that :

$$\nabla_{K\cup L}\phi - \nabla'_{K\cup L}\phi = \frac{1}{|K\cup L|}\sum_{\sigma\in\partial K\cup L}\int_\sigma\sum_{K\in V_\sigma}\alpha_{K,\sigma}(\phi_K - \phi(\boldsymbol{x}))d\boldsymbol{x}.$$

A Taylor expansion leads to :

$$\nabla_{K\cup L}\phi - \nabla'_{K\cup L}\phi = \frac{1}{|K\cup L|}\left\{\sum_{\sigma\in\partial K\cup L}\int_\sigma\sum_{K\in V_\sigma}\alpha_{K,\sigma}\nabla\phi(\boldsymbol{x}_K)\cdot(\boldsymbol{x}_K - \boldsymbol{x})d\boldsymbol{x}\right\} + O(h_{\mathcal{M}}).$$

Another Taylor expansion reads :

$$\nabla_{K\cup L}\phi - \nabla'_{K\cup L}\phi = \frac{1}{|K\cup L|}\Big\{\sum_{\sigma\in\partial K\cup L}\Big[\nabla\phi(\boldsymbol{x}_\sigma) + D^2\phi(\boldsymbol{x}_\sigma)(\boldsymbol{x}_K - \boldsymbol{x}_\sigma)\Big]\cdot \int_\sigma\sum_{K\in V_\sigma}\alpha_{K,\sigma}(\boldsymbol{x}_K - \boldsymbol{x})d\boldsymbol{x}\Big\} + O(h_{\mathcal{M}}).$$

$$\int_\sigma\sum_{K\in V_\sigma}\alpha_{K,\sigma}(\boldsymbol{x}_K - \boldsymbol{x})d\boldsymbol{x} = \int_\sigma(\boldsymbol{x}_\sigma - \boldsymbol{x})d\boldsymbol{x}.$$

By definition of the mass center the last integral is equal to zero. Consequently,

$$\nabla_{K\cup L}\phi - \nabla'_{K\cup L}\phi = O(h_{\mathcal{M}})$$

Finally, gathering the results for the two components, we get that :

$$(\nabla_{\mathcal{E}}\phi)_\sigma = \Big(\nabla\phi(\boldsymbol{x}_\sigma)\cdot\boldsymbol{n}_{K,\sigma}\Big)\boldsymbol{n}_{K,\sigma} + \Big(\nabla\phi(\boldsymbol{x}_\sigma)\cdot\boldsymbol{n}_{//\sigma}\Big)\boldsymbol{n}_{//\sigma} + O(h_{\mathcal{M}}).$$

$(\boldsymbol{n}_{K,\sigma},\ \boldsymbol{n}_{//\sigma})$ is an orthonormal basis of $\mathbb{R}^2$, so :

$$(\nabla_{\mathcal{E}}\phi)_\sigma = \nabla\phi(\boldsymbol{x}_\sigma) + O(h_{\mathcal{M}}).$$

which concludes the proof. ■

A stronger result can be obtained with the scheme on Cartesian grids. It is pointed out in the next proposition

Proposition 4.3

The spatial operator in the Cartesian case, given by, for $G_{\mathcal{M}} \in H_{\mathcal{M}}$:

$$F_{\mathcal{M}}(G_{\mathcal{M}}) = \sum_{K\in\mathcal{M}}\left[\sum_{\sigma=K|L\in\mathcal{E}(K)}\frac{|\sigma|}{d|D_\sigma|}\frac{(G_L - G_K)}{\sqrt{(G_L - G_K)^2 + d_\sigma^2|\nabla_{//\sigma}G_{\mathcal{M}}|^2}}(G_\sigma - G_K)\right]\chi_K, \tag{4.25}$$

is strongly consistent with $|\nabla G|$.

Proof : Let $\phi \in C_c^\infty(\Omega)$ and $\phi_\mathcal{M} \in H_\mathcal{M}$ its interpolation on the mesh. Consider $K \in \mathcal{M}$ and $\boldsymbol{v}$ a constant vector. Let $\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v})$ be :

$$\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v}) = \sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{|\sigma|}{d|D_\sigma|}(\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma})(\phi_\sigma - \phi_K).$$

With the upwind interpolation, we get that :

$$\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v}) = -\sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma}(\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma})^-(\phi_L - \phi_K).$$

A simple Taylor expansion leads to :

$$\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v}) = -\sum_{\sigma=K|L\in\mathcal{E}(K)} (\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma})^-\nabla\phi(\boldsymbol{x}_K).\boldsymbol{n}_{K,\sigma} + O(h_\mathcal{M}),$$

so

$$\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v}) = \nabla\phi(\boldsymbol{x}_K)\cdot\sum_{\sigma=K|L\in\mathcal{E}(K)} (\boldsymbol{v}\cdot\boldsymbol{n}_{L,\sigma})^+\boldsymbol{n}_{L,\sigma} + O(h_\mathcal{M}).$$

Thanks to the Cartesian grid, we have :

$$\sum_{\sigma=K|L\in\mathcal{E}(K)} (\boldsymbol{v}\cdot\boldsymbol{n}_{L,\sigma})^+\boldsymbol{n}_{L,\sigma} = \sum_{i=1}^{d}(\boldsymbol{v}\cdot\boldsymbol{e}^{(i)})\boldsymbol{e}^{(i)} = \boldsymbol{v},$$

so we have :

$$\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v}) = \boldsymbol{v}\cdot\nabla\phi(\boldsymbol{x}_K) + O(h_\mathcal{M}).$$

Concerning the MUSCL interpolation, we have :

$$\begin{aligned}\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v}) = \frac{1}{2}\sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma}(\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma})^+ \min\left(\phi_K - \phi_{M_K^\sigma}\frac{d_\sigma}{d_{K|M_K^\sigma}}, \phi_L - \phi_K\right)\\ - \sum_{\sigma=K|L\in\mathcal{E}(K)} (\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma})^-(\phi_L - \phi_K)\\ - \frac{1}{2}\sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma}(\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma})^- \min\left(\phi_L - \phi_{M_L^\sigma}\frac{d_\sigma}{d_{L|M_L^\sigma}}, \phi_K - \phi_L\right),\end{aligned}$$

where M_K^σ refers to the opposite cell to σ in K. It is easy to see that :

$$\frac{1}{d_\sigma}\min\left(\phi_K - \phi_{M_K^\sigma}\frac{d_\sigma}{d_{K|M_K^\sigma}}, \phi_L - \phi_K\right) = \nabla\phi(\boldsymbol{x}_K)\cdot\boldsymbol{n}_{K,\sigma} + O(h_\mathcal{M}),$$

and,

$$\frac{1}{d_\sigma}\min\left(\phi_L - \phi_{M_L^\sigma}\frac{d_\sigma}{d_{L|M_L^\sigma}}, \phi_K - \phi_L\right) = \nabla\phi(\boldsymbol{x}_K)\cdot\boldsymbol{n}_{L,\sigma} + O(h_\mathcal{M}).$$

Therefore,

$$\begin{aligned}\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v}) = \frac{1}{2}\sum_{\sigma\in\mathcal{E}(K)} (\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma})^+\nabla\phi(\boldsymbol{x}_K)\cdot\boldsymbol{n}_{K,\sigma} +\\ \frac{1}{2}\sum_{\sigma\in\mathcal{E}(K)} (\boldsymbol{v}\cdot\boldsymbol{n}_{L,\sigma})^+\nabla\phi(\boldsymbol{x}_K)\cdot\boldsymbol{n}_{L,\sigma} + O(h_\mathcal{M}),\end{aligned}$$

which leads to :

$$\begin{aligned}\tilde{F}_K(\phi_\mathcal{M}, \boldsymbol{v}) &= \nabla\phi(\boldsymbol{x}_K)\cdot\sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{2}\left((\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma})^+\boldsymbol{n}_{K,\sigma} + (\boldsymbol{v}\cdot\boldsymbol{n}_{L,\sigma})^+\boldsymbol{n}_{L,\sigma}\right) + O(h_\mathcal{M})\\ &= \nabla\phi(\boldsymbol{x}_K)\cdot\boldsymbol{v} + O(h_\mathcal{M}).\end{aligned}$$

Noticing, thanks to the strong consistency of $\nabla_{\mathcal{E}}$, that :

$$F_{\mathcal{M}}(\phi_{\mathcal{M}}) = \sum_{K\in\mathcal{M}} \tilde{F}_K\left(\phi_{\mathcal{M}}, \frac{\nabla\phi(\boldsymbol{x}_K)}{|\nabla\phi(\boldsymbol{x}_K)|}\right)\chi_K + O(h_{\mathcal{M}}),$$

we can conclude that :

$$\lim_{m\to\infty} F_{\mathcal{M}}(\phi_{\mathcal{M}}) = |\nabla\phi|,$$

which concludes the proof. ■

4.4.3 Invariance under translation

Let us formulate the scheme as follows :

$$\forall n \in [0, N-1], \qquad G_{\mathcal{M}}^{n+1} = SCH(G_{\mathcal{M}}^n), \tag{4.26}$$

where

$$SCH(G_{\mathcal{M}}^n) = G_{\mathcal{M}}^n - \delta t F_{\mathcal{M}}(G_{\mathcal{M}}^n).$$

The scheme satisfies the following property :

Proposition 4.4 (*Invariance under Translation with constants*)

$\forall\lambda \in \mathbb{R}$, and $\forall\phi_{\mathcal{M}} \in H_{\mathcal{M}}$,

$$SCH(\phi_{\mathcal{M}} + \lambda) = \lambda + SCH(\phi_{\mathcal{M}}) \tag{4.27}$$

Proof : Let $\lambda \in \mathbb{R}$ and $\phi_{\mathcal{M}} \in H_{\mathcal{M}}$. Thanks to the definition of SCH, we need to prove that :

$$F_{\mathcal{M}}(\phi_{\mathcal{M}} + \lambda) = F_{\mathcal{M}}(\phi_{\mathcal{M}}).$$

Looking at (4.14) and (4.18), we need to check that $\nabla_{K\cup L}\left(\phi_{\mathcal{M}} + \lambda\right) = \nabla_{K\cup L}\phi_{\mathcal{M}}$. We remind that :

$$\nabla_{K\cup L}\left(\phi_{\mathcal{M}} + \lambda\right) = \sum_{\sigma\in\partial K\cup L} \frac{|\sigma|}{|K\cup L|}\left(\phi_\sigma + \lambda\right)\boldsymbol{n}_{K\cup L,\sigma}$$

We have :

$$\nabla_{K\cup L}\left(\phi_{\mathcal{M}} + \lambda\right) = \nabla_{K\cup L}\phi_{\mathcal{M}} + \lambda \sum_{\sigma\in\partial K\cup L} \frac{|\sigma|}{|K\cup L|}\boldsymbol{n}_{K\cup L,\sigma}.$$

Using the divergence theorem, we get that :

$$\sum_{\sigma\in\partial K\cup L} \frac{|\sigma|}{|K\cup L|}\boldsymbol{n}_{K\cup L,\sigma} = \int_{K\cup L} \nabla(1) = 0,$$

which concludes the proof. ■

The fact that the Cartesian version of the scheme satisfy the same property is immediate. The remaining property of the scheme is only valid for the Cartesian scheme.

4.4.4 Monotonicity

Let $(\phi_{\mathcal{M}}, \psi_{\mathcal{M}}) \in H_{\mathcal{M}}$. Let us define the following partial order

$$\phi_{\mathcal{M}} \le \psi_{\mathcal{M}} \qquad \Longleftrightarrow \qquad \forall K \in \mathcal{M}, \quad \phi_K \le \psi_K. \tag{4.28}$$

Then we get the following result with the Cartesian upwind scheme only.

Proposition 4.5 (*Monotonicity of the Upwind Cartesian scheme*)

Suppose that the following CFL condition is satisfied

$$\delta t \leq \frac{1}{\sum_{\sigma\in\mathcal{E}(K)} \frac{1+\frac{1}{2}\sqrt{1+r^2}}{d_\sigma}}, \qquad r = \max_{(\sigma,\sigma')\in\mathcal{E}(K)} \frac{d_\sigma}{d_{\sigma'}}. \tag{4.29}$$

Then he have the following result :

$$\forall(\phi_\mathcal{M},\psi_\mathcal{M}) \in H_\mathcal{M}, \qquad \phi_\mathcal{M} \leq \psi_\mathcal{M} \quad \Longrightarrow \quad F_\mathcal{M}(\phi_\mathcal{M}) \leq F_\mathcal{M}(\psi_\mathcal{M}).$$

Proof : For the sake of clarity we prove the result in 2D. The extension to all dimension can be done at the cost of heavier notations and CFL conditions. We can equivalently check that SCH is a non decreasing function of each variable. Let $K \in \mathcal{M}$ and $\phi_\mathcal{M} \in H_\mathcal{M}$. We have :

$$SCH(\phi_\mathcal{M})|_K = \phi_K + \delta t \sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma} f_{K,\sigma}\left(\phi_\mathcal{M}\right),$$

with,

$$f_{K,\sigma}\left(\phi_\mathcal{M}\right) = \frac{(\phi_L-\phi_K)^-}{\sqrt{(\phi_L-\phi_K)^2 + d_\sigma^2|\nabla_{//\sigma}\phi_\mathcal{M}|^2}}(\phi_L-\phi_K)$$

The monotonicity of $f_{K,\sigma}$ in ϕ_L is equivalent to the monotonicity of the function :

$$f : x \longmapsto \frac{x^- x}{|x|} = -x^-, \quad \forall x \in \mathbb{R}$$

because $\nabla_{//\sigma}\phi_\mathcal{M}$ does not depend on ϕ_L in the Cartesian case (see (4.20)). We can conclude that $f_{K,\sigma}$ is a non decreasing function of ϕ_L. Concerning the monotonicity in ϕ_{K^-} and ϕ_{K^+} it is equivalent to the variations of :

$$f : x \longmapsto -\frac{1}{x^+},$$

which is a non decreasing function. We can conclude that $SCH(\phi_\mathcal{M})|_K$ is an increasing function of each $\left(\phi_M\right)_{\substack{M\in\mathcal{M}\\ M\neq K}}$. Concerning ϕ_K, we have :

$$SCH(\phi_\mathcal{M})|_K = g(\phi_K) = \phi_K - \delta t \sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma} \frac{(\phi_K-\phi_L)^+}{\sqrt{(\phi_L-\phi_K)^2 + d_\sigma^2|\nabla_{//\sigma}\phi_\mathcal{M}|^2}}(\phi_K-\phi_L)$$

The analysis of this function can be splitted into three cases. If, $\forall\sigma \in \mathcal{E}(K)$, $\phi_K \leq \phi_L$, then $g(\phi) = \phi_K$ which is non decreasing. The second case is when, $\forall\sigma \in \mathcal{E}(K)$, $\phi_K \geq \max(\phi_{K^+}, \phi_{K^-}, \phi_L)$. We have :

$$g(\phi_K) = \phi_K - \sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{\delta t}{d_\sigma}\left(\phi_K-\phi_L\right).$$

which is non decreasing if,

$$\delta t \leq \frac{1}{\sum_{\sigma\in\mathcal{E}(K)} d_\sigma^{-1}}.$$

Finally, suppose that $\forall\sigma \in \mathcal{E}(K)$, $\phi_L \leq \phi_K \leq \phi_{K^+}$ (or ϕ_{K^-}), we have, denoting by $r_\sigma = \frac{d_\sigma}{d_{\sigma^+}}$:

$$g(\phi_K) = \phi_K - \sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma} \frac{\phi_K-\phi_L}{\sqrt{(\phi_K-\phi_L)^2 + r_\sigma^2(\phi_K-\phi_{K^+})^2}}(\phi_K-\phi_L).$$

Let us derive this function :

$$g'(\phi_K) = 1 - \sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma} \frac{\phi_K-\phi_L}{\sqrt{(\phi_K-\phi_L)^2 + r_\sigma^2(\phi_K-\phi_{K^+})^2}}$$
$$- \sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma} \frac{r_\sigma^2(\phi_{K^+}-\phi_K)(\phi_K-\phi_L)(\phi_{K^+}-\phi_K)}{\left((\phi_K-\phi_L)^2 + r_\sigma^2(\phi_K-\phi_{K^+})^2\right)^{3/2}}$$

One can notice directly that :

$$\sum_{\sigma=K|L\in\mathcal{E}(K)} \frac{1}{d_\sigma} \frac{\phi_K - \phi_L}{\sqrt{(\phi_K-\phi_L)^2 + r_\sigma^2(\phi_K-\phi_{K^+})^2}} \leq 1.$$

In order to upper-bound the second sum, we analyze the function

$$h : x \longmapsto \frac{r^2 x(a-x)a}{(x^2+d^2(a-x)^2)^{3/2}},$$

where a, r are strictly positive constants. We split the function in two parts $h(x) = h_1(x)h_2(x)$ with :

$$\begin{aligned} h_1(x) =& \frac{r^2 x(a-x)}{x^2+r^2(a-x)^2}, \\ h_2(x) =& \frac{a}{\sqrt{x^2+r^2(a-x)^2}}. \end{aligned}$$

Concerning h_1 we can equivalently consider the function defined on $\mathbb{R}^+$ by :

$$y \longmapsto \frac{r^2}{y+\frac{r^2}{y}} = \frac{r^2 y}{y^2+r^2}.$$

A quick study of the function shows that,

$$\max_{y\in\mathbb{R}^+} \frac{r^2 y}{y^2+r^2} = \frac{r}{2} = \max_{x\in[0,a]} h_1(x).$$

The same work is performed with h_2 and leads to :

$$\max_{x\in[0,a]} h_2(x) = \frac{\sqrt{1+r^2}}{r}$$

Gathering the results, we get that :

$$\forall x \in [0,a], \quad h(x) \leq \frac{1}{2}\sqrt{1+r^2}$$

As a result, writing out $r = \max_{(\sigma,\sigma')\in\mathcal{E}(K)} \frac{d_\sigma}{d_{\sigma'}}$, we get that $g'(\phi_K) \geq 0$ provided that (4.29) is satisfied. This CFL condition ensures that $SCH(\phi_\mathcal{M})|_K$ is a non decreasing function of ϕ_K, which concludes the proof. ■

Remark 4.4 (*Discrete weak form*)

The combination of the monotonicity and the invariance under translation leads to a property which is a discrete counterpart of (4.4) (and (4.5) respectively). Let $G^{(T)}_{\mathcal{M}^{(m)}} = \sum_{n=0}^{N-1} G^{n+1}_{\mathcal{M}^{(m)}} \mathcal{X}_{[t^n,t^{n+1}]}$ be the solution of the scheme

$$G^{n+1}_\mathcal{M} = SCH(G^n_\mathcal{M}), \quad \forall n \in [0, N-1],$$

which we suppose to be monotonous and invariant under translations. We introduce the space $L_\mathcal{M}(\Omega\times[0,T))$ of functions constant on every $K\times[t^n,t^{n+1})$ in $\mathcal{M}\times[0,T)$. Let $\varphi \in L_\mathcal{M}(\Omega\times[0,T))$ such that $G^{(T)}_{\mathcal{M}^{(m)}} - \varphi$ has a local maximum at (K_0, t^{n_0}). We have obviously, $\forall K \in \mathcal{M}$:

$$G^{n_0}_{K_0} - \varphi^{n_0}_{K_0} \geq G^{n_0-1}_K - \varphi^{n_0-1}_K,$$

so

$$G^{n_0-1}_K \leq G^{n_0}_{K_0} - \varphi^{n_0}_{K_0} + \varphi^{n_0-1}_K.$$

Thanks to the monotonicity of the scheme we have

$$SCH(G_K^{n_0-1}) \leq SCH(G_{K_0}^{n_0} - \varphi_{K_0}^{n_0} + \varphi_K^{n_0-1}).$$

We have $SCH(G_K^{n_0-1}) = G_K^{n_0}$ and $G_{K_0}^{n_0} - \varphi_{K_0}^{n_0}$ is a constant so the invariance under translation of the scheme leads to

$$G_K^{n_0} \leq SCH(\varphi_K^{n_0-1}) + G_{K_0}^{n_0} - \varphi_{K_0}^{n_0}.$$

Taking the component K_0 induces :

$$\varphi_{K_0}^{n_0} \leq SCH(\varphi_K^{n_0-1})_{K_0}.$$

In other words,

$$\eth t \varphi_{K_0}^{n_0-1} + F_{\mathcal{M}}(\varphi^{n_0-1})_{K_0} \leq 0,$$

which is clearly a discrete counterpart of (4.4). The result is similar with a local minimum.

Remark 4.5

All the results proved here are still valid using a convective velocity and a front propagation speed. It only changes the different CFL conditions. However the monotonicity results cannot be extended to the MUSCL interpolation, and more generally to the non Cartesian case.

4.5 A convergence result

This section states the main theoretical result of this paper, namely the uniform convergence of the Upwind scheme on Cartesian grids only. We first recall the convergence theorem developped in [6], adapted to our scheme and notations.

Theorem 4.6

Let $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}, \delta t^{(m)}\right\}$ be a sequence of discretizations such that the space and time steps tend to zero as $m \to \infty$. Let $\bar{G}$ be the viscosity solution of (4.6). Consider the following explicit scheme, for $n \in [0, N-1]$:

$$\eth t G_m^n + F_{\mathcal{M}}(G_m^n) = 0,$$

and the complete solution defined by $G_m^{(T)} = \sum_{n=0}^{N-1} G_m^{n+1} \mathcal{X}_{[t^n, t^{n+1}]}$. We suppose that :

- *The spatial operator $F_{\mathcal{M}}$ is strongly consistent with the continuous operator $G \longmapsto |\nabla G|$.*
- *The scheme is invariant under translations : $F_{\mathcal{M}}(G_{\mathcal{M}} + v) = F_{\mathcal{M}}(G_{\mathcal{M}})$.*
- *The scheme is monotone.*

Then,

$$G_{\mathcal{M}^{(m)}} \longrightarrow \bar{G} \text{ uniformly as } m \to \infty.$$

Since we have shown the required properties in Theorem (4.6), we can thus conclude to the convergence of the scheme, which we state in the following corollary.

Corollary 4.7

Let $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}, \delta t^{(m)}\right\}$ be a sequence of discretizations such that the space and time steps tend to zero as $m \to \infty$. Now suppose there exists $r > 0$, such that $\forall m \in \mathbb{N}$,

$\forall\,(\sigma,\sigma')\in\mathcal{E}^{(m)}$,

$$\frac{d_\sigma}{d_{\sigma'}}\leq r.$$

Suppose that, for any $m\in\mathbb{N}$,

$$\delta t^{(m)}\leq\max_{K\in\mathcal{M}^{(m)}}\frac{1}{\sum_{\sigma\in\mathcal{E}(K)}\frac{1+\frac{1}{2}\sqrt{1+r^2}}{d_\sigma}}.$$

Then the solution of the upwind Cartesian scheme (4.9)-(4.25) $G_m^{(T)}$ *converges uniformly towards* $\bar{G}$.

4.6 Numerical results

4.6.1 One dimension

The domain is $\Omega=(0,1)$. We use homegenous Neumann boundary conditions in $x=0$ and $x=1$. We suppose that the time and space steps are constant for simplicity. Consider the following initial data :

$$G_0(x)=|\sin(4\pi x)| \tag{4.30}$$

We give the solution at $T=0.05s$, with an upwind interpolation for the spatial operator, and a fixed CFL equal to 1/10.

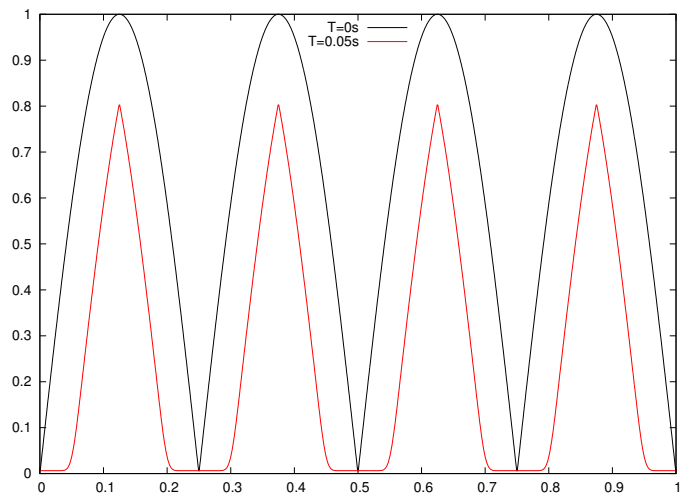


Figure 4.3 – Solution of the G-equation with the upwind scheme at $T=0.05s$.

It is possible to determine in 1D the unique viscosity solution for a given initial data, provided the variations of the initial data are known. With this initial data, provided that $T\leq\frac{1}{8}$,

the viscosity solution is :

$$G_{\text{visc}}(r,\theta,T) = \begin{cases} 0, & \forall r \in [0,T] \cup [\frac{1}{4}-T, \frac{1}{4}+T] \cup [\frac{1}{2}-T, \frac{1}{2}+T] \cup [\frac{3}{4}-T, \frac{3}{4}+T] \cup [1-T, 1], \\ |\sin(4\pi(r-T))|, & \forall r \in [T, \frac{1}{8}] \cup [\frac{1}{4}+T, \frac{3}{8}] \cup [\frac{1}{2}+T, \frac{5}{8}] \cup [\frac{3}{4}+T, \frac{7}{8}], \\ |\sin(4\pi(r+T))|, & \forall r \in [\frac{1}{8}, \frac{1}{4}-T] \cup [\frac{3}{8}, \frac{1}{2}-T] \cup [\frac{5}{8}, \frac{3}{4}-T] \cup [\frac{7}{8}, 1-T]. \end{cases} \tag{4.31}$$

The proof can be found in the appendix. Consequently we can highlight numerically the theoritical result about the convergence of the solution of our scheme towards the viscosity solution. The figure below gives the error in L^1 norm according to the space step, for a fixed CFL equal to $\frac{1}{10}$.

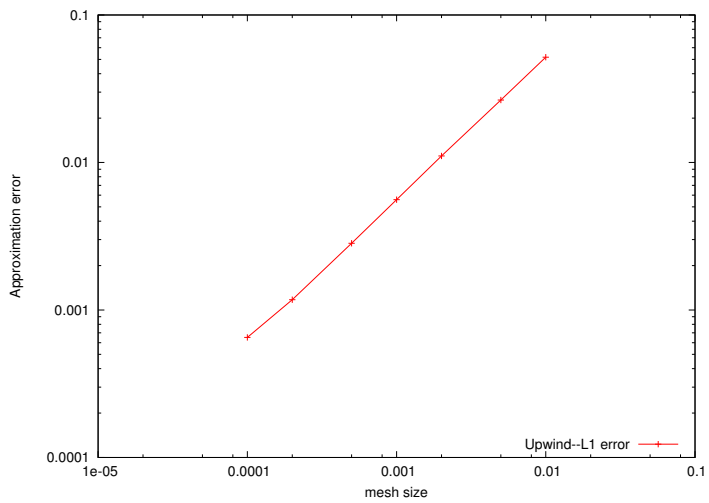


Figure 4.4 – L1 norm error at T=0.05s and CFL= $\frac{1}{10}$ – Upwind interpolation.

We can also see the behaviour of the scheme if we use discontinuous initial data. We consider the following :

$$G_0(x) = \begin{cases} 0, & \text{if } x \leq 0.5 \\ 1, & \text{otherwise.} \end{cases}$$

The result at time $T = 0.2s$ is given below, for the upwind scheme and the MUSCL scheme.

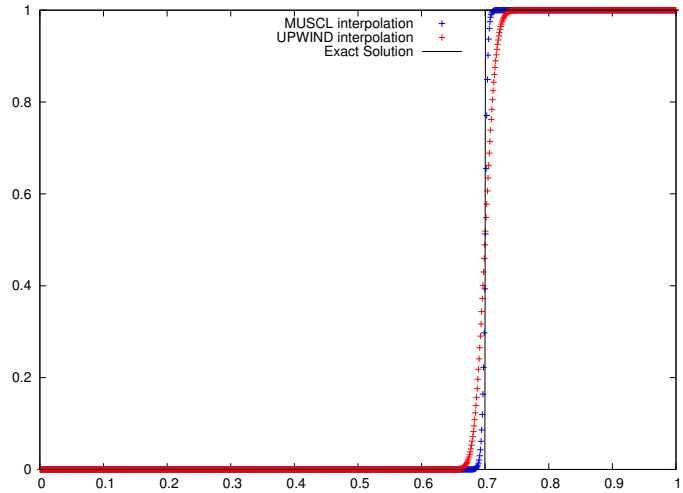


Figure 4.5 – solution at $T = 0.05s$ and CFL= $\frac{1}{10}$ with $h = 10^{-3}$.

The MUSCL scheme brings less numerical diffusion, as expected. Normally one can not define a viscosity solution for discontinuous initial data. However one expects the solution to be the same as the general viscosity solution given for BUC initial data (see (B.2) in the Appendix).

4.6.2 Two Dimensions

Unstructured grid

The computational domain is $\Omega = [-\frac{1}{2}, \frac{1}{2}]^2$. The mesh consists in convex quadrilaterals. We give an example of the discretization below. These grids are built from a regular Cartesian grid

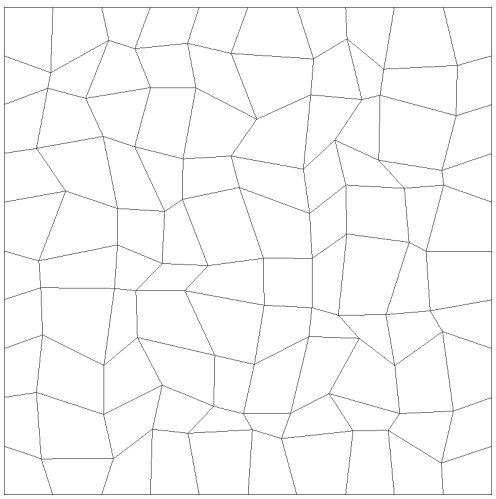


Figure 4.6 – Example of a 10×10 unstructured grid

for which a random displacement of length ϵh is applied to each node where h is the space step. We consider homogeneous Neumann conditions at the boundaries. The initial data are given in the polar coordinates (r, θ) :

$$G_0(r, \theta) = r\left(1 + \frac{1}{2}\cos(4\theta)\right).$$

Results obtained at different times are given below. The scheme used is the UPWIND version for unstructured meshes, with a space step $h = \dfrac{1}{200}$, and a constant CFL equal to $\frac{1}{10}$.

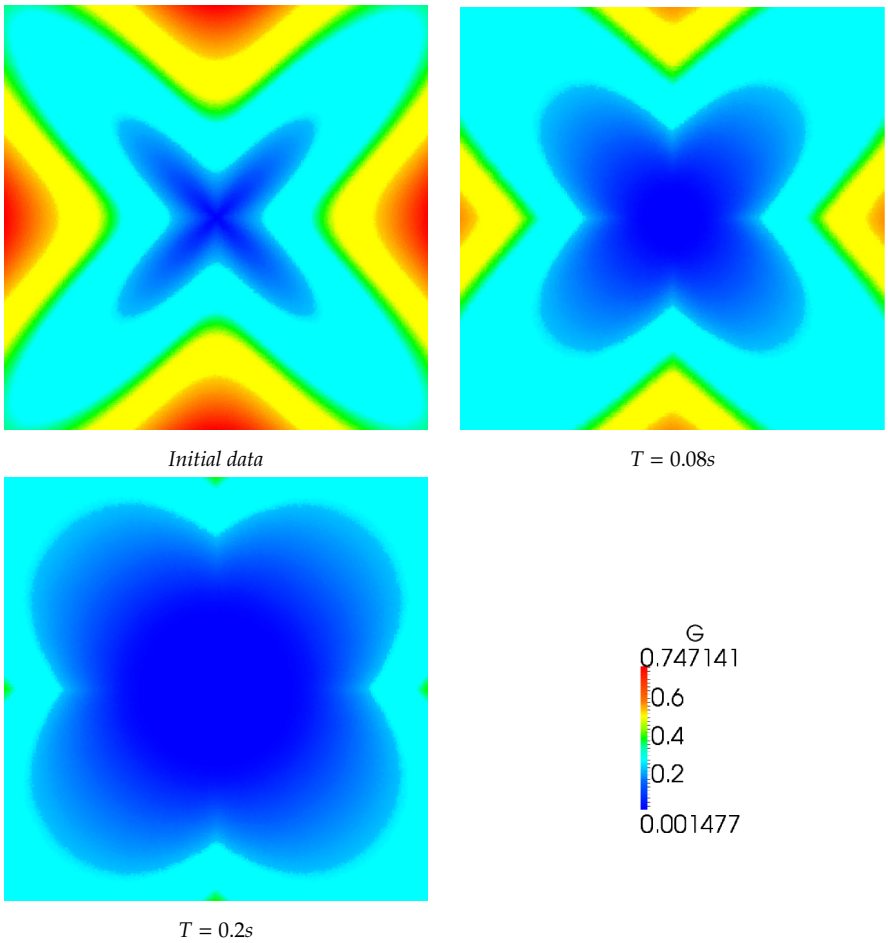


Figure 4.7 – G at different times with the upwind scheme on an unstructured mesh – $h = \dfrac{1}{200}$ – $CFL = \dfrac{1}{10}$

An other possible test case is the following one :

$$G_0(r, \theta) = |\sin(4\pi r)|. \tag{4.32}$$

Results obtained with different meshes are displayed just below. The scheme used is the UPWIND version for unstructured meshes, with a space step $h = \dfrac{1}{400}$, a constant CFL equal to $\frac{1}{10}$ and a final time equal to $T = 0.04s$.

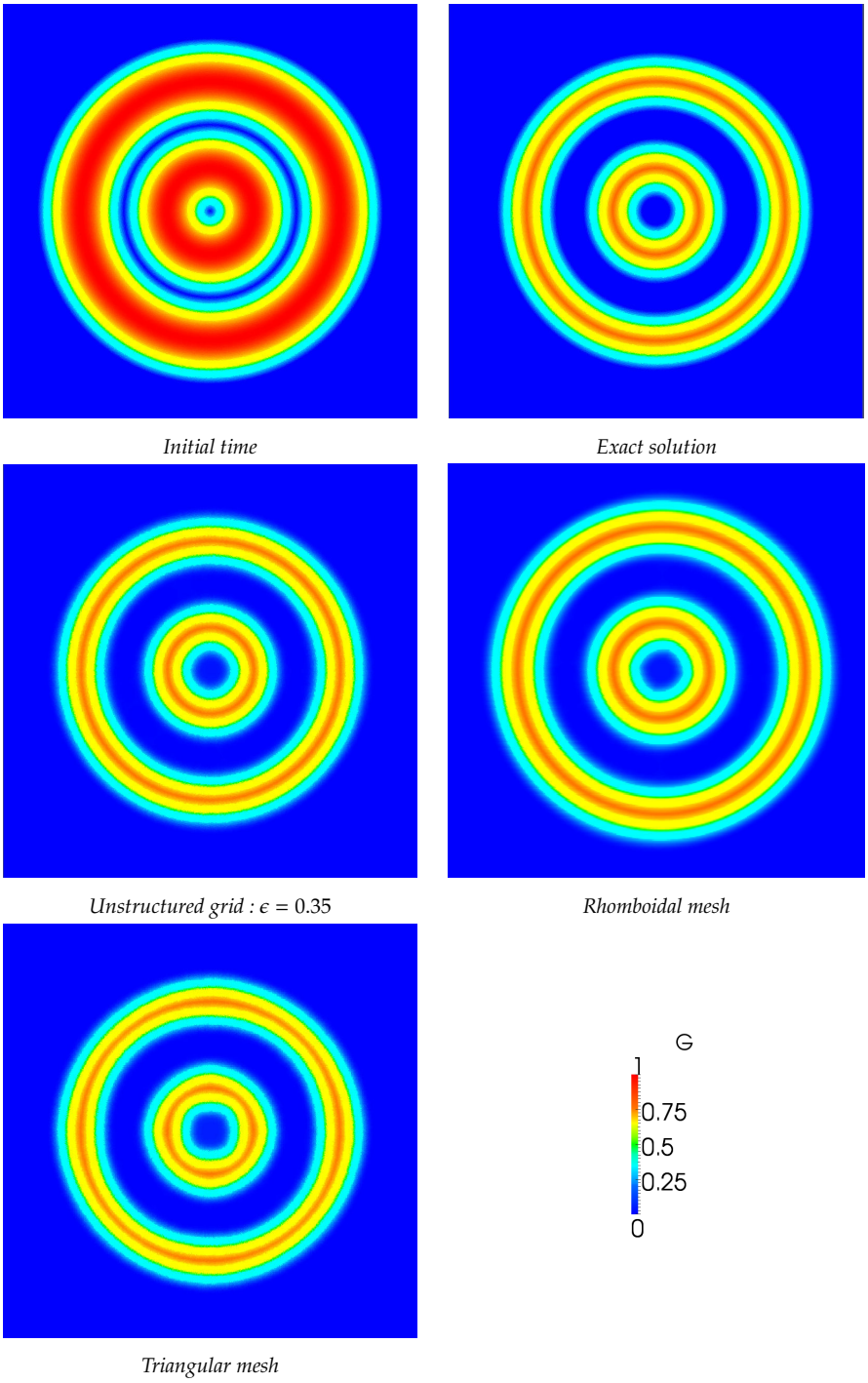


Figure 4.8 – G on different meshes –$T = 0.04s$ – $h = \dfrac{1}{400}$ – $CFL = \dfrac{1}{10}$

Finally we plot some convergence results. First of all the viscosity solution of the G-equation associated with the initial data (4.32), is given below :

$$G_{\text{visc}}(r,\theta,T) = \begin{cases} 0, & \forall r \in [0,T] \cup [\frac{1}{4}-T, \frac{1}{4}+T] \cup [\frac{1}{2}-T, \frac{1}{2}+T] \cup [\frac{3}{4}-T, \frac{3}{4}+T] \cup [1-T, 1], \\ |\sin(4\pi(r-T))|, & \forall r \in [T, \frac{1}{8}] \cup [\frac{1}{4}+T, \frac{3}{8}] \cup [\frac{1}{2}+T, \frac{5}{8}] \cup [\frac{3}{4}+T, \frac{7}{8}], \\ |\sin(4\pi(r+T))|, & \forall r \in [\frac{1}{8}, \frac{1}{4}-T] \cup [\frac{3}{8}, \frac{1}{2}-T] \cup [\frac{5}{8}, \frac{3}{4}-T] \cup [\frac{7}{8}, 1-T]. \end{cases}$$

We take $G_{\text{visc}}(r,\theta,T = 0.01s)$ as the initial data. The final time is set to $T = 0.04s$. The results

are given below, with a constant CFL equal to $\frac{1}{10}$, using three different meshes : an unstructured mesh with a deformation ratio equal to $\epsilon = 0.1$, a triangular mesh which consists of a square grid where each square is cut in half following the same diagonal, and a Rhomboidal mesh composed of parallelogrames with a large angle equal to $\frac{2\pi}{3}$.

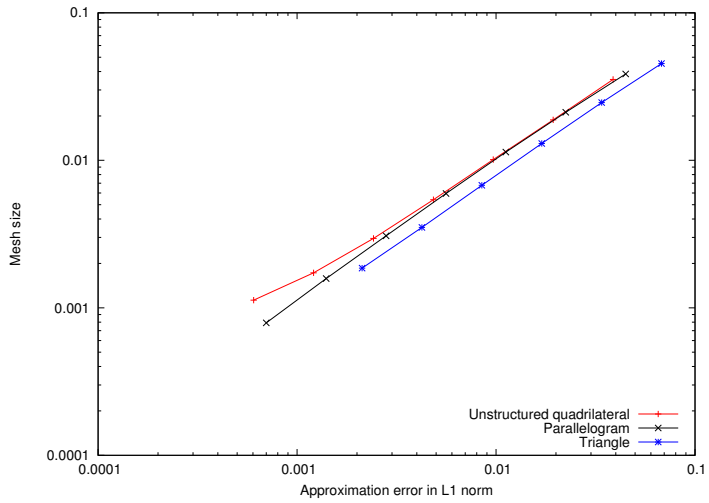


Figure 4.9 – L1 norm error at T=0.05s and CFL= $\dfrac{1}{10}$ – Upwind interpolation.

Cartesian grids

We use the same test to compare the convergence of the MUSCL scheme, the Upwind scheme, and an upwind finite difference scheme described in [20] designed for the Hamilton-Jacobi equations. In order to properly observe a difference in the convergence rate we use a Runge-Kutta time discretization of order two.

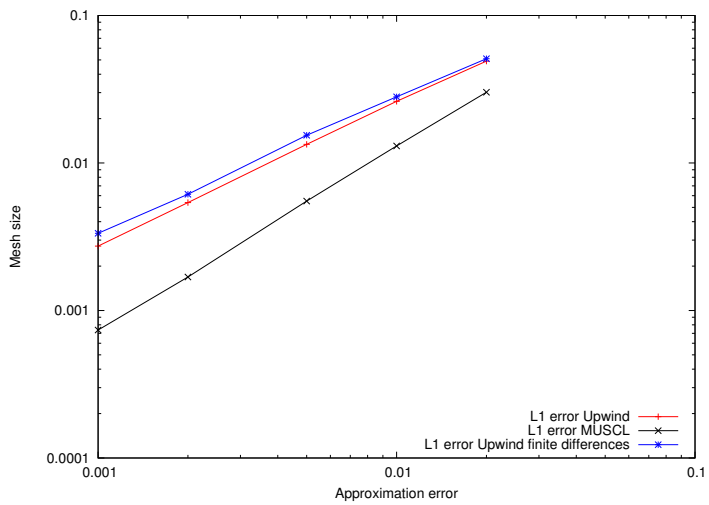


Figure 4.10 – L1 norm error at $T = 0.05s$ and CFL= $\dfrac{1}{10}$

Annexe A

Euler equations

A.1 Some results concerning explicit finite volume convection operators

The convection operator appearing in the mass balance equation reads, in the continuous problem, $\rho \to C(\rho) = \partial_t \rho + \mathrm{div}(\rho \boldsymbol{u})$, where $\boldsymbol{u}$ stands for a given velocity field, which is not assumed to satisfy any divergence constraint. We recall [36, Appendix A] that if ψ is a regular function from $(0, +\infty)$ to $\mathbb{R}$; then :

$$\psi'(\rho)\, C(\rho) = \partial_t\Big(\psi(\rho)\Big) + \mathrm{div}\Big(\psi(\rho)\boldsymbol{u}\Big) + \Big(\rho\psi'(\rho) - \psi(\rho)\Big)\, \mathrm{div}\boldsymbol{u}. \tag{A.1}$$

This computation is of course completely formal and only valid for regular functions ρ and $\boldsymbol{u}$. The following lemma states a discrete analogue to (A.1) for the explicit scheme studied in this thesis (see.[36, Appendix A] for an implicit scheme).

Lemma A.1

Let P be a polygonal (resp. polyhedral) bounded set of $\mathbb{R}^2$ (resp. $\mathbb{R}^3$), and let $\mathcal{E}(P)$ be the set of its edges (resp. faces). Let ψ be a twice continuously differentiable function defined over $(0, +\infty)$. Let $\rho_P^* > 0$, $\rho_P > 0$, $\delta t > 0$; consider three families $(\rho_\eta^*)_{\eta\in\mathcal{E}(P)} \subset \mathbb{R}_+ \setminus \{0\}$, $(V_\eta^*)_{\eta\in\mathcal{E}(P)} \subset \mathbb{R}$ and $(F_\eta^*)_{\eta\in\mathcal{E}(P)} \subset \mathbb{R}$ such that

$$\forall \eta \in \mathcal{E}(P), \qquad F_\eta^* = \rho_\eta^*\, V_\eta^*.$$

Let $R_{P,\delta t}$ be defined by :

$$R_{P,\delta t} = \Big[\frac{|P|}{\delta t}\,(\rho_P - \rho_P^*) + \sum_{\eta\in\mathcal{E}(P)} F_\eta^*\Big]\,\psi'(\rho_P) \\ - \frac{|P|}{\delta t}\,[\psi(\rho_P) - \psi(\rho_P^*)] + \sum_{\eta\in\mathcal{E}(P)} \psi(\rho_\eta^*)V_\eta^* + [\rho_P^*\psi'(\rho_P^*) - \psi(\rho_P^*)] \sum_{\eta\in\mathcal{E}(P)} V_\eta^*.$$

Then this quantity may be expressed as follows :

$$R_{P,\delta t} = \frac{1}{2}\frac{|P|}{\delta t}\,(\rho_P - \rho_P^*)^2\,\psi''(\overline{\rho}_P^{(1)}) - \frac{1}{2}\sum_{\eta\in\mathcal{E}(P)} V_\eta^*\,(\rho_P^* - \rho_\eta^*)^2\,\psi''(\overline{\rho}_\eta^*) + \sum_{\eta\in\mathcal{E}(P)} V_\eta^*\rho_\eta^*\,(\rho_P - \rho_P^*)\,\psi''(\overline{\rho}_P^{(2)}),$$

where $\overline{\rho}_P^{(1)}$, $\overline{\rho}_P^{(2)} \in [\![\rho_P, \rho_P^*]\!]$ and $\forall \eta \in \mathcal{E}(P)$, $\overline{\rho}_\eta^* \in [\![\rho_P^*, \rho_\eta^*]\!]$. We recall that, for $a,\ b \in \mathbb{R}$, we denote by $[\![a, b]\!]$ the interval $[\![a, b]\!] = \{\theta a + (1-\theta)b,\ \theta \in [0, 1]\}$.

Proof : By the definition of F^*_η, we have :

$$\Big[\frac{|P|}{\delta t}(\rho_P - \rho^*_P) + \sum_{\eta\in\mathcal{E}(P)} F^*_\eta\Big]\,\psi'(\rho_P) = \frac{|P|}{\delta t}(\rho_P - \rho^*_P)\,\psi'(\rho_P) + \sum_{\eta\in\mathcal{E}(P)} \rho^*_\eta V^*_\eta \psi'(\rho^*_P) + \sum_{\eta\in\mathcal{E}(P)} \rho^*_\eta V^*_\eta\Big[\psi'(\rho_P) - \psi'(\rho^*_P)\Big]. \quad \text{(A.2)}$$

By Taylor expansions of ψ, there exist two real numbers $\overline{\rho}^{(1)}_P$ and $\overline{\rho}^{(2)}_P \in [\![\rho^*_P, \rho_P]\!]$ and a family of real numbers $(\overline{\rho}^*_\eta)_{\eta\in\mathcal{E}(P)}$ satisfying, $\forall\eta\in\mathcal{E}(P)$, $\overline{\rho}^*_\eta \in [\![\rho^*_P, \rho^*_\eta]\!]$, and such that :

$$\begin{aligned}
&(\rho_P - \rho^*_P)\,\psi'(\rho_P) = \psi(\rho_P) - \psi(\rho^*_P) + \frac{1}{2}\,(\rho_P - \rho^*_P)^2\,\psi''(\overline{\rho}^{(1)}_P),\\
&\rho^*_\eta\psi'(\rho^*_P) = \psi(\rho^*_\eta) + [\rho^*_P\psi'(\rho^*_P) - \psi(\rho^*_P)] - \frac{1}{2}\,(\rho^*_\eta - \rho^*_P)^2\,\psi''(\overline{\rho}^*_\eta),\\
&\psi'(\rho_P) - \psi'(\rho^*_P) = (\rho_P - \rho^*_P)\,\psi''(\overline{\rho}^{(2)}_P).
\end{aligned}$$

Substituting in (A.2) yields the result we are seeking. ■

We now turn to the convection operator appearing in the momentum balance equation, which reads, in the continuous setting, $z \to C_\rho(z) = \partial_t(\rho z) + \mathrm{div}(\rho z \boldsymbol{u})$, where ρ (resp. $\boldsymbol{u}$) stands for a given scalar (resp. vector) field ; we wish to obtain some property of C_ρ under the assumption that ρ and $\boldsymbol{u}$ satisfy the mass balance equation, *i.e.* $\partial_t\rho + \mathrm{div}(\rho\boldsymbol{u}) = 0$. Formally, using twice the mass balance yields :

$$\begin{aligned}
\psi'(z)\,C_\rho(z) = \psi'(z)\Big[\partial_t(\rho\,z) + \mathrm{div}(\rho\,z\,\boldsymbol{u})\Big] &= \psi'(z)\rho\Big[\partial_t z + \boldsymbol{u}\cdot\nabla z\Big]\\
&= \rho\Big[\partial_t\psi(z) + \boldsymbol{u}\cdot\nabla\psi(z)\Big] = \partial_t\Big(\rho\,\psi(z)\Big) + \mathrm{div}\Big(\rho\,\psi(z)\,\boldsymbol{u}\Big).
\end{aligned}$$

Taking for z a component of the velocity field, this relation is the central argument used to derive the kinetic energy balance. The following lemma states a discrete counterpart of this identity, for a finite volume first-order explicit convection operator.

Lemma A.2

Let P be a polygonal (resp. polyhedral) bounded set of $\mathbb{R}^2$ (resp. $\mathbb{R}^3$) and let $\mathcal{E}(P)$ be the set of its edges (resp. faces). Let $\rho^*_P > 0$, $\rho_P > 0$, $\delta t > 0$, and $(F^*_\eta)_{\eta\in\mathcal{E}(P)} \subset \mathbb{R}$ be such that

$$\frac{|P|}{\delta t}\,(\rho_P - \rho^*_P) + \sum_{\eta\in\mathcal{E}(P)} F^*_\eta = 0. \quad \text{(A.3)}$$

Let ψ be a twice continuously differentiable function defined over $(0, +\infty)$. For $u^*_P \in \mathbb{R}$, $u_P \in \mathbb{R}$ and $(u^*_\eta)_{\eta\in\mathcal{E}(P)} \subset \mathbb{R}$ let us define :

$$R_{P,\delta t} = \Big[\frac{|P|}{\delta t}\Big(\rho_P\,u_P - \rho^*_P\,u^*_P\Big) + \sum_{\eta\in\mathcal{E}(P)} F^*_\eta\,u^*_\eta\Big]\,\psi'(u_P) - \Big[\frac{|P|}{\delta t}\Big[\rho_P\,\psi(u_P) - \rho^*_P\,\psi(u^*_P)\Big] + \sum_{\eta\in\mathcal{E}(P)} F^*_\eta\,\psi(u^*_\eta)\Big].$$

Then :

(*i*) *the remainder term $R_{P,\delta t}$ reads :*

$$R_{P,\delta t} = \frac{1}{2}\,\frac{|P|}{\delta t}\,\rho_P\,(u_P - u^*_P)^2\psi''(\overline{u}^{(1)}_P) - \frac{1}{2}\sum_{\eta\in\mathcal{E}(P)} F^*_\eta\,(u^*_\eta - u^*_P)^2\psi''(\overline{u}^*_\eta) + \sum_{\eta\in\mathcal{E}(P)} F^*_\eta\,(u^*_\eta - u^*_P)\,(u_P - u^*_P)\,\psi''(\overline{u}^{(2)}_P) \quad \text{(A.4)}$$

with $\overline{u}_P^{(1)}, \overline{u}_P^{(2)} \in [\![u_P, u_P^*]\!]$, *and* $\forall \eta \in \mathcal{E}(P)$, $\overline{u}_\eta^* \in [\![u_P^*, u_\eta^*]\!]$.

(*ii*) *If we suppose that the function* ψ *is convex and that* $u_\eta^* = u_P^*$ *as soon as* $F_\eta^* \geq 0$, *then* $R_{P,\delta t}$ *is non-negative under the CFL condition :*

$$\delta t \leq \frac{|P|\, \rho_P\, \underline{\psi}_P''}{\sum_{\eta \in \mathcal{E}(P)} (F_\eta^*)^- \, (\overline{\psi}_P'')^2 / \underline{\psi}_\eta''}, \tag{A.5}$$

where $\underline{\psi}_P'' = \min\limits_{s \in [\![u_P, u_P^*]\!]} \psi''(s)$, $\overline{\psi}_P'' = \max\limits_{s \in [\![u_P, u_P^*]\!]} \psi''(s)$ *and* $\underline{\psi}_\eta'' = \min\limits_{s \in [\![u_P^*, u_\eta^*]\!]} \psi''(s)$.

(*iii*) *In the case* $\psi(s) = \frac{s^2}{2}$ *(and therefore* $\psi''(s) = 1$, $\forall s \in (0, +\infty)$*), which is used to establish the discrete kinetic inequality, the remainder term reads*

$$R_{P,\delta t} = \frac{1}{2} \frac{|P|}{\delta t} \rho_P \, (u_P - u_P^*)^2 - \frac{1}{2} \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \, (u_\eta^* - u_P^*)^2 + \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \, (u_\eta^* - u_P^*) \, (u_P - u_P^*)$$

and is non-negative under the following simple CFL condition :

$$\delta t \leq \frac{|P|\, \rho_P}{\sum_{\eta \in \mathcal{E}(P)} (F_\eta^*)^-}. \tag{A.6}$$

Proof : Let T_P be defined by :

$$T_P = \Big[\frac{|P|}{\delta t} \left(\rho_P \, u_P - \rho_P^* \, u_P^* \right) + \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \, u_\eta^* \Big] \, \psi'(u_P).$$

Using equation (A.3) multiplied by u_P^*, we obtain :

$$T_P = \Big[\frac{|P|}{\delta t} \, \rho_P \left(u_P - u_P^* \right) + \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \, (u_\eta^* - u_P^*) \Big] \, \psi'(u_P).$$

We now define the remainder terms r_P and $(r_\eta^*)_{\eta \in \mathcal{E}(P)}$ by :

$$r_P = (u_P - u_P^*) \, \psi'(u_P) - \Big[\psi(u_P) - \psi(u_P^*) \Big], \quad r_\eta^* = (u_P^* - u_\eta^*) \, \psi'(u_P^*) - \Big[\psi(u_P^*) - \psi(u_\eta^*) \Big].$$

With these notations, we get :

$$\begin{aligned} T_P = \frac{|P|}{\delta t} \, \rho_P \, \Big[\psi(u_P) - \psi(u_P^*) \Big] + \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \Big[\psi(u_\eta^*) - \psi(u_P^*) \Big] \\ + \frac{|P|}{\delta t} \, \rho_P \, r_P - \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \, r_\eta^* + \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \, (u_\eta^* - u_P^*) \left(\psi'(u_P) - \psi'(u_P^*) \right). \end{aligned}$$

Using once again equation (A.3), this time multiplied by $\psi(u_P^*)$, we obtain :

$$\begin{aligned} T_P = \frac{|P|}{\delta t} \, \Big[\rho_P \psi(u_P) - \rho_P^* \psi(u_P^*) \Big] + \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \psi(u_\eta^*) \\ + \frac{|P|}{\delta t} \, \rho_P \, r_P - \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \, r_\eta^* + \sum_{\eta \in \mathcal{E}(P)} F_\eta^* \, (u_\eta^* - u_P^*) \left(\psi'(u_P) - \psi'(u_P^*) \right). \end{aligned}$$

The expression (A.4) of the remainder term $R_{P,\delta t}$ follows by remarking that, by a Taylor expansion, there exist $\overline{u}_P^{(1)}, \overline{u}_P^{(2)} \in [\![u_P, u_P^*]\!]$, and $\forall \eta \in \mathcal{E}(P)$, $\overline{u}_\eta^* \in [\![u_P^*, u_\eta^*]\!]$ such that :

$$r_P = \frac{1}{2} \, \psi''(\overline{u}_P^{(1)}) \, (u_P - u_P^*)^2, \quad r_\eta^* = \frac{1}{2} \, \psi''(\overline{u}_\eta^*) \, (u_\eta^* - u_p^*)^2$$

and

$$\psi'(u_P) - \psi'(u_P^*) = \psi''(\overline{u}_P^{(2)})\,(u_P - u_P^*).$$

If ψ is convex, r_P is non-negative. If, in addition, $u_P^* - u_\eta^*$ vanishes for any $\eta \in \mathcal{E}(P)$ when F_η^* is non-negative, $-r_\eta^*$ is non-negative. By Young's inequality, the last term in $R_{P,\delta t}$ may be bounded as follows :

$$\Big|\sum_{\eta\in\mathcal{E}(P)} (F_\eta^*)^-\,(u_\eta^* - u_P^*)\,(u_P - u_P^*)\,\psi''(\overline{u}_P^{(2)})\Big|$$
$$\leq \frac{\psi''(\overline{u}_P^{(2)})^2}{2}\Big[\sum_{\eta\in\mathcal{E}(P)} (F_\eta^*)^-\,\frac{1}{\psi''(\overline{u}_\eta^*)}\Big](u_P - u_P^*)^2 + \frac{1}{2}\sum_{\eta\in\mathcal{E}(P)} (F_\eta^*)^-\,(u_\eta^* - u_P^*)^2\,\psi''(\overline{u}_\eta^*),$$

so this term may be absorbed in the first two ones under the CFL condition (A.5). ■

A.2 Explicit formulas of the WLRs in the MAC case

The purpose of this section is to give explicit formulae for the WLRs W_K^n, defined in (1.42). For the sake of clarity, we suppose that the size of each cell is constant and denoted by dx. The time step is denoted by dt. Let us call $\Delta = \max(dx,\, dt)$. We set $x_i = x_0 + idx$, $y_i = y_0 + idx$. For the sake of understanding, we will numerate the elements of the mesh. Let us denote by $K_{i+\frac{1}{2}}$ the element $[x_i, x_{i+1}]$ of the mesh in 1D and $K_{i+\frac{1}{2},j+\frac{1}{2}}$ the element $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ in two dimensions.From now on, we drop the notation $K_{j+\frac{1}{2}}$, and we only keep the cardinal, in order to lighten the expressions.

A.2.1 One-dimensionnal case

$$W_{i+\frac{1}{2}}^n = \int_0^T \int_\Omega \rho^\Delta(x,t)\phi_{i+\frac{1}{2},t}^n(x,t) + \rho^\Delta(x,t)u^\Delta(x,t)\phi_{i+\frac{1}{2},x}^n(x,t)dxdt \tag{A.7}$$

We need to define the local test functions $\phi_{i+\frac{1}{2}}^n$, in order to be able to compute the viscosity.As in [42] we use quadratic and linear B-splines :

$$\phi_{j+\frac{1}{2}}^n(x,t) = B_{j+\frac{1}{2}}(x)B^n(t) \tag{A.8}$$

with :

$$B_{j+\frac{1}{2}}(x) = \begin{cases} \frac{1}{2}\left(\frac{x - x_{j-1}}{dx}\right)^2, & \text{if } x_{j-1} \leq x \leq x_j \\ \frac{3}{4} - \left(\frac{x - x_{j+\frac{1}{2}}}{dx}\right)^2, & \text{if } x_j \leq x \leq x_{j+1} \\ \frac{1}{2}\left(\frac{x - x_{j+2}}{dx}\right)^2, & \text{if } x_{j+1} \leq x \leq x_{j+2} \\ 0, & \text{otherwise} \end{cases} \tag{A.9}$$

and

$$B^n(t) = \begin{cases} \left(\frac{t - t^{n-1}}{dt}\right), & \text{if } t^{n-1} \leq t \leq t^n \\ \left(\frac{t^{n+1} - t}{dt}\right), & \text{if } t^n \leq t \leq t^{n+1} \\ 0, & \text{otherwise} \end{cases} \tag{A.10}$$

It can be easily checked that for every smooth test function ϕ, there exists $b^n_{j+\frac{1}{2}} \in \mathcal{R}$ independent of dx and dt such that :

$$\phi(x,t) = \sum_j \sum_n b^n_{j+\frac{1}{2}} \phi^n_{j+\frac{1}{2}}(x,t) + O(\Delta^2). \tag{A.11}$$

A straightforward calculation leads to :

$$\begin{aligned} W^n_{j+\frac{1}{2}} &= \frac{1}{6}\left[\rho^{n+1}_{j+\frac{3}{2}} - \rho^n_{j+\frac{3}{2}} + 4\left(\rho^{n+1}_{j+\frac{1}{2}} - \rho^n_{j+\frac{1}{2}}\right) + \rho^{n+1}_{j-\frac{1}{2}} - \rho^n_{j-\frac{1}{2}}\right] dx \\ &+ \frac{1}{4}\left[\rho^{n+1}_{j+\frac{3}{2}} u^{n+1}_{j+\frac{3}{2}} - \rho^{n+1}_{j-\frac{1}{2}} u^{n+1}_{j-\frac{1}{2}} + \rho^n_{j+\frac{3}{2}} u^n_{j+\frac{3}{2}} - \rho^n_{j-\frac{1}{2}} u^n_{j-\frac{1}{2}}\right] dt. \end{aligned} \tag{A.12}$$

A.2.2 Two dimensionnal case

We give the new definition of the local test functions in the two-dimensionnal case :

$$\phi^n_{j+\frac{1}{2},k+\frac{1}{2}}(x,y,t) = B_{j+\frac{1}{2}}(x) B_{k+\frac{1}{2}}(x) B^n(t). \tag{A.13}$$

$B_{j+\frac{1}{2}}$ and B^n have been defined in the previous part and we define $B_{k+\frac{1}{2}}$ in a similar way (see [42]) :

$$B_{k+\frac{1}{2}}(y) = \begin{cases} \frac{1}{2}\left(\frac{y - y_{k-1}}{dy}\right)^2, & \text{if } y_{k-1} \le x \le y_k \\ \frac{3}{4} - \left(\frac{y - y_{k+\frac{1}{2}}}{dy}\right)^2, & \text{if } y_k \le y \le y_{k+1} \\ \frac{1}{2}\left(\frac{y - y_{k+2}}{dy}\right)^2, & \text{if } y_{k+1} \le y \le y_{k+2} \\ 0, & \text{otherwise} \end{cases} \tag{A.14}$$

Let $\Delta = \max(dx, dy, dt)$. Let u and v be the x and y components of the velocity respectively. After computations, the 2-D version of the WLR is given by :

$$W^n_{i+\frac{1}{2},j+\frac{1}{2}} = \frac{1}{36\Delta} dx dy U^n_{i+\frac{1}{2},j+\frac{1}{2}} + \frac{1}{12\Delta}\left(dy dt V^n_{i+\frac{1}{2},j+\frac{1}{2}} + dx dt Y^n_{i+\frac{1}{2},j+\frac{1}{2}}\right) \tag{A.15}$$

where

$$\begin{aligned} U^n_{i+\frac{1}{2},j+\frac{1}{2}} = & \Big[\rho^{n+1}_{i+\frac{3}{2},j+\frac{3}{2}} - \rho^n_{i+\frac{3}{2},j+\frac{3}{2}} + \rho^{n+1}_{i-\frac{1}{2},j+\frac{3}{2}} - \rho^n_{i-\frac{1}{2},j+\frac{3}{2}} + \rho^{n+1}_{i+\frac{3}{2},j-\frac{1}{2}} - \rho^n_{i+\frac{3}{2},j-\frac{1}{2}} \\ & + \rho^{n+1}_{i-\frac{1}{2},j-\frac{1}{2}} - \rho^n_{i-\frac{1}{2},j-\frac{1}{2}}\Big] + 4\Big[\rho^{n+1}_{i+\frac{3}{2},j+\frac{1}{2}} - \rho^n_{i+\frac{3}{2},j+\frac{1}{2}} + \rho^{n+1}_{i-\frac{1}{2},j+\frac{1}{2}} \\ & - \rho^n_{i-\frac{1}{2},j+\frac{1}{2}} + \rho^{n+1}_{i+\frac{1}{2},j-\frac{1}{2}} - \rho^n_{i+\frac{1}{2},j-\frac{1}{2}} + \rho^{n+1}_{i+\frac{1}{2},j+\frac{3}{2}} - \rho^n_{i+\frac{1}{2},j+\frac{3}{2}}\Big] \\ & + 16\Big[\rho^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}} - \rho^n_{i+\frac{1}{2},j+\frac{1}{2}}\Big] \end{aligned} \tag{A.16}$$

$$\begin{aligned} V^n_{i+\frac{1}{2},j+\frac{1}{2}} = & \Big[(\rho u)^{n+1}_{i+\frac{3}{2},j+\frac{3}{2}} - (\rho u)^{n+1}_{i+\frac{3}{2},j-\frac{1}{2}} + (\rho u)^{n+1}_{i-\frac{1}{2},j+\frac{3}{2}} - (\rho u)^{n+1}_{i-\frac{1}{2},j-\frac{1}{2}} \\ & + (\rho u)^n_{i+\frac{3}{2},j+\frac{3}{2}} - (\rho u)^n_{i-\frac{1}{2},j+\frac{3}{2}} + (\rho u)^n_{i+\frac{3}{2},j-\frac{1}{2}} - (\rho u)^n_{i-\frac{1}{2},j-\frac{1}{2}}\Big] \\ & + 4\Big[(\rho u)^{n+1}_{i+\frac{3}{2},j+\frac{1}{2}} - (\rho u)^{n+1}_{i-\frac{1}{2},j+\frac{1}{2}} + (\rho u)^n_{i+\frac{3}{2},j+\frac{1}{2}} - (\rho u)^n_{i-\frac{1}{2},j+\frac{1}{2}}\Big] \end{aligned} \tag{A.17}$$

$$\begin{aligned} Y^n_{i+\frac{1}{2},j+\frac{1}{2}} = & \Big[(\rho v)^{n+1}_{i+\frac{3}{2},j+\frac{3}{2}} - (\rho v)^{n+1}_{i-\frac{1}{2},j+\frac{3}{2}} + (\rho v)^{n+1}_{i-\frac{1}{2},j+\frac{3}{2}} - (\rho v)^{n+1}_{i-\frac{1}{2},j-\frac{1}{2}} \\ & + (\rho v)^n_{i+\frac{3}{2},j+\frac{3}{2}} - (\rho v)^n_{i+\frac{3}{2},j-\frac{1}{2}} + (\rho v)^n_{i-\frac{1}{2},j+\frac{3}{2}} - (\rho v)^n_{i-\frac{1}{2},j-\frac{1}{2}}\Big] \\ & + 4\Big[(\rho v)^{n+1}_{i+\frac{1}{2},j+\frac{3}{2}} - (\rho v)^{n+1}_{i+\frac{1}{2},j-\frac{1}{2}} + (\rho v)^n_{i+\frac{1}{2},j+\frac{3}{2}} - (\rho v)^n_{i+\frac{1}{2},j-\frac{1}{2}}\Big] \end{aligned} \tag{A.18}$$

A.3 2D Riemann problems

We present in this section the 19 possible configurations presented in [44]. Upwind and MUSCL interpolations are compared through density. Some parameters are the same for all the test cases. We consider a cartesian grid made of $400 * 400$ uniform cells. The CFL is supposed to be constant equal to $dt/dx = 1/10$.

These test cases are designed so that there exists only one single wave on each interface. They are of three different kind : rarefaction waves ($\vec{R}$), contact discontinuities ($J^{\pm}$), and shock waves ($\vec{S}$). Right ($\vec{.}$) and left arrows ($\overleftarrow{.}$) stand for forward and backward waves respectively. Exponents over J refer to positive and negative contacts. For test containing shock waves, we use WLR viscosity with the MUSCL interpolation and the calibration parameter c_m is equal to 1.

The initial data, for $\boldsymbol{u} = (u, v)$ and $\boldsymbol{x} = (x, y)$ are given by :

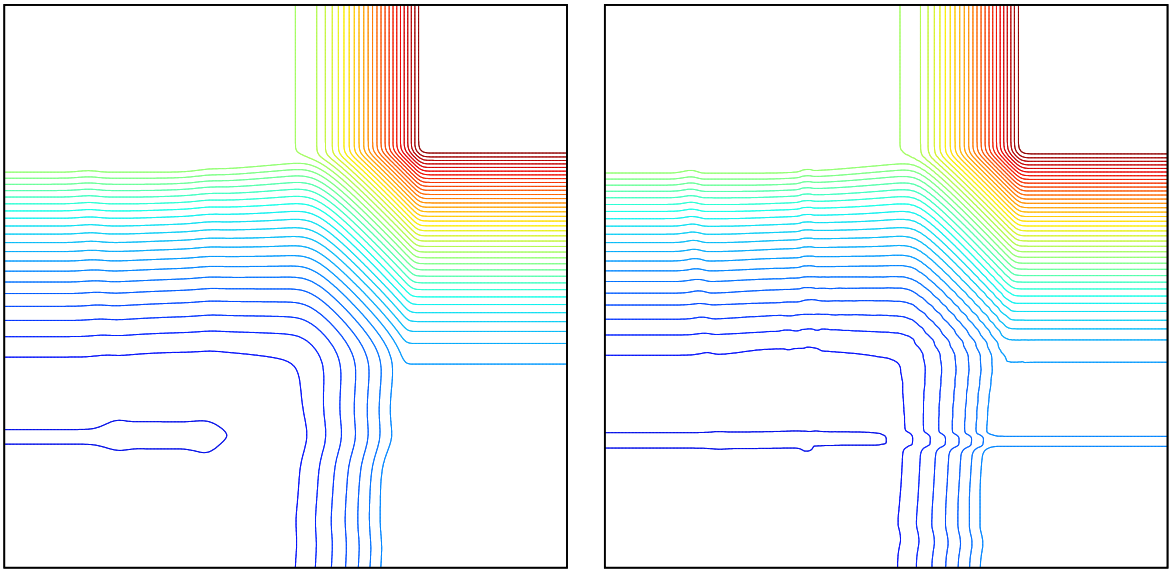
$$(\rho, p, u, v) \begin{cases} (\rho_1, p_1, u_1, v_1) & \text{if } x > 0.5 \text{ and } y > 0.5 \\ (\rho_2, p_2, u_2, v_2) & \text{if } x < 0.5 \text{ and } y > 0.5 \\ (\rho_3, p_3, u_3, v_3) & \text{if } x < 0.5 \text{ and } y < 0.5 \\ (\rho_4, p_4, u_4, v_4) & \text{if } x > 0.5 \text{ and } y < 0.5 \end{cases}$$

Configuration 1–

$$\begin{array}{ccc} & \vec{R}_{2,1} & \\ \vec{R}_{3,2} & & \vec{R}_{4,1} \\ & \vec{R}_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0 \\ v_1 = 0 \end{bmatrix} \begin{bmatrix} \rho_2 = 0.5197 \\ p_2 = 0.4 \\ u_2 = -0.7259 \\ v_2 = 0 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.1072 \\ p_3 = 0.0439 \\ u_3 = -0.7259 \\ v_3 = -1.4045 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.2579 \\ p_4 = 0.15 \\ u_4 = 0 \\ v_4 = -1.4045 \end{bmatrix}.$$



Upwind *MUSCL*

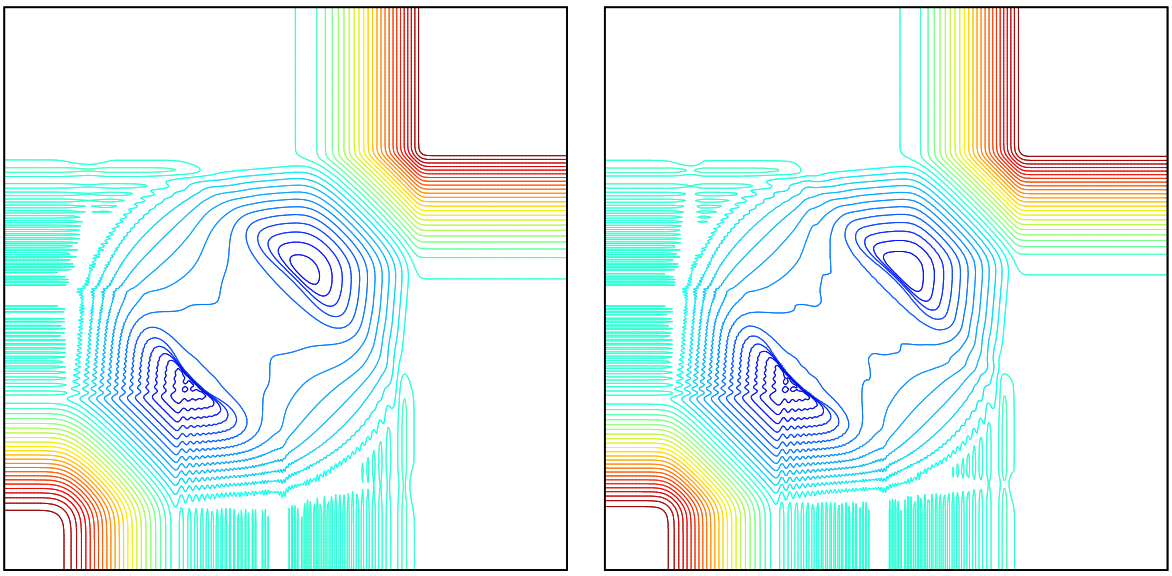
Figure A.1 – A two-dimensional Riemann problem : Configuration 1 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.2$.

Configuration 2–

$$\begin{array}{ccc} & \vec{R}_{2,1} & \\ \overleftarrow{R}_{3,2} & & \vec{R}_{4,1} \\ & \overleftarrow{R}_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0 \\ v_1 = 0 \end{bmatrix} \begin{bmatrix} \rho_2 = 0.5197 \\ p_2 = 0.4 \\ u_2 = -0.7259 \\ v_2 = 0 \end{bmatrix} \begin{bmatrix} \rho_3 = 1 \\ p_3 = 1 \\ u_3 = -0.7259 \\ v_3 = -0.7259 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5197 \\ p_4 = 0.4 \\ u_4 = 0 \\ v_4 = -0.7259 \end{bmatrix}.$$



Upwind *MUSCL*

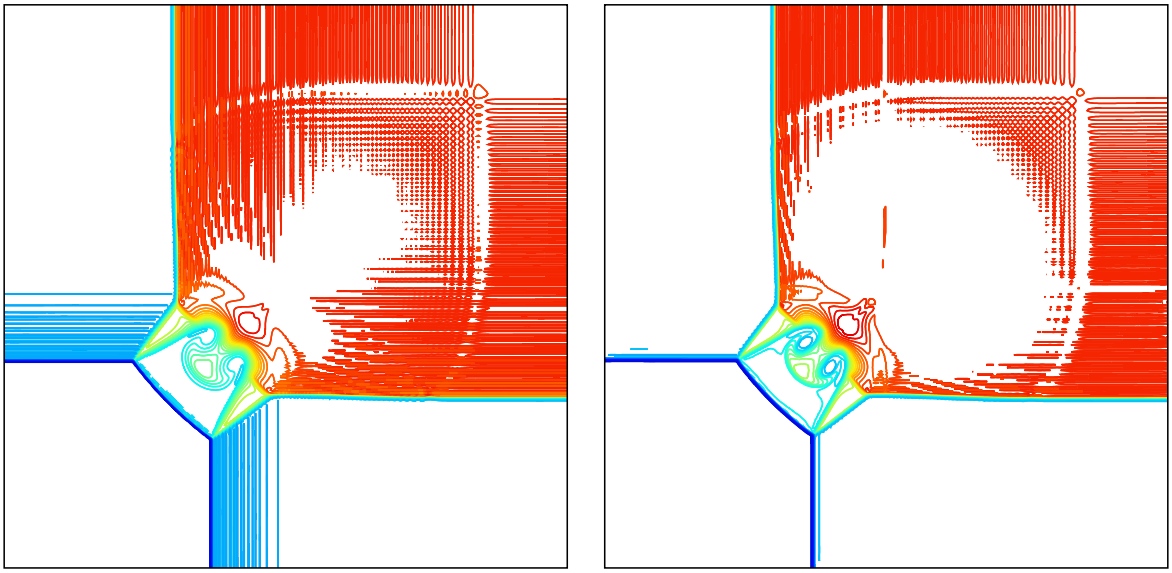
Figure A.2 – A two-dimensional Riemann problem : Configuration 2 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.2$.

Configuration 3–

$$\begin{array}{ccc} & \overleftarrow{S}_{2,1} & \\ \overleftarrow{S}_{3,2} & & \overleftarrow{S}_{4,1} \\ & \overleftarrow{S}_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1.5 \\ p_1 = 1.5 \\ u_1 = 0 \\ v_1 = 0 \end{bmatrix} \begin{bmatrix} \rho_2 = 0.5323 \\ p_2 = 0.3 \\ u_2 = 1.206 \\ v_2 = 0 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.138 \\ p_3 = 0.029 \\ u_3 = 1.206 \\ v_3 = 1.206 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5323 \\ p_4 = 0.3 \\ u_4 = 0 \\ v_4 = 1.206 \end{bmatrix}.$$



Upwind *MUSCL*

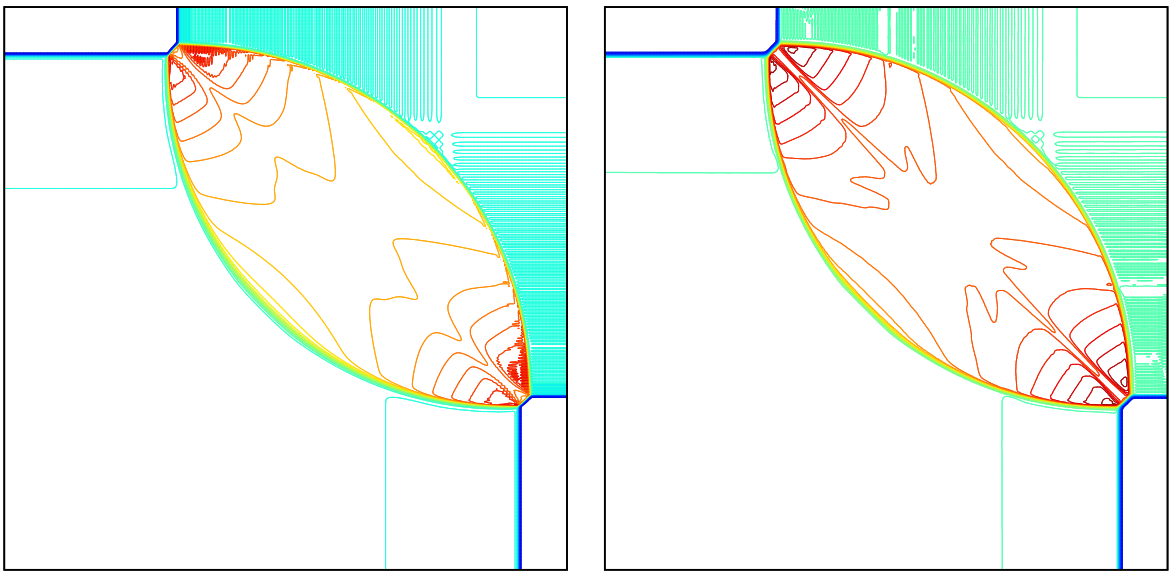
Figure A.3 – A two-dimensional Riemann problem : Configuration 3 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.3$.

Configuration 4–

$$\begin{array}{ccc} & \overleftarrow{S}_{2,1} & \\ \vec{S}_{3,2} & & \overleftarrow{S}_{4,1} \\ & \vec{S}_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1.1 \\ p_1 = 1.1 \\ u_1 = 0 \\ v_1 = 0 \end{bmatrix} \begin{bmatrix} \rho_2 = 0.5065 \\ p_2 = 0.35 \\ u_2 = 0.8939 \\ v_2 = 0 \end{bmatrix} \begin{bmatrix} \rho_3 = 1.1 \\ p_3 = 1.1 \\ u_3 = 0.8939 \\ v_3 = 0.8939 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5065 \\ p_4 = 0.35 \\ u_4 = 0 \\ v_4 = 0.8939 \end{bmatrix}.$$



Upwind *MUSCL*

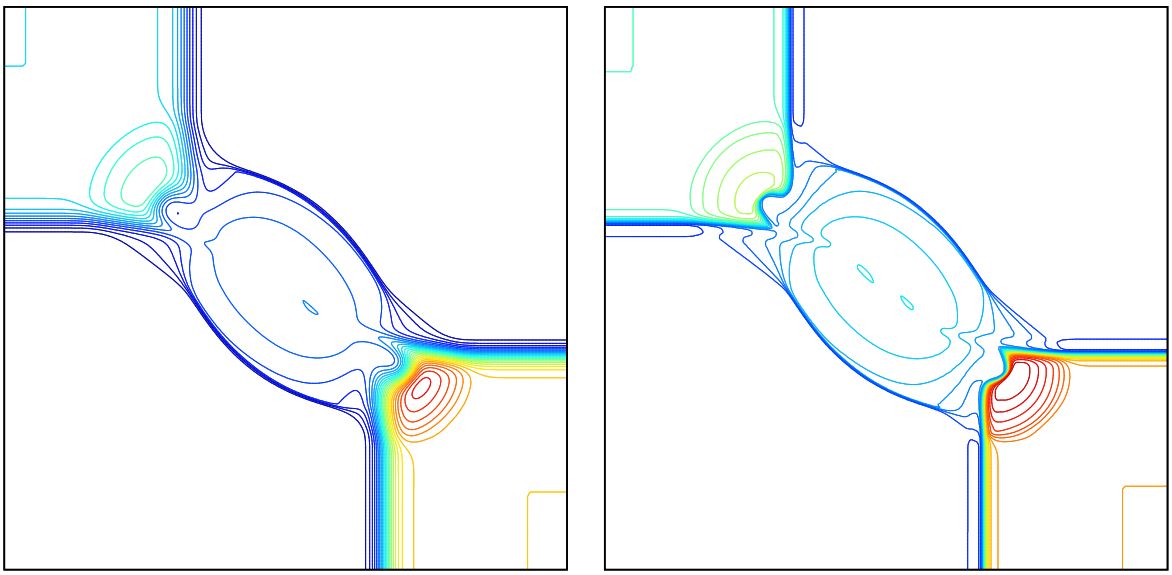
Figure A.4 – A two-dimensional Riemann problem : Configuration 4 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.25$.

Configuration 5–

$$\begin{array}{ccc} & J^-_{2,1} & \\ J^-_{3,2} & & J^-_{4,1} \\ & J^-_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = -0.75 \\ v_1 = -0.5 \end{bmatrix} \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = -0.75 \\ v_2 = 0.5 \end{bmatrix} \begin{bmatrix} \rho_3 = 1 \\ p_3 = 1 \\ u_3 = 0.75 \\ v_3 = 0.5 \end{bmatrix} \begin{bmatrix} \rho_4 = 3 \\ p_4 = 1 \\ u_4 = 0.75 \\ v_4 = -0.5 \end{bmatrix}.$$



Upwind *MUSCL*

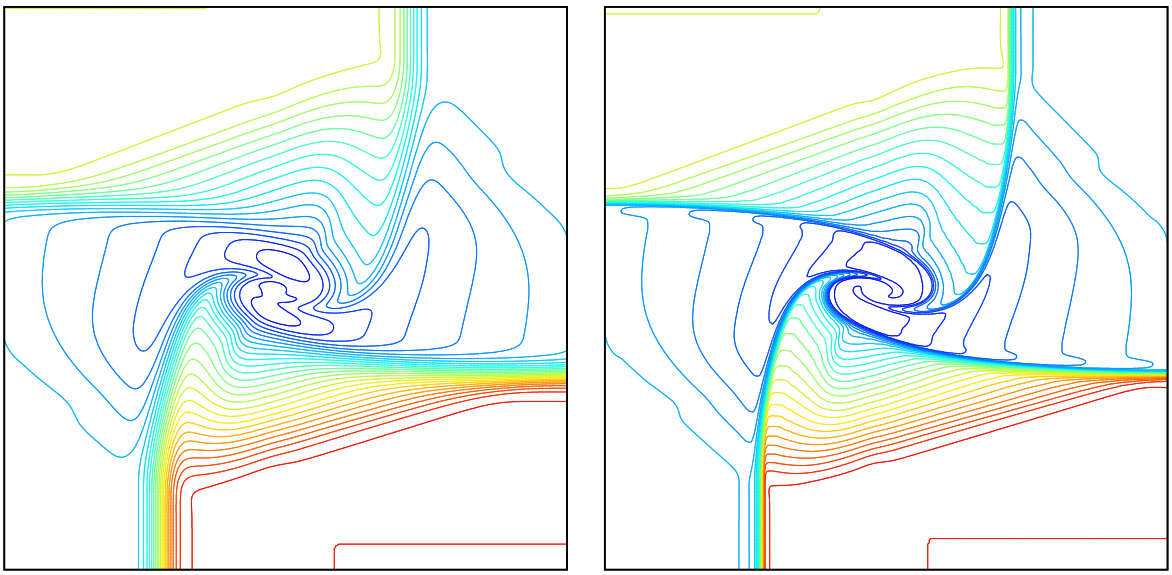
Figure A.5 – A two-dimensional Riemann problem : Configuration 5 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.23$.

Configuration 6–

$$\begin{array}{ccc} & J^-_{2,1} & \\ J^+_{3,2} & & J^+_{4,1} \\ & J^-_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0.75 \\ v_1 = -0.5 \end{bmatrix} \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = 0.75 \\ v_2 = 0.5 \end{bmatrix} \begin{bmatrix} \rho_3 = 1 \\ p_3 = 1 \\ u_3 = -0.75 \\ v_3 = 0.5 \end{bmatrix} \begin{bmatrix} \rho_4 = 3 \\ p_4 = 1 \\ u_4 = -0.75 \\ v_4 = -0.5 \end{bmatrix}.$$



Upwind *MUSCL*

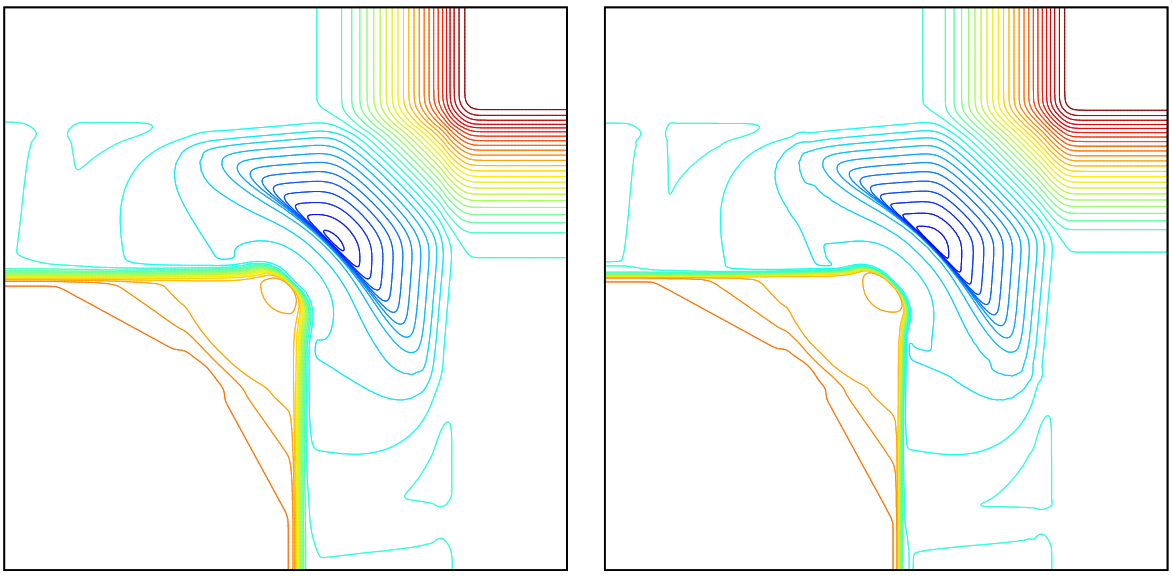
Figure A.6 – A two-dimensional Riemann problem : Configuration 6 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.3$.

Configuration 7–

$$\begin{array}{ccc} & \vec{R}_{2,1} & \\ J^-_{3,2} & & \vec{R}_{4,1} \\ & J^-_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0.1 \\ v_1 = 0.1 \end{bmatrix} \quad \begin{bmatrix} \rho_2 = 0.5197 \\ p_2 = 0.4 \\ u_2 = -0.6259 \\ v_2 = 0.1 \end{bmatrix} \quad \begin{bmatrix} \rho_3 = 0.8 \\ p_3 = 0.4 \\ u_3 = 0.1 \\ v_3 = 0.1 \end{bmatrix} \quad \begin{bmatrix} \rho_4 = 0.5197 \\ p_4 = 0.4 \\ u_4 = 0.1 \\ v_4 = -0.6259 \end{bmatrix}.$$



Upwind *MUSCL*

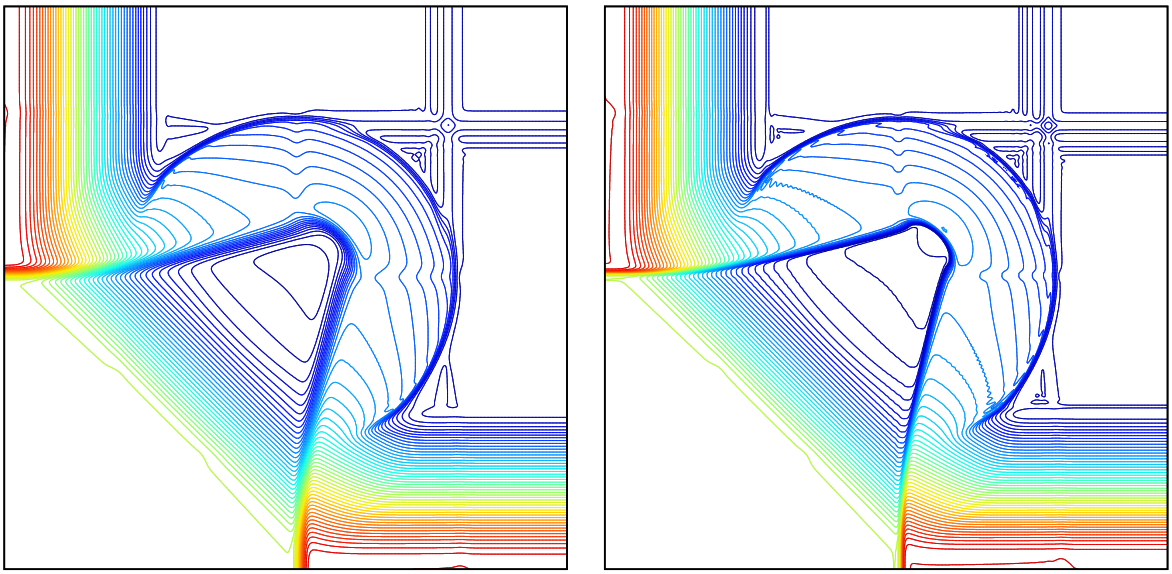
Figure A.7 – A two-dimensional Riemann problem : Configuration 7 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.25$.

Configuration 8–

$$\begin{array}{ccc} & \vec{R}_{2,1} & \\ J^-_{3,2} & & \overleftarrow{R}_{4,1} \\ & J^-_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 0.5197 \\ p_1 = 0.4 \\ u_1 = 0.1 \\ v_1 = 0.1 \end{bmatrix} \begin{bmatrix} \rho_2 = 1 \\ p_2 = 1 \\ u_2 = -0.6259 \\ v_2 = 0.1 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.8 \\ p_3 = 1 \\ u_3 = 0.1 \\ v_3 = 0.1 \end{bmatrix} \begin{bmatrix} \rho_4 = 1 \\ p_4 = 1 \\ u_4 = 0.1 \\ v_4 = -0.6259 \end{bmatrix}.$$



Upwind *MUSCL*

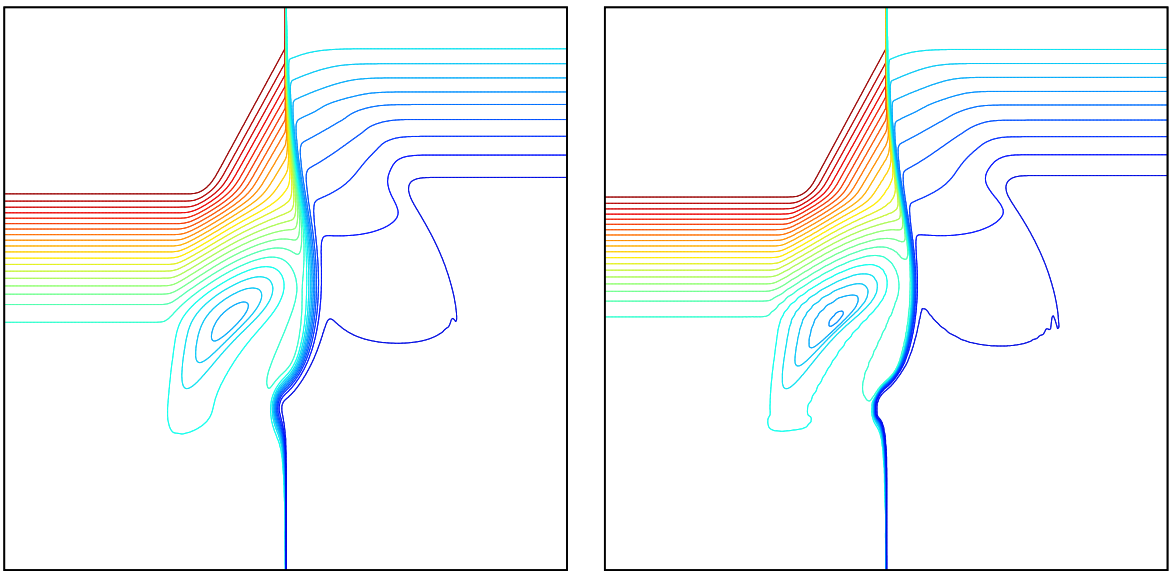
Figure A.8 – A two-dimensional Riemann problem : Configuration 8 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.25$.

Configuration 9–

$$\begin{array}{ccc} & J^+_{2,1} & \\ \vec{R}_{3,2} & & \vec{R}_{4,1} \\ & J^+_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0 \\ v_1 = 0.3 \end{bmatrix} \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = 0 \\ v_2 = -0.3 \end{bmatrix} \begin{bmatrix} \rho_3 = 1.039 \\ p_3 = 0.4 \\ u_3 = 0 \\ v_3 = -0.8133 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5197 \\ p_4 = 0.4 \\ u_4 = 0 \\ v_4 = -0.4259 \end{bmatrix}.$$



Upwind *MUSCL*

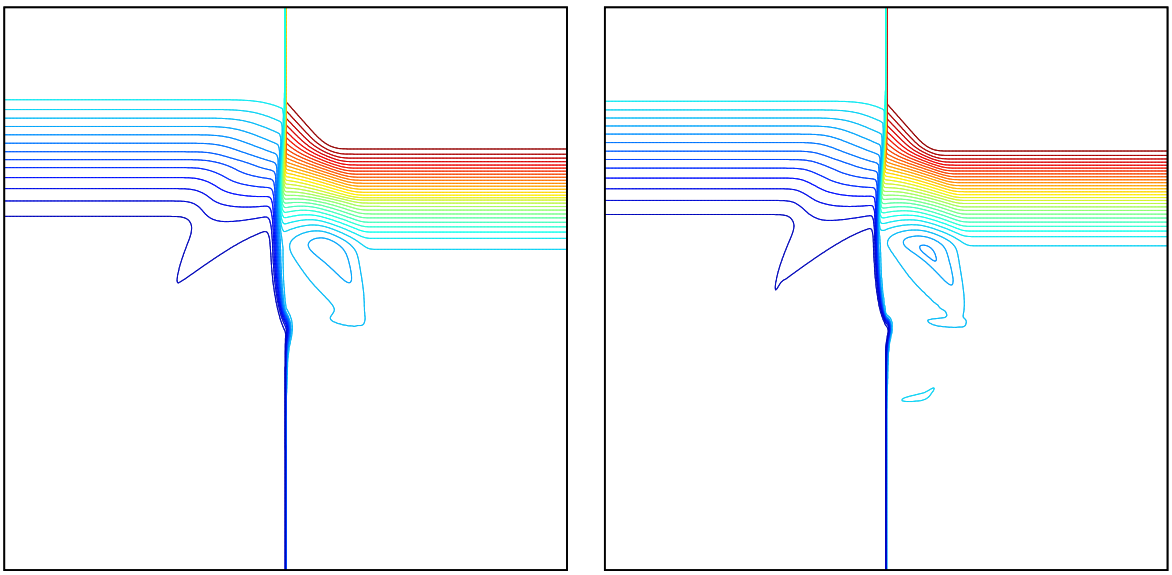
Figure A.9 – A two-dimensional Riemann problem : Configuration 9 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.3$.

Configuration 10–

$$\begin{array}{ccc} & J^-_{2,1} & \\ \vec{R}_{3,2} & & \vec{R}_{4,1} \\ & J^+_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0 \\ v_1 = 0.4297 \end{bmatrix} \begin{bmatrix} \rho_2 = 0.5 \\ p_2 = 1 \\ u_2 = 0 \\ v_2 = 0.6076 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.2281 \\ p_3 = 0.3333 \\ u_3 = 0 \\ v_3 = -0.6076 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.4562 \\ p_4 = 0.3333 \\ u_4 = 0 \\ v_4 = -0.4297 \end{bmatrix}.$$



Upwind *MUSCL*

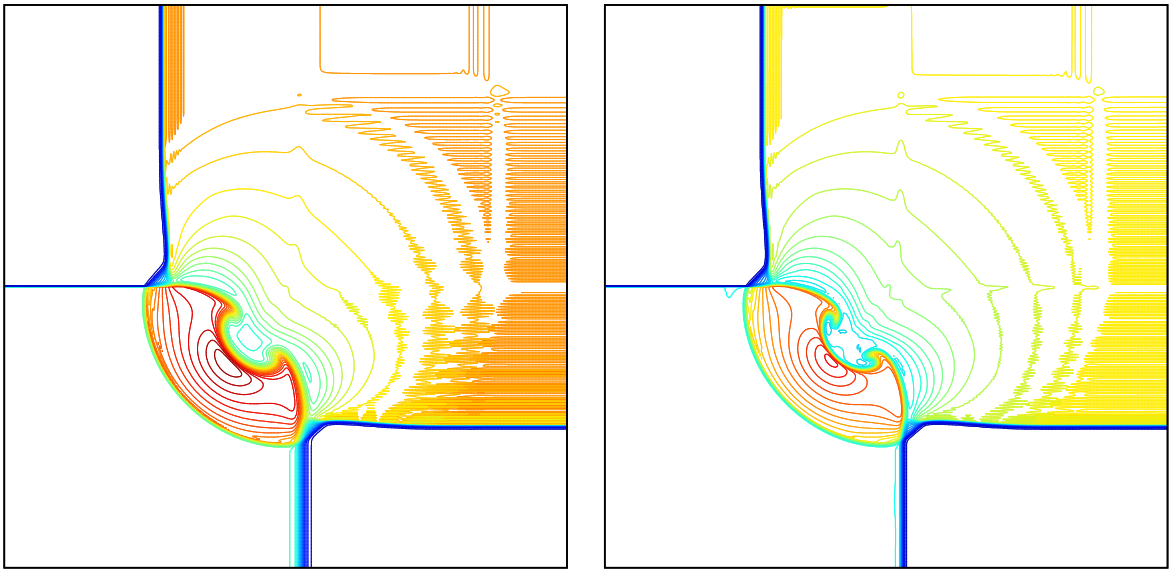
Figure A.10 – A two-dimensional Riemann problem : Configuration 10 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.15$.

Configuration 11–

$$\begin{array}{ccc} & \overleftarrow{S}_{2,1} & \\ J^+_{3,2} & & \overleftarrow{S}_{4,1} \\ & J^+_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0.1 \\ v_1 = 0 \end{bmatrix} \begin{bmatrix} \rho_2 = 0.5313 \\ p_2 = 0.4 \\ u_2 = 0.8276 \\ v_2 = 0 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.8 \\ p_3 = 0.4 \\ u_3 = 0.1 \\ v_3 = 0 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5313 \\ p_4 = 0.4 \\ u_4 = 0.1 \\ v_4 = 0.7276 \end{bmatrix}.$$



Upwind *MUSCL*

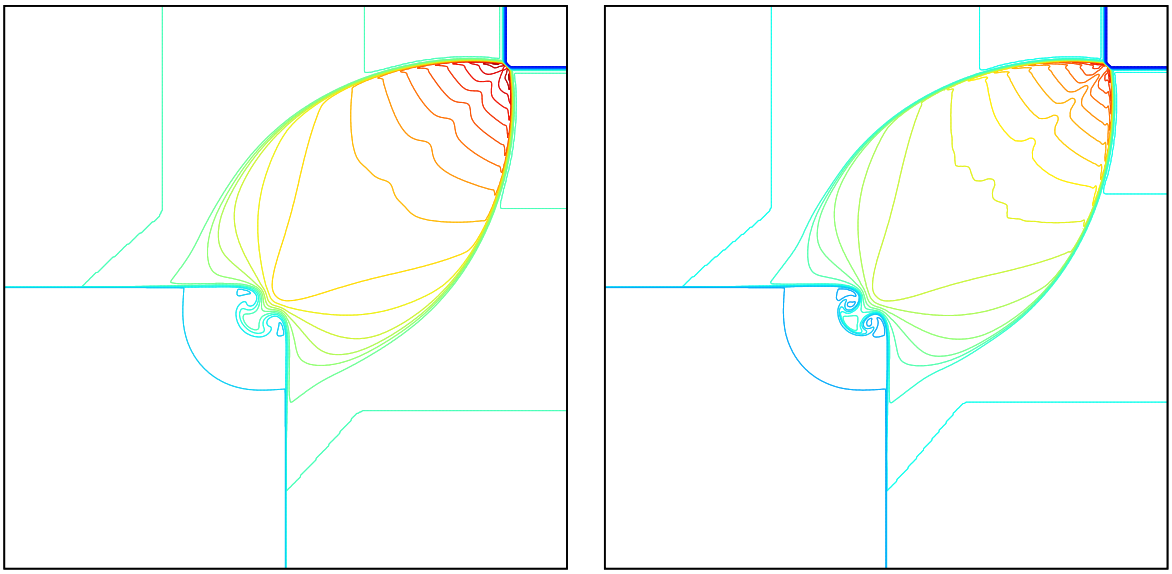
Figure A.11 – A two-dimensional Riemann problem : Configuration 11 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.15$.

Configuration 12–

$$\begin{array}{ccc} & \vec{S}_{2,1} & \\ J^+_{3,2} & & \vec{S}_{4,1} \\ & J^+_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 0.5313 \\ p_1 = 0.4 \\ u_1 = 0 \\ v_1 = 0 \end{bmatrix} \begin{bmatrix} \rho_2 = 1 \\ p_2 = 1 \\ u_2 = 0.7276 \\ v_2 = 0 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.8 \\ p_3 = 1 \\ u_3 = 0 \\ v_3 = 0 \end{bmatrix} \begin{bmatrix} \rho_4 = 1 \\ p_4 = 1 \\ u_4 = 0 \\ v_4 = 0.7276 \end{bmatrix}.$$



Upwind *MUSCL*

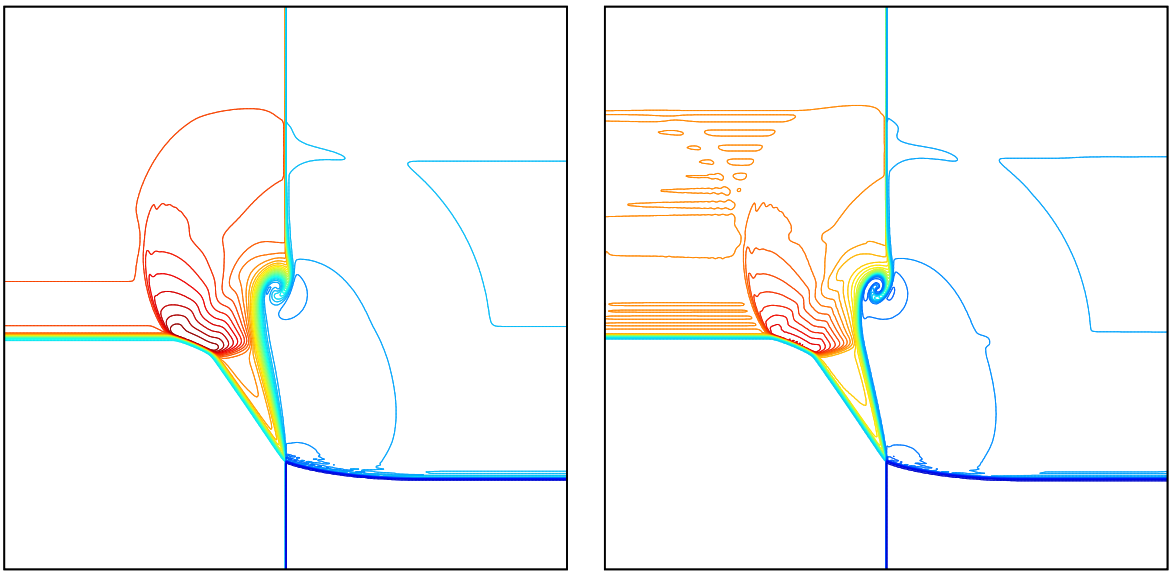
Figure A.12 – A two-dimensional Riemann problem : Configuration 12 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.25$.

Configuration 13–

$$\begin{array}{ccc} & J^-_{2,1} & \\ \overleftarrow{S}_{3,2} & & \overleftarrow{S}_{4,1} \\ & J^-_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0 \\ v_1 = -0.3 \end{bmatrix} \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = 0 \\ v_2 = 0.3 \end{bmatrix} \begin{bmatrix} \rho_3 = 1.0625 \\ p_3 = 0.4 \\ u_3 = 0 \\ v_3 = 0.8145 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5313 \\ p_4 = 0.4 \\ u_4 = 0 \\ v_4 = 0.4276 \end{bmatrix}.$$



Upwind *MUSCL*

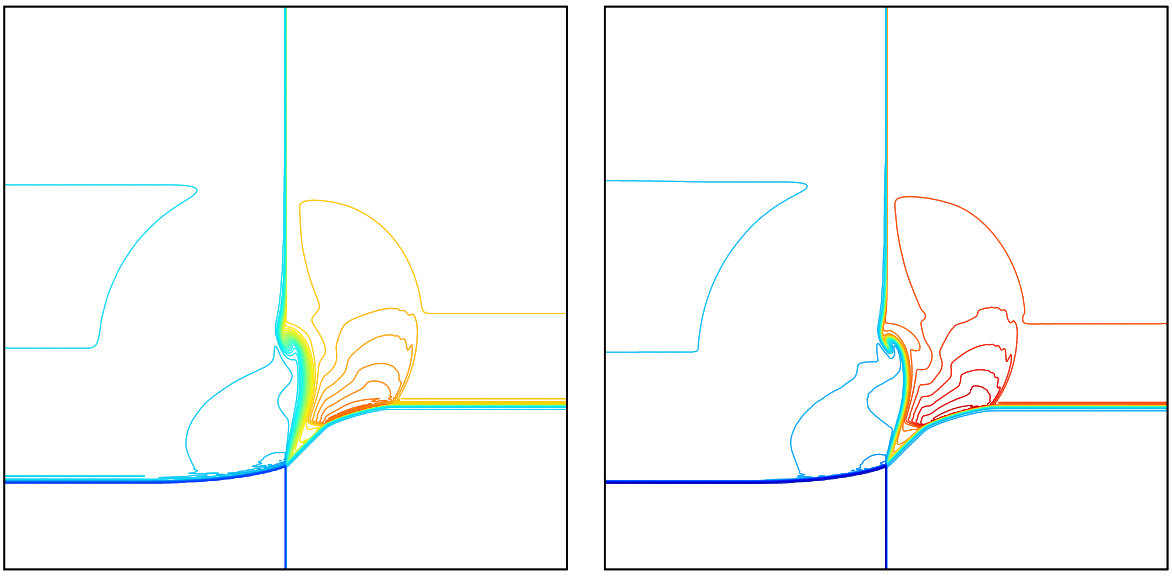
Figure A.13 – A two-dimensional Riemann problem : Configuration 13 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.3$.

Configuration 14–

$$\begin{array}{ccc} & J^+_{2,1} & \\ \overleftarrow{S}_{3,2} & & \overleftarrow{S}_{4,1} \\ & J^-_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 2 \\ p_1 = 8 \\ u_1 = 0 \\ v_1 = -0.5606 \end{bmatrix} \begin{bmatrix} \rho_2 = 1 \\ p_2 = 8 \\ u_2 = 0 \\ v_2 = -1.2172 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.4736 \\ p_3 = 2.6667 \\ u_3 = 0 \\ v_3 = 1.2172 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.9474 \\ p_4 = 2.6667 \\ u_4 = 0 \\ v_4 = 1.1606 \end{bmatrix}.$$



Upwind *MUSCL*

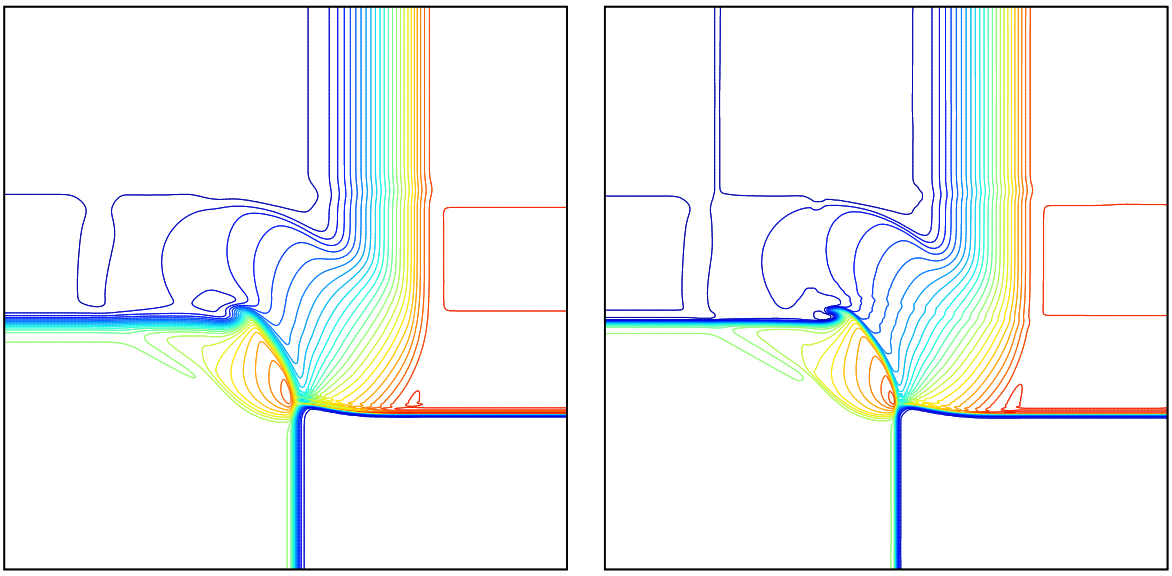
Figure A.14 – A two-dimensional Riemann problem : Configuration 14 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.1$.

Configuration 15–

$$\begin{array}{ccc} & \vec{R}_{2,1} & \\ J^-_{3,2} & & \overleftarrow{S}_{4,1} \\ & J^+_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0.1 \\ v_1 = -0.3 \end{bmatrix} \begin{bmatrix} \rho_2 = 0.5197 \\ p_2 = 0.4 \\ u_2 = -0.6259 \\ v_2 = -0.3 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.8 \\ p_3 = 0.4 \\ u_3 = 0.1 \\ v_3 = -0.3 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5313 \\ p_4 = 0.4 \\ u_4 = 0.1 \\ v_4 = 0.4276 \end{bmatrix}.$$



Upwind *MUSCL*

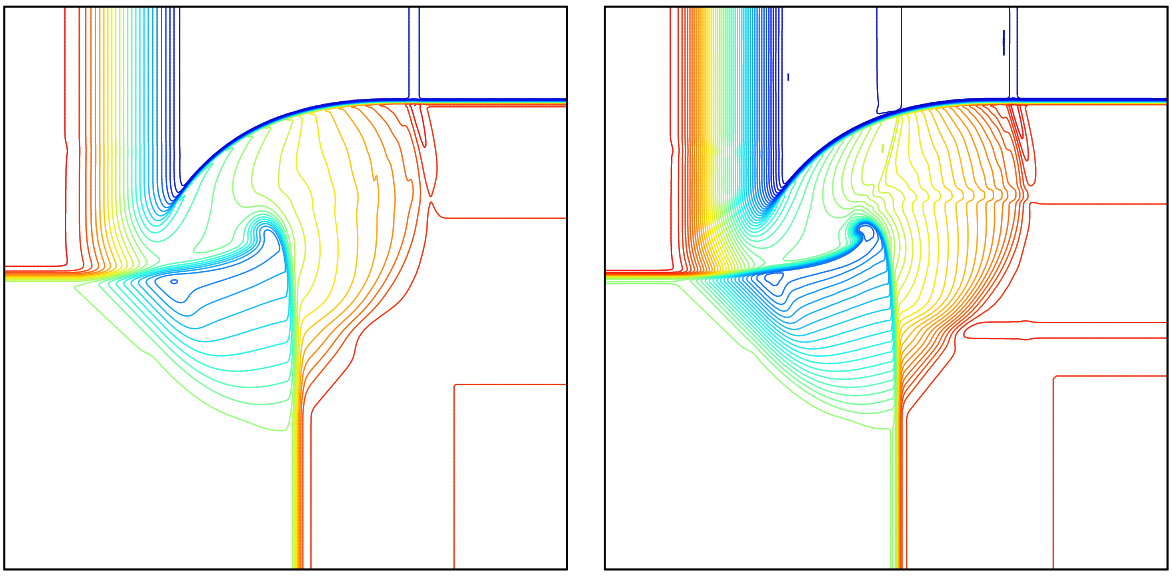
Figure A.15 – A two-dimensional Riemann problem : Configuration 15 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.2$.

Configuration 16–

$$\begin{array}{ccc} & \overleftarrow{R}_{2,1} & \\ J^-_{3,2} & & \vec{S}_{4,1} \\ & J^+_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 0.5313 \\ p_1 = 0.4 \\ u_1 = 0.1 \\ v_1 = 0.1 \end{bmatrix} \begin{bmatrix} \rho_2 = 1.0222 \\ p_2 = 1 \\ u_2 = -0.6179 \\ v_2 = 0.1 \end{bmatrix} \begin{bmatrix} \rho_3 = 0.8 \\ p_3 = 1 \\ u_3 = 0.1 \\ v_3 = 0.1 \end{bmatrix} \begin{bmatrix} \rho_4 = 1 \\ p_4 = 1 \\ u_4 = 0.1 \\ v_4 = 0.8276 \end{bmatrix}.$$



Upwind *MUSCL*

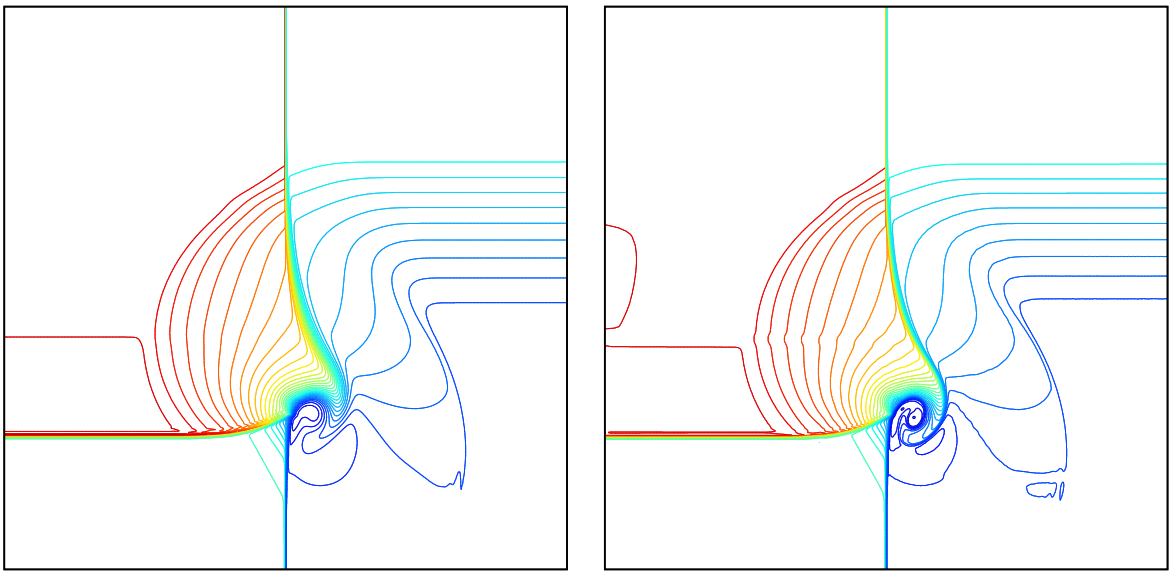
Figure A.16 – A two-dimensional Riemann problem : Configuration 16 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.2$.

Configuration 17–

$$\begin{array}{ccc} & J^-_{2,1} & \\ \overleftarrow{S}_{3,2} & & \vec{R}_{4,1} \\ & J^-_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0 \\ v_1 = -0.4 \end{bmatrix} \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = 0 \\ v_2 = -0.3 \end{bmatrix} \begin{bmatrix} \rho_3 = 1.0625 \\ p_3 = 0.4 \\ u_3 = 0 \\ v_3 = 0.2145 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5197 \\ p_4 = 0.4 \\ u_4 = 0 \\ v_4 = -1.1259 \end{bmatrix}.$$



Upwind *MUSCL*

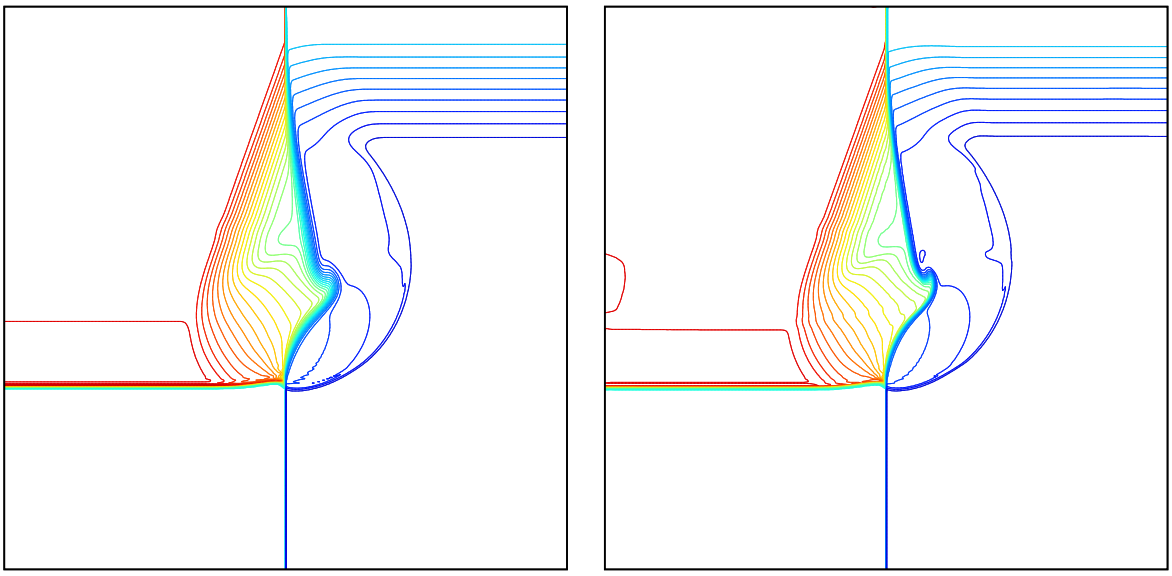
Figure A.17 – A two-dimensional Riemann problem : Configuration 17 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.3$.

Configuration 18–

$$\begin{array}{ccc} & J^+_{2,1} & \\ \overleftarrow{S}_{3,2} & & \vec{R}_{4,1} \\ & J^+_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0 \\ v_1 = 1 \end{bmatrix} \quad \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = 0 \\ v_2 = -0.3 \end{bmatrix} \quad \begin{bmatrix} \rho_3 = 1.0625 \\ p_3 = 0.4 \\ u_3 = 0 \\ v_3 = 0.2145 \end{bmatrix} \quad \begin{bmatrix} \rho_4 = 0.5197 \\ p_4 = 0.4 \\ u_4 = 0 \\ v_4 = 0.2741 \end{bmatrix}.$$



Upwind *MUSCL*

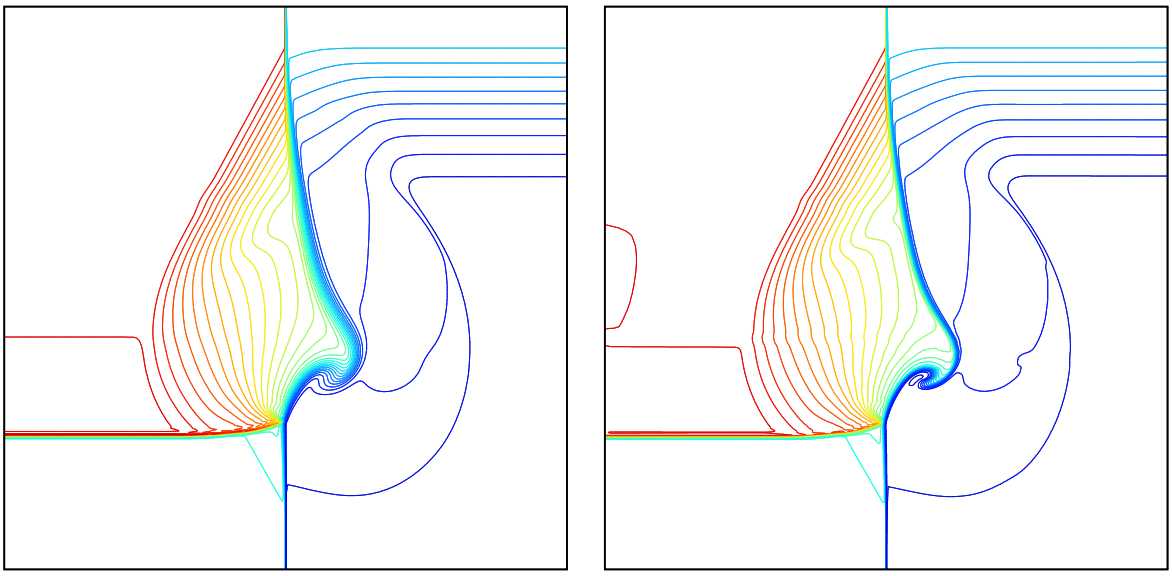
Figure A.18 – A two-dimensional Riemann problem : Configuration 18 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.2$.

Configuration 19–

$$\begin{array}{ccc} & J^+_{2,1} & \\ \overleftarrow{S}_{3,2} & & \vec{R}_{4,1} \\ & J^-_{3,4} & \end{array}$$

Corresponding initial data are

$$\begin{bmatrix} \rho_1 = 1 \\ p_1 = 1 \\ u_1 = 0 \\ v_1 = 0.3 \end{bmatrix} \begin{bmatrix} \rho_2 = 2 \\ p_2 = 1 \\ u_2 = 0 \\ v_2 = -0.3 \end{bmatrix} \begin{bmatrix} \rho_3 = 1.0625 \\ p_3 = 0.4 \\ u_3 = 0 \\ v_3 = 0.2145 \end{bmatrix} \begin{bmatrix} \rho_4 = 0.5197 \\ p_4 = 0.4 \\ u_4 = 0 \\ v_4 = -0.4259 \end{bmatrix}.$$



Upwind *MUSCL*

Figure A.19 – A two-dimensional Riemann problem : Configuration 19 in [44] – Comparison of the Upwind and MUSCL schemes – density at $t = 0.3$.

Annexe B

G-equation

B.1 Viscosity solutions of the eikonal equation

Consider the 1D problem (4.6) with the initial data

$$G_0(x) = \sin(\pi x), \quad x \in \Omega$$

where $\Omega = (0, 1)$. The boundary conditions remain Newmann homogenous. Consider a final time $T \leq \frac{1}{2}$. Then the viscosity solution at $t \leq T$ is given below :

$$G_{\text{visc}}(x,t) = \begin{cases} 0, & \forall x \in [0,t] \cup [1-t, 1], \\ \sin(\pi(x-t)), & \forall x \in [t, \frac{1}{2}], \\ \sin(\pi(x+t)), & \forall x \in [\frac{1}{2}, 1-t]. \end{cases} \tag{B.1}$$

Proof : First we consider a point x_0, t_0 on which the function G_{visc} is C^1. For example suppose that $x_0 \in (t_0, \frac{1}{2})$. In this case we have,

$$|\nabla G_{\text{visc}}(x_0, t_0)| = \pi \cos(\pi(x_0 - t_0)) = -\partial_t G_{\text{visc}}(x_0, t_0) \qquad \text{(Strong solution).}$$

Consequently if we consider $\phi \in C^1(\Omega)$ such that $G_{\text{visc}} - \phi$ has a local maximum on (x_0, t_0), then we have, thanks to the regularity of G_{visc} :

$$\partial_t G_{\text{visc}}(x_0, t_0) = \partial_T \phi(x_0, t_0) \qquad \nabla G_{\text{visc}}(x_0, t_0) = \nabla \phi(x_O, t_0),$$

so we get directly that

$$\partial_T \phi(x_0, t_0) + |\nabla \phi(x_O, t_0)| = 0.$$

We have the same result taking $\phi \in C^1$ such that $G_{\text{visc}} - \phi$ has a local minimum on (x_0, t_0).

In a similar way, if we take $x_0 \in (\frac{1}{2}, 1 - t_0)$, we have,

$$|\nabla G_{\text{visc}}(x_0, t_0)| = -\pi \cos(\pi(x_0 + t_0)) = -\partial_t G_{\text{visc}}(x_0, t_0) \qquad \text{(Strong solution).}$$

and the results remain true. We need to consider the discontinuity points of the derivatives of G_{visc}.

First consider that $x_0 = \frac{1}{2}$. At this point,

$$G_{\text{visc}}(\frac{1}{2}, t) = \sin(\pi(\frac{1}{2} - t)) = f(t),$$

which can be derived

$$f'(t) = -\pi \cos(\pi(\frac{1}{2} - t))$$

Let $\phi \in C^1$ such that $G_{\text{visc}} - \phi$ has a local maximum on $(\frac{1}{2}, t_0)$. We have, thanks to the regularity in t of the solution at $x_0 = \frac{1}{2}$,

$$\partial_t G_{\text{visc}}(\frac{1}{2}, t_0) = f'(t) = \partial_t \phi(\frac{1}{2}, t_0).$$

The function $[G_{\text{visc}} - \phi](., t_0)$ is $C1$ by piece and has a local maximum on $\frac{1}{2}$. As a result,

$$\partial_{x,g}\left(G_{\text{visc}} - \phi\right)(\frac{1}{2}, t_0) \geq 0, \qquad \partial_{x,d}\left(G_{\text{visc}} - \phi\right)(\frac{1}{2}, t_0) \leq 0.$$

ϕ being a function $C^1(\Omega)$, we have $\partial_{x,d}\phi(x_0, t_0) = \partial_{x,g}\phi(x_0, t_0) = \nabla\phi(x_0, t_0)$. We then have

$$\nabla\phi(x_0, t_0) \leq \partial_{x,g} G_{\text{visc}}(x_0, t_0) \text{ and } -\nabla\phi(x_0, t_0) \leq -\partial_{x,d} G_{\text{visc}}(x_0, t_0).$$

Now we use the fact that :

$$\partial_t G_{\text{visc}}(\frac{1}{2}, t_0) + \partial_{x,g} G_{\text{visc}}(\frac{1}{2}, t_0) = 0,$$
$$\partial_t G_{\text{visc}}(\frac{1}{2}, t_0) - \partial_{x,d} G_{\text{visc}}(\frac{1}{2}, t_0) = 0,$$

to get that :

$$\partial_t \phi(x_0, t_0) + \nabla\phi(x_0, t_0) \leq 0,$$
$$\partial_t \phi(x_0, t_0) - \nabla\phi(x_0, t_0) \leq 0,$$

so we have :

$$\partial_t \phi(x_0, t_0) + |\nabla\phi(x_0, t_0)| \leq 0.$$

Considering the case where $G_{\text{visc}} - \phi$ has a local minimum in $(\frac{1}{2}, t_0)$ is pointless as it is impossible to find such regular function ϕ. Indeed we would have $\partial_x\phi(x_0, t_0) \geq \partial_{x,g} G_{\text{visc}}(x_0, t_0)$ and $\partial_x\phi(x_0, t_0) \leq \partial_{x,d} G_{\text{visc}}(x_0, t_0)$. If G_{visc} is not C^1 then the two inequalities cannot be true at the same time. The other points of discontinuity are (t, t) and $(1 - t, t)$, $0 \leq t \leq T$. Let us focus on the first family of points, the results for the second coming out directly. They are local minimum of the function $G_{\text{visc}}(., t)$. Therefore we only consider $\phi \in C^1(\Omega \times [0, T])$ such that $G_{\text{visc}} - \phi$ has a local minimum on (t_0, t_0). The function of the single variable $f : t \longmapsto G_{\text{visc}}(t, t) - \phi(t, t)$ reaches a local minimum at t_0. As $G_{\text{visc}}(t, t) = 0, \forall t \in [0, T]$, f is a C^1 function and we have :

$$f'(t_0) = 0 = \partial_t \phi(t_0, t_0) + \nabla\phi(t_0, t_0).$$

Furthermore, considering $\epsilon > 0$, we have $G_{\text{visc}}(t_0, t_0 + \epsilon) = 0$ and then :

$$\lim_{\epsilon \to 0} \frac{G_{\text{visc}}(t_0, t_0 + \epsilon) - G_{\text{visc}}(t_0, t_0)}{\epsilon} = 0 = \partial_{t,d} G_{\text{visc}}(t_0, t_0).$$

We have $\partial_{t,d}(G_{\text{visc}} - \phi)(t_0, t_0) \geq 0$ (local minimum property), so we get that $\partial_t \phi(t_0, t_0) \leq 0$. Therefore $\nabla\phi(t_0, t_0) \geq 0$ and we can deduce that

$$\partial_t \phi(t_0, t_0) + |\nabla\phi(t_0, t_0)| = 0,$$

which concludes the proof. ■

One can see that this proof can be easily extended to the test case (4.30). Furthermore our 2D numerical example is actually a false 2D problem as it only involve one variable, the radius, in polar coordinates. One can proove that the solution given by (4.31) replacing x by r is the viscosity solution of the problem with the initial data (4.32).

Remark B.1

One can see that we considered the case where the final time T was less than $\frac{1}{2}$. It seems

obvious that the viscosity solution G_{visc} is equal to the null function if $t > \frac{1}{2}$.

The example given here for the sake of understanding can be extended to every $G_O \in BUC(\mathbb{R}^d)$. The viscosity solution is then defined on $\mathbb{R}^d \times (0, +\infty)$ by :

$$G(\boldsymbol{x}, t) = \inf_{|\boldsymbol{x}-\boldsymbol{y}|\leq t} G_0(\boldsymbol{y}). \tag{B.2}$$

The proof of this result can be found in [7], and it is based on the following lemma

Lemma B.1

Let us set

$$S(t)G(\boldsymbol{x}) = \inf_{|\boldsymbol{x}-\boldsymbol{y}|\leq t} G(\boldsymbol{y}).$$

Then S is a monotonous semigroup on $C(\mathbb{R}^d)$.

Proof : The proof is rather simple as

$$S(t) \circ S(s)G(\boldsymbol{x}) = \inf_{|\boldsymbol{x}-\boldsymbol{y}|\leq t} \left(\inf_{|\boldsymbol{z}-\boldsymbol{y}|\leq s} G(\boldsymbol{z}) \right).$$

This computation is equivalent to seek the infimum in the set

$$\{\boldsymbol{z} \text{ such that } \exists \boldsymbol{y} \text{ such that } |\boldsymbol{x}-\boldsymbol{y}| \leq t \text{ and } |\boldsymbol{z}-\boldsymbol{y}| \leq s\}.$$

Now, this set is equal to the set

$$\{\boldsymbol{z}\} \text{ such that } |\boldsymbol{x}-\boldsymbol{z}| \leq t+s,$$

so the infimum are equal and $S(t+s) = S(t) \circ S(s)$. Now consider G_1 and G_2 two functions of $C(\mathbb{R}^d)$ such that $G_1 \leq G_2$ and let $t > 0$. Thanks to the continuity of G_2, $\exists \boldsymbol{y}_{\boldsymbol{x},t} \in B(\boldsymbol{x}, t)$ such that $S(t)G_2(\boldsymbol{x}) = G_2(\boldsymbol{y}_{\boldsymbol{x},t})$. Consequently $G_2(\boldsymbol{y}_{\boldsymbol{x},t}) \geq G_1(\boldsymbol{y}_{\boldsymbol{x},t}) \geq S(t)G_1(\boldsymbol{x})$, which concludes the proof. ■

Now let $\phi \in C^1(\mathbb{R}^d \times (0, +\infty))$ and suppose that $(\boldsymbol{x}, t)$ is a local maximum of $G - \phi$. Thanks to the semigroup property of S we get that :

$$G(\boldsymbol{x}, t) = S(t)G_0(\boldsymbol{x}) = S(h)S(t-h)G_0(\boldsymbol{x}) = S(h)G(\boldsymbol{x}, t-h).$$

Therefore, for all $0 < h < t$, we have

$$G(\boldsymbol{x}, t) = \inf_{|\boldsymbol{x}-\boldsymbol{y}|\leq h} G(\boldsymbol{y}, t-h). \tag{B.3}$$

$(\boldsymbol{x}, t)$ being a local maximum of $G - \phi$, we have, if h is sufficiently small, and $|\boldsymbol{x}-\boldsymbol{y}| \leq h$:

$$G(\boldsymbol{y}, t-h) - \phi(\boldsymbol{y}, t-h) \leq G(\boldsymbol{x}, t) - \phi(\boldsymbol{x}, t),$$

which is equivalent to

$$G(\boldsymbol{y}, t-h) \leq G(\boldsymbol{x}, t) - \phi(\boldsymbol{x}, t) + \phi(\boldsymbol{y}, t-h).$$

Injecting this in (B.3) leads to

$$\phi(\boldsymbol{x}, t) \leq \inf_{|\boldsymbol{x}-\boldsymbol{y}|\leq h} \phi(\boldsymbol{y}, t-h).$$

A first order Taylor expansion at the point $(\boldsymbol{x}, t)$ leads to

$$0 \leq \inf_{|\boldsymbol{x}-\boldsymbol{y}|\leq h} \left[-\partial_t \phi(\boldsymbol{x}, t) + \nabla\phi(\boldsymbol{x}, t) \cdot \frac{\boldsymbol{y}-\boldsymbol{x}}{h} + o(1) \right].$$

Using that fact that $-\inf(-) = \sup()$, we have

$$\partial_t \phi(\boldsymbol{x}, t) + \sup_{|\boldsymbol{x}-\boldsymbol{y}|\leq h} \nabla\phi(\boldsymbol{x}, t) \cdot \frac{\boldsymbol{x}-\boldsymbol{y}}{h} + o(1) \leq 0.$$

Thanks to the Cauchy-Schwarz inequality :

$$|\nabla\phi(x,t)\cdot\frac{x-y}{h}| \leq |\nabla\phi(x,t)|.$$

By taking $y = x - \frac{\nabla\phi(x,t)}{|\nabla\phi(x,t)|}h$ we see that the previous upper-bound is reached. Therefore ,

$$\partial_t\phi(x,t) + |\nabla\phi(x,t)| + o(1) \leq 0,$$

and passing to the limit when $h \to 0$ leads to the desired result.

B.2 Additional Properties of the scheme

The convergence results are obtained for the upwind scheme in the Cartesian case. Issues come from the monotonicity of the scheme and the strong consistency of the discrete spatial operator on unstructured meshes. However it satisfies a weaker consistency result.

Definition B.1 (*Weak consistency*)

Let $F(G)$ be an operator approximated by $F_{\mathcal{M}}(G)$. Let $h_{\mathcal{M}} = \max_{K\in\mathcal{M}} \text{diam}(K)$. Let $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}\right\}$ be a sequence of discretizations such that the size $h_{\mathcal{M}}^{(m)}$ tends to zero as $m \to \infty$. The discrete spatial operator $F_{\mathcal{M}}$ is said to be weakly consistent with F if for every $\phi, \psi \in C_c^\infty(\Omega)$:

$$\lim_{m\to\infty} \int_\Omega F_{\mathcal{M}^{(m)}}(\psi_{\mathcal{M}^{(m)}})\phi_{\mathcal{M}^{(m)}} = \int_\Omega F(\nabla\psi)\phi$$

Lemma B.2

Let $\nabla_{\mathcal{E}}^w$ be a discrete gradient operator defined as follows :

$$\text{For } \phi_{\mathcal{M}} \in H_{\mathcal{M}}, \qquad \nabla_{\mathcal{E}}^w\phi_{\mathcal{M}} = \sum_{\sigma\in\mathcal{E}} \frac{|\sigma|}{|D_\sigma|}(\phi_L - \phi_K)\boldsymbol{n}_{K,\sigma}\mathcal{X}_{D_\sigma}.$$

Let $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}\right\}$ be a sequence of discretizations such that the size $h^{(m)}$ of the mesh $\mathcal{M}^{(m)}$ tend to zero as $m \to \infty$. For $m \in \mathbb{N}$, let $q^{(m)} \in H_{\mathcal{M}^{(m)}}$ and assume that there exists $C \in \mathbb{R}$ such that, for all $m \in \mathbb{N}$, $\|\nabla_{\mathcal{E}^{(m)}}^w q^{(m)}\|_{L^p(\Omega)^d} \leq C$ for some $1 \leq p < \infty$. Assume also that there exists $\overline{q} \in L^p(\Omega)$ such that $q^{(m)}$ converges strongly in $L^p(\Omega)$ towards $\overline{q}$ as $m \to \infty$. Then $\overline{q} \in W_0^{1,p}(\Omega)$ and $\nabla_{\mathcal{E}^{(m)}}^w q^{(m)}$ converges weakly in $L^p(\Omega)^d$ towards $\nabla\overline{q}$ as $m \to \infty$.

The proof of the lemma can be found here [43].

We can now formulate the main proposition of this section.

Proposition B.3

Let $F_{\mathcal{M}}$ be the spatial operator of our scheme defined in (4.14). It is weakly consistent with $F(G) = |\nabla G|$.

Proof : We will proove this proposition with the MUSCL interpolation, the upwind interpolation just being a particular case. Let $\phi, \psi \in C_c^\infty(\Omega)$ and $\phi_{\mathcal{M}}, \psi_{\mathcal{M}}$ their interpolation. Let $\mathcal{D}^{(m)} = \left\{\mathcal{M}^{(m)}, \mathcal{E}^{(m)}, \mathcal{P}^{(m)}\right\}$ be a sequence of discretization such that $h_{\mathcal{M}}^{(m)} \to 0$ as $m \to \infty$. For the sake of simplicity we will omit the subscript m. We have :

$$\int_\Omega F_{\mathcal{M}}(\phi_{\mathcal{M}})\psi_{\mathcal{M}} = \sum_{K\in\mathcal{M}} \sum_{\sigma=K|L\in\mathcal{E}(K)} |\sigma| \frac{(\phi_L - \phi_K)}{\sqrt{(\phi_L - \phi_K)^2 + d_\sigma^2|\nabla_{//\sigma}\phi_{\mathcal{M}}|^2}}(\phi_\sigma - \phi_K)\psi_K.$$

Reordering the sums leads to :

$$\int_\Omega F_\mathcal{M}(\phi_\mathcal{M})\psi_\mathcal{M} = \sum_{\sigma\in\mathcal{E}_{\text{int}}} |\sigma| \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot \boldsymbol{n}_{K,\sigma}\phi_\sigma(\psi_K - \psi_L) - \sum_{\sigma\in\mathcal{E}_{\text{int}}} |\sigma| \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot \boldsymbol{n}_{K,\sigma}(\phi_K\psi_K - \phi_L\psi_L).$$

We can rewrite these sums :

$$\int_\Omega F_\mathcal{M}(\phi_\mathcal{M})\psi_\mathcal{M} = -\sum_{\sigma\in\mathcal{E}_{\text{int}}} |D_\sigma|\phi_\sigma \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot (\nabla^w_\mathcal{E}\psi_\mathcal{M})_\sigma + \sum_{\sigma\in\mathcal{E}_{\text{int}}} |D_\sigma| \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot (\nabla^w_\mathcal{E}\psi_\mathcal{M}\phi_\mathcal{M})_\sigma.$$

We set :

$$T_1^{(m)} = -\sum_{\sigma\in\mathcal{E}_{\text{int}}} |D_\sigma|\phi_\sigma \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot (\nabla^w_\mathcal{E}\psi_\mathcal{M})_\sigma,$$

$$T_2^{(m)} = \sum_{\sigma\in\mathcal{E}_{\text{int}}} |D_\sigma| \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot (\nabla^w_\mathcal{E}\psi_\mathcal{M}\phi_\mathcal{M})_\sigma.$$

The convergence of the second term comes directly. Indeed we use the strong consistency of $\nabla_\mathcal{E}$ together with the weak convergence of $\nabla^w_\mathcal{E}$. Consequently we get that :

$$\lim_{m\to\infty} T_2^{(m)} = \lim_{m\to\infty} \int_\Omega \frac{\nabla_\mathcal{E}\phi_\mathcal{M}}{|\nabla_\mathcal{E}\phi_\mathcal{M}|} \cdot \nabla^w_\mathcal{E}\psi_\mathcal{M}\phi_\mathcal{M} = \int_\Omega \frac{\nabla\phi}{|\nabla\phi|} \cdot \nabla(\psi\phi).$$

We split $T_1^{(m)}$ in two terms $T_1^{(m)} = T_{1,1}^{(m)} + R_{1,1}^{(m)}$ with :

$$T_{1,1}^{(m)} = -\sum_{\sigma\in\mathcal{E}_{\text{int}}} \left(|D_{K,\sigma}|\phi_K + |D_{L,\sigma}|\phi_L\right) \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot (\nabla^w_\mathcal{E}\psi_\mathcal{M})_\sigma,$$

$$R_{1,1}^{(m)} = -\sum_{\sigma\in\mathcal{E}_{\text{int}}} \left[|D_{K,\sigma}|(\phi_\sigma - \phi_K) + |D_{L,\sigma}|(\phi_\sigma - \phi_L)\right] \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot (\nabla^w_\mathcal{E}\psi_\mathcal{M})_\sigma.$$

For the first term $T_{1,1}^{(m)}$, we have :

$$T_{1,1}^{(m)} = -\int_\Omega \phi_\mathcal{M} \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})}{|(\nabla_\mathcal{E}\phi_\mathcal{M})|} \cdot (\nabla^w_\mathcal{E}\psi_\mathcal{M}).$$

Passing to the limit in the term leads to :

$$\lim_{m\to\infty} T_{1,1}^{(m)} = -\int_\Omega \phi \frac{\nabla\phi}{|\nabla\phi|} \cdot \nabla\psi$$

Thanks to the MUSCL interpolation, we have :

$$\text{For } \sigma = K|L \in \mathcal{E}_{\text{int}}, \quad \phi_\sigma = \beta_{K,\sigma}\phi_K + (1-\beta_{K,\sigma})\phi_L, \quad \beta_{K,\sigma} \in [0,1]$$

Consequently we get :

$$R_{1,1}^{(m)} = -\sum_{\sigma\in\mathcal{E}_{\text{int}}} \left[|D_{K,\sigma}|(1-\alpha_{K,\sigma})(\phi_L - \phi_K) + |D_{L,\sigma}|\alpha_{K,\sigma}(\phi_K - \phi_L)\right] \frac{(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma}{|(\nabla_\mathcal{E}\phi_\mathcal{M})_\sigma|} \cdot (\nabla^w_\mathcal{E}\psi_\mathcal{M})_\sigma.$$

We can bound the term as follows :

$$|R_{1,1}^{(m)}| \le h_\mathcal{M} C_{\phi,\psi},$$

where $C_{\phi,\psi}$ is a constant depending only on ϕ and ψ. As a result :

$$\lim_{m\to\infty} R_{1,1}^{(m)} = 0.$$

If we gather the results :

$$\lim_{m\to\infty} \int_\Omega F_\mathcal{M}(\phi_\mathcal{M})\psi_\mathcal{M} = -\int_\Omega \phi \frac{\nabla\phi}{|\nabla\phi|} \cdot \nabla\psi + \int_\Omega \frac{\nabla\phi}{|\nabla\phi|} \cdot \nabla(\psi\phi)$$

We recall that :

$$\nabla(\psi\phi) = \psi\nabla\phi + \phi\nabla\psi,$$

so we get that :

$$\lim_{m\to\infty} \int_\Omega F_{\mathcal{M}}(\phi_{\mathcal{M}})\psi_{\mathcal{M}} = \int_\Omega \psi\nabla\phi \cdot \frac{\nabla\phi}{|\nabla\phi|} = \int_\Omega |\nabla\phi|\psi,$$

which concludes the proof. ■

Bibliographie

[1] R. Abgrall. Numerical discretization of the first-order Hamilton-Jacobi equation on triangular meshes. *Comm. Pure Appl. Math.*, 49(12) :1339–1373, 1996.

[2] Rémi Abgrall and Sophie Dallet. An asymptotic preserving scheme for the barotropic Baer-Nunziato model. In *Finite volumes for complex applications. VII. Elliptic, parabolic and hyperbolic problems*, volume 78 of *Springer Proc. Math. Stat.*, pages 749–757. Springer, Cham, 2014.

[3] G. Ansanay-Alex, F. Babik, J.-C. Latché, and D. Vola. An L^2-stable approximation of the Navier-Stokes convection operator for low-order non-conforming finite elements. *International Journal for Numerical Methods in Fluids*, 66 :555–580, 2011.

[4] G. Ansanay-Alex, F. Babik, J.-C. Latché, and D. Vola. An L^2-stable approximation of the Navier-Stokes convection operator for low-order non-conforming finite elements. *International Journal for Numerical Methods in Fluids*, 66 :555–580, 2011.

[5] Steeve Augoula and Rémi Abgrall. High order numerical discretization for Hamilton-Jacobi equations on triangular meshes. *J. Sci. Comput.*, 15(2) :197–229, 2000.

[6] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Anal.*, 4(3) :271–283, 1991.

[7] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Paris, 1994.

[8] T. Barth and M. Ohlberger. Finite volume methods : foundation and analysis. In E. Stein, R. de Borst, and T.J.R. Hughes, editors, *Encyclopedia of Computational Mechanics, Volume I*, chapter 15. John Wiley & Sons, 2004.

[9] Timothy J. Barth and James A. Sethian. Numerical schemes for the Hamilton-Jacobi and level set equations on triangulated domains. *J. Comput. Phys.*, 145(1) :1–40, 1998.

[10] F. Berthelin, T. Goudon, and S. Minjeaud. Kinetic schemes on staggered grids for barotropic Euler models : entropy-stability analysis. *Mathematics of Computation*, 84 :2221–2262, 2015.

[11] C. Berthon, Y. Coudière, and V. Desveaux. Second-order MUSCL schemes based on Dual Mesh Gradient Reconstruction (DMGR). *Mathematical Modelling and Numerical Analysis*, 48 :583–602, 2014.

[12] F. Bouchut. *Nonlinear Stability of finite volume methods for hyperbolic conservation laws*. Birkhauser, 2004.

[13] Steve Bryson and Doron Levy. High-order central WENO schemes for multidimensional Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 41(4) :1339–1369 (electronic), 2003.

[14] T. Buffard and S. Clain. Monoslope and multislope MUSCL methods for unstructured meshes. *Journal of Computational Physics*, 229 :3745–3776, 2010.

[15] C. Calgaro, E. Chane-Kane, E. Creusé, and T. Goudon. L^∞-stability of vertex-based MUSCL finite volume schemes on unstructured grids : Simulation of incompressible flows with high density ratios. *Journal of Computational Physics*, 229 :6027–6046, 2010.

[16] CALIF^{3}S. A software components library for the computation of reactive turbulent flows. `https://gforge.irsn.fr/gf/project/isis`.

[17] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An all-regime Lagrange-Projection like scheme for the gas dynamics equations on unstructured meshes. June 2014.

[18] A.J. Chorin. Numerical solution of the Navier-Stokes equations. *Mathematics of Computation*, 22 :745–762, 1968.

[19] P. G. Ciarlet. Basic error estimates for elliptic problems. In P. Ciarlet and J.L. Lions, editors, *Handbook of Numerical Analysis, Volume II*, pages 17–351. North Holland, 1991.

[20] M. G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Math. Comp.*, 43(167) :1–19, 1984.

[21] Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 277(1) :1–42, 1983.

[22] M. Crouzeix and P.A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. *RAIRO Série Rouge*, 7 :33–75, 1973.

[23] R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI : a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 30(4) :1009–1043, 2010.

[24] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. In *Handbook of numerical analysis, Vol. VII*, Handb. Numer. Anal., VII, pages 713–1020. North-Holland, Amsterdam, 2000.

[25] L. Gastaldo, R. Herbin, W. Kheriji, C. Lapuerta, and J.-C. Latché. Staggered discretizations, pressure correction schemes and all speed barotropic flows. In *Finite Volumes for Complex Applications VI - Problems and Perspectives - Prague, Czech Republic*, volume 2, pages 39–56, 2011.

[26] L. Gastaldo, R. Herbin, J.-C. Latché, and N. Therme. Consistency of some staggered schemes for the Euler equations. *in preparation*, 2015.

[27] Laura Gastaldo, Raphaèle Herbin, and Jean-Claude Latché. A discretization of the phase mass balance in fractional step algorithms for the drift-flux model. *IMA J. Numer. Anal.*, 31(1) :116–146, 2011.

[28] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*. Springer, 1996.

[29] D. Grapsas, R. Herbin, W. Kheriji, and J.-C. Latché. An unconditionally stable staggered pressure correction scheme for the compressible Navier-Stokes equations. *under revision*, 2015.

[30] J.L Guermond and R. Pasquetti. Entropy-based nonlinear viscosity for Fourier approximations of conservation laws. *Comptes Rendus de l'Académie des Sciences de Paris – Série I – Analyse Numérique*, 346 :801–806, 2008.

[31] J.L Guermond, R. Pasquetti, and B. Popov. Entropy viscosity method for nonlinear conservation laws. *Journal of Computational Physics*, 230 :4248–4267, 2011.

[32] F.H. Harlow and A.A. Amsden. Numerical calculation of almost incompressible flow. *Journal of Computational Physics*, 3 :80–93, 1968.

[33] F.H. Harlow and A.A. Amsden. A numerical fluid dynamics calculation method for all flow speeds. *Journal of Computational Physics*, 8 :197–213, 1971.

[34] F.H. Harlow and J.E. Welsh. Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Physics of Fluids*, 8 :2182–2189, 1965.

[35] R. Herbin, W. Kheriji, and J.-C. Latché. An unconditionally stable pressure correction scheme for compressible Navier-Stokes equations. *submitted*, 2013.

[36] R. Herbin, W. Kheriji, and J.-C. Latché. On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations. *ESAIM : Mathematical Modelling and Numerical Analysis*, 2014. published on line.

[37] R. Herbin and J.-C. Latché. Kinetic energy control in the MAC discretization of the compressible Navier-Stokes equations. *International Journal of Finites Volumes*, 7, 2010.

[38] R. Herbin, J.-C. Latché, and T.T. Nguyen. Explicit staggered schemes for the compressible Euler equations. *ESAIM : Proceedings*, 40 :83–102, 2013.

[39] R. Herbin, J.-C. Latché, and T.T. Nguyen. On some consistent explicit staggered schemes for the shallow water and Euler equations. *under revision*, 2015.

[40] ISIS. A CFD computer code for the simulation of reactive turbulent flows. `https://gforge.irsn.fr/gf/project/isis`.

[41] G. Kossioris, Ch. Makridakis, and P. E. Souganidis. Finite volume schemes for Hamilton-Jacobi equations. *Numer. Math.*, 83(3) :427–442, 1999.

[42] A. Kurganov and Y. Liu. New adaptative artificial viscosity method for hyperbolic systems of conservation laws. *Journal of Computational Physics*, 231 :8114–8132, 2012.

[43] J.-C. Latché and K. Saleh. A convergent staggered scheme for variable density incompressible Navier-Stokes equations. *submitted*, 2014.

[44] P.D. Lax and X.-D. Liu. Solution of two dimensional riemann problem of gas dynamics by positive schemes. *SIAM J. Sci. Comput*, 19 :319–340, 1995.

[45] C. Le Touze, A. Murrone, and H. Guillard. Multislope MUSCL method for general unstructured meshes. *Journal of Computational Physics*, 284 :389–418, 2015.

[46] Pierre-Louis Lions. *Generalized solutions of Hamilton-Jacobi equations*, volume 69 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, Mass.-London, 1982.

[47] M.-S. Liou. A sequel to AUSM, part II : AUSM+-up. *Journal of Computational Physics*, 214 :137–170, 2006.

[48] M.-S. Liou and C.J. Steffen. A new flux splitting scheme. *Journal of Computational Physics*, 107 :23–39, 1993.

[49] K. Nerinckx, J. Vierendeels, and E. Dick. Mach-uniformity through the coupled pressure and temperature correction algorithm. *Journal of Computational Physics*, 206 :597–623, 2005.

[50] K. Nerinckx, J. Vierendeels, and E. Dick. A Mach-uniform algorithm : coupled versus segregated approach. *Journal of Computational Physics*, 224 :314–331, 2007.

[51] Stanley Osher and James A. Sethian. Fronts propagating with curvature-dependent speed : algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.*, 79(1) :12–49, 1988.

[52] Stanley Osher and Chi-Wang Shu. High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 28(4) :907–922, 1991.

[53] L. Piar, F. Babik, R. Herbin, and J.-C. Latché. A formally second order cell centered scheme for convection-diffusion equations on general grids. *International Journal for Numerical Methods in Fluids*, 71 :873–890, 2013.

[54] R. Rannacher and S. Turek. Simple nonconforming quadrilateral Stokes element. *Numerical Methods for Partial Differential Equations*, 8 :97–111, 1992.

[55] M. Schäfer and S. Turek. Benchmark computations of laminar flow around a cylinder. *Notes on Numerical Fluid Mechanics*, 52, 1996.

[56] Susana Serna and Jianliang Qian. Fifth-order weighted power-ENO schemes for Hamilton-Jacobi equations. *J. Sci. Comput.*, 29(1) :57–81, 2006.

[57] J. A. Sethian and A. Vladimirsky. Fast methods for the eikonal and related Hamilton-Jacobi equations on unstructured meshes. *Proc. Natl. Acad. Sci. USA*, 97(11) :5699–5703, 2000.

[58] Panagiotis E. Souganidis. Approximation schemes for viscosity solutions of Hamilton-Jacobi equations. *J. Differential Equations*, 59(1) :1–43, 1985.

[59] J.L. Steger and R.F. Warming. Flux vector splitting of the inviscid gaz dynamics equations with applications to finite difference methods. *Journal of Computational Physics*, 40 :263–293, 1981.

[60] R. Temam. Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires II. *Arch. Rat. Mech. Anal.*, 33 :377–385, 1969.

[61] N. Therme, L. Gastaldo, R. Herbin, and J.-C. Latché. Stable explicit staggered schemes with MUSCL and artificial viscosity techniques for the Euler equations. *in preparation*, 2015.

[62] E. Toro. *Riemann solvers and numerical methods for fluid dynamics – A practical introduction (third edition)*. Springer, 2009.

[63] E.F. Toro and M.E. Vázquez-Cendón. Flux splitting schemes for the Euler equations. *Computers & Fluids*, 70 :1–12, 2012.

[64] P. Woodward and P. Colella. The numerical simulation of two-dimensional fluid flow with strong shocks. *Journal of Computational Physics*, 54 :115–173, 1984.

[65] Jue Yan and Stanley Osher. A local discontinuous Galerkin method for directly solving Hamilton-Jacobi equations. *J. Comput. Phys.*, 230(1) :232–244, 2011.

[66] G.-C. Zha and E. Bilgen. Numerical solution of Euler equations by a new flux vector splitting scheme. *International Journal for Numerical Methods in Fluids*, 17 :115–144, 1993.