Discrétisation en maillage non structuré général et applications LES

Florian Haider

Thèse de Doctorat
de l'Université Pierre et Marie Curie – Paris 6

École doctorale de Sciences Mécaniques, Acoustique et Électronique de Paris

Spécialité : Mécanique

Présentée par

Florian HAIDER

Pour obtenir le grade de

Docteur de l'Université Pierre et Marie Curie

Sujet de la thèse :

Discrétisation en maillage non structuré général et applications LES

soutenue le 29 mai 2009 devant le jury composé de :

M. R. Abgrall	Univ. Bordeaux I	*Rapporteur*
M. P. Brenner	Eads	*Membre invité*
M. B. Courbet	Onera/Dsna	*Examinateur*
M. J.-P. Croisille	Univ. Paul Verlaine Metz	*Directeur de thèse*
M. Ch. Hirsch	Vrije Universiteit Brussel	*Rapporteur*
M. J.-M. Le Gouez	Onera/Dsna	*Membre invité*
M. B. Perthame	Univ. Pierre et Marie Curie	*Examinateur*
M. P. Sagaut	Univ. Pierre et Marie Curie	*Directeur de thèse*

Version du 23 juin 2009

Discrétisation en maillage non structuré général et applications LES

L'objectif est d'améliorer la stabilité et la précision de la discrétisation spatiale de type volumes finis sur des maillages non structurés. L'intérêt réside dans l'application croissante des volumes finis à la simulation des grandes échelles (LES) qui exige une discrétisation précise. Un autre objectif est le développement d'algorithmes permettant de reconstruire les polynômes de degré élevé en maillage non structuré sur de petits voisinages (stencils).

L'étude commence par une analyse générale de la reconstruction des polynômes de degré k en maillage non structuré et une étude numérique de la convergence en maillage pour la reconstruction des polynômes de degré 2 et 3. L'étude présente plusieurs algorithmes permettant de reconstruire des polynômes sur de petits voisinages. Des expériences numériques confirment l'ordre d'approximation de ces méthodes pour les polynômes de degré 2 en dimension 2.

L'étude théorique de la stabilité dégage des principes généraux pour concevoir des méthodes de reconstruction stables. L'étude théorique de la précision caractérise les erreurs induites par le maillage non structuré à l'aide de l'approche de l'équation modifiée. Ces études sont complétées par des expériences numériques.

L'étude formule également des algorithmes de limitation en maillage non structuré basés sur une approche géométrique.

Les calculs LES d'un écoulement subsonique au-dessus d'une cavité et d'un jet supersonique permettent de valider et comparer plusieurs options de discrétisation spatiale implémentées dans le code Cedre de l'Onera. Les résultats de l'étude de stabilité permettent d'améliorer le calcul du jet en maillage de tétraèdres.

Mots-clés : discrétisation spatiale, méthode de volumes finis, Muscl, maillage non structuré général, lois de conservation, mécanique des fluides compressibles, simulation des grandes échelles, SGE, turbulence ;

Discretization on general unstructured grids and applications to LES

The objective is to improve the stability and accuracy of finite volume spatial discretization on unstructured grids. The interest lies in the growing use of finite volumes for large eddy simulation (LES) that requires accurate discretization methods. Another goal is the design of algorithms capable of reconstructing polynomials of higher degree on unstructured grids using only small and compact stencils.

The study starts with a general analysis of the reconstruction of polynomials of degree k on unstructured grids, completed by numerical measurements of the convergence rate of the reconstruction error for polynomials of degree 2 and 3. The study presents algorithms for the reconstruction of polynomials on small stencils. Numerical experiments confirm the order of the approximation of these reconstruction methods for quadratic polynomials in 2 dimensions.

A theoretical stability analysis exhibits general principles for the design of stable reconstruction methods. A theoretical accuracy analysis, based on the modified equation approach, highlights the errors induced by unstructured grids. The theoretical investigations are completed and confirmed by numerical experiments.

The study of slope limiters on unstructured grids formulates algorithms based on a geometric approach.

Large eddy simulations of a subsonic flow over a cavity and of a supersonic jet allow the validation and comparison of several discretization features implemented in the code Cedre of Onera.The results of the theoretical stability analysis make it possible to obtain better results for the jet computation on tetrahedral grids.

Keywords: spatial discretization, finite volume method, Muscl, general unstructured grids, conservation laws, compressible fluid dynamics, large eddy simulation, LES, turbulence;

Les travaux de recherche ont été effectués au
Département de Simulation Numérique et Aéroacoustique de l'Onera
BP72 – 29 avenue de la Division Leclerc 92322 Châtillon Cedex
Tel +33 1 46 73 40 40 Fax +33 1 46 73 41 41
http://www.onera.fr/dsna/index.php

Remerciements

Je souhaite adresser ici tous mes remerciements aux personnes qui m'ont apporté leur aide et qui ont ainsi contribué à l'élaboration de ce mémoire. Je remercie très chaleureusement Bernard Courbet dont les idées sont à l'origine d'une grande partie de ce travail. Je remercie mes deux directeurs de thèse, Pierre Sagaut et Jean-Pierre Croisille, pour leur soutien et leurs encouragements. J'exprime également ma gratitude à tous les membres du jury et aux deux rapporteurs, Rémi Abgrall et Charles Hirsch, pour leurs remarques et commentaires dont j'essaierai de tenir compte dans mes futurs travaux.

Les longues discussions avec Jean-Pierre Croisille et Bernard Courbet ont été très importantes pour le travail sur les méthodes numériques. Un très grand merci à Pierre Brenner pour le nombre incroyable d'idées et de conseils, tout particulièrement pour la méthode des corrections successives analysée et testée dans le cadre de cette thèse. Je tiens à souligner ici que la méthode des moindres carrés couplés repose sur des idées de Bernard Courbet qui sont le produit de sa riche expérience de numéricien à l'ONERA.

Je remercie très chaleureusement tous mes collègues de l'unité CIME de l'ONERA pour leur collaboration et leur aide pendant ces années de thèse. J'ai rencontré des personnes très compétentes et engagées avec lesquelles c'est un plaisir de travailler. J'exprime également toute ma reconnaissance à l'ensemble du personnel du DSNA pour l'accueil chaleureux et particulièrement à son directeur Jean-Marie Le Gouez pour l'intérêt qu'il a porté à cette thèse. Je n'oublie pas les développeurs et les utilisateurs de CEDRE au sein du DEFA, du DMAE et des autres départements pour leur contribution au projet CEDRE qui est d'une certaine façon à l'origine des questions traitées ici.

Je suis redevable au personnel administratif du DSNA, et en particulier à Ghislaine Denis, pour leur disponibilité et leur gentillesse. Un facteur absolument crucial pour la réussite de cette thèse a été l'aide apportée par le personnel informatique du DSNA et de l'ONERA, sans lequel je n'aurais pu effectuer ni les développements dans CEDRE ni les calculs de simulation.

Je suis également reconnaissant envers le personnel de la documentation, de l'imprimerie et des autres services de l'ONERA qui m'ont aidé dans la réalisation de ce travail.

Un grand merci à Élise Callet pour les corrections d'orthographe. Toutes les erreurs restantes sont de ma responsabilité car j'ai retouché le texte après ses corrections.

Il est particulièrement important pour moi de remercier toutes les personnes – au niveau de l'unité, des départements et de la Direction Générale de l'ONERA – qui se sont investies en faveur de mon embauche à l'ONERA. Celle-ci me permettra de poursuivre mes travaux de recherche et d'implémenter les méthodes développées ici dans CEDRE. L'objectif final est de disposer de méthodes suffisamment robustes et précises pour pouvoir effectuer des calculs de simulation des grandes échelles de qualité satisfaisante sur des maillages non structurés généraux.

Je remercie Benny, mes parents, ma sœur et mon grand-père pour leurs encouragements. Pour finir, je remercie tout particulièrement mon épouse Rosa pour son soutien solide et sa patience sans limites tout au long de cette aventure.

Table des matières

Glossaire

$\mathcal{A}_{\alpha\beta}$	Face entre les cellules numéro α et β, page 44
$\mathcal{T}_\alpha$	Cellule numéro α, page 43
$\boldsymbol{a}_{\alpha\beta}$	Vecteur surface de la face $\mathcal{A}_{\alpha\beta}$, page 45
h	Diamètre maximal des cellules, page 43
$\boldsymbol{\nu}_{\alpha\beta}$	Vecteur normal unitaire de la face $\mathcal{A}_{\alpha\beta}$, page 45
CS	Méthode des corrections successives, page 115
CSE	Méthode des corrections successives sur un voisinage élargi, page 127
LES	Large eddy simulation (Simulation des grandes échelles), page 29
MCCIE	Méthodes des moindres carrés couplés par itération sur un voisinage élargi, page 126
MCCI	Méthode des moindres carrés couplés par itération, page 108
MUSCL	Monotone Upstream-centered Scheme for Conservation Laws, page 13
SGE	Simulation des grandes échelles, page 29

CHAPITRE 1

Introduction

1.1. Contexte et objectif

Les schémas volumes finis sont aujourd'hui très répandus pour résoudre des systèmes de lois de conservation telles que les équations de Navier-Stokes. Les travaux de Bram van Leer dans [**105, 106, 107, 108, 109**] ont sensiblement augmenté la précision de l'approche initiale de Sergei K. Godounov [**53**]. L'idée principale est l'introduction de reconstructions polynomiales qui permettent une évaluation plus précise des flux aux interfaces entre les cellules. L'utilisation de limiteurs assure une reconstruction monotone afin de supprimer les oscillations artificielles typiques des schémas d'ordre élevé. La famille de méthodes qui utilise une reconstruction linéaire par cellule est aujourd'hui connue sous le sigle MUSCL, pour *Monotone Upstream-centered Scheme for Conservation Laws*.

La simplicité de ces méthodes a favorisé leur utilisation dans des applications industrielles qui impliquent souvent des écoulements dans des géométries complexes. Pour cette raison, il est devenu nécessaire d'étendre les méthodes de type MUSCL aux maillages non structurés généraux qui sont mieux adaptés pour mailler ce type de géométrie. Un autre avantage de ces maillages est de faciliter le découpage du maillage en domaines pour la parallélisation des traitements.

Un exemple d'application est le logiciel d'énergétique CEDRE développé par l'ONERA. La chaîne de calculs comprend le code CEDRE proprement dit, lui-même composé de plusieurs solveurs (fluide, particules, conduction, rayonnement) ainsi que de pré- et post-traitements. Ce code résout les équations de Navier-Stokes ou d'Euler pour des mélanges d'espèces décrites par des lois d'état générales.

La turbulence est un phénomène physique important en mécanique des fluides et il est par conséquent important qu'un code industriel et scientifique comme CEDRE prenne en compte les effets de la turbulence sur l'écoulement. Puisque le coût de la simulation directe de la turbulence est encore trop élevé, des modèles simplifiés de la turbulence, comme l'approche RANS, sont largement utilisés en pratique. Ces modèles sont surtout dédiés aux écoulements stationnaires et leur faible coût de calcul s'obtient par une résolution beaucoup moins fine des échelles de l'écoulement.

La demande croissante pour des simulations instationnaires précises a encouragé le développement de modèles qui se situent entre la simulation directe de la turbulence et l'approche RANS. Une famille importante de modèles est connue sous le nom de *simulation des grandes échelles*, appelée LES ou *Large Eddy Simulation* en anglais. Une description détaillée de la plupart des modèles de la simulation des grandes échelles se trouve dans l'ouvrage [**92**]. L'augmentation des performances de calcul a commencé à rendre cette méthode intéressante pour les applications industrielles et pour le code CEDRE.

Les modèles de simulation des grandes échelles nécessitent cependant des méthodes précises de discrétisation spatiale. Certains travaux scientifiques, cf. [**17**] et [**81**], ont démontré que les méthodes de discrétisation spatiale de CEDRE ne sont pas suffisamment précises pour effectuer des simulations des grandes échelles de façon satisfaisante en maillage non structuré. Selon les options de la discrétisation spatiale, les simulations rencontrent les problèmes suivants :

(1) La méthode de discrétisation spatiale dégrade la précision de façon excessive.

(2) Le schéma montre des problèmes de stabilité et le calcul s'arrête, en général en raison d'une température négative.

Afin de remédier aux problèmes cités ci-dessus, la présente étude vise à améliorer la robustesse et la précision des schémas classiques de volumes finis en maillage non structuré. Plus précisément, cette thèse poursuit deux objectifs complémentaires :

(1) Le premier objectif, détaillé dans la section 1.2, est double : il s'agit d'abord d'identifier les méthodes de reconstruction qui maximisent la stabilité du schéma. L'idée derrière cette approche est d'éviter que la stabilité du schéma ne dépende uniquement des limiteurs. Les limiteurs dégradent par définition la précision de la discrétisation spatiale et, pour cette raison, leur utilisation doit être restreinte au minimum. Cela suppose en contrepartie que la stabilité du schéma ne soit pas uniquement assurée par les limiteurs. Ensuite, il faut améliorer les méthodes de limitation en maillage non structuré général afin qu'elles dégradent le moins possible la précision du schéma.

(2) Le deuxième objectif, détaillé dans la section 1.3, consiste à explorer la montée en ordre du schéma numérique par des reconstructions de polynômes de degré plus élevé. Il est très important de trouver des algorithmes de reconstruction suffisamment rapides pour une implémentation dans des grands codes de calculs parallèles et vectoriels comme Cedre.

Pour atteindre ces objectifs, l'étude reprend de façon générale les aspects suivants des schémas volumes finis :

(1) La théorie de la reconstruction des polynômes en maillage non structuré général.

(2) L'intégration des polynômes sur les faces des cellules.

(3) L'influence de la reconstruction sur la stabilité des schémas volumes finis.

(4) L'influence de la reconstruction sur la précision des schémas volumes finis.

(5) Les méthodes de limitation en maillage non structuré général.

Finalement, les évolutions concernant le premier objectif seront testées sur deux calculs tridimensionnels utilisant un modèle de la simulation des grandes échelles. Le plan détaillé et le déroulement de l'étude sont présentés dans la section 1.4.

1.2. Étude de la stabilité et précision du schéma MUSCL

L'expérience avec Cedre montre que la relation entre robustesse et précision des méthodes de type Muscl est particulièrement délicate en maillage non structuré. Deux exemples sont le calcul d'un écoulement subsonique au-dessus d'une cavité profonde, cf. [**17**], et le calcul d'un jet supersonique chaud, cf. [**81**]. Tous les deux sont des calculs instationnaires avec une modélisation de la turbulence par le modèle de Smagorinski, cf. [**92**].

Le cas de la cavité montre que le schéma numérique s'avère moins robuste sur un maillage de tétraèdres que sur un maillage structuré. Plus précisément, le calcul sur les tétraèdres nécessite une limitation plus forte des gradients reconstruits pour se dérouler correctement. Le même calcul, effectué sur un maillage structuré, peut se contenter d'une limitation moins restrictive, voir [**17**, p. 144] pour les détails. On constate alors que la précision du calcul est moins bonne en maillage non structuré. Cela permet de tirer deux conclusions :

(1) La reconstruction du gradient en maillage non structuré rend le calcul moins robuste.

(2) Les limiteurs ont une influence forte sur la précision en maillage non structuré.

Le cas du jet montre que l'écoulement ne devient pas turbulent sur un maillage de tétraèdres. Il faut insérer un maillage structuré dans la zone de mélange pour que la turbulence se développe, voir [**81**] pour les détails. Cela laisse présumer que la limitation de gradient génère trop de dissipation numérique en maillage non structuré.

Ces constats suggèrent d'étudier deux questions particulières :

(1) La relation entre la reconstruction du gradient et la stabilité du schéma numérique.

(2) L'impact des limiteurs de gradient sur la précision en maillage non structuré.

La littérature recense plusieurs résultats de stabilité sur des maillages non structurés généraux. Même si la plupart des preuves sont valables pour des lois de conservation scalaires, elles restent intéressantes pour le cas des systèmes. Le respect du principe du maximum, prouvé dans [**10**], assure que la solution d'un schéma de type volumes finis reste confinée entre les minima et maxima locaux au voisinage d'une cellule. Une méthode qui respecte le principe du maximum garantit donc non seulement que la solution reste bornée mais aussi qu'elle reste positive si la condition initiale était positive. D'autres résultats de stabilité font partie de théorèmes de convergence, comme, par exemple [**73, 27, 18, 88**], et sont donc importants pour les bases théoriques des méthodes des volumes finis en maillage non structuré.

Il faut cependant noter qu'une grande partie des résultats recourent de façon explicite à la limitation de gradient. Les expériences avec Cedre citées plus haut montrent que la dissipation numérique apportée par les limiteurs est souvent difficile à contrôler. Pour des calculs industriels, l'objectif est donc plutôt de relâcher le plus possible les contraintes de limitation. Cette approche exige par contre de contrôler l'influence de la reconstruction du gradient sur la stabilité du schéma numérique afin de ne pas compromettre la robustesse. Pour cette raison, la présente étude se concentre sur la stabilité de la méthode Muscl *sans limiteurs*. Cela permet d'explorer l'impact de la méthode de reconstruction du gradient, de la taille du voisinage de reconstruction et du type de maillage sur la stabilité.

L'objectif suivant est alors d'améliorer les limiteurs qui servent à supprimer des oscillations artificielles et qui assurent la stabilité au voisinage de chocs et dans des zones avec de forts gradients. Ces algorithmes de limitation ne doivent pas trop dégrader la précision des calculs. Les expériences avec Cedre citées ci-dessus montrent cependant que les méthodes actuelles de limitation dans Cedre sont à cet égard insuffisantes.

Dans la littérature, il existe un certain nombre de résultats sur les limiteurs en maillage non structuré. Un exemple est le principe du maximum déjà mentionné ci-dessus, cf. [**8, 10**]. Ce résultat spécifie des intervalles admissibles pour les valeurs reconstruites aux interfaces entre les cellules. Si ces intervalles sont respectés, la solution ne peut pas dépasser les minima et maxima locaux au prochain pas de temps. D'autres travaux comme [**68, 86**] couvrent plutôt l'aspect géométrique de la limitation en maillage non structuré. Il s'agit là du problème de trouver le gradient qui est le plus proche du gradient reconstruit mais qui respecte les intervalles prescrits aux faces. Il faut examiner si ces méthodes sont assez efficaces pour trouver leur place dans un code industriel et scientifique comme Cedre. Il est également nécessaire de tenter de les améliorer.

Les méthodes de type ENO/WENO évitent l'utilisation de limiteurs par une sélection dynamique des voisinages de reconstruction. Un algorithme de type ENO reconstruit plusieurs solutions sur différents voisinages et sélectionne celle qui provoque le moins d'oscillations. Néanmoins, ces reconstructions multiples nécessitent un effort de calcul qui semble encore trop élevé pour permettre l'utilisation dans un code tel que Cedre. Pour cette raison, la présente étude ne poursuit pas cette piste.

1.3. Étude de la montée en ordre du schéma

La deuxième étape de la présente étude a comme objectif la montée en ordre des méthodes des volumes finis en maillage non structuré. Cet objectif passe en général par la reconstruction de polynômes de degré plus élevé. La littérature contient notamment un ensemble de résultats sur la reconstruction des polynômes de degré deux en maillage non structuré, voir par exemple [**62, 63, 8, 10, 44**]. Ces résultats démontrent que la reconstruction des polynômes de degré deux fournit de meilleurs résultats que la reconstruction linéaire par cellule.

Cependant, plus le degré des polynômes est élevé, plus les voisinages de reconstruction doivent être grands. Comme un code de calcul doit calculer, trier et stocker les données de connectivité entre les cellules, l'élargissement des voisinages nécessite des ressources significatives de calcul. Ce problème constitue un sérieux obstacle à l'utilisation de reconstructions polynomiales de degré élevé.

Les travaux cités ci-dessus se concentrent principalement sur l'aspect des méthodes numériques et ne parlent que très peu de la réalisation concrète des algorithmes sur ordinateur. Un

récent exemple est la présentation d'un schéma volumes finis basé sur une reconstruction des polynômes de degré trois [**71**].

Le problème de la taille des voisinages de reconstruction était l'une des raisons pour le développement d'alternatives aux méthodes des volumes finis. Parmi ces alternatives, il faut citer de façon non exhaustive les approches de type *Discontinuous Galerkin* (DG) qui évitent les reconstructions sur des voisinages larges par une discrétisation différente. Elles font, par conséquent, l'objet d'un intérêt croissant. D'autres approches importantes sont la *Spectral Volume Method* (SVM) et les *Residual Distribution* (RD) *Schemes*. La section 2.7 fournit une liste de références pour ces méthodes.

Il semble cependant important d'explorer si des algorithmes de reconstruction précis mais efficaces sont possibles dans le cadre des méthodes classiques de type volumes finis car un certain investissement a déjà été fait dans la programmation de ces méthodes. De tels algorithmes doivent reconstruire des solutions sur des voisinages larges tout en utilisant des algorithmes informatiques qui travaillent sur de petits voisinages. La recherche et l'étude de tels algorithmes constituent la deuxième étape du présent travail.

Un exemple pour une reconstruction de fonctions non polynomiales sont les travaux de T. Sonar, cf. [**99, 97**], qui utilise une interpolation par *radial basis functions* pour éviter les problèmes de l'interpolation polynomiale. Ces techniques semblent néanmoins encore trop lourdes pour une implémentation dans un code tel que Cedre. Pour cette raison, cette piste n'a pas été poursuivie.

1.4. Plan de l'étude

En résumé, la présente étude se pose comme objectif d'améliorer les méthodes classiques de type volumes finis en poursuivant plusieurs axes :

- explorer la stabilité des méthodes de discrétisation en absence des limiteurs pour augmenter la robustesse des calculs et réduire la limitation au minimum indispensable ;
- augmenter la précision en explorant des méthodes de reconstruction de polynômes de degré élevé, dont l'algorithme informatique travaille sur de très petits voisinages ;
- améliorer les méthodes de limitation afin qu'elles dégradent le moins possible la précision des calculs ;
- tester et valider les évolutions qui ont pu être programmées dans le logiciel Cedre à l'aide de calculs tridimensionnels en maillage non structuré général.

Pour atteindre ces objectifs, l'étude concentre l'effort principal sur les aspect suivants des méthodes des volumes finis :

- la reconstruction polynomiale en maillage non structuré général ;
- l'intégration des fonctions reconstruites sur les faces des cellules ;
- les algorithmes de limitation en maillage non structuré général.

La démarche de l'étude a été la suivante. Une étape préliminaire a consisté à passer en revue la géométrie des maillages non structurés et à dériver certaines identités géométriques importantes pour ces maillages. Cette étape a aussi permis d'introduire une notation géométrique efficace qui facilite l'écriture des formules. Une deuxième étape préliminaire a consisté à poser de façon rigoureuse les bases de la discrétisation spatiale par la méthode des volumes finis. Cette étape a servi à clarifier toutes les notions indispensables pour les étapes suivantes. Elle a consisté à introduire des définitions importantes comme celles des flux numériques, de la reconstruction et de l'intégration des flux sur les faces des cellules. Ce travail a aussi permis de définir de façon rigoureuse l'erreur de troncature induite par la discrétisation spatiale. L'erreur de troncature permet ensuite de classer la précision des schémas selon l'ordre de troncature. Après ces travaux préparatoires, il a été possible d'aborder l'étude proprement dite. Elle a d'abord consisté à explorer, de la façon la plus générale, la reconstruction polynomiale en maillage non structuré. Plus précisément, le travail a consisté à définir pour chaque ordre de troncature la famille la plus large de reconstructions permettant d'obtenir l'ordre souhaité. Une partie importante de cette étape a été la recherche d'algorithmes informatiques pour réaliser de façon efficace des reconstructions polynomiales sur ordinateur. L'étape suivante a été la recherche de méthodes permettant de calculer de façon précise les intégrales des flux numériques sur les faces des

cellules. Le travail sur la reconstruction a ensuite servi à analyser en profondeur la relation de la reconstruction avec la stabilité de la discrétisation spatiale. L'étape suivante a consisté à étudier la précision de la discrétisation spatiale de façon théorique, à l'aide de l'approche de l'équation modifiée et de façon numérique, à l'exemple de la convection linéaire. Certaines évolutions ont ensuite été validées à l'aide de deux simulations des grandes échelles en dimension trois, notamment d'un écoulement subsonique au-dessus d'une cavité profonde et d'un jet chaud supersonique.

L'organisation du présent document suit le plan de travail de l'étude. L'introduction est suivie d'une brève bibliographie sur les méthodes numériques de type volumes finis. Le chapitre 3 présente les bases de la simulation des grandes échelles des équations de Navier-Stokes compressibles. Le chapitre 4 présente brièvement le logiciel Cedre. Le chapitre 5 définit la géométrie des maillages non structurés et présente une notation efficace et confortable pour faciliter l'écriture des formules dans les chapitres suivants. Le chapitre 6 présente le cadre de la discrétisation spatiale des lois de conservation par la méthode des volumes finis. L'étude proprement dite commence par le chapitre 7 dont le but est d'explorer de la façon la plus générale la reconstruction polynomiale en maillage non structuré. Le chapitre 8 est dédié à l'étude d'algorithmes efficaces pour la reconstruction polynomiale et le chapitre 9 est dédié aux méthodes d'intégration de fonctions sur les faces des cellules. Le chapitre 10 présente le travail sur la stabilité et le chapitre 11, celui sur la précision des méthodes. Le chapitre 12 est dédié aux limiteurs en maillage non structuré. Les chapitres 13 et 14 décrivent les calculs tridimensionnels effectués pour valider et tester les méthodes qui ont pu être programmées dans Cedre.

CHAPITRE 2

Bilan bibliographique du schéma MUSCL

2.1. Objectif du chapitre

Ce chapitre donne un aperçu de l'importante bibliographie du schéma MUSCL sur maillage non structuré. Cette bibliographie se trouve dans des journaux de différents types :

- journaux de mécanique des fluides appliquée ou théorique : *Computers and Fluids, International Journal for Numerical Methods in Fluids... ;*
- journaux de calcul scientifique généralistes : *Journal of Computational Physics, SIAM Journal on Scientific Computing ...* ;
- journaux de mathématiques appliquées : *SIAM Journal on Numerical Analysis...*

Il y a également les actes de nombreux congrès et quelques cours, par exemple [**10**]. Une monographie très utile est [**103**]. Les ouvrages spécifiquement consacrés au schéma MUSCL sont par contre très peu nombreux. Cela est dû à la diversité des pratiques effectives de mise en œuvre de la méthode, voir par exemple [**62, 73, 79, 52**]. La suite de ce chapitre met l'accent sur quelques jalons bibliographiques qui paraissent historiquement importants.

2.2. L'approche de Godounov[1]

Le problème général auquel on s'intéresse dans ce travail est la simulation numérique de la dynamique des fluides compressibles avec une attention particulière à la simulation des grandes échelles. Depuis plus de soixante ans, cette question a donné lieu à des développements importants dans les mathématiques pures et appliquées ainsi que dans la physique. Sur le plan théorique, les questions mathématiques de l'existence et de l'unicité des solutions des équations de la mécanique des fluides compressibles sont restées essentiellement ouvertes. Sur le plan physique, très peu de solutions analytiques ont un intérêt réel. Cependant, il est important de disposer de simulations précises dans de très nombreux secteurs scientifiques : aéronautique, astronautique, énergétique, cosmologie, climatologie, etc.

L'une des difficultés numériques consiste à calculer des solutions pouvant comporter des discontinuités de différents types. Ces discontinuités, essentiellement les ondes de choc et les discontinuités de contact, sont caractéristiques des systèmes de lois de conservation. Un moment important a été la prise de conscience que la discrétisation des équations de la mécanique des fluides compressibles devait préserver au plus près la structure conservative des équations. Ceci a mis en évidence le besoin de schémas conservatifs, c'est-à-dire de schémas numériques avec une propriété de conservativité locale, afin d'assurer une vitesse correcte de propagation des discontinuités, cf. [**85**].

Une étape importante a été franchie par S.K. Godounov dans les années 50, cf. [**53**]. Les schémas classiques aux différences utilisent des développements de Taylor locaux, valables dans les zones de régularité de la solution. Ces schémas ont donc besoin de réglages particuliers pour se comporter correctement aux voisinages de discontinuités. Godounov, au contraire, choisit de représenter partout le fluide par des moyennes de cellule, même dans les zones où la solution est régulière. Il profite ensuite de cette représentation discontinue pour définir l'évolution en temps des moyennes de cellule.

Considérons une loi de conservation hyperbolique en dimension un, donnée par

$$\left.\begin{aligned} \frac{\partial u\,(x,t)}{\partial t} + \frac{\partial f(u\,(x,t))}{\partial x} = 0, \qquad t>0, \quad x\in\mathbb{R} \\ u(x,0) = u_0(x) \end{aligned}\right\}. \tag{2.2.1}$$

[1] Cette section s'inspire en partie de l'ouvrage M. Ben-Artzi et J. Falcovitz [**12**].

On appelle *problème de Riemann* le problème de Cauchy (2.2.1), au sens faible, pour la condition initiale discontinue

$$u_0(x) = \begin{cases} U_{\mathrm{g}} & \text{si } x < x_0 \\ U_{\mathrm{d}} & \text{si } x \geq x_0 \end{cases} . \tag{2.2.2}$$

Appelons $\widetilde{u}\,(x,t\,;\,U_{\mathrm{g}},U_{\mathrm{d}},x_0)$ la solution de (2.2.1) au sens faible pour la condition initiale (2.2.2). Le schéma de Godounov tire alors parti d'une propriété intéressante de la loi de conservation (2.2.1). Il s'agit du fait que la valeur au point $x = x_0$ de la solution exacte $\widetilde{u}\,(x,t\,;\,U_{\mathrm{g}},U_{\mathrm{d}},x_0)$ du problème de Riemann ne dépend pas du temps. On note cette valeur constante $\widetilde{u}_R\,(x_0,U_{\mathrm{g}},U_{\mathrm{d}})$.

Supposons donné un maillage régulier unidimensionnel de diamètre $h > 0$. Le barycentre x_j de la cellule numéro j est $x_j = jh$ pour $j \in \mathbb{Z}$ et les deux faces de cette cellule sont respectivement les points $x_{j+1/2} = \left(j + \frac{1}{2}\right) h$ et $x_{j-1/2} = \left(j - \frac{1}{2}\right) h$.

Alors la moyenne

$$\overline{u}_j^n = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u\,(x,t^n)\; dx$$

de la solution exacte $u\,(x,t)$ de (2.2.1) vérifie

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \frac{\Delta t}{h} \left\{ \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f\left(u\left(x_{j+1/2},t\right)\right)\; dt - \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f\left(u\left(x_{j-1/2},t\right)\right)\; dt \right\} . \tag{2.2.3}$$

La relation (2.2.3) est vraie pour la solution exacte de (2.2.1). Elle signifie que la moyenne exacte $\overline{u}_j(t)$ au temps t^{n+1} s'exprime en fonction de la moyenne exacte au temps t^n et des moyennes des flux d'interface $f\left(u\left(x_{j\pm1/2},t\right)\right)$ sur $t \in [t^n, t^n + \Delta t[$.

L'équation (2.2.1) motive l'introduction de la

Définition 2.2.1 (Schéma de Godounov). Le schéma de Godounov pour l'approximation du problème (2.2.1) s'écrit

$$\left.\begin{aligned} U_j^{n+1} &= U_j^n - \frac{\Delta t}{h} \left[f\left(\widetilde{u}_R\left(x_{j+1/2},U_j^n,U_{j+1}^n\right)\right) - f\left(\widetilde{u}_R\left(x_{j-1/2},U_{j-1}^n,U_j^n\right)\right)\right] \\ U_j^0 &= \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u_0\,(x)\; dx \end{aligned}\right\} \tag{2.2.4}$$

où $\widetilde{u}_R\left(x_{j+1/2},U_j^n,U_{j+1}^n\right)$ et $\widetilde{u}_R\left(x_{j-1/2},U_{j-1}^n,U_j^n\right)$ sont les valeurs des solutions du problème de Riemann aux points $x_{j+1/2}$ et $x_{j-1/2}$.

Le *flux numérique* durant $[t^n, t^n + \Delta t[$ est une fonction de U_j^n et U_{j+1}^n, donnée par la solution de Riemann

$$f_{j+1/2}^{n+1/2}\left(U_j^n,U_{j+1}^n\right) \triangleq f\left(\widetilde{u}_R\left(x_{j+1/2},U_j^n,U_{j+1}^n\right)\right) . \tag{2.2.5}$$

Contrairement aux schémas aux différences, tels que les schémas de Lax-Friedrichs ou de Lax-Wendroff, le schéma (2.2.4) ne repose pas sur une approximation de type différences finies des opérateurs différentiels, mais sur une approche volumes finis. Autrement dit, ce schéma fait évoluer les moyennes de cellule au moyen d'un *flux numérique*. Le schéma de Godounov possède donc une base relativement abstraite pour un schéma numérique. Rappelons qu'une des versions du schéma, le schéma de Glimm, cf. [**36**], est à la base de l'un des seuls résultats théoriques d'existence de solutions des équations d'Euler compressibles. Il possède des propriétés importantes sur le plan mathématique, notamment :

(1) la monotonie et

(2) la consistance avec la condition d'entropie.

Ces propriétés permettent de montrer la convergence du schéma (2.2.4) vers la solution exacte lorsque $h \to 0$. La base théorique de l'algorithme de calcul est par conséquent très solide. Insistons sur le fait que le schéma (2.2.4) ne dépend pas de paramètres numériques arbitraires. Il est le seul schéma entièrement défini par le fait qu'il donne, à partir d'une condition initiale au temps t^n, la moyenne de la solution exacte au temps t^{n+1}.

Le schéma (2.2.4) consiste en deux étapes :

(1) *Étape de projection* : une approximation $U^n(x)$ étant connue au temps t^n, on effectue la projection sur les fonctions constantes par morceaux, ce qui donne U_j^n.

(2) *Étape de propagation :* on résout le problème de Cauchy (2.2.1) durant $[t^n, t^{n+1}[$ en calculant $\widetilde{u}_R(x_{j+1/2}, U_j^n, U_{j+1}^n)$ pour chaque interface $x_{j+1/2}$, ce qui donne $U^{n+1}(x)$ au temps t^{n+1}.

En dimension un, la démarche ci-dessus a l'avantage de pouvoir s'appliquer directement aux systèmes de dynamique des fluides compressibles, à savoir les équations d'Euler en coordonnées d'Euler ou de Lagrange.

2.3. L'approche de Van Leer

Bien qu'il constitue un point de repère traditionnel pour toute méthode des volumes finis en mécanique des fluides compressibles, le schéma de Godounov (2.2.4) présente de nombreux inconvénients sur le plan pratique[2].

(1) Le schéma de Godounov [**53**] est d'ordre un car il est monotone. Le théorème de Godounov montre en effet qu'un schéma qui préserve la monotonie de la solution est d'ordre inférieur ou égal à un. Cela signifie que le schéma de Godounov est peu précis dans les régions où la solution est régulière. Dans la pratique, la notion de monotonie est donc trop contraignante pour les besoins de la simulation numérique industrielle. Il est cependant important de noter que le théorème de Godounov est valable uniquement pour les schémas qui fournissent une discrétisation linéaire de l'équation de convection linéaire. Cela permet de contourner le problème soulevé par Godounov à l'aide de techniques non linéaires, notamment d'algorithmes utilisant des limiteurs.

(2) Le schéma ne peut pas s'interpréter comme l'application d'un schéma en temps à un schéma semi-discret en espace. Dans la pratique, on voudrait utiliser différentes méthodes d'intégration explicites et implicites en temps au lieu du simple schéma d'Euler de (2.2.4).

B. Van Leer a introduit la méthode dite MUSCL, l'acronyme pour *Monotonic Upstream-centered Scheme for Conservation Laws*, dans une communication [**105**] pour la *Third International Conference on Numerical Methods in Fluids*, en 1972 à Paris, et une série d'articles [**106, 107, 108, 109**], parus entre 1974 et 1979. Le but était d'obtenir des schémas d'ordre élevé à partir d'un schéma de type Godounov. Alors que [**108**] décrit des schémas pour l'équation de convection linéaire, [**109**] introduit un schéma pour l'écoulement idéal compressible d'un gaz en dimension un et en formulation de Lagrange. Il inclut le cas de la symétrie cylindrique et sphérique, un terme source F pour la quantité de mouvement et un terme source G pour l'énergie. Les variables d'état sont le volume spécifique V, l'énergie totale spécifique E, la vitesse u et la pression p. Le système hyperbolique pour l'écoulement idéal compressible s'écrit en formulation lagrangienne

$$\left.\begin{array}{l} \dfrac{\partial V}{\partial t} - \dfrac{\partial\,(x^\alpha u)}{\partial \xi} = 0 \\ \dfrac{\partial u}{\partial t} + \dfrac{\partial\,(x^\alpha p)}{\partial \xi} = F \\ \dfrac{\partial E}{\partial t} + \dfrac{\partial\,(x^\alpha p u)}{\partial \xi} = uF + G \end{array}\right\} \tag{2.3.1}$$

où ξ est la coordonnée matérielle d'une particule et x est la coordonnée spatiale. Dans le système (2.3.1), le choix $\alpha = 0$ décrit la symétrie du plan, le choix $\alpha = 1$ décrit la symétrie du cylindre et le choix $\alpha = 2$ décrit la symétrie sphérique.

Par la suite, on résume seulement les éléments les plus importants de l'algorithme que Van Leer propose pour résoudre le système (2.3.1). Il considère un maillage de type différences finies ξ_i, $1 \leq i \leq N$. Les volumes de contrôle sont les intervalles $[\xi_i, \xi_{i+1}]$ dont les barycentres sont notés $\xi_{i+\frac{1}{2}}$. Les interfaces entre les volumes de contrôle sont donc les points ξ_i. L'une des modifications

[2]Cette section s'inspire en partie de l'article de M. Holt [**64**].

les plus importantes que B. Van Leer apporte au schéma de Godounov consiste à calculer dans chaque cellule des pentes ΔV_j, Δu_j et ΔE_j à partir des moyennes de cellule V_j, u_j et E_j. À l'aide de ces pentes, l'algorithme construit, dans chaque cellule, des valeurs $V_{i,+}$, $u_{i,+}$ et $E_{i,+}$ à droite de ξ_i et des valeurs $V_{i,-}$, $u_{i,-}$ et $E_{i,-}$ à gauche de ξ_i. Ensuite, l'algorithme applique le flux numérique de Godounov au point $\xi_{i+1/2}$ aux valeurs $[V_{i,+}, u_{i,+}, E_{i,+}]$ et $[V_{i+1,-}, u_{i+1,-}, E_{i+1,-}]$, ce qui permet de calculer les moyennes de cellule V_i, u_i et E_i au prochain pas de temps.

Enfin, les limiteurs de pente, présentés aujourd'hui comme le cœur de l'approche MUSCL, s'appellent « algorithmes de monotonie » et apparaissent dans la section « Techniques accessoires ». Ils sont indispensables car le schéma peut occasionner une perte de monotonie, ce qui peut provoquer une densité ou une pression négative dans certaines cellules. Dans [**108**], Van Leer propose donc, dans un certain nombre de situations bien identifiées, de limiter l'amplitude des valeurs de Δv, Δu et ΔE. Dans sa version initiale, cette procédure est largement empirique.

Par la suite, Woodward et Colella (1982) font évoluer le schéma de Van Leer dans [**122**]. Les auteurs effectuent une investigation très détaillée des différents aspects du schéma MUSCL :

(1) Ils remplacent le solveur de Riemann exact par un solveur approché.

(2) Une interpolation quadratique remplace l'interpolation linéaire.

(3) Ils clarifient et précisent la procédure de limitation de Van Leer.

Dans la suite, plusieurs solveurs de Riemann approchés ont été introduits, dont les plus connus sont ceux de Roe [**91**] et de Osher [**87**]. De nombreuses variantes sont comparées dans le contexte des équations d'Euler en dimension un. L'article [**122**] présente le cas test sur l'interaction de deux ondes explosives, « two interacting blast waves » dans l'original. Les auteurs généralisent ensuite la procédure au cas bidimensionnel sur maillage cartésien à l'aide d'un splitting directionnel.

Signalons, pour terminer, les travaux de Ben-Artzi et Falcovitz sur les schémas de type *Generalized Riemann Problem* (GRP). Il s'agit d'une généralisation de la méthode de Van Leer par un solveur de Riemann local plus précis. Le schéma GRP peut être considéré comme une généralisation naturelle du schéma de Godounov à l'ordre deux. L'ouvrage récent [**12**] fait le point sur cette approche.

2.4. Les schémas TVD

À la suite de ces premiers travaux, décrits dans la section 2.3, il est rapidement devenu indispensable de clarifier les bases de la méthode MUSCL. Les différentes étapes du schéma MUSCL soulèvent plusieurs questions :

(1) Est-il possible de remplacer la résolution exacte du problème de Riemann par une formule de flux numérique plus simple ?

(2) Comment peut-on généraliser la notion de monotonie, qui s'applique aux lois de conservation scalaires, à la dynamique des gaz ?

(3) Quels critères faut-il imposer à la limitation du gradient ?

À la suite des travaux de Van Leer, différents chercheurs ont fourni un effort important pour axiomatiser le schéma MUSCL dans un contexte plus simple que celui de la dynamique des gaz. Cette démarche a été initiée par Harten et par Roe. Ces travaux partent du principe qu'une méthode de discrétisation des lois de conservation scalaires a besoin de trois ingrédients essentiels :

(1) Un flux numérique sous la forme d'un solveur approché du problème de Riemann.

(2) Une formule d'interpolation de gradient.

(3) Une formule de limitation de gradient.

Ce point de vue a donné lieu à des études variées qui ont toutefois conduit à une analyse fragmentée de ces différentes questions. Ainsi, beaucoup d'articles ont été consacrés à la conception de limiteurs de pente optimaux parmi certaines classes de formules paramétrées. Parmi les travaux les plus importants, on peut citer ceux de Harten [**59**] qui a proposé la notion de *variation totale décroissante* (TVD).

Harten définit, pour la solution exacte d'une loi de conservation, une propriété de monotonie qui s'exprime par deux conditions :

(1) La solution ne fait apparaître aucun nouvel extremum local.

(2) Les minima locaux ne peuvent pas diminuer et les maxima locaux ne peuvent pas augmenter.

Harten prouve que les schémas monotones possèdent la propriété TVD et que les schémas TVD préservent la propriété de monotonie définie ci-dessus. Les schémas TVD forment donc une classe plus large que les schémas monotones et possèdent la propriété de monotonie qui empêche a priori des oscillations artificielles d'apparaître. Les schémas TVD peuvent être du second ordre. Il est possible de spécifier des conditions explicites pour qu'un schéma soit TVD, cf. par exemple Sweby [**102**], ce qui a contribué à leur succès.

La situation autour de 1987 se trouve, par exemple, dans le rapport de H.C. Yee [**125**] qui mène une comparaison extensive d'un ensemble de variantes des schémas TVD. Pour une présentation de la notion de schéma TVD, voir [**84, 103, 83**]. L'une des questions les plus importantes reste cependant ouverte : il n'est pas clair comment généraliser, de façon univoque, les schémas obtenus pour une équation scalaire au cas des systèmes d'équations de la mécanique des fluides compressibles.

Naturellement, l'extension au cas multidimensionnel est encore plus compliquée. Elle se fait sur maillage structuré, avec un splitting directionnel, cf. [**124**] et [**123**]. Il est important de rappeler le résultat de Goodman et Leveque [**55**] qui prouvent que l'ordre des schémas TVD en dimension $d > 1$ ne peut excéder un. La notion TVD est donc, dans un sens strict, une notion unidimensionnelle. Cela a motivé le développement de conditions plus faibles pour la monotonie, comme le principe du maximum de Barth [**8, 10**] qui est utilisé dans cette étude pour définir des limiteurs en maillage non structuré. L'avantage de ces critères plus récents réside également dans le fait qu'ils s'adaptent aux maillages non structurés généraux, ce qui n'est pas le cas du critère TVD en dimension $d > 1$. D'autres critères de limitation de pente spécifiques à la dynamique des gaz (domaines de variables physiques invariants, consistance avec la condition d'entropie) ont également été explorés [**37, 39, 14, 16, 15, 19**].

Les travaux cités ci-dessus ont permis de clarifier l'algorithme du schéma MUSCL mais n'ont pas touché au fait que le schéma se compose de trois éléments fondamentaux :

(1) Évaluation des flux numériques aux valeurs interpolées aux interfaces.

(2) Interpolation de gradient.

(3) Limitation de gradient.

Par la suite, de nombreuses recherches expérimentales ont été menées pour l'extension du principe des limiteurs en dimensions deux et trois sur maillage structuré. Cela a donné lieu à une littérature très importante, que ce soit sur des exemples académiques, comme l'équation de transport scalaire, e.g. [**101**], ou dans le contexte des équations de la mécanique des fluides.

2.5. Les schémas ENO

En présence d'un extremum, les schémas TVD sont seulement d'ordre un, ce qui montre que la propriété TVD est trop restrictive. Cela a conduit Harten et Osher [**61**] à introduire les schémas ENO avec l'objectif d'assouplir la propriété TVD. ENO est l'acronyme pour *essentiellement non oscillant*. Un schéma ENO ne fait pas apparaître d'extrema locaux autres que ceux présents dans la condition initiale. La conception des schémas ENO se présente sous la forme d'un problème d'interpolation spatiale. Le principe de base est de calculer une reconstruction de polynômes sur plusieurs voisinages et de sélectionner la reconstruction qui génère le moins d'oscillations. Un travail important a été la mise au point de schémas en temps précis et compatibles avec la discrétisation spatiale ENO, dans deux articles de Shu et Osher, [**95**] et [**96**]. Dans [**94**], les auteurs testent de façon intensive des applications à la mécanique des fluides compressibles. Différents types d'amélioration ont été apportés afin de remédier au problème du coût de calcul élevé des schémas ENO, voir par exemple [**74**].

2.6. Extension aux maillages non structurés

2.6.1. Introduction. Les années 1986-1988 ont vu apparaître le besoin de généraliser l'algorithme MUSCL à des maillages non structurés et à des modèles thermodynamiques plus complexes. Ce développement était motivé par des besoins applicatifs spécifiques comme des géométries d'écoulement compliquées, les mélanges de gaz, des gaz réels ainsi que des réactions chimiques. Dans de nombreux cas, cette généralisation s'est faite de façon empirique par l'adaptation de l'algorithme sur une base formelle. Cela veut dire que les trois ingrédients du schéma, c'est-à-dire les flux numériques, l'interpolation et les limiteurs ont été adaptés aux maillages non structurés et à la thermodynamique complexe.

Ces schémas n'utilisent plus de solveur de Riemann exact. La littérature recense plusieurs formules de flux numériques : on cite la généralisation du flux de Roe à des lois d'état de gaz réels [**1**], les flux cinétiques [**88**] et [**41**], des variantes de flux mêlant considérations algébriques et thermodynamiques, des flux hybrides [**38**]. Par ailleurs, de nombreuses recherches expérimentales sur l'extension des limiteurs de pente aux maillages non structurés ont été menées, que ce soit sur des exemples académiques, comme l'équation de transport scalaire, ou dans le contexte des équations de la mécanique des fluides, par exemple [**82**], [**24**] et [**8**]. Le succès du schéma MUSCL a donné lieu à une activité très importante en mécanique des fluides numérique. Le schéma MUSCL est à la base d'un ensemble de codes commerciaux, par exemple FLUENT, et permet d'obtenir des résultats satisfaisants dans beaucoup d'applications. Il est flexible, se prête bien à la décomposition de domaines et peut être utilisé avec divers types d'interpolation.

2.6.2. Reconstruction du gradient de cellules et interpolation. Des travaux pour généraliser la méthode ENO aux maillages non structurés ont vu le jour relativement tôt, notamment sous l'impulsion de A. Harten, voir par exemple [**60**] et [**2**]. Divers procédés de reconstruction de polynômes ont été introduits et testés numériquement pour les équations scalaires. Ce travail a été très utile pour revoir le formalisme de l'interpolation polynomiale locale sur des maillages non structurés. Ces questions sont très développées dans la recherche sur les fonctions spline, notamment dans le domaine du traitement d'images. Cependant, les besoins spécifiques d'une interpolation locale conservative ont nécessité de reprendre des études en amont. De plus, l'application de ces techniques aux équations de la mécanique des fluides pose de nouvelles difficultés. Par la suite, cette voie de recherche a été poursuivie par plusieurs auteurs, ce qui a engendré une littérature très abondante dans les années 2000, voir par exemple [**26**]. De nombreuses variantes ont été explorées, que ce soit dans le cadre de schémas cell-vertex ou cell-center.

Ces travaux n'ont cependant pas permis de faire émerger de méthode commune permettant d'obtenir des schémas MUSCL précis et stables au-delà de l'ordre deux. Ceci a amené beaucoup d'équipes d'horizons très divers (aérodynamique, énergétique, milieux poreux, hydrologie, astronomie...) à développer de façon plus ou moins empirique leur propre méthode d'interpolation locale et leur propre version de schéma MUSCL [**42**]. Ce type d'activité se poursuit aussi bien dans le contexte académique [**22**] que dans celui de l'industrie logicielle, voir par exemple [**71**], qui présente une interpolation des polynômes d'ordre trois, testée avec succès.

2.6.3. Généralisation des schémas ENO par les RBF. Une série d'études mérite d'être mentionnée en particulier, celle de Sonar [**99, 98, 97**], et Iske et Sonar [**70**], qui ont proposé une généralisation systématique de l'approche ENO au cas multidimensionnel en utilisant une interpolation par des fonctions appelées *radial basis functions* (RBF). Ce type d'interpolation, introduite dès les années 70, voir par exemple [**43**], a été étudié de façon systématique par Powell [**23**]. Il s'agit de méthodes pour interpoler des données spatiales localisées de façon dispersée. Une excellente référence est l'ouvrage récent [**120**]. Ce type d'interpolation est très proche de celles que l'on rencontre, par exemple, en géostatistique [**113**].

Sonar propose dans [**99**] de remplacer l'interpolation polynomiale par une interpolation de type RBF. En effet, l'interpolation par des polynômes pose des problèmes dans le cas multidimensionnel. Une étude systématique des RBF pour les volumes finis sur maillage triangulaire est menée dans [**97**], [**99**], [**98**] et [**70**]. Pour le moment, cette idée n'a pas encore trouvé son chemin dans le milieu industriel.

2.7. Méthodes apparentées : DG, SVM, RD et GRP

Les schémas volumes finis présentés dans les sections précédentes ne sont pas les seules méthodes utilisées pour les lois de conservation hyperboliques. Outre la méthode GRP [**12**], citée à la fin de la section 2.3, il faut mentionner de façon non exhaustive d'autres méthodes importantes qui suscitent un intérêt grandissant :

(1) La méthode *Discontinuous Galerkin* (DG) date du début des années 70 lorsque Reed et Hill l'ont proposée pour la simulation de problèmes de transport de neutrons [**89**]. La méthode DG a ensuite été adaptée à la mécanique des fluides dans les années 80 et 90, par Cockburn, Shu et al. Des références pour cette méthode sont par exemple [**30, 32, 31, 28, 33, 29, 21**].

(2) À partir de 2002, Wang et al. ont introduit l'approche de la *Spectral Volume Method* (SVM). Des références sont par exemple [**115, 116, 117, 118, 80, 100, 45, 58**].

(3) L'approche des *Residual Distribution (RD) Schemes* constitue une alternative à la méthode classique des volumes finis pour obtenir des schémas d'ordre élevé. Des références récentes sont [**90, 3, 4, 5, 7, 6**].

Une description détaillée de ces méthodes dépasserait le cadre de cette thèse car elles ont donné lieu à un volume important de travaux dans la littérature. On note seulement que l'implémentation des méthodes DG, SVM et RD dans Cedre nécessiterait d'importantes modifications dans l'architecture logicielle. C'est une des raisons principales pour laquelle la présente étude tente d'améliorer la méthode des volumes finis dans le contexte du schéma Muscl classique.

2.8. Positionnement de la présente étude

Le bilan bibliographique montre clairement qu'il n'existe pas une seule et unique méthode Muscl, mais plutôt un ensemble de schémas plus ou moins analogues qui partagent cependant une structure commune : ils sont constitués de trois composantes essentielles :

(1) Des flux numériques évalués aux interfaces des cellules.

(2) L'interpolation de valeurs aux faces des cellules pour augmenter la précision.

(3) Des algorithmes de limitation pour supprimer des oscillations artificielles.

Avant de continuer, il est utile de clarifier la terminologie par la

Remarque 2.8.1 (Reconstruction et interpolation). Cette étude utilise le terme de *reconstruction* de préférence à celui d'*interpolation*. Lors de l'interpolation d'une fonction à partir d'un nombre fini de valeurs, on demande en général que la fonction interpolée passe par les points prescrits. Dans le cadre de cette étude, lorsqu'on interpole une fonction à partir d'un nombre fini de valeurs ponctuelles ou moyennes de cellule, on ne demande en général pas que l'interpolant passe exactement par les points donnés ni qu'elle ait exactement les moyennes de cellule prescrites. L'interpolation au sens de cette étude ressemble plutôt à l'approximation d'une fonction. C'est pourquoi le terme reconstruction semble plus adapté dans le contexte présent : on entend ici par *reconstruction* l'opération consistant à approcher une fonction inconnue à partir d'un nombre fini de valeurs ponctuelles ou moyennes de cellule de la fonction inconnue. □

Dans la plupart des travaux publiés, on remarque que la reconstruction (interpolation) se fait souvent au moyen d'un ensemble restreint de méthodes : soit on utilise la formule des moindres carrés, soit on utilise des formules établies à partir du théorème de Green, comme dans [**44**]. Cela contraste avec les études sur les limiteurs et les flux numériques qui présentent une multitude de méthodes variées. Il ne semble pas y avoir d'étude qui considère une famille générale de reconstructions afin d'étudier quelles sont les reconstructions (interpolations) les plus adaptées pour la stabilité et la précision en maillage non structuré. Ceci constitue la motivation principale de mener la présente étude sur la reconstruction dans les chapitres 7 et 8 et son impact sur la stabilité et la précision dans les chapitres 10 et 11.

Concernant les limiteurs, il existe certains travaux dédiés aux maillages non structurés, comme par exemple [**8, 10, 68, 86**]. Quelques questions restent cependant ouvertes :

(1) Les algorithmes publiés s'implémentent-ils facilement dans un logiciel comme CEDRE sous les contraintes de la parallélisation et de la vectorisation des traitements ?

(2) Il est difficile d'évaluer l'impact des méthodes de limitation sur des grands calculs instationnaires en trois dimensions, notamment des calculs de la simulation des grandes échelles.

(3) Certaines méthodes peuvent-elles être améliorées ?

Ces points justifient une étude sur les limiteurs en maillage non structuré, présentée dans le chapitre 12.

CHAPITRE 3

Introduction à la simulation des grandes échelles des équations de Navier-Stokes compressibles

3.1. Objectif du chapitre

L'objectif de ce chapitre est de présenter les bases de la simulation des grandes échelles pour les équations de Navier-Stokes compressibles. La section 3.2 explique la modélisation des écoulements compressibles par les équations de Navier-Stokes. La section 3.3 résume de façon très succincte les éléments de la simulation des grandes échelles qui sont nécessaires pour les calculs décrits dans les chapitres 13 et 14.

3.2. Introduction aux équations de Navier-Stokes compressibles

Une application importante de la méthode des volumes finis est l'approximation des équations de Navier-Stokes qui gouvernent la dynamique de fluides compressibles. La modélisation suit l'approche de la mécanique des milieux continus. Cette approche permet une description satisfaisante de la dynamique des fluides si le libre parcours moyen des particules dans le fluide est beaucoup plus petit que l'échelle de l'écoulement en considération. On peut alors décrire l'état du fluide en un point de l'espace par des grandeurs moyennées comme la densité $\varrho(\boldsymbol{x},t)$, la vitesse $\boldsymbol{v}(\boldsymbol{x},t)$ et des variables thermodynamiques telles que la pression $p(\boldsymbol{x},t)$, la température $T(\boldsymbol{x},t)$, la densité de l'énergie totale $e_{\text{tot}}(\boldsymbol{x},t)$ et la densité de l'énergie interne $e_{\text{int}}(\boldsymbol{x},t)$. Des références pour le matériel présenté ici sont par exemple les livres [**11**], [**34**] et [**25**].

Les équations de Navier-Stokes sur un domaine physique $\Omega \subset \mathbb{R}^d$ se formulent à partir des principes de la conservation de la masse, de la quantité de mouvement et de l'énergie du fluide. Soit $\mathcal{O} \subset \Omega$ un volume suffisamment régulier dont la surface $\partial\Omega$ admet une normale $\boldsymbol{\nu}(\boldsymbol{x})$ en chaque point $\boldsymbol{x} \in \partial\Omega$ en dehors d'un ensemble de mesure nulle. Pour alléger l'écriture des formules, on omet dans la suite la dépendance des inconnues par rapport à $\boldsymbol{x}$ et t. Si l'on suppose $\mathcal{O}$ fixe, le principe de conservation de la masse du fluide contenu dans $\mathcal{O}$ s'écrit comme

$$\frac{d}{dt}\int_{\mathcal{O}} \varrho\, dx = -\int_{\partial\mathcal{O}} \varrho\,(\boldsymbol{\nu}\cdot\boldsymbol{v})\; d\sigma \tag{3.2.1}$$

où $d\sigma$ est l'élément surface du bord $\partial\mathcal{O}$ de $\mathcal{O}$. Dans la suite, on néglige les forces volumiques. Le principe de conservation de la quantité de mouvement est alors donné par l'équation

$$\frac{d}{dt}\int_{\mathcal{O}} \varrho\boldsymbol{v}\, dx = -\int_{\partial\mathcal{O}} (\boldsymbol{\nu}\cdot\boldsymbol{v})\,\boldsymbol{v}\varrho\, d\sigma - \int_{\partial\mathcal{O}} \boldsymbol{\nu} p\, d\sigma + \int_{\partial\mathcal{O}} \boldsymbol{\nu}\cdot\boldsymbol{\tau}\, d\sigma \tag{3.2.2}$$

où $\boldsymbol{\tau}(\boldsymbol{x},t)$ est le tenseur des contraintes visqueuses. Le principe de conservation de l'énergie se formule comme

$$\frac{d}{dt}\int_{\mathcal{O}} \varrho e_{\text{tot}}\, dx = -\int_{\partial\mathcal{O}} (\boldsymbol{\nu}\cdot\boldsymbol{v})\, e_{\text{tot}}\varrho\, d\sigma - \int_{\partial\mathcal{O}} (\boldsymbol{\nu}\cdot\boldsymbol{v})\, p\, d\sigma + \int_{\partial\mathcal{O}} \boldsymbol{\nu}\cdot\boldsymbol{\tau}\cdot\boldsymbol{v}\, d\sigma - \int_{\partial\mathcal{O}} \boldsymbol{\nu}\cdot\boldsymbol{q}_{th}\, d\sigma \tag{3.2.3}$$

où $\boldsymbol{q}_{th}(\boldsymbol{x},t)$ est le flux de chaleur. La densité d'énergie totale est la somme de la densité d'énergie interne et de la densité d'énergie cinétique

$$e_{\text{tot}}(\boldsymbol{x},t) = e_{\text{int}}(\boldsymbol{x},t) + \frac{1}{2}\left\|\boldsymbol{v}(\boldsymbol{x},t)\right\|^2 .$$

Sous l'hypothèse d'une régularité suffisante de la solution, il est possible de formuler les équations (3.2.1), (3.2.2) et (3.2.3) comme le système d'équations aux dérivées partielles

$$\frac{\partial \varrho}{\partial t} + \boldsymbol{\nabla} \cdot (\varrho \boldsymbol{v}) = 0 \tag{3.2.4}$$

$$\frac{\partial (\varrho \boldsymbol{v})}{\partial t} + \boldsymbol{\nabla} \cdot (\varrho \boldsymbol{v} \otimes \boldsymbol{v}) + \boldsymbol{\nabla} p - \boldsymbol{\nabla} \cdot \boldsymbol{\tau} = 0 \tag{3.2.5}$$

$$\frac{\partial (\varrho e_{\text{tot}})}{\partial t} + \boldsymbol{\nabla} \cdot (\varrho \boldsymbol{v} e_{\text{tot}} + p\boldsymbol{v} - \boldsymbol{\tau} \cdot \boldsymbol{v} + \boldsymbol{q}_{th}) = 0\,. \tag{3.2.6}$$

Il faut compléter les trois équations (3.2.4), (3.2.5) et (3.2.6) par des relations thermodynamiques et mécaniques qui décrivent le comportement du fluide. Dans un sens strict, les relations thermodynamiques sont uniquement valables pour un système à l'équilibre. L'hypothèse adoptée ici est que le fluide contenu dans un petit volume autour de chaque point $\boldsymbol{x} \in \Omega$ peut être considéré comme étant localement à l'équilibre. On formule donc toutes les relations thermodynamiques pour les quantités qui dépendent de l'espace et du temps comme la pression $p(\boldsymbol{x}, t)$, la température $T(\boldsymbol{x}, t)$ et la densité de masse $\varrho(\boldsymbol{x}, t)$.

Les écoulements d'air considérés dans les chapitres 13 et 14 permettent d'utiliser l'hypothèse des gaz parfaits. Pour un gaz parfait, la pression $p(\boldsymbol{x}, t)$, la température $T(\boldsymbol{x}, t)$ et la densité de masse $\varrho(\boldsymbol{x}, t)$ sont reliées par la formule

$$p(\boldsymbol{x}, t) = \frac{\mathrm{R_m}}{\mathrm{M_{gaz}}} \varrho(\boldsymbol{x}, t)\, T(\boldsymbol{x}, t)$$

où

$$R_{\mathrm{m}} = 8,314472\, \frac{\mathrm{J}}{\mathrm{K\,mol}}$$

est la *constante universelle des gaz parfaits* et $\mathrm{M_{gaz}}$ est la *masse molaire* du gaz considéré, c'est-à-dire la masse d'une mole du gaz exprimée en kg. La constante $\mathrm{R_m}$ est reliée à la *constante de Boltzmann*

$$\mathrm{k_B} = 1,3806 \times 10^{-23}\, \frac{\mathrm{J}}{\mathrm{K}}$$

et le *nombre d'Avogadro*

$$\mathrm{N_A} = 6,022 \times 10^{23}\, \frac{1}{\mathrm{mol}}$$

par

$$\mathrm{R_m} = \mathrm{N_A k_B}\,.$$

Puisque la masse molaire de l'air sec vaut

$$\mathrm{M_{air}} = 0,0289644\, \frac{\mathrm{kg}}{\mathrm{mol}}\,,$$

la constante spécifique de l'air sec est

$$\mathrm{r_{air}} = \frac{\mathrm{R_m}}{\mathrm{M_{air}}} = 287,058\, \frac{\mathrm{J}}{\mathrm{K\,kg}}\,.$$

La loi d'état de l'air sec sous l'hypothèse d'un gaz parfait est alors donnée par

$$p(\boldsymbol{x}, t) = \mathrm{r_{air}} \varrho(\boldsymbol{x}, t)\, T(\boldsymbol{x}, t)\ .$$

En général, l'énergie interne e_{int} du fluide dépend de la pression p et de la température T. Bien que Cedre prenne en charge la modélisation des gaz réels, il suffit pour les calculs effectués dans le cadre de cette thèse d'adopter l'hypothèse des gaz parfaits, pour lesquels l'énergie interne ne dépend que de la température. L'air est principalement un mélange de gaz diatomiques composé à 78% d'azote $\mathrm{N_2}$ et à 21% d'oxygène $\mathrm{O_2}$. D'après la théorie cinétique des gaz, la densité d'énergie interne d'une mole de gaz parfait diatomique est égale à

$$e_{\text{int}}(\boldsymbol{x}, t) = \frac{5}{2} \frac{\mathrm{R_m}}{\mathrm{M_{gaz}}} T(\boldsymbol{x}, t) = \mathrm{c_V} T(\boldsymbol{x}, t)$$

où $\mathrm{c_V}$ est la capacité thermique massique à volume constant. La capacité massique à pression constante $\mathrm{c_p}$ d'un gaz parfait peut être déterminée à partir de la capacité massique à volume constant selon la formule

$$\mathrm{c}_p = \mathrm{c_V} + \frac{\mathrm{R_m}}{\mathrm{M_{gaz}}} = \frac{7}{2}\frac{\mathrm{R_m}}{\mathrm{M_{gaz}}}$$

où la deuxième égalité est vraie pour un gaz idéal diatomique.

Ensuite, il est nécessaire d'exprimer le tenseur des contraintes visqueuses en fonction des autres variables. Pour les écoulements d'air simulés dans les chapitres 13 et 14, on considère l'air comme un fluide newtonien. Cela signifie que le tenseur des contraintes visqueuses est donné par

$$\boldsymbol{\tau}\left(\boldsymbol{x},t\right) = \lambda_{\mathrm{L}}\left(\boldsymbol{\nabla}\cdot\boldsymbol{v}\left(\boldsymbol{x},t\right)\right)\boldsymbol{\delta}^{(2)} + 2\mu_{\mathrm{L}}\,\boldsymbol{\nabla}\odot\boldsymbol{v}\left(\boldsymbol{x},t\right)$$

où $\odot$ est le produit tensoriel symétrique défini par (5.2.19) et $\boldsymbol{\delta}^{(2)}$ est le tenseur identité défini par (5.2.23). Les grandeurs λ_{L} et μ_{L} sont respectivement le premier et le deuxième *coefficient de Lamé*. Le deuxième coefficient de Lamé est aussi appelé *viscosité dynamique* ou *viscosité moléculaire*. Dans la suite, on adopte *l'hypothèse de Stokes*

$$\lambda_{\mathrm{L}} = -\frac{2}{3}\mu_{\mathrm{L}}\,, \tag{3.2.7}$$

ce qui donne le tenseur des contraintes visqueuses

$$\boldsymbol{\tau}\left(\boldsymbol{x},t\right) = \mu_{\mathrm{L}}\left[2\,\boldsymbol{\nabla}\odot\boldsymbol{v}\left(\boldsymbol{x},t\right) - \frac{2}{3}\left(\boldsymbol{\nabla}\cdot\boldsymbol{v}\left(\boldsymbol{x},t\right)\right)\boldsymbol{\delta}^{(2)}\right]. \tag{3.2.8}$$

La viscosité dynamique μ_{L} n'est pas constante. Aux températures considérées, μ_{L} dépend uniquement de la température. Cette dépendance peut être modélisée par la *formule de Sutherland*

$$\mu_{\mathrm{L}} = \mu_{\mathrm{ref}}\left(\frac{T}{T_{\mathrm{ref}}}\right)^{3/2}\frac{T_{\mathrm{ref}} + 110,4\,\mathrm{K}}{T + 110,4\,\mathrm{K}} \tag{3.2.9}$$

où $T_{\mathrm{ref}} = 273,15\,\mathrm{K}$ et $\mu_{\mathrm{ref}} = 1,716\,10^{-5}\,\frac{\mathrm{kg}}{\mathrm{ms}}$. La viscosité cinématique ν_{cin} s'obtient par la division de μ_{L} par la densité

$$\nu_{\mathrm{cin}} = \frac{\mu_{\mathrm{L}}}{\varrho}\,.$$

Il reste encore à définir le flux de chaleur $\boldsymbol{q}_{th}\left(\boldsymbol{x},t\right)$. La loi de Fourier permet de l'exprimer comme fonction de la température

$$\boldsymbol{q}_{th}\left(\boldsymbol{x},t\right) = -\kappa_{\mathrm{th}}\boldsymbol{\nabla}\left(T\left(\boldsymbol{x},t\right)\right) \tag{3.2.10}$$

où la grandeur κ_{th} est appelée la *conductivité thermique*. La conductivité thermique caractérise le comportement des matériaux lors du transfert thermique par conduction. Elle est étroitement liée à la *diffusivité thermique* α_{th} par la formule

$$\alpha_{\mathrm{th}} = \frac{\kappa_{\mathrm{th}}}{\varrho\mathrm{c}_p}$$

où c_p est la chaleur spécifique à pression constante. La diffusivité thermique caractérise la facilité avec laquelle la chaleur diffuse dans le matériau. Les valeurs de la viscosité dynamique et de la diffusivité thermique sont comparées entre elles au moyen du *nombre de Prandtl*

$$\mathrm{Pr} = \frac{\nu_{\mathrm{cin}}}{\alpha_{\mathrm{th}}} = \frac{\mu_{\mathrm{L}}\mathrm{c}_p}{\kappa_{\mathrm{th}}}\,.$$

3.3. Éléments de la simulation des grandes échelles

Cette section présente les éléments de la simulation des grandes échelles (SGE) qui sont nécessaires pour modéliser la turbulence dans les simulations d'écoulement décrites dans les chapitres 13 et 14. La description donnée ici suit les lignes du livre [**92**] qui est un ouvrage de référence pour la simulation des grandes échelles dans le cas des écoulements incompressibles. L'adaptation des modèles au cas des écoulements compressibles reprend l'approche utilisée dans [**75**] et [**17**]. Le terme anglais pour simulation des grandes échelles est *large eddy simulation*, abrégé LES.

Considérons la simulation d'un écoulement de fluide. Pour obtenir un résultat très proche de la réalité physique, la simulation doit résoudre toutes les échelles de la dynamique de l'écoulement. Cela signifie que les échelles de la discrétisation, c'est-à-dire le diamètre des mailles et le pas de temps, doivent être petites par rapport à toutes les échelles de l'écoulement. Notons L_{sc} l'échelle de la dynamique qui porte le plus d'énergie et η_{sc} l'échelle de Kolmogorov, c'est-à-dire la plus petite échelle de la dynamique de l'écoulement. Pour un écoulement turbulent homogène et isotrope, il est possible d'estimer le rapport entre L_{sc} et η_{sc} par

$$\frac{\mathrm{L}_{sc}}{\eta_{\mathrm{sc}}} = \mathrm{O}\left(\mathrm{Re}^{\frac{3}{4}}\right)$$

où Re est le nombre de Reynolds, cf. [**92**]. Pour simuler l'écoulement dans un cube de volume L_{sc}^3 , il faut donc simuler $\mathrm{O}\left(\mathrm{Re}^{9/4}\right)$ degrés de liberté. Une estimation similaire vaut pour le rapport des échelles temporelles.

À l'heure actuelle, les ordinateurs sont encore trop faibles pour pouvoir simuler un tel nombre de degrés de liberté. Dans les écoulements qui intéressent l'aéronautique, les nombres de Reynolds peuvent atteindre des ordres de grandeur de 10^8. Une solution consiste à simuler uniquement une partie des degrés de liberté de façon directe. L'interaction entre ces degrés de liberté et ceux qui ne sont pas simulés directement doit être prise en compte par un modèle. Cela se fait en général par l'ajout de termes spécifiques dans les équations qui gouvernent les degrés de liberté simulés. Ces termes décrivent alors uniquement une moyenne statistique de l'action des degrés de liberté négligés.

La simulation des grandes échelles réduit le nombre de degrés de liberté par une séparation entre grandes et petites échelles à l'aide d'une longueur de coupure $\overline{\Delta}$. Dans l'espace spectral, la longueur de coupure correspond à une fréquence de coupure. Les fréquences inférieures à la fréquence de coupure sont résolues par la simulation alors que les fréquences supérieures sont modélisées.

Pour effectuer explicitement cette séparation des échelles, on applique un filtre aux équations de Navier-Stokes. Pour des raisons de simplicité, on suppose ce filtre isotrope et homogène, c'est-à-dire indépendant de la position et de l'orientation dans l'espace. Dans l'espace physique, un filtre homogène et isotrope peut être représenté par un opérateur de convolution. La partie résolue d'une variable $\phi(\boldsymbol{x}, t)$ est donnée par

$$\overline{\phi}(\boldsymbol{x}, t) = \int_{\mathbb{R}^3} \int_{-\infty}^{\infty} G\left(\boldsymbol{x} - \boldsymbol{x}', t - t'\right) \phi\left(\boldsymbol{x}', t'\right) \, dt' \, dx' \tag{3.3.1}$$

où G est le noyau de convolution du filtre. Dans l'espace spectral, les transformés de Fourier obéissent à la relation

$$\widehat{\overline{\phi}}(\boldsymbol{k}, \omega) = \widehat{G}(\boldsymbol{k}, \omega) \, \widehat{\phi}(\boldsymbol{k}, \omega) \ .$$

La partie non résolue de ϕ est donnée par

$$\phi'(\boldsymbol{x}, t) = \phi(\boldsymbol{x}, t) - \overline{\phi}(\boldsymbol{x}, t) \ .$$

Afin de pouvoir manipuler les équations de Navier-Stokes après filtrage, on exige que le filtre préserve les constantes

$$\overline{\phi}_0 = \phi_0 \, , \tag{3.3.2}$$

qu'il soit linéaire

$$\overline{\phi}(\boldsymbol{x}, t) + \overline{\psi}(\boldsymbol{x}, t) = \overline{(\phi + \psi)}(\boldsymbol{x}, t) \tag{3.3.3}$$

et qu'il commute avec la dérivation

$$\overline{\frac{\partial \phi}{\partial t}}(\boldsymbol{x}, t) = \frac{\partial \overline{\phi}}{\partial t}(\boldsymbol{x}, t) \ , \ \overline{\boldsymbol{\nabla} \phi}(\boldsymbol{x}, t) = \boldsymbol{\nabla} \overline{\phi}(\boldsymbol{x}, t) \ . \tag{3.3.4}$$

Il faut noter que le filtrage décrit ci-dessus n'est pas adapté aux équations de Navier-Stokes compressibles car il donne un nombre important de termes inconnus. Pour adapter la simulation des grandes échelles aux écoulements compressibles, on poursuit ici l'approche utilisée dans [**75**] et [**17**].

On commence par introduire un filtrage pondéré par la masse volumique qui s'inspire d'une démarche proposée par A. Favre et al., cf. [**49**]. Pour ce filtre, appelé *filtre de Favre* ou *opérateur de Favre*, la partie résolue d'une variable $\phi(\boldsymbol{x},t)$ est notée $\widetilde{\phi}(\boldsymbol{x},t)$ et définie par

$$\widetilde{\phi}(\boldsymbol{x},t) = \frac{\overline{(\varrho\phi)}(\boldsymbol{x},t)}{\overline{\varrho}(\boldsymbol{x},t)}\,. \tag{3.3.5}$$

La partie non résolue de la variable $\phi(\boldsymbol{x},t)$ est définie par

$$\phi"(\boldsymbol{x},t) \triangleq \phi(\boldsymbol{x},t) - \widetilde{\phi}(\boldsymbol{x},t)\ .$$

La partie non résolue $\phi"(\boldsymbol{x},t)$ s'appelle également la *partie sous-maille* de $\phi(\boldsymbol{x},t)$ car elle représente les échelles inférieures à la longueur de coupure $\overline{\Delta}$.

L'opérateur de Favre satisfait (3.3.2) et (3.3.3) mais ne satisfait plus (3.3.4). On peut considérer ce filtrage comme un changement de variable qui est bien défini car $\varrho(\boldsymbol{x},t) > 0$. Son intérêt réside dans le fait qu'il donne des équations dont la structure est proche des équations de Navier-Stokes, cf. [**75**].

Afin d'obtenir les équations de Navier-Stokes filtrées, on applique d'abord le filtrage homogène défini par (3.3.1) aux équations de conservation de la masse (3.2.4), de la quantité de mouvement (3.2.5) et de l'énergie (3.2.6). Les termes immédiatement calculables sont placés dans le membre de gauche et les termes sous-maille à modéliser dans le membre de droite. On omet la dépendance des variables par rapport à $\boldsymbol{x}$ et t pour rendre les formules moins encombrantes.

L'équation de conservation de la masse devient, avec le changement de variables de Favre (3.3.5)

$$\frac{\partial\overline{\varrho}}{\partial t} + \boldsymbol{\nabla}\cdot(\overline{\varrho}\widetilde{\boldsymbol{v}}) = 0\,. \tag{3.3.6}$$

L'équation de la quantité de mouvement pour la simulation des grandes échelles s'écrit sous la forme

$$\frac{\partial(\overline{\varrho}\widetilde{\boldsymbol{v}})}{\partial t} + \boldsymbol{\nabla}\cdot(\overline{\varrho}\widetilde{\boldsymbol{v}}\otimes\widetilde{\boldsymbol{v}}) + \boldsymbol{\nabla}\overline{p} - \boldsymbol{\nabla}\cdot\widetilde{\boldsymbol{\tau}} = -\boldsymbol{a}_1 + \boldsymbol{a}_2 \tag{3.3.7}$$

où les termes dans le membre de droite sont donnés par

$$\boldsymbol{a}_1 = \boldsymbol{\nabla}\cdot\left(\overline{\varrho}\widetilde{\boldsymbol{v}\otimes\boldsymbol{v}} - \overline{\varrho}\widetilde{\boldsymbol{v}}\otimes\widetilde{\boldsymbol{v}}\right) \tag{3.3.8}$$

$$\boldsymbol{a}_2 = \boldsymbol{\nabla}\cdot(\overline{\boldsymbol{\tau}} - \widetilde{\boldsymbol{\tau}})\ . \tag{3.3.9}$$

Ces termes s'appellent *termes sous-maille* car ils représentent l'influence des parties sous-maille sur les échelles résolues.

Le terme $\boldsymbol{a}_1$ contient le tenseur

$$\boldsymbol{\tau}^{(\mathrm{sgs})} \triangleq \overline{\varrho}\left(\widetilde{\boldsymbol{v}\otimes\boldsymbol{v}} - \widetilde{\boldsymbol{v}}\otimes\widetilde{\boldsymbol{v}}\right) \tag{3.3.10}$$

qui s'appelle *tenseur sous-maille*. La décomposition de Leonard, cf. [**77**], permet d'écrire $\boldsymbol{\tau}^{(\mathrm{sgs})}$ sous la forme

$$\boldsymbol{\tau}^{(\mathrm{sgs})} = \overline{\varrho}\left(\widetilde{\widetilde{\boldsymbol{v}}\otimes\widetilde{\boldsymbol{v}}} - \widetilde{\boldsymbol{v}}\otimes\widetilde{\boldsymbol{v}}\right) + \overline{\varrho}\left(\widetilde{\boldsymbol{v}"\otimes\widetilde{\boldsymbol{v}}} + \widetilde{\widetilde{\boldsymbol{v}}\otimes\boldsymbol{v}"}\right) + \overline{\varrho}\left(\widetilde{\boldsymbol{v}"\otimes\boldsymbol{v}"}\right)\ . \tag{3.3.11}$$

Le premier tenseur dans le membre de droite de (3.3.11) s'appelle *tenseur de Leonard* et décrit les interactions entre les grandes échelles. Il est directement calculable car il dépend uniquement des parties résolues de $\boldsymbol{v}$ et ϱ. Le deuxième tenseur dans le membre de droite de (3.3.11) s'appelle *tenseur des contraintes croisées* et le troisième est le *tenseur de Reynolds* qui décrit l'interaction entre les échelles sous-maille.

Le terme $\boldsymbol{a}_2$ est composé du tenseur des contraintes visqueuses des grandes échelles $\widetilde{\boldsymbol{\tau}}$, défini par

$$\widetilde{\boldsymbol{\tau}} \triangleq \mu_{\mathrm{L}}\left(\widetilde{T}\right)\left\{2\,\nabla\odot\widetilde{\boldsymbol{v}} - \frac{2}{3}\,(\nabla\cdot\widetilde{\boldsymbol{v}})\,\boldsymbol{\delta}^{(2)}\right\}\ , \tag{3.3.12}$$

et du tenseur $\overline{\boldsymbol{\tau}}$ issu du filtrage du tenseur des contraintes visqueuses $\boldsymbol{\tau}$. Il est généralement admis que la viscosité dynamique μ_{L} et le gradient de la vitesse $\boldsymbol{v}$ sont décorrélés, ce qui permet d'écrire le tenseur $\overline{\boldsymbol{\tau}}$ sous la forme

$$\overline{\boldsymbol{\tau}} = \overline{\mu_{\mathrm{L}}\left(T\right)}\overline{\left\{2\nabla \odot \boldsymbol{v} - \frac{2}{3}\left(\nabla \cdot \boldsymbol{v}\right)\boldsymbol{\delta}^{(2)}\right\}}. \tag{3.3.13}$$

Le tenseur $\boldsymbol{\delta}^{(2)}$ dans (3.3.12) et (3.3.13) est le tenseur (5.2.23) dont les composantes sont le symbole de Kronecker.

Pour établir une équation pour la conservation de l'énergie des grandes échelles, on adopte la notion d'énergie calculable, proposée par Vreman dans [**111**] et définie par

$$\widehat{\varrho e_{\mathrm{tot}}} \triangleq \frac{1}{\gamma - 1}\overline{p} + \frac{1}{2}\overline{\varrho}\,\widetilde{\boldsymbol{v}} \cdot \widetilde{\boldsymbol{v}} \tag{3.3.14}$$

pour un gaz parfait à c_p constant. Les grandeurs filtrées de la pression $\overline{p}$, de la densité de masse $\overline{\varrho}$ et de la température $\widetilde{T}$ sont reliées par la loi d'état filtrée

$$\overline{p} = \mathrm{r}_{\mathrm{air}}\,\overline{\varrho}\,\widetilde{T}\,. \tag{3.3.15}$$

La dérivation en temps de (3.3.14) donne

$$\frac{\partial\widehat{\varrho e_{\mathrm{tot}}}}{\partial t} = \frac{1}{\gamma - 1}\frac{\partial\overline{p}}{\partial t} + \frac{1}{2}\left(\widetilde{\boldsymbol{v}} \cdot \widetilde{\boldsymbol{v}}\right)\frac{\partial\overline{\varrho}}{\partial t} + \overline{\varrho}\,\widetilde{\boldsymbol{v}} \cdot \frac{\partial\widetilde{\boldsymbol{v}}}{\partial t}\,. \tag{3.3.16}$$

Il est possible de reformuler les deux derniers termes dans le membre de droite de (3.3.16) en utilisant (3.3.6)

$$\begin{aligned}\frac{1}{2}\left(\widetilde{\boldsymbol{v}} \cdot \widetilde{\boldsymbol{v}}\right)\frac{\partial\overline{\varrho}}{\partial t} + \overline{\varrho}\widetilde{\boldsymbol{v}} \cdot \frac{\partial\widetilde{\boldsymbol{v}}}{\partial t} &= -\frac{1}{2}\left(\widetilde{\boldsymbol{v}} \cdot \widetilde{\boldsymbol{v}}\right)\frac{\partial\overline{\varrho}}{\partial t} + \widetilde{\boldsymbol{v}} \cdot \left(\widetilde{\boldsymbol{v}}\frac{\partial\overline{\varrho}}{\partial t}\right) + \widetilde{\boldsymbol{v}} \cdot \left(\frac{\partial\widetilde{\boldsymbol{v}}}{\partial t}\overline{\varrho}\right) = \\ &= -\frac{1}{2}\left(\widetilde{\boldsymbol{v}} \cdot \widetilde{\boldsymbol{v}}\right)\frac{\partial\overline{\varrho}}{\partial t} + \widetilde{\boldsymbol{v}} \cdot \frac{\partial\left(\overline{\varrho}\widetilde{\boldsymbol{v}}\right)}{\partial t} = \frac{1}{2}\left(\widetilde{\boldsymbol{v}} \cdot \widetilde{\boldsymbol{v}}\right)\nabla\left(\overline{\varrho}\widetilde{\boldsymbol{v}}\right) + \widetilde{\boldsymbol{v}} \cdot \frac{\partial\left(\overline{\varrho}\widetilde{\boldsymbol{v}}\right)}{\partial t}\,.\end{aligned} \tag{3.3.17}$$

L'utilisation de (3.3.17) dans (3.3.16) donne une équation d'évolution pour $\widehat{\varrho e_{\mathrm{tot}}}$

$$\frac{\partial\widehat{\varrho e_{\mathrm{tot}}}}{\partial t} = \frac{1}{\gamma - 1}\frac{\partial\overline{p}}{\partial t} + \widetilde{\boldsymbol{v}} \cdot \frac{\partial\left(\overline{\varrho}\widetilde{\boldsymbol{v}}\right)}{\partial t} + \frac{1}{2}\left(\widetilde{\boldsymbol{v}} \cdot \widetilde{\boldsymbol{v}}\right)\nabla\left(\overline{\varrho}\widetilde{\boldsymbol{v}}\right)\,. \tag{3.3.18}$$

L'insertion de l'équation de la quantité de mouvement (3.3.7) dans (3.3.18) permet alors d'établir l'équation d'évolution de l'énergie calculable

$$\frac{\partial\widehat{\varrho e_{\mathrm{tot}}}}{\partial t} + \nabla \cdot \left\{\widetilde{\boldsymbol{v}}\left(\widehat{\varrho e_{\mathrm{tot}}} + \overline{p}\right)\right\} - \nabla \cdot \left(\widetilde{\boldsymbol{\tau}} \cdot \widetilde{\boldsymbol{v}}\right) + \nabla\widetilde{\boldsymbol{q}}_{\mathrm{th}} = -\boldsymbol{b}_1 - \boldsymbol{b}_2 - \boldsymbol{b}_3 + \boldsymbol{b}_4 + \boldsymbol{b}_5 + \boldsymbol{b}_6 - \boldsymbol{b}_7 \tag{3.3.19}$$

où $\widetilde{\boldsymbol{\tau}}$ est donné par (3.3.12) et $\widetilde{\boldsymbol{q}}_{th}$ est le flux de chaleur calculable

$$\widetilde{\boldsymbol{q}}_{\mathrm{th}} = -\kappa_{\mathrm{th}}\boldsymbol{\nabla}\widetilde{T}\,. \tag{3.3.20}$$

La somme dans le membre de droite de (3.3.19) regroupe les termes sous-maille donnés par

$$\boldsymbol{b}_1 = \frac{1}{\gamma - 1}\nabla \cdot \left(\overline{p\boldsymbol{v}} - \overline{p}\widetilde{\boldsymbol{v}}\right) \tag{3.3.21}$$

$$\boldsymbol{b}_2 = \overline{p\nabla \cdot \boldsymbol{v}} - \overline{p}\nabla \cdot \widetilde{\boldsymbol{v}} \tag{3.3.22}$$

$$\boldsymbol{b}_3 = \nabla \cdot \left(\boldsymbol{\tau}^{(\mathrm{sgs})} \cdot \widetilde{\boldsymbol{v}}\right) \tag{3.3.23}$$

$$\boldsymbol{b}_4 = \boldsymbol{\tau}^{(\mathrm{sgs})} \bullet \left(\nabla \otimes \widetilde{\boldsymbol{v}}\right) \tag{3.3.24}$$

$$\boldsymbol{b}_5 = \overline{\boldsymbol{\tau} \bullet \left(\nabla \otimes \widetilde{\boldsymbol{v}}\right)} - \overline{\boldsymbol{\tau}} \bullet \left(\nabla \otimes \widetilde{\boldsymbol{v}}\right) \tag{3.3.25}$$

$$\boldsymbol{b}_6 = \nabla \cdot \left(\overline{\boldsymbol{\tau}} \cdot \widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{\tau}} \cdot \widetilde{\boldsymbol{v}}\right) \tag{3.3.26}$$

$$\boldsymbol{b}_7 = \nabla \cdot \left(\overline{\boldsymbol{q}} - \widetilde{\boldsymbol{q}}\right)\,. \tag{3.3.27}$$

Afin de pouvoir résoudre le système des équations (3.3.6), (3.3.7) et (3.3.19), il est nécessaire de modéliser les termes sous-maille $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ et $\boldsymbol{b}_1$à $\boldsymbol{b}_7$.

Le terme le plus important est le terme $\boldsymbol{a}_1$ qui est le seul terme présent dans le cas des écoulements incompressibles. Il contient le tenseur sous-maille $\boldsymbol{\tau}^{(\mathrm{sgs})}$ qui n'est pas directement calculable car il contient les échelles non résolues $\boldsymbol{v}$" de la vitesse $\boldsymbol{v}$. Un classement introduit par Sagaut [**92**, p. 72] permet de distinguer deux catégories de modèles.

(1) Les *modèles structurels* essaient d'approcher la structure du tenseur $\boldsymbol{\tau}^{(\mathrm{sgs})}$. Dans cette approche, l'hypothèse de modélisation consiste à exprimer $\boldsymbol{v}$" ou $\boldsymbol{\tau}^{(\mathrm{sgs})}$ en fonction des échelles résolues $\widetilde{\boldsymbol{v}}$, ce qui nécessite une bonne connaissance de la structure des petites échelles. La dynamique des petites échelles doit être indépendante de l'évolution des échelles résolues ou dépendre de façon suffisamment simple de cette dernière.

(2) Au lieu d'approcher le tenseur $\boldsymbol{\tau}^{(\mathrm{sgs})}$, les *modèles fonctionnels* tentent de modéliser l'action des échelles sous-maille $\boldsymbol{v}$" sur les échelles résolues $\widetilde{\boldsymbol{v}}$ par l'introduction de termes dissipatifs ou dispersifs dans l'équation de la quantité de mouvement. Dans cette approche, l'hypothèse de modélisation consiste essentiellement à exprimer $\nabla \cdot \boldsymbol{\tau}^{(\mathrm{sgs})}$ en fonction des échelles résolues $\widetilde{\boldsymbol{v}}$. Cette approche nécessite une bonne connaissance des mécanismes d'échange interéchelle. Il faut que la dynamique des petites échelles soit universelle et indépendante des échelles résolues de l'écoulement.

Les calculs effectués dans le cadre de cette thèse reposent sur l'approche fonctionnelle.

Le cadre naturel pour l'approche fonctionnelle est la théorie de la turbulence développée par Kolmogorov, cf. [**72**]. Le processus de transfert d'énergie entre les différentes échelles de la turbulence peut être présenté sous une forme simplifiée, la cascade de Kolmogorov.

(1) L'écoulement moyen transmet de l'énergie cinétique aux grosses structures tourbillonnaires.

(2) L'énergie cinétique est ensuite transférée d'une échelle supérieure vers l'échelle immédiatement inférieure par des phénomènes d'étirement tourbillonnaire.

(3) Aux plus petites échelles, la viscosité moléculaire dissipe l'énergie cinétique sous forme de chaleur.

La cascade de Kolmogorov donne lieu à la notion de viscosité sous-maille qui repose sur l'hypothèse que le mécanisme de transfert d'énergie des échelles résolues vers les échelles sous-maille ressemble aux mécanismes moléculaires de diffusion. Cela permet de représenter le tenseur sous-maille $\boldsymbol{\tau}^{(\mathrm{sgs})}$ comme un tenseur de contraintes visqueuses

$$\boldsymbol{\tau}^{(\mathrm{sgs})} = \mu_{\mathrm{L}}^{(\mathrm{sgs})} \left[2\, \boldsymbol{\nabla} \odot \widetilde{\boldsymbol{v}} - \frac{2}{3}\, (\boldsymbol{\nabla} \cdot \widetilde{\boldsymbol{v}})\, \boldsymbol{\delta}^{(2)} \right] \tag{3.3.28}$$

où la viscosité sous-maille $\mu_{\mathrm{L}}^{(\mathrm{sgs})}$ doit être modélisée.

Un exemple important de modèle sous-maille est le modèle de Smagorinsky [**92**, p. 107] qui exprime la viscosité sous-maille sous la forme

$$\mu_{\mathrm{L}}^{(\mathrm{sgs})} = \left(\mathrm{C_S}\overline{\Delta}\right)^2 \sqrt{2\, (\boldsymbol{\nabla} \odot \widetilde{\boldsymbol{v}}) \bullet (\boldsymbol{\nabla} \odot \widetilde{\boldsymbol{v}})} \tag{3.3.29}$$

où $\boldsymbol{\nabla} \odot \widetilde{\boldsymbol{v}}$ est le tenseur des taux de déformation, $\overline{\Delta}$ est la longueur de coupure et $\mathrm{C_S}$ est une constante. Sous l'hypothèse que le spectre d'énergie de l'écoulement reste constant dans le temps [**92**, p. 98] la constante $\mathrm{C_S}$ s'évalue à

$$\mathrm{C_S} \approx 0,18\,.$$

Pour les besoins de la simulation, $\overline{\Delta}$ peut être identifié au diamètre des mailles. Les équations (3.3.28) et (3.3.29) permettent donc de modéliser le terme $\boldsymbol{a}_1$ en fonction des échelles résolues.

Vreman et al., cf. [**112**] ont effectué des simulations numériques directes d'une couche de mélange pour évaluer l'importance des différents termes sous-maille. Le résultat indique qu'il est possible de négliger la contribution du terme $\boldsymbol{a}_2$ par rapport à celle de $\boldsymbol{a}_1$. Parmi les termes $\boldsymbol{b}_1$ à $\boldsymbol{b}_7$ dans l'équation de l'énergie, les termes $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ et $\boldsymbol{b}_3$ sont jugés prépondérants et les termes $\boldsymbol{b}_4$ à $\boldsymbol{b}_7$ peuvent donc être négligés.

Dans le modèle de Smagorinsky, le terme $\boldsymbol{b}_3$ devient directement calculable grâce aux équations (3.3.28) et (3.3.29). L'introduction d'une *conductivité thermique sous-maille* $\kappa_{\text{th}}^{(\text{sgs})}$ et l'analogie de Prandtl, cf. [**112, 76**], permettent de modéliser la somme de $\boldsymbol{b}_1$ et $\boldsymbol{b}_2$ comme

$$\boldsymbol{b}_1 + \boldsymbol{b}_2 = -\boldsymbol{\nabla} \cdot \left(\kappa_{\text{th}}^{(\text{sgs})} \boldsymbol{\nabla} \widetilde{T} \right) .$$

La conductivité thermique sous-maille $\kappa_{\text{th}}^{(\text{sgs})}$ peut être calculée au moyen du *nombre de Prandtl sous-maille* $\text{Pr}^{(\text{sgs})}$ par la formule

$$\kappa_{\text{th}}^{(\text{sgs})} = \frac{\mu_{\text{L}}^{(\text{sgs})} \text{c}_p}{\text{Pr}^{(\text{sgs})}}$$

où c_p est la chaleur spécifique à pression constante. Dans le cadre de cette étude, le nombre de Prandtl sous-maille a été fixé à

$$\text{Pr}^{(\text{sgs})} = 0,9$$

qui est la valeur de cette constante pour les modèles RANS, cf. [**17**].

Le modèle de Smagorinsky a été implémenté dans Cedre dans le cadre de la thèse de N. Bertier [**17**]. Concernant la présente étude, le modèle de Smagorinsky a servi pour les calculs tridimensionnels décrits dans les chapitres 13 et 14.

CHAPITRE 4

La chaîne de calculs CEDRE

4.1. Domaines d'application et modèles physiques

La chaîne de calculs CEDRE est un outil logiciel développé par l'ONERA depuis la fin des années 1990 pour la simulation numérique dans le domaine de l'énergétique et la propulsion, pour des applications industrielles et de recherche.

Les disciplines scientifiques concernées sont diverses, la mécanique des fluides étant en pratique au centre de la plupart des applications. Cependant, les transferts de chaleur par rayonnement et par conduction dans les parois solides doivent souvent être pris en compte dans une interaction très étroite avec le fluide. D'autres sous-systèmes physiques modélisés par des solveurs en cours d'intégration dans la chaîne ne seront pas mentionnés ici.

4.1.1. Le milieu fluide. La nature du milieu fluide est très variable suivant les applications :

- dans certains cas, le fluide est assimilable à un mélange de gaz parfaits à chaleurs massiques dépendant de la température ; les versions initiales du logiciel étaient limitées à cette hypothèse ;
- le modèle thermodynamique de CEDRE a été récemment profondément remanié et généralisé pour pouvoir prendre en compte des mélanges dont chaque composant est gouverné par une loi d'état $\rho = \rho(p, T)$ quelconque, ce qui rend désormais accessibles de très nombreuses applications nouvelles (conditions éloignées du gaz parfait, liquides, mélanges à l'équilibre chimique etc.).

Les phénomènes à prendre en compte sont variés, et CEDRE comprend un grand nombres de modèles :

- dans de nombreuse applications, le fluide est le siège de réactions chimiques se traduisant par des sources de masses pour les différentes espèces ;
- le fluide transporte souvent une phase dispersée (gouttelettes liquides ou particules solides) que l'on peut simuler dans l'approche lagrangienne ou eulérienne. Dans les deux cas, des modèles décrivent les transferts de masses, quantité de mouvement et énergie entre phases.
- l'écoulement est en général turbulent : dans certains cas, l'approche RANS (Reynolds Averaged Navier-Stokes Equations) est considérée comme suffisante, mais la simulation des grandes échelles et les approches voisines sont de plus en plus accessibles. CEDRE comprend donc un ensemble de modèles RANS (modèles à 2 équations ou plus, variantes ASM etc.) et LES (modèle de Smagorinsky...), la turbulence ayant également des effets croisés avec les modèles précédents (combustion turbulente, interaction avec la phase dispersée) ;
- enfin, les applications industrielles exigent souvent la modélisation de phénomènes complexes au niveau des frontières (entrées et sorties avec conditions particulières, parois débitantes ou poreuses etc.)

4.1.2. Conduction dans les parois solides. Les échanges de chaleur dans les parois solides interviennent de manière significative dans de nombreux cas : le code CEDRE comprend donc un module permettant de traiter la conduction thermique dans un milieu au repos. Les modèles physiques actuels pour ce solide (milieu isotrope de chaleur massique et conductivité thermique dépendant de la température) et ses interactions avec le fluide sont relativement élémentaires mais sont appelés à être généralisés.

4.1.3. Rayonnement thermique. Deux approches sont disponibles pour la simulation du rayonnement thermique entre les parois et en volume :

– une méthode pour la résolution de l'équation de transfert radiatif gouvernant la luminance monochromatique, intégrée sur les longueurs d'onde, les directions de propagation et le volume physique ;
– une méthode de type *Monte Carlo* pour la simulation directe des trajectoires des photons.

4.2. Méthodes numériques

Les modèles cités dans la section 4.1 font appel à des maillages dont les cellules sont utilisées :
– comme volumes de contrôle pour tous les solveurs reposant sur des équations de bilan (approche eulérienne pour le fluide et la phase dispersée, conduction thermique) ;
– comme bases de localisation et de trajectographie pour les méthodes particulaires (approche lagrangienne pour la phase dispersée, méthode de *Monte Carlo* pour le rayonnement).

Cedre admet un maillage constitué de cellules polyédriques :
– chaque cellule est limitée par un nombre quelconque de faces ;
– chaque face repose sur un nombre quelconque de sommets.

Ce modèle géométrique très général permet de traiter les familles de maillages classiques (tétraèdres, prismes, pyramides, hexaèdres) ou issus de méthodes novatrices (dodécaèdres, hexaèdres coupés aux limites, duals de tétraèdres, maillages raccordés le long de surfaces communes ou par chevauchement, cellules complexes issues du raffinement etc.). Il convient de signaler que le traitement des différents types de cellules est entièrement générique dans la mesure où toutes les contributions surfaciques aux grandeurs de mailles sont calculées par face : quelles que soient les cellules auxquelles elle appartient, une face interne met toujours en communication deux cellules exactement.

Tous les solveurs reposant sur des équations de bilan utilisent une méthodologie commune :
– approche « volumes finis centrés sur les cellules » ;
– stricte distinction entre discrétisation spatiale et intégration en temps. Cette approche, connue généralement sous le nom de *méthode des lignes*, permet au niveau logiciel de regrouper l'ensemble de la discrétisation spatiale dans un module commun appelable à volonté par les différentes méthodes d'intégration en temps. Sur le plan théorique, cette séparation simplifie considérablement l'étude du schéma global, qui apparaît comme la composition d'un opérateur spatial et d'un opérateur temporel ;
– la discrétisation spatiale reprend les principaux éléments de l'approche Muscl : interpolation algébrique à partir des variables de base (moyennes de mailles), limitations éventuelles, application de flux numériques décentrés pour la partie convective. Les méthodes d'interpolation pour les flux diffusifs font appel aux mêmes éléments constitutifs (moyennes et gradients moyens de maille) ;
– les divers solveurs proposent de nombreuses méthodes explicites (Runge-Kutta de différents ordres) et implicites (Euler implicite, méthodes implicites d'inspiration Runge-Kutta). Chaque méthode d'intégration en temps constitue une couche de plus haut niveau appelant le module de discrétisation spatiale chargé de fournir les résidus pour les cellules internes et limites (ainsi que leurs matrices jacobiennes par rapport aux quantités conservées dans le cas des méthodes implicites). Pour toutes les méthodes implicites, la résolution du système implicite linéarisé fait appel à une méthode itérative adaptée. Pour le fluide, une technique de gradient conjugué non symétrique (Gmres) est utilisée par défaut dans la mesure où elle est extrêmement polyvalente et efficace dans la plupart des situations.

Les méthodes numériques sont l'objet de contraintes provenant d'une part de la complexité des modèles physiques, d'autre part des difficultés propres aux maillages non structurés généraux.

4.3. Aspects logiciels

4.3.1. La chaîne CEDRE. Dès l'origine du projet, l'objectif a été de fournir une chaîne de calculs complète entre mailleurs et visualiseurs, voir figure 4.3.1 :

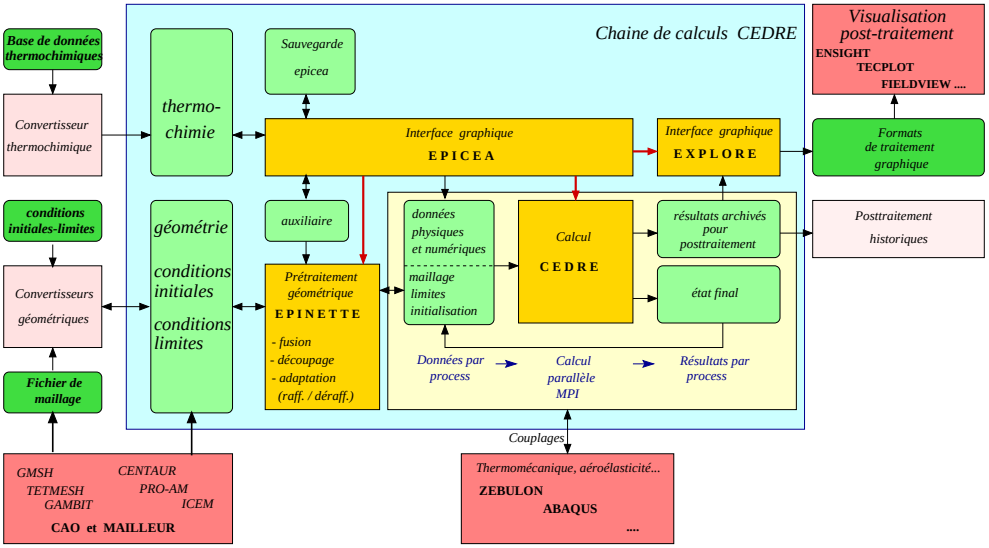


FIG. 4.3.1: La chaîne de calculs CEDRE

– certains mailleurs industriels fournissent directement des fichiers de maillage au format CEDRE, et un convertisseur permet dans le cas contraire de transformer les formats « par élément » au format « par face » utilisé par CEDRE ;
– l'interface graphique EPICEA centralise la mise en données des modèles physiques et méthodes numériques, et contrôlera à terme l'ensemble de la chaîne ;
– l'utilitaire EPINETTE regroupe divers pré-traitements géométriques, en particulier le partitionnement en vue du calcul parallèle ;
– un autre utilitaire permet le passage des bases données thermochimiques externes aux données CEDRE ;
– le code CEDRE proprement dit assure le calcul ;
– enfin, l'interface graphique EXPLORE assure des post-traitements externes et prépare les données pour un grand nombre de visualiseurs ;
– pour les modèles physiques non pris en compte par CEDRE (thermomécanique, aéroélasticité etc.) une possibilité de couplage existe avec des codes externes.

4.3.2. Calcul parallèle. Le calcul proprement dit, code CEDRE, voir section 4.3.1, est parallèle :
– Les maillages associés aux différents solveurs sont partitionnés en sous-domaines répartis sur les processeurs disponibles de manière à équilibrer les charges ;
– Les échanges d'informations entre sous-domaines appartenant à des processeurs différents se font par des fonctions de la bibliothèque parallèle MPI.

En pratique, la scalabilité des calculs est excellente, avec une accélération pratiquement proportionnelle au nombre de processeurs utilisés.

CHAPITRE 5

Introduction à la géométrie des maillages non structurés

5.1. Objectif du chapitre

Le premier objectif du chapitre est de fixer les notations mathématiques générales. Le deuxième objectif est de définir la structure et la géométrie des maillages non structurés généraux. Le chapitre explique toutes les notions géométriques nécessaires à la compréhension de la discrétisation spatiale des lois de conservation sur de tels maillages :

- définition des faces et cellules en dimensions deux et trois ;
- définition des normales, des barycentres et des moments des faces ;
- définition des barycentres et des moments des cellules.

Le troisième objectif est de définir une notation adaptée aux maillages non structurés. Cette notation doit permettre d'alléger l'écriture, en particulier en ce qui concerne les sommes sur les cellules du maillage. Le quatrième objectif consiste à définir un certain nombre de tenseurs géométriques, dérivés du maillage non structuré. Finalement, le dernier objectif est d'établir, pour ces tenseurs, des identités géométriques utilisées dans les chapitres suivants.

Il faut par ailleurs souligner que les notations et les résultats de ce chapitre peuvent s'avérer utiles pour d'autres méthodes numériques que les schémas volumes finis.

5.2. Notations mathématiques

Cette section fixe les notations mathématiques générales. Ces notations sont indispensables dans les chapitres suivants. Le premier objectif est d'introduire une notation cohérente et homogène. Le deuxième objectif est d'éviter le plus possible l'écriture explicite de sommes afin de rendre les formules moins lourdes.

5.2.1. Convention typographiques. Les conventions de notation suivantes sont utilisées dans la suite du document.

- Les vecteurs dans $\mathbb{R}^d$ sont écrits en gras comme $\boldsymbol{c}$.
- Les tenseurs dans $\mathbb{R}^d$ sont également notés en gras et leur ordre est indiqué entre parenthèses. Par exemple, $\boldsymbol{a}^{(k)}$ est un tenseur d'ordre k et de composantes $a_{i_1 \cdots i_k}$. Si le tenseur est un scalaire ou un vecteur, l'indication de l'ordre peut être omise. Pour certains tenseurs bien définis, l'indication de l'ordre est également omise.
- Les matrices sont notées par des majuscules comme H lorsqu'elles ne sont pas interprétées comme des tenseurs d'ordre deux dans $\mathbb{R}^d$. La matrice unité dans $\mathbb{R}^d$ est I_d et la transposée d'une matrice H est notée H^t.
- Les ensembles des matrices réelles et complexes avec d lignes et m colonnes sont respectivement notés par $\mathbb{M}_{d,m}(\mathbb{R})$ et $\mathbb{M}_{d,m}(\mathbb{C})$.
- Les vecteurs qui désignent des solutions semi-discrètes ou discrètes sont écrites en écriture *fraktur* comme $\mathfrak{u}$.
- Les entités géométriques comme les cellules et les faces sont notées en fonte calligraphique comme $\mathcal{A}$.
- Les ensembles sont notés en *blackboard gras* comme $\mathbb{V}$.
- Le symbole $\triangleq$ indique la définition d'un nouvel objet alors que le symbole $=$ est utilisé pour les identités qui sont des résultats.

5.2.2. Vecteurs et tenseurs. Dans cette section et la suite du document, on travaille en coordonnées cartésiennes $\{x_1, \ldots, x_d\}$ de $\mathbb{R}^d$.

- La base canonique de $\mathbb{R}^d$ est notée $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$.

– Le produit scalaire de deux vecteurs dans $\mathbb{R}^d$ est noté

$$\boldsymbol{x} \cdot \boldsymbol{y} \triangleq \sum_{j=1}^{d} x_j y_j \,. \tag{5.2.1}$$

– Pour des vecteurs dans $\mathbb{C}^N$, on introduit un produit scalaire hermitien par la définition

$$(\boldsymbol{u}, \boldsymbol{v}) \triangleq \sum_{j=1}^{N} u_j^* v_j \tag{5.2.2}$$

où u_j^* désigne le conjugué du nombre complexe u_j.

– La norme vectorielle associée à (5.2.1) est

$$\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}} \,. \tag{5.2.3}$$

– La norme (5.2.3) induit une norme pour les matrices $A \in \mathbb{M}_{d,m}(\mathbb{R})$. Cette norme est appelée *norme spectrale* et donnée par

$$\|A\|_2 = \sup_{\|\boldsymbol{x}\|_2 \leq 1} \|Ax\|_2 \,. \tag{5.2.4}$$

Une autre norme matricielle importante est la norme de Frobenius, définie par

$$\|A\|_F = \sqrt{\mathrm{trace}\,(A^t A)} \tag{5.2.5}$$

– Les tenseurs $\boldsymbol{a}^{(k)}$ d'ordre $k \geq 1$ sont définis comme des applications k-linéaires

$$\boldsymbol{a}^{(k)} : \left\{ \begin{array}{ccc} \left(\mathbb{R}^d\right)^k & \longrightarrow & \mathbb{R} \\ (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) & \longmapsto & \boldsymbol{a}^{(k)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \end{array} \right. .$$

Dans la base canonique $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$ de $\mathbb{R}^d$, le tenseur $\boldsymbol{a}^{(k)}$ est caractérisé par ses composantes

$$a_{i_1 \cdots i_k} \triangleq \boldsymbol{a}^{(k)}(\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_k}) \;,\; 1 \leq i_l \leq d \,,\; 1 \leq l \leq k \,.$$

Par convention, un tenseur $\boldsymbol{a}^{(k)}$ d'ordre $k = 0$ a une seule composante a qui est un nombre réel.

– Soit $\mathfrak{S}_k$ le groupe des permutations de l'ensemble $\{1, \ldots, k\}$. Le tenseur $\boldsymbol{a}^{(k)}$ est appelé *symétrique* si pour toute permutation $\pi \in \mathfrak{S}_k$ de l'ensemble $\{1, \ldots, k\}$

$$a_{i_1 \cdots i_k} = a_{i_{\pi(1)} \cdots i_{\pi(k)}} \,. \tag{5.2.6}$$

La *partie symétrique* du tenseur $\boldsymbol{a}^{(k)}$ est définie par

$$\mathrm{sym}\left(\boldsymbol{a}^{(k)}\right)_{i_1 \cdots i_k} \triangleq \frac{1}{k!} \sum_{\pi \in \mathfrak{S}_k} a_{i_{\pi(1)} \cdots i_{\pi(k)}} \,. \tag{5.2.7}$$

Tout tenseur symétrique $\boldsymbol{a}^{(k)}$ est égal à sa partie symétrique. Si $\boldsymbol{a}^{(2)}$ est un tenseur d'ordre deux, la définition (5.2.7) donne

$$\mathrm{sym}\left(\boldsymbol{a}^{(2)}\right)_{i_1 i_2} = \frac{1}{2}\left(a_{i_1 i_2} + a_{i_2 i_1}\right) .$$

– Le symbole $\cdot$ est aussi utilisé pour le produit d'un vecteur et d'un tenseur d'ordre deux

$$\left(\boldsymbol{x} \cdot \boldsymbol{a}^{(2)}\right)_i \triangleq \sum_{j=1}^{d} x_j a_{ji} \tag{5.2.8}$$

$$\left(\boldsymbol{a}^{(2)} \cdot \boldsymbol{x}\right)_i \triangleq \sum_{j=1}^{d} a_{ij} x_j \tag{5.2.9}$$

$$\boldsymbol{x} \cdot \boldsymbol{a}^{(2)} \cdot \boldsymbol{y} \triangleq \sum_{i=1}^{d} \sum_{j=1}^{d} x_i a_{ij} y_j \,. \tag{5.2.10}$$

– Le produit $\cdot$ se généralise aux tenseurs d'ordre plus élevé. Si $\boldsymbol{a}^{(k)}$ est un tenseur d'ordre k avec des éléments $a_{i_1\cdots i_k}$ et $\boldsymbol{b}^{(m)}$ est un tenseur d'ordre $m \leq k$ avec des éléments $b_{i_1\cdots i_m}$, on définit

$$\left(\boldsymbol{a}^{(k)} \cdot \boldsymbol{b}^{(m)}\right)_{i_1\cdots i_{k-m}} \triangleq \sum_{j_1=1}^{d} \cdots \sum_{j_m=1}^{d} a_{i_1\cdots i_{k-m}j_1\cdots j_m} b_{j_1\cdots j_m} \tag{5.2.11}$$

et

$$\left(\boldsymbol{b}^{(m)} \cdot \boldsymbol{a}^{(k)}\right)_{i_1\cdots i_{k-m}} \triangleq \sum_{j_1=1}^{d} \cdots \sum_{j_m=1}^{d} b_{j_1\cdots j_m} a_{j_1\cdots j_m i_1\cdots i_{k-m}} \,. \tag{5.2.12}$$

– Dans le cas particulier où les tenseurs sont du même ordre, $m = k$, (5.2.11) et (5.2.12) sont égales à la *contraction* de $\boldsymbol{a}^{(k)}$ et $\boldsymbol{b}^{(k)}$ qui est notée par

$$\boldsymbol{a}^{(k)} \bullet \boldsymbol{b}^{(k)} \triangleq \sum_{i_1=1}^{d} \cdots \sum_{i_k=1}^{d} a_{i_1\cdots i_k} b_{i_1\cdots i_k} \,. \tag{5.2.13}$$

Pour deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ les écritures suivantes sont alors équivalentes

$$\boldsymbol{x} \bullet \boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{y} = \sum_{j=1}^{d} x_j y_j \tag{5.2.14}$$

– Le produit tensoriel $\otimes$ rend beaucoup de formules plus légères. On utilise la définition la plus simple du produit tensoriel. Si $\boldsymbol{a}^{(k)}$ est un tenseur d'ordre k avec des éléments $a_{i_1\cdots i_k}$ et $\boldsymbol{b}^{(m)}$ est un tenseur d'ordre m avec des éléments $b_{i_1\cdots i_m}$, leur produit tensoriel $c^{(m+k)} = \boldsymbol{a}^{(k)} \otimes \boldsymbol{b}^{(m)}$ est défini par les composantes

$$c_{i_1\cdots i_{k+m}} \triangleq a_{i_1\cdots i_k} b_{i_{k+1}\cdots i_{k+m}} \,. \tag{5.2.15}$$

Un exemple fréquemment utilisé est le produit tensoriel de deux vecteurs

$$(\boldsymbol{x} \otimes \boldsymbol{y})_{ij} = x_i y_j \,.$$

Si $\boldsymbol{a}^{(2)}$ est un tenseur d'ordre deux et $\boldsymbol{x}$ et $\boldsymbol{y}$ deux vecteurs, (5.2.15), (5.2.13) et (5.2.10) permettent d'écrire de façon équivalente

$$\sum_{i=1}^{d}\sum_{j=1}^{d} x_i a_{ij} y_j = \boldsymbol{x} \cdot \boldsymbol{a}^{(2)} \cdot \boldsymbol{y} = \boldsymbol{a}^{(2)} \bullet (x \otimes y) \,. \tag{5.2.16}$$

Si $\boldsymbol{a}^{(0)}$ est un tenseur d'ordre $k = 0$, on définit

$$\boldsymbol{a}^{(0)} \otimes \boldsymbol{b}^{(m)} \triangleq a\boldsymbol{b}^{(m)} \tag{5.2.17}$$

où a est la seule composante de $\boldsymbol{a}^{(0)}$. De façon analogue, on définit

$$\boldsymbol{a}^{(k)} \otimes \boldsymbol{b}^{(0)} \triangleq b\boldsymbol{a}^{(k)} \tag{5.2.18}$$

pour un tenseur $\boldsymbol{b}^{(0)}$ d'ordre $m = 0$ ayant la seule composante b.

– Le produit symétrique d'un tenseur $\boldsymbol{a}^{(k)}$ d'ordre k et d'un tenseur $\boldsymbol{b}^{(m)}$ d'ordre m est le tenseur symétrique d'ordre $k + m$ noté par le symbole $\odot$ et défini par

$$\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)} = \operatorname{sym}\left(\boldsymbol{a}^{(k)} \otimes \boldsymbol{b}^{(m)}\right) \,. \tag{5.2.19}$$

Ses composantes sont :

$$\left(\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)}\right)_{i_1\cdots i_{k+m}} \triangleq \frac{1}{(k+m)!} \sum_{\pi\in\mathfrak{S}_{k+m}} a_{i_{\pi(1)}\cdots i_{\pi(k)}} b_{i_{\pi(k+1)}\cdots i_{\pi(k+m)}} \,.$$

Si $\boldsymbol{a}^{(k)}$ et $\boldsymbol{b}^{(m)}$ sont des tenseurs symétriques d'ordre respectif k et m, leur produit tensoriel $c^{(m+k)} = \boldsymbol{a}^{(k)} \otimes \boldsymbol{b}^{(m)}$ n'est pas nécessairement un tenseur symétrique mais leur produit symétrique $\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)}$ l'est. Pour des tenseurs d'ordre $k = 0$, on définit

$$\boldsymbol{a}^{(0)} \odot \boldsymbol{b}^{(m)} = a\boldsymbol{b}^{(m)} \tag{5.2.20}$$

et

$$\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(0)} = b\boldsymbol{a}^{(k)} \,. \tag{5.2.21}$$

– Le tenseur $\boldsymbol{\delta}^{(2k)}$ d'ordre $2k$ est défini par

$$\left(\boldsymbol{\delta}^{(2k)}\right)_{i_1\cdots i_k\, j_1\cdots j_k} \triangleq \frac{1}{k!} \sum_{\pi\in\mathfrak{S}_k} \delta_{i_1 j_{\pi(1)}} \cdots \delta_{i_k j_{\pi(k)}} \tag{5.2.22}$$

où δ_{ij} est le symbole de Kronecker. En particulier, $\boldsymbol{\delta}^{(2)}$ est un tenseur d'ordre deux dont les composantes sont le symbole de Kronecker

$$\left(\boldsymbol{\delta}^{(2)}\right)_{ij} = \delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \,. \tag{5.2.23}$$

Par conséquent, $\boldsymbol{\delta}^{(2)}$ peut être identifié à la matrice unité. On a pour tout tenseur $\boldsymbol{a}^{(k)}$ d'ordre k avec la notation (5.2.11)

$$\left(\boldsymbol{\delta}^{(2k)} \cdot \boldsymbol{a}^{(k)}\right)_{i_1\cdots i_k} = \sum_{j_1=1}^{d} \cdots \sum_{j_k=1}^{d} \frac{1}{k!} \sum_{\pi\in\mathfrak{S}_k} \delta_{i_1 j_{\pi(1)}} \cdots \delta_{i_k j_{\pi(k)}} a_{j_1\cdots j_k} = \frac{1}{k!} \sum_{\pi\in\mathfrak{S}_k} a_{i_{\pi^{-1}(1)}\cdots i_{\pi^{-1}(k)}} \tag{5.2.24}$$

où π^{-1} est la permutation inverse de π. Puisque l'ensemble des permutations $\mathfrak{S}_k$ de $\{1,\ldots,k\}$ est un groupe, chaque élément π a un élément inverse unique. La somme dans la dernière ligne de (5.2.24) peut donc aussi bien parcourir les éléments $\pi^{-1} \in \mathfrak{S}_k$ que les éléments $\pi \in \mathfrak{S}_k$. Cela permet d'écrire l'identité

$$\left(\boldsymbol{\delta}^{(2k)} \cdot \boldsymbol{a}^{(k)}\right)_{i_1\cdots i_k} = \frac{1}{k!} \sum_{\pi^{-1}\in\mathfrak{S}_k} a_{i_{\pi^{-1}(1)}\cdots i_{\pi^{-1}(k)}} = \frac{1}{k!} \sum_{\pi\in\mathfrak{S}_k} a_{i_{\pi(1)}\cdots i_{\pi(k)}} = \operatorname{sym}\left(\boldsymbol{a}^{(k)}\right)_{i_1\cdots i_k} \tag{5.2.25}$$

où $\operatorname{sym}\left(\boldsymbol{a}^{(k)}\right)$ est la partie symétrique de $\boldsymbol{a}^{(k)}$ définie par (5.2.7). En notation tensorielle, (5.2.25) devient

$$\operatorname{sym}\left(\boldsymbol{a}^{(k)}\right) = \boldsymbol{\delta}^{(2k)} \cdot \boldsymbol{a}^{(k)} \,. \tag{5.2.26}$$

Le tenseur $\boldsymbol{\delta}^{(2k)}$ définit donc par la relation (5.2.26) une application linéaire qui envoie les tenseurs $\boldsymbol{a}^{(k)}$ d'ordre k sur leur partie symétrique.

– Si $\boldsymbol{a}$ est un vecteur, il est utile de définir

$$(\boldsymbol{a})^n \triangleq \underbrace{\boldsymbol{a} \otimes \cdots \otimes \boldsymbol{a}}_{n\times} \tag{5.2.27}$$

qui est un tenseur symétrique d'ordre n de composantes

$$\left((\boldsymbol{a})^n\right)_{i_1\cdots i_n} = a_{i_1} \cdots a_{i_n} \,. \tag{5.2.28}$$

En particulier, si $\boldsymbol{b}^{(n)}$ est un tenseur symétrique d'ordre n, il est possible d'écrire de façon équivalente

$$\boldsymbol{b}^{(n)} \bullet (\boldsymbol{a})^n = \boldsymbol{b}^{(n)} \bullet \left[\underbrace{\boldsymbol{a} \otimes \cdots \otimes \boldsymbol{a}}_{n\times}\right] = \sum_{i_1=1}^{d} \cdots \sum_{i_n=1}^{d} b_{i_1\cdots i_n} a_{i_1} \cdots a_{i_n} \,.$$

Pour $n = 0$, on définit

$$(\boldsymbol{a})^0 \triangleq 1 \,. \tag{5.2.29}$$

– La dérivée d'ordre j d'une fonction v est notée $D^{(j)}v$. Pour des raisons de simplicité, la dérivée première est également notée par le symbole du gradient

$$\boldsymbol{\nabla} v \triangleq D^{(1)} v \,.$$

Si v est j fois continûment différentiable, $D^{(j)}v$ est un tenseur symétrique d'ordre j de composantes

$$\left(D^{(j)}v\right)_{i_1\cdots i_j} \triangleq \frac{\partial^j v}{\partial x_{i_1} \cdots \partial x_{i_j}} \,,\; 1 \leq i_1, \cdots, i_j \leq d \,.$$

Son évaluation en un point $\boldsymbol{x} \in \mathbb{R}^d$ est

$$\left(D^{(j)} v \Big|_{\boldsymbol{x}} \right)_{i_1 \cdots i_j} \triangleq \frac{\partial^j v(\boldsymbol{x})}{\partial x_{i_1} \cdots \partial x_{i_j}} .$$

Avec cette notation, le développement de Taylor de la fonction v à l'ordre k en $\boldsymbol{x}_0$ s'écrit

$$v(x) = \sum_{j=0}^{k} \frac{1}{j!} D^{(j)} u \Big|_{\boldsymbol{x}_0} \bullet (\boldsymbol{x} - \boldsymbol{x}_0)^j + \mathrm{O}\left(h^{k+1}\right) . \tag{5.2.30}$$

L'équation (5.2.30) utilise les notations

$$D^{(n)} u \Big|_{\boldsymbol{x}_0} \bullet (\boldsymbol{x} - \boldsymbol{x}_0)^n = D^{(n)} u \Big|_{\boldsymbol{x}_0} \bullet [(\boldsymbol{x} - \boldsymbol{x}_0) \otimes \cdots \otimes (\boldsymbol{x} - \boldsymbol{x}_0)] =$$
$$= \sum_{i_1=1}^{d} \cdots \sum_{i_n=1}^{d} \frac{\partial^j u(\boldsymbol{x}_0)}{\partial x_{i_1} \cdots \partial x_{i_n}} (x_{i_1} - x_{0,i_1}) \cdots (x_{i_n} - x_{0,i_n}) \tag{5.2.31}$$

ainsi que la convention

$$D^{(0)} u \Big|_{\boldsymbol{x}_0} \bullet (\boldsymbol{x} - \boldsymbol{x}_0)^0 \triangleq u(\boldsymbol{x}_0) . \tag{5.2.32}$$

– L'intégrale d'une fonction intégrable v sur un volume $\Omega \subseteq \mathbb{R}^d$ est notée

$$\int_{\Omega} v(\boldsymbol{x}) \, dx$$

où dx désigne la mesure de Lebesgue de $\mathbb{R}^d$.

– L'intégrale de surface de v sur le bord $\partial\Omega$ de Ω est notée

$$\int_{\partial\Omega} v(\boldsymbol{x}) \, d\sigma$$

où $d\sigma$ désigne l'élément surface de $\partial\Omega$.

5.3. Notion de maillage non structuré général

L'objet du présent travail est la discrétisation de lois de conservation hyperboliques

$$\partial_t u(\boldsymbol{x}, t) + \boldsymbol{\nabla} \cdot \boldsymbol{f}(u(\boldsymbol{x}, t)) = 0 , \, \boldsymbol{x} \in \Omega \subset \mathbb{R}^d , \, t \geq t_0$$

sur maillage non structuré général où Ω est le domaine de calcul et $d = 1, 2, 3$ est la dimension de l'espace. Le premier pas pour résoudre cette loi de conservation sur ordinateur consiste à découper le domaine physique en N polyèdres généraux

$$\Omega = \bigcup_{\alpha=1}^{N} \mathcal{T}_\alpha .$$

Les polyèdres sont supposés disjoints dans le sens que le d-volume de leur intersection est zéro. Des polyèdres adjacents partagent donc des sommets. En trois dimensions, ils peuvent également avoir des arêtes communes. Dans la suite, ces polyèdres seront appelés *cellules* ou *mailles*. Les cellules sont numérotées par des lettres grecques afin de conserver les lettres latines pour d'autres besoins d'indexation. Le symbole $\mathcal{T}_\alpha$ désigne la cellule numéro α. Son barycentre est défini par

$$\boldsymbol{x}_\alpha \triangleq \frac{1}{|\mathcal{T}_\alpha|} \int_{\mathcal{T}_\alpha} \boldsymbol{x} \, dx$$

et son d-volume est noté $|\mathcal{T}_\alpha|$.

Les méthodes numériques développées dans les chapitres suivants sont conçues pour fournir de meilleures approximations lorsque le diamètre des cellules diminue. Pour caractériser la résolution du maillage, il est utile d'introduire le *diamètre du maillage*, noté h, comme le diamètre maximal des cellules

$$h \triangleq \sup_{\alpha} \sup_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{T}_\alpha} \|\boldsymbol{x} - \boldsymbol{y}\|_2 . \tag{5.3.1}$$

La face entre les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ est appelée $\mathcal{A}_{\alpha\beta}$. Cette face est orientée de la cellule $\mathcal{T}_\alpha$ vers la cellule $\mathcal{T}_\beta$ et la face orientée en sens inverse est appelée $\mathcal{A}_{\beta\alpha}$. Le barycentre de la face $\mathcal{A}_{\alpha\beta}$ est appelé $\boldsymbol{x}_{\alpha\beta}$. Il coïncide avec le barycentre de la face $\mathcal{A}_{\beta\alpha}$, $\boldsymbol{x}_{\alpha\beta} = \boldsymbol{x}_{\beta\alpha}$. Chaque face $\mathcal{A}_{\alpha\beta}$ possède un vecteur surface $\boldsymbol{a}_{\alpha\beta}$ qui a la même orientation que $\mathcal{A}_{\alpha\beta}$.

Il est nécessaire de donner une définition plus précise des faces. La cellule $\mathcal{T}_\alpha$ est délimitée par les faces $\mathcal{A}_{\alpha\beta}$ qui la séparent de ses voisines immédiates. En dimension deux, deux cellules adjacentes partagent exactement deux sommets du maillage et la face entre les deux cellules est définie comme le segment joignant ces deux sommets. En dimension trois, il peut arriver que les sommets partagés par deux cellules ne soient pas coplanaires. Pour tenir compte de ce cas, il est nécessaire d'introduire une définition plus générale de face entre deux cellules. La définition s'appuie sur l'ensemble des sommets partagés par les deux cellules adjacentes $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$. Cet ensemble est noté $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l\}$ et supposé ordonné tel que les segments $\{\overline{\boldsymbol{v}_1\boldsymbol{v}_2}, \overline{\boldsymbol{v}_2\boldsymbol{v}_3}, \ldots, \overline{\boldsymbol{v}_l\boldsymbol{v}_1}\}$ forment un contour fermé. On note $\boldsymbol{v}_{l+1} \triangleq \boldsymbol{v}_1$ pour faciliter l'écriture. Le choix d'un point arbitraire $\boldsymbol{p}$ permet de définir une face $\mathcal{A}'_{\alpha\beta}$ comme la réunion des l triangles $\mathcal{A}^{(i)}_{\alpha\beta} = \overline{\boldsymbol{p}\boldsymbol{v}_i\boldsymbol{v}_{i+1}}$. La relation

$$\left|\mathcal{A}'_{\alpha\beta}\right| (\boldsymbol{x}_{\alpha\beta} - \boldsymbol{p}) = \int_{\mathcal{A}'_{\alpha\beta}} (\boldsymbol{x} - \boldsymbol{p})\, d\sigma = \sum_{i=1}^{l} \int_{\mathcal{A}^{(i)}_{\alpha\beta}} (\boldsymbol{x} - \boldsymbol{p})\, d\sigma = \sum_{i=1}^{l} \left|\mathcal{A}^{(i)}_{\alpha\beta}\right| \frac{1}{3} \left[(\boldsymbol{v}_i - \boldsymbol{p}) + (\boldsymbol{v}_{i+1} - \boldsymbol{p})\right]$$

relie le point $\boldsymbol{p}$, les sommets et le barycentre $\boldsymbol{x}_{\alpha\beta}$ de $\mathcal{A}'_{\alpha\beta}$. Le choix $\boldsymbol{x}_{\alpha\beta} = \boldsymbol{p}$ donne une équation implicite pour $\boldsymbol{x}_{\alpha\beta}$ qui est

$$\sum_{i=1}^{l} \left|\mathcal{A}^{(i)}_{\alpha\beta}\right| \frac{1}{3} \left[(\boldsymbol{v}_i - \boldsymbol{x}_{\alpha\beta}) + (\boldsymbol{v}_{i+1} - \boldsymbol{x}_{\alpha\beta})\right] = 0\,.$$

Elle peut être résolue itérativement pour $\boldsymbol{x}_{\alpha\beta}$. La face résultante $\mathcal{A}_{\alpha\beta}$ est l'union des l triangles $\mathcal{A}^{(i)}_{\alpha\beta} = \overline{\boldsymbol{x}_{\alpha\beta}\boldsymbol{v}_i\boldsymbol{v}_{i+1}}$. Pour des faces planes, comme par exemple les faces triangulaires de tétraèdres ou les faces de parallélépipèdes, cette définition des faces coïncide avec la définition habituelle.

Une définition alternative d'une face existe pour le cas spécifique de quatre sommets qui ne sont pas situés dans le même plan [**20**]. Dans cette situation, si $\{\overline{\boldsymbol{v}_1\boldsymbol{v}_2}, \overline{\boldsymbol{v}_2\boldsymbol{v}_3}, \overline{\boldsymbol{v}_3\boldsymbol{v}_4}, \overline{\boldsymbol{v}_4\boldsymbol{v}_1}\}$ sont les quatre segments formant le contour de la face, il est possible de définir une face comme la réunion des deux triangles $\overline{\boldsymbol{v}_1\boldsymbol{v}_2\boldsymbol{v}_4}$ et $\overline{\boldsymbol{v}_3\boldsymbol{v}_2\boldsymbol{v}_4}$ ou comme réunion des deux triangles $\overline{\boldsymbol{v}_2\boldsymbol{v}_1\boldsymbol{v}_3}$ et $\overline{\boldsymbol{v}_4\boldsymbol{v}_1\boldsymbol{v}_3}$.

On souhaite ensuite définir une notion univoque de vecteur surface associé à chaque face. Pour rappel, en dimension deux, la face $\mathcal{A}_{\alpha\beta}$ est définie comme un segment de droite. Le vecteur surface $\boldsymbol{a}_{\alpha\beta}$ de la face est dans ce cas un vecteur perpendiculaire à ce segment. Il est orienté de la cellule $\mathcal{T}_\alpha$ vers la cellule $\mathcal{T}_\beta$ et sa longueur est égale à la longueur de la face. En dimension trois, il est possible de définir un vecteur surface pour chaque face qui ne dépend que du contour formé par les sommets $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l\}$. Soient $\mathcal{A}'_{\alpha\beta}$ et $\mathcal{A}''_{\alpha\beta}$ deux choix différents pour la face entre les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$. Selon la définition donnée dans le paragraphe ci-dessus, les deux faces sont alors délimitées par le même contour qui est noté $\partial\mathcal{A}_{\alpha\beta}$. Dans ce cas, l'union des deux faces est une surface fermée qui délimite un d-volume fermé appelé $\mathcal{B}_{\alpha\beta}$. Les deux faces ont la même orientation et leurs normales unitaires pointent de la cellule $\mathcal{T}_\alpha$ vers la cellule $\mathcal{T}_\beta$. On peut supposer que la normale de $\mathcal{A}'_{\alpha\beta}$ pointe vers l'extérieur de $\mathcal{B}_{\alpha\beta}$ et celle de $\mathcal{A}''_{\alpha\beta}$ vers l'intérieur de $\mathcal{B}_{\alpha\beta}$. Sous la condition d'une régularité suffisante des deux surfaces, l'application du théorème de Green à une fonction constante implique

$$\int_{\mathcal{A}'_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x})\, d\sigma - \int_{\mathcal{A}''_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x})\, d\sigma = \int_{\mathcal{B}_{\alpha\beta}} \nabla(1)\, dx = 0\,. \tag{5.3.2}$$

Dans (5.3.2), $\boldsymbol{\nu}(\boldsymbol{x})$ est le vecteur normal unitaire dans chaque point $\boldsymbol{x}$ de la face et $d\sigma$ et dx désignent respectivement l'élément de surface et l'élément de d-volume. Le signe négatif dans l'intégrale sur $\mathcal{A}''_{\alpha\beta}$ dans (5.3.2) vient du fait que l'orientation de la normale doit être inversé

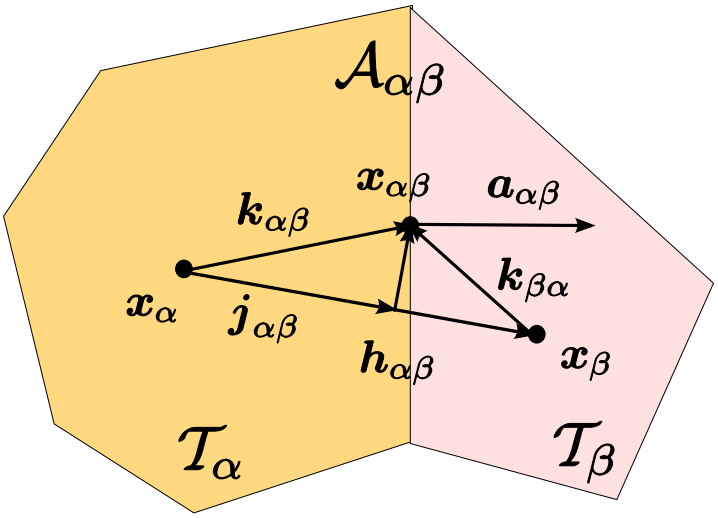


FIG. 5.3.1: Notations en maillage non structuré

pour cette face. Cette relation prouve que la définition du vecteur surface par

$$\boldsymbol{a}_{\alpha\beta} \triangleq \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}\,(\boldsymbol{x})\; d\sigma \tag{5.3.3}$$

dépend uniquement du contour $\partial\mathcal{A}_{\alpha\beta}$ formé par les sommets de la face.

Les vecteurs surface de $\mathcal{A}_{\alpha\beta}$ et de $\mathcal{A}_{\beta\alpha}$ sont de directions opposées et obéissent donc à la relation

$$\boldsymbol{a}_{\alpha\beta} = -\boldsymbol{a}_{\beta\alpha}\,. \tag{5.3.4}$$

La définition des faces et de leur vecteur surface implique que

$$\|\boldsymbol{a}_{\alpha\beta}\| \leq \int_{\mathcal{A}_{\alpha\beta}} d\sigma = |\mathcal{A}_{\alpha\beta}|$$

avec égalité si la face est plane. Le symbole $\boldsymbol{\nu}_{\alpha\beta}$ désigne le vecteur normal unitaire de la face $\mathcal{A}_{\alpha\beta}$

$$\boldsymbol{\nu}_{\alpha\beta} \triangleq \|\boldsymbol{a}_{\alpha\beta}\|^{-1}\,\boldsymbol{a}_{\alpha\beta}\,. \tag{5.3.5}$$

Enfin, pour la définition des méthodes numériques, il est utile d'introduire les vecteurs géométriques $\boldsymbol{h}_{\alpha\beta}$ et $\boldsymbol{k}_{\alpha\beta}$ définis par

$$\boldsymbol{h}_{\alpha\beta} \triangleq \boldsymbol{x}_{\beta} - \boldsymbol{x}_{\alpha}\,;\ \text{pour toutes les cellules}\,\mathcal{T}_{\alpha},\mathcal{T}_{\beta} \tag{5.3.6}$$

$$\boldsymbol{k}_{\alpha\beta} \triangleq \boldsymbol{x}_{\alpha\beta} - \boldsymbol{x}_{\alpha}\,;\ \text{pour toutes les cellules adjacentes}\ \mathcal{T}_{\alpha},\mathcal{T}_{\beta}\,. \tag{5.3.7}$$

Le vecteur $\boldsymbol{j}_{\alpha\beta}$ est la projection orthogonale de $\boldsymbol{k}_{\alpha\beta}$ sur $\boldsymbol{h}_{\alpha\beta}$

$$\boldsymbol{j}_{\alpha\beta} \triangleq \frac{\boldsymbol{h}_{\alpha\beta}\cdot\boldsymbol{k}_{\alpha\beta}}{\|\boldsymbol{h}_{\alpha\beta}\|^{2}}\boldsymbol{h}_{\alpha\beta}\,. \tag{5.3.8}$$

Le vecteur $\boldsymbol{b}_{\alpha\beta}$ est défini par $\boldsymbol{b}_{\alpha\beta} \triangleq \boldsymbol{k}_{\alpha\beta} - \boldsymbol{j}_{\alpha\beta}$. La figure 5.3.1 montre ces notations à l'aide de deux cellules en maillage non structuré bidimensionnel. Elle permet d'illustrer les relations

$$\boldsymbol{h}_{\alpha\beta} = \boldsymbol{k}_{\alpha\beta} - \boldsymbol{k}_{\beta\alpha} \tag{5.3.9}$$

$$\boldsymbol{h}_{\alpha\beta} = \boldsymbol{j}_{\alpha\beta} - \boldsymbol{j}_{\beta\alpha} \tag{5.3.10}$$

$$\boldsymbol{b}_{\alpha\beta} = \boldsymbol{b}_{\beta\alpha}\,. \tag{5.3.11}$$

5.4. Définition des voisinages

Les méthodes de type volumes finis utilisent dans chaque cellule des informations provenant de cellules voisines plus ou moins proches. Pour cette raison, il s'avère nécessaire de définir différents types de voisinages pour chaque cellule.

Le voisinage le plus simple de la cellule $\mathcal{T}_\alpha$ est l'ensemble des mailles qui partagent une face avec $\mathcal{T}_\alpha$. Ce voisinage s'appelle le *premier voisinage* de $\mathcal{T}_\alpha$ et le nombre de premiers voisins de $\mathcal{T}_\alpha$ est appelé l_α. Si $\mathcal{T}_\alpha$ est un simplexe de $\mathbb{R}^d$, il vient $l_\alpha = d+1$. On désigne par

$$\mathbb{V}_\alpha \triangleq \{\beta_1, \ldots, \beta_{l_\alpha}\} \tag{5.4.1}$$

l'ensemble des indices des cellules du premier voisinage de $\mathcal{T}_\alpha$. Par définition $\alpha \notin \mathbb{V}_\alpha$. Le *premier voisinage augmenté* est défini comme

$$\widehat{\mathbb{V}}_\alpha \triangleq \mathbb{V}_\alpha \cup \{\alpha\} = \{\alpha, \beta_1, \ldots, \beta_{l_\alpha}\} \,. \tag{5.4.2}$$

Le deuxième voisinage de la cellule $\mathcal{T}_\alpha$ se définit comme l'ensemble des premiers voisins des premiers voisins de $\mathcal{T}_\alpha$ sans la cellule $\mathcal{T}_\alpha$ elle-même

$$\mathbb{V}_\alpha^{(2)} \triangleq \bigcup_{\beta \in \mathbb{V}_\alpha} \mathbb{V}_\beta \setminus \{\alpha\} \,. \tag{5.4.3}$$

Le *deuxième voisinage augmenté* de $\mathcal{T}_\alpha$ est noté

$$\widehat{\mathbb{V}}_\alpha^{(2)} \triangleq \mathbb{V}_\alpha^{(2)} \cup \{\alpha\} \,. \tag{5.4.4}$$

Il est possible de définir de façon récursive le k-ième voisinage de la cellule par

$$\mathbb{V}_\alpha^{(k)} \triangleq \bigcup_{\beta \in \mathbb{V}_\alpha^{(k-1)}} \mathbb{V}_\beta \setminus \{\alpha\} \tag{5.4.5}$$

ainsi que le *k-ième voisinage augmenté* par

$$\widehat{\mathbb{V}}_\alpha^{(k)} \triangleq \mathbb{V}_\alpha^{(k)} \cup \{\alpha\} \,. \tag{5.4.6}$$

Un autre voisinage utile est l'ensemble des cellules qui partagent un sommet avec la cellule $\mathcal{T}_\alpha$. Ce voisinage est désigné par le symbole $\mathbb{V}_\alpha^{(s)}$. Dans des maillages de tétraèdres, le voisinage $\mathbb{V}_\alpha^{(s)}$ est en général plus grand que le deuxième voisinage.

5.5. Notation géométrique allégée

L'introduction d'une notation adaptée simplifie l'écriture et la compréhension des formules. Si deux cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ n'ont pas d'interface commune, les vecteurs $\boldsymbol{a}_{\alpha\beta}$ and $\boldsymbol{k}_{\alpha\beta}$ sont par convention nuls. De plus, on définit leur face commune $\mathcal{A}_{\alpha\beta}$ comme l'ensemble vide. De cette façon, toute intégrale de surface sur $\mathcal{A}_{\alpha\beta}$ est automatiquement zéro. Les vecteurs $\boldsymbol{a}_{\alpha\alpha}$ et $\boldsymbol{k}_{\alpha\alpha}$ sont définis comme les vecteurs nuls. Le vecteur $\boldsymbol{h}_{\alpha\alpha} = \boldsymbol{x}_\alpha - \boldsymbol{x}_\alpha = 0$ en raison de la définition (5.3.6). Ces conventions de notation permettent de supprimer l'indexation explicite dans toutes les sommes sur le premier voisinage et d'écrire $\sum_\alpha$ au lieu de $\sum_{\beta \in \mathbb{V}_\alpha}$. Un exemple de l'utilité de cette convention est l'application du théorème de Green à une fonction constante. La définition des vecteurs surface (5.3.3) entraîne l'identité

$$0 = \int_{\mathcal{T}_\alpha} \boldsymbol{\nabla}\,(1)\; dx = \sum_{\beta \in \mathbb{V}_\alpha} \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}\,(\boldsymbol{x})\; d\sigma = \sum_{\beta \in \mathbb{V}_\alpha} \boldsymbol{a}_{\alpha\beta} \,. \tag{5.5.1}$$

Grâce à la convention ci-dessus, la somme dans (5.5.1) s'écrit simplement comme

$$\sum_\beta \boldsymbol{a}_{\alpha\beta} = 0 \,. \tag{5.5.2}$$

Dans la suite du document, cette convention sera appliquée à d'autres vecteurs et tenseurs associés aux faces $\mathcal{A}_{\alpha\beta}$, c'est-à-dire que le vecteur ou tenseur est par convention nul si les deux cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ n'ont pas d'interface commune. *Ceci sera toutefois mentionné de façon explicite dans la définition du vecteur ou tenseur en question.* On a par exemple $\boldsymbol{h}_{\alpha\beta} = \boldsymbol{x}_\beta - \boldsymbol{x}_\alpha \neq 0$ même si les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ n'ont pas d'interface commune.

5.6. Propriétés du produit tensoriel symétrique

L'objectif de cette section est d'établir quelques propriétés du produit symétrique $\odot$ défini par (5.2.19). Ces propriétés seront très utiles pour le chapitre 8. On commence par la

PROPOSITION 5.6.1 (Propriétés du produit $\odot$). *Soient $\boldsymbol{a}^{(k)}$, $\boldsymbol{b}^{(m)}$, $\boldsymbol{c}^{(l)}$, des tenseurs dans $\mathbb{R}^d$. Le produit tensoriel symétrique défini par (5.2.19) a les propriétés suivantes :*

(i) *Commutativité :* $\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m)} \odot \boldsymbol{a}^{(k)}$.

(ii) *Associativité :* $\left(\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)}\right) \odot \boldsymbol{c}^{(l)} = \boldsymbol{a}^{(k)} \odot \left(\boldsymbol{b}^{(m)} \odot \boldsymbol{c}^{(l)}\right)$.

(iii) *Distributivité :* $\boldsymbol{a}^{(k)} \odot \left(\boldsymbol{b}^{(m)} + \boldsymbol{c}^{(m)}\right) = \boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)} + \boldsymbol{a}^{(k)} \odot \boldsymbol{c}^{(m)}$.

DÉMONSTRATION. Les trois propriétés découlent des définitions (5.2.20) et (5.2.21) si l'un des tenseurs est d'ordre 0. On suppose dans la suite $k, l, m \geq 1$ et on prouve d'abord la commutativité. La définition (5.2.19) implique que les composantes de $\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)}$ sont données par

$$\left(\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)}\right)_{i_1 \cdots i_{k+m}} = \frac{1}{(k+m)!} \sum_{\pi \in \mathfrak{S}_{k+m}} a_{i_{\pi(1)} \cdots i_{\pi(k)}} b_{i_{\pi(k+1)} \cdots i_{\pi(k+m)}} \,. \tag{5.6.1}$$

Comme la somme dans (5.6.1) se fait sur toutes les permutations $\pi \in \mathfrak{S}_{k+m}$ et puisque $\mathfrak{S}_{k+m}$ est un groupe, il est possible de sommer sur toutes les permutations de la forme $\pi \circ \widetilde{\pi}$ où $\widetilde{\pi}$ est un élément fixe du groupe $\mathfrak{S}_{k+m}$, π parcourt tous les éléments du groupe $\mathfrak{S}_{k+m}$ et $\pi \circ \widetilde{\pi}$ désigne la composition des permutations π et $\widetilde{\pi}$ définie par

$$\pi \circ \widetilde{\pi}\,(j) = \pi\left(\widetilde{\pi}\,(j)\right) \,,\; 1 \leq j \leq k+m \,.$$

Avec la définition de $\widetilde{\pi}$ par

$$\widetilde{\pi}\,(j) \triangleq \begin{cases} j+m & \text{si } 1 \leq j \leq k \\ j-k & \text{si } k+1 \leq j \leq k+m \end{cases} ,$$

(5.6.1) devient

$$\begin{aligned} &\frac{1}{(k+m)!} \sum_{\pi \in \mathfrak{S}_{k+m}} a_{i_{\pi(\widetilde{\pi}(1))} \cdots i_{\pi(\widetilde{\pi}(k))}} b_{i_{\pi(\widetilde{\pi}(k+1))} \cdots i_{\pi(\widetilde{\pi}(k+m))}} = \\ &\qquad = \frac{1}{(k+m)!} \sum_{\pi \in \mathfrak{S}_{k+m}} b_{i_{\pi(1)} \cdots i_{\pi(m)}} a_{i_{\pi(m+1)} \cdots i_{\pi(m+k)}} = \left(\boldsymbol{b}^{(m)} \odot \boldsymbol{a}^{(k)}\right)_{i_1 \cdots i_{k+m}} , \end{aligned} \tag{5.6.2}$$

ce qui prouve la commutativité.

L'associativité se démontre de la façon suivante. On décompose le produit

$$\begin{aligned} &\left(\left(\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)}\right) \odot \boldsymbol{c}^{(l)}\right)_{i_1 \cdots i_{k+m+l}} = \\ &\qquad = \frac{1}{(k+m+l)!} \sum_{\pi \in \mathfrak{S}_{k+m+l}} \left(\boldsymbol{a}^{(k)} \odot \boldsymbol{b}^{(m)}\right)_{i_1 \cdots i_{k+m}} c_{i_{\pi(k+m+1)} \cdots i_{\pi(k+m+l)}} = \\ &\qquad = \sum_{\pi \in \mathfrak{S}_{k+m+l}} \sum_{\varsigma \in \mathfrak{S}_{k+m}} \frac{a_{i_{\pi(\varsigma(1))} \cdots i_{\pi(\varsigma(k))}} b_{i_{\pi(\varsigma(k+1))} \cdots i_{\pi(\varsigma(k+m))}} c_{i_{\pi(k+m+1)} \cdots i_{\pi(k+m+l)}}}{(k+m+l)!\,(k+m)!} \,. \end{aligned} \tag{5.6.3}$$

Il est possible de changer l'ordre des sommes sur π et ς dans la dernière ligne de (5.6.3). Pour une permutation $\varsigma \in \mathfrak{S}_{k+m}$ fixe, la somme sur $\pi \in \mathfrak{S}_{k+m+l}$ peut être remplacée par la somme sur $\pi \circ \widetilde{\pi}$ où π parcourt $\mathfrak{S}_{k+m+l}$ et $\widetilde{\pi}$ est définie par la permutation inverse ς^{-1} de ς comme

$$\widetilde{\pi}\,(j) \triangleq \begin{cases} \varsigma^{-1}(j) & \text{si } 1 \leq j \leq k+m \\ j & \text{si } k+m+1 \leq j \leq k+m+l \end{cases} .$$

Cela permet d'écrire (5.6.3) comme

$$\sum_{\varsigma\in\mathfrak{S}_{k+m}}\sum_{\pi\in\mathfrak{S}_{k+m+l}}\frac{a_{i_{\pi(\varsigma(1))}\cdots i_{\pi(\varsigma(k))}}b_{i_{\pi(\varsigma(k+1))}\cdots i_{\pi(\varsigma(k+m))}}c_{i_{\pi(k+m+1)}\cdots i_{\pi(k+m+l)}}}{(k+m+l)!\,(k+m)!}=$$

$$=\sum_{\varsigma\in\mathfrak{S}_{k+m}}\sum_{\pi\in\mathfrak{S}_{k+m+l}}\frac{a_{i_{\pi(\varsigma^{-1}(\varsigma(1)))}\cdots i_{\pi(\varsigma^{-1}(\varsigma(k)))}}b_{i_{\pi(\varsigma^{-1}(\varsigma(k+1)))}\cdots i_{\pi(\varsigma^{-1}(\varsigma(k+m)))}}c_{i_{\pi(k+m+1)}\cdots i_{\pi(k+m+l)}}}{(k+m+l)!\,(k+m)!}=$$

$$=\sum_{\varsigma\in\mathfrak{S}_{k+m}}\sum_{\pi\in\mathfrak{S}_{k+m+l}}\frac{a_{i_{\pi(1)}\cdots i_{\pi(k)}}b_{i_{\pi(k+1)}\cdots i_{\pi(k+m)}}c_{i_{\pi(k+m+1)}\cdots i_{\pi(k+m+l)}}}{(k+m+l)!\,(k+m)!}\tag{5.6.4}$$

où la relation $\varsigma^{-1}(\varsigma(j))=j$ a permis de simplifier la dernière ligne de (5.6.4). La permutation ς n'apparaît d'ailleurs plus dans la dernière ligne de (5.6.4). Cela permet d'effectuer la somme sur $\varsigma\in\mathfrak{S}_{k+m}$ ce qui occasionne un facteur $(k+m)!$ et donne le résultat final

$$\left(\left(\boldsymbol{a}^{(k)}\odot\boldsymbol{b}^{(m)}\right)\odot\boldsymbol{c}^{(l)}\right)_{i_1\cdots i_{k+m+l}}=$$

$$=\frac{1}{(k+m+l)!}\sum_{\pi\in\mathfrak{S}_{k+m+l}}a_{i_{\pi(1)}\cdots i_{\pi(k)}}b_{i_{\pi(k+1)}\cdots i_{\pi(k+m)}}c_{i_{\pi(k+m+1)}\cdots i_{\pi(k+m+l)}}\,.\tag{5.6.5}$$

L'application du même raisonnement au produit $\boldsymbol{a}^{(k)}\odot\left(\boldsymbol{b}^{(m)}\odot\boldsymbol{c}^{(l)}\right)$ donne

$$\left(\boldsymbol{a}^{(k)}\odot\left(\boldsymbol{b}^{(m)}\odot\boldsymbol{c}^{(l)}\right)\right)_{i_1\cdots i_{k+m+l}}=$$

$$=\frac{1}{(k+m+l)!}\sum_{\pi\in\mathfrak{S}_{k+m+l}}a_{i_{\pi(1)}\cdots i_{\pi(k)}}b_{i_{\pi(k+1)}\cdots i_{\pi(k+m)}}c_{i_{\pi(k+m+1)}\cdots i_{\pi(k+m+l)}}\,,\tag{5.6.6}$$

ce qui prouve l'associativité.

La distributivité découle du fait que

$$\left(\boldsymbol{a}^{(k)}\odot\left(\boldsymbol{b}^{(m)}+\boldsymbol{c}^{(m)}\right)\right)_{i_1\cdots i_{k+m}}=$$

$$=\frac{1}{(k+m)!}\sum_{\pi\in\mathfrak{S}_{k+m}}a_{i_{\pi(1)}\cdots i_{\pi(k)}}\left(b_{i_{\pi(k+1)}\cdots i_{\pi(k+m)}}+c_{i_{\pi(k+1)}\cdots i_{\pi(k+m)}}\right)=$$

$$=\frac{1}{(k+m)!}\sum_{\pi\in\mathfrak{S}_{k+m}}a_{i_{\pi(1)}\cdots i_{\pi(k)}}b_{i_{\pi(k+1)}\cdots i_{\pi(k+m)}}+\frac{1}{(k+m)!}\sum_{\pi\in\mathfrak{S}_{k+m}}a_{i_{\pi(1)}\cdots i_{\pi(k)}}c_{i_{\pi(k+1)}\cdots i_{\pi(k+m)}}\,,\tag{5.6.7}$$

ce qui montre

$$\boldsymbol{a}^{(k)}\odot\left(\boldsymbol{b}^{(m)}+\boldsymbol{c}^{(m)}\right)=\boldsymbol{a}^{(k)}\odot\boldsymbol{b}^{(m)}+\boldsymbol{a}^{(k)}\odot\boldsymbol{c}^{(m)}\,.$$

□

Pour des vecteurs, on peut démontrer le

LEMME 5.6.2. *Pour tout vecteur $\boldsymbol{a}$ dans $\mathbb{R}^d$, on peut écrire la k-ième puissance de $\boldsymbol{a}$ comme*

$$\boldsymbol{a}^k\triangleq\underbrace{\boldsymbol{a}\otimes\cdots\otimes\boldsymbol{a}}_{k}=\underbrace{\boldsymbol{a}\odot\cdots\odot\boldsymbol{a}}_{k}\,.\tag{5.6.8}$$

DÉMONSTRATION. L'associativité du produit symétrique permet d'omettre les parenthèses dans (5.6.8) et d'écrire

$$\boldsymbol{a}\odot\cdots\odot\boldsymbol{a}=\frac{1}{k!}\sum_{\pi\in\mathfrak{S}_k}a_{i_{\pi(1)}}\cdots a_{i_{\pi(k)}}$$

L'identité (5.6.8) vient alors du fait que

$$a_{i_{\pi(1)}}\cdots a_{i_{\pi(k)}}=a_{i_1}\cdots a_{i_k}$$

pour toute permutation $\pi \in \mathfrak{S}_k$. □

La proposition 5.6.1 permet de démontrer la

PROPOSITION 5.6.3 (Formule du binôme). *Soient $\boldsymbol{a}$ et $\boldsymbol{b}$ des vecteurs dans $\mathbb{R}^d$. Le produit symétrique satisfait la formule du binôme*

$$(\boldsymbol{a}+\boldsymbol{b})^k = \sum_{l=0}^{k} \binom{k}{l} \boldsymbol{a}^l \odot \boldsymbol{b}^{k-l} . \tag{5.6.9}$$

DÉMONSTRATION. Le lemme 5.6.2 permet d'écrire

$$(\boldsymbol{a}+\boldsymbol{b})^k = (\boldsymbol{a}+\boldsymbol{b}) \odot \cdots \odot (\boldsymbol{a}+\boldsymbol{b}) .$$

La proposition 5.6.1 montre que le produit $\odot$ est associatif, commutatif et distributif. L'identité (5.6.9) se démontre donc comme pour des nombres réels par induction sur k. □

5.7. Tenseurs géométriques en maillage non structuré

Cette section définit des tenseurs importants construits à partir des cellules et des faces définies dans la section 5.3.

Le *moment d'ordre k* de la cellule $\mathcal{T}_\alpha$ est défini comme le tenseur symétrique d'ordre k

$$\boldsymbol{x}_\alpha^{(k)} \triangleq \frac{1}{|\mathcal{T}_\alpha|} \int_{\mathcal{T}_\alpha} (\boldsymbol{x}-\boldsymbol{x}_\alpha)^k \, dx = \frac{1}{|\mathcal{T}_\alpha|} \int_{\mathcal{T}_\alpha} \underbrace{(\boldsymbol{x}-\boldsymbol{x}_\alpha) \otimes \cdots \otimes (\boldsymbol{x}-\boldsymbol{x}_\alpha)}_{k\times} \, dx . \tag{5.7.1}$$

Le moment $\boldsymbol{x}_\alpha^{(k)}$ est un tenseur de composantes

$$x_{\alpha,i_1\cdots i_k} = \frac{1}{|\mathcal{T}_\alpha|} \int_{\mathcal{T}_\alpha} (x_{i_1} - x_{\alpha,i_1}) \cdots (x_{i_k} - x_{\alpha,i_k}) \, dx . \tag{5.7.2}$$

La définition (5.7.1) implique les identités

$$\boldsymbol{x}_\alpha^{(0)} = 1 \tag{5.7.3}$$

$$\boldsymbol{x}_\alpha^{(1)} = 0 . \tag{5.7.4}$$

Comme $\boldsymbol{x}-\boldsymbol{x}_\alpha = \mathrm{O}(h)$ pour $\boldsymbol{x} \in \mathcal{T}_\alpha$ il vient

$$\boldsymbol{x}_\alpha^{(k)} = \mathrm{O}\left(h^k\right) .$$

Le moment $\boldsymbol{l}_{\alpha\beta}^{(j)}$ d'ordre j de la face $\mathcal{A}_{\alpha\beta}$ est défini comme le tenseur symétrique d'ordre j

$$\boldsymbol{l}_{\alpha\beta}^{(j)} \triangleq \frac{1}{|\mathcal{A}_{\alpha\beta}|} \int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{x}-\boldsymbol{x}_{\alpha\beta})^j \, d\sigma = \frac{1}{|\mathcal{A}_{\alpha\beta}|} \int_{\mathcal{A}_{\alpha\beta}} \underbrace{(\boldsymbol{x}-\boldsymbol{x}_{\alpha\beta}) \otimes \cdots \otimes (\boldsymbol{x}-\boldsymbol{x}_{\alpha\beta})}_{j\times} \, d\sigma . \tag{5.7.5}$$

Le moment $\boldsymbol{l}_{\alpha\beta}^{(j)}$ est un tenseur de composantes

$$l_{i_1\cdots i_j}^{\alpha\beta} = \frac{1}{|\mathcal{A}_{\alpha\beta}|} \int_{\mathcal{A}_{\alpha\beta}} (x_{i_1} - x_{\alpha\beta,i_1}) \cdots \left(x_{i_j} - x_{\alpha\beta,i_j}\right) \, d\sigma . \tag{5.7.6}$$

Puisque le barycentre $\boldsymbol{x}_{\alpha\beta}$ de la face $\mathcal{A}_{\alpha\beta}$ est indépendant de l'orientation de la face $\mathcal{A}_{\alpha\beta}$, on a $\boldsymbol{l}_{\alpha\beta}^{(j)} = \boldsymbol{l}_{\beta\alpha}^{(j)}$. Par convention $\boldsymbol{l}_{\alpha\alpha}^{(j)} \triangleq 0$ et on définit $\boldsymbol{l}_{\alpha\beta}^{(j)} \triangleq 0$ si les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ n'ont pas d'interface commune. Le comportement de $\boldsymbol{l}_{\alpha\beta}^{(j)}$ lorsque h tend vers zéro est

$$\boldsymbol{l}_{\alpha\beta}^{(j)} = \mathrm{O}\left(h^j\right)$$

car $\boldsymbol{x}-\boldsymbol{x}_{\alpha\beta} = \mathrm{O}(h)$ pour $\boldsymbol{x} \in \mathcal{A}_{\alpha\beta}$.

Il est intéressant de généraliser la définition (5.3.7) des vecteurs $\boldsymbol{k}_{\alpha\beta}$ par celle des tenseurs $\boldsymbol{k}_{\alpha\beta}^{(j)}$ suivants

$$\boldsymbol{k}_{\alpha\beta}^{(j)} \triangleq \frac{1}{|\mathcal{A}_{\alpha\beta}|} \int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{x}-\boldsymbol{x}_\alpha)^j \, d\sigma = \frac{1}{|\mathcal{A}_{\alpha\beta}|} \int_{\mathcal{A}_{\alpha\beta}} \underbrace{(\boldsymbol{x}-\boldsymbol{x}_\alpha) \otimes \cdots \otimes (\boldsymbol{x}-\boldsymbol{x}_\alpha)}_{j\times} \, d\sigma . \tag{5.7.7}$$

Par convention $\boldsymbol{k}_{\alpha\alpha}^{(j)} \triangleq 0$ et on définit $\boldsymbol{k}_{\alpha\beta}^{(j)} \triangleq 0$ si les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ n'ont pas d'interface commune.

En général, on a $\boldsymbol{k}_{\alpha\beta}^{(j)} \neq \boldsymbol{k}_{\beta\alpha}^{(j)}$ car

$$\boldsymbol{k}_{\beta\alpha}^{(j)} \triangleq \frac{1}{|\mathcal{A}_{\alpha\beta}|}\int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{x}-\boldsymbol{x}_\beta)^j \, d\sigma\,.$$

Les tenseurs $\boldsymbol{k}_{\alpha\beta}^{(j)}$ et $\boldsymbol{k}_{\beta\alpha}^{(j)}$ sont cependant reliés par l'identité

$$\boldsymbol{k}_{\alpha\beta}^{(j)} = \sum_{i=0}^{j}\binom{j}{i}\boldsymbol{k}_{\beta\alpha}^{(i)} \odot \boldsymbol{h}_{\alpha\beta}^{j-i} \tag{5.7.8}$$

qui découle du lemme 5.6.2 et de la proposition 5.6.3 sur la formule du binôme

$$\boldsymbol{k}_{\alpha\beta}^{(j)} = \frac{1}{|\mathcal{A}_{\alpha\beta}|}\int_{\mathcal{A}_{\alpha\beta}} \underbrace{(\boldsymbol{x}-\boldsymbol{x}_\beta+\boldsymbol{h}_{\alpha\beta})\odot\cdots\odot(\boldsymbol{x}-\boldsymbol{x}_\beta+\boldsymbol{h}_{\alpha\beta})}_{j}\, d\sigma =$$

$$= \frac{1}{|\mathcal{A}_{\alpha\beta}|}\int_{\mathcal{A}_{\alpha\beta}} \sum_{i=0}^{j}\binom{j}{i}(\boldsymbol{x}-\boldsymbol{x}_\beta)^i \odot \boldsymbol{h}_{\alpha\beta}^{j-i}\, d\sigma = \sum_{i=0}^{j}\binom{j}{i}\boldsymbol{k}_{\beta\alpha}^{(i)}\odot\boldsymbol{h}_{\alpha\beta}^{j-i}\,,$$

ce qui prouve (5.7.8).

Pour les développements de Taylor de la section 5.8, il est intéressant d'introduire la notion de tenseurs composés suivante.

Définition 5.7.1 (Tenseurs composés du maillage).

(i) Les tenseurs symétriques $\boldsymbol{y}_{\alpha\beta}^{(l|j)}$ d'ordre $j+l$ sont définis par

$$\boldsymbol{y}_{\alpha\beta}^{(l|j)} \triangleq \boldsymbol{h}_{\alpha\beta}^{l}\odot\boldsymbol{x}_\beta^{(j)}\,. \tag{5.7.9}$$

(ii) Les tenseurs symétriques $\boldsymbol{z}_{\alpha\beta}^{(k)}$ sont définis par

$$\boldsymbol{z}_{\alpha\beta}^{(k)} \triangleq \sum_{l=0}^{k}\binom{k}{l}\boldsymbol{y}_{\alpha\beta}^{(k-l|l)} = \sum_{l=0}^{k}\binom{k}{l}\boldsymbol{h}_{\alpha\beta}^{k-l}\odot\boldsymbol{x}_\beta^{(l)}\,. \tag{5.7.10}$$

La définition 5.7.1 est motivée par le

Lemme 5.7.2 (Représentation des tenseurs composés). *Les tenseurs composés $\boldsymbol{z}_{\alpha\beta}^{(k)}$ satisfont l'identité*

$$\boldsymbol{z}_{\alpha\beta}^{(k)} = \frac{1}{|\mathcal{T}_\beta|}\int_{\mathcal{T}_\beta}(\boldsymbol{x}-\boldsymbol{x}_\alpha)^k\, dx \tag{5.7.11}$$

et en particulier l'identité

$$\boldsymbol{z}_{\alpha\alpha}^{(k)} = \boldsymbol{x}_\alpha^{(k)}\,. \tag{5.7.12}$$

Démonstration. L'identité

$$\boldsymbol{x}-\boldsymbol{x}_\alpha = \boldsymbol{x}-\boldsymbol{x}_\beta+\boldsymbol{h}_{\alpha\beta}$$

permet d'écrire

$$\frac{1}{|\mathcal{T}_\beta|}\int_{\mathcal{T}_\beta}(\boldsymbol{x}-\boldsymbol{x}_\alpha)^k\, dx = \frac{1}{|\mathcal{T}_\beta|}\int_{\mathcal{T}_\beta}(\boldsymbol{x}-\boldsymbol{x}_\beta+\boldsymbol{h}_{\alpha\beta})^k\, dx\,.$$

L'application de la formule du binôme de la proposition 5.6.3 prouve alors que

$$\frac{1}{|\mathcal{T}_\beta|}\int_{\mathcal{T}_\beta}(\boldsymbol{x}-\boldsymbol{x}_\beta+\boldsymbol{h}_{\alpha\beta})^k\, dx = \frac{1}{|\mathcal{T}_\beta|}\int_{\mathcal{T}_\beta}\sum_{l=0}^{k}\binom{k}{l}\boldsymbol{h}_{\alpha\beta}^{k-l}\odot(\boldsymbol{x}-\boldsymbol{x}_\beta)^l\, dx =$$

$$= \sum_{l=0}^{k}\binom{k}{l}\boldsymbol{h}_{\alpha\beta}^{k-l}\odot\boldsymbol{x}_\beta^{(l)} = \boldsymbol{z}_{\alpha\beta}^{(k)}\,,$$

ce qui démontre l'identité (5.7.11). L'identité (5.7.12) découle alors de la définition (5.7.1) des moments $\boldsymbol{x}_\alpha^{(k)}$. □

Si $\boldsymbol{a}^{(j+l)}$ est un tenseur symétrique d'ordre $j+l$, la symétrie du tenseur $\boldsymbol{h}_{\alpha\beta}^l \odot \boldsymbol{x}_\beta^{(j)}$ dans (5.7.9) entraîne l'identité

$$\boldsymbol{a}^{(j+l)} \bullet \left[\boldsymbol{h}_{\alpha\beta}^l \otimes \boldsymbol{x}_\beta^{(j)}\right] = \boldsymbol{a}^{(j+l)} \bullet \boldsymbol{y}_{\alpha\beta}^{(l|j)} . \tag{5.7.13}$$

Le comportement des tenseurs $\boldsymbol{y}_{\alpha\beta}^{(l|j)}$ et $\boldsymbol{z}_{\alpha\beta}^{(k)}$ lorsque $h \to 0$ est caractérisé par

$$\begin{aligned} \boldsymbol{y}_{\alpha\beta}^{(l|j)} &= \mathrm{O}\left(h^{l+j}\right) \\ \boldsymbol{z}_{\alpha\beta}^{(k)} &= \mathrm{O}\left(h^{k}\right) . \end{aligned}$$

L'identité (5.7.4) entraîne

$$\boldsymbol{y}_{\alpha\beta}^{(l|1)} = 0$$

pour tout $l \in \mathbb{N}$. Les tenseurs $\boldsymbol{z}_{\alpha\beta}^{(k)}$ pour $0 \leq k \leq 4$ sont, en raison de $\boldsymbol{x}_\beta^{(1)} = 0$, donnés par

$$\begin{aligned} \boldsymbol{z}_{\alpha\beta}^{(0)} &= 1 \\ \boldsymbol{z}_{\alpha\beta}^{(1)} &= \boldsymbol{h}_{\alpha\beta} \\ \boldsymbol{z}_{\alpha\beta}^{(2)} &= \boldsymbol{h}_{\alpha\beta}^2 + \boldsymbol{x}_\beta^{(2)} \\ \boldsymbol{z}_{\alpha\beta}^{(3)} &= \boldsymbol{h}_{\alpha\beta}^3 + 3\boldsymbol{h}_{\alpha\beta} \odot \boldsymbol{x}_\beta^{(2)} + \boldsymbol{x}_\beta^{(3)} \\ \boldsymbol{z}_{\alpha\beta}^{(4)} &= \boldsymbol{h}_{\alpha\beta}^4 + 6\boldsymbol{h}_{\alpha\beta}^2 \odot \boldsymbol{x}_\beta^{(2)} + 4\boldsymbol{h}_{\alpha\beta} \odot \boldsymbol{x}_\beta^{(3)} + \boldsymbol{x}_\beta^{(4)} . \end{aligned}$$

Il est parfois nécessaire de décomposer les tenseurs $\boldsymbol{y}_{\alpha\gamma}^{(k-l|l)}$ et $\boldsymbol{z}_{\alpha\gamma}^{(k)}$ à l'aide de l'identité $\boldsymbol{h}_{\alpha\gamma} = \boldsymbol{h}_{\alpha\beta} + \boldsymbol{h}_{\beta\gamma}$.

PROPOSITION 5.7.3 (Identités composées du maillage). *Soient $\mathcal{T}_\alpha$, $\mathcal{T}_\beta$ et $\mathcal{T}_\gamma$ trois cellules quelconques du maillage. Les tenseurs composés $\boldsymbol{y}_{\alpha\gamma}^{(l|j)}$ et $\boldsymbol{z}_{\alpha\gamma}^{(k)}$ satisfont alors*

$$\boldsymbol{y}_{\alpha\gamma}^{(l|j)} = \sum_{m=0}^{l} \binom{l}{m} \boldsymbol{h}_{\alpha\beta}^m \odot \boldsymbol{y}_{\beta\gamma}^{(l-m|j)} \tag{5.7.14}$$

et

$$\boldsymbol{z}_{\alpha\gamma}^{(k)} = \sum_{m=0}^{k} \binom{k}{m} \boldsymbol{h}_{\alpha\beta}^m \odot \boldsymbol{z}_{\beta\gamma}^{(k-m)} . \tag{5.7.15}$$

DÉMONSTRATION. L'application de la formule du binôme au vecteur $\boldsymbol{h}_{\alpha\gamma} = \boldsymbol{h}_{\alpha\beta} + \boldsymbol{h}_{\beta\gamma}$ et l'associativité du produit $\odot$ donnent l'identité

$$\boldsymbol{y}_{\alpha\gamma}^{(l|j)} = \boldsymbol{h}_{\alpha\gamma}^l \odot \boldsymbol{x}_\gamma^{(j)} = \sum_{m=0}^{l} \binom{l}{m} \boldsymbol{h}_{\alpha\beta}^m \odot \boldsymbol{h}_{\beta\gamma}^{l-m} \odot \boldsymbol{x}_\gamma^{(j)} = \sum_{m=0}^{l} \binom{l}{m} \boldsymbol{h}_{\alpha\beta}^m \odot \boldsymbol{y}_{\beta\gamma}^{(l-m|j)}$$

ce qui prouve (5.7.14). L'application de la même formule aux tenseurs $\boldsymbol{z}_{\alpha\gamma}^{(k)}$ définis par (5.7.10) donne

$$\boldsymbol{z}_{\alpha\gamma}^{(k)} = \sum_{l=0}^{k} \binom{k}{l} \boldsymbol{h}_{\alpha\gamma}^l \odot \boldsymbol{x}_\gamma^{(k-l)} = \sum_{l=0}^{k} \sum_{m=0}^{l} \binom{k}{l} \binom{l}{m} \boldsymbol{h}_{\alpha\beta}^m \odot \boldsymbol{h}_{\beta\gamma}^{l-m} \odot \boldsymbol{x}_\gamma^{(k-l)} . \tag{5.7.16}$$

Le domaine de la double somme est donné par

$$\begin{aligned} &\{\, (l,m) \in \mathbb{Z} \times \mathbb{Z} |\, 0 \leq l\,,\, l \leq k\,,\, 0 \leq m\,,\, m \leq l\} = \\ &\qquad = \{\, (l,m) \in \mathbb{Z} \times \mathbb{Z} |\, m \leq k\,,\, l \leq k\,,\, 0 \leq m\,,\, m \leq l\} = \\ &\qquad\quad = \{\, (i+m,m) \in \mathbb{Z} \times \mathbb{Z} |\, m \leq k\,,\, i \leq k-m\,,\, 0 \leq m\,,\, 0 \leq i\} . \end{aligned} \tag{5.7.17}$$

Dans la dernière ligne de (5.7.17), la variable l a été remplacée par $l = i+m$. L'équation (5.7.16) devient, après le réarrangement de la double somme et avec l'insertion $l = i + m$, l'identité

$$\sum_{m=0}^{k}\sum_{i=0}^{k-m}\frac{k!}{(k-m-i)!}\frac{1}{m!i!}\boldsymbol{h}^m_{\alpha\beta}\odot\left(\boldsymbol{h}^i_{\beta\gamma}\odot\boldsymbol{x}^{(k-m-i)}_{\gamma}\right)=$$
$$=\sum_{m=0}^{k}\frac{k!}{m!\,(k-m)!}\sum_{i=0}^{k-m}\frac{(k-m)!}{(k-m-i)!i!}\boldsymbol{h}^m_{\alpha\beta}\odot\left(\boldsymbol{h}^i_{\beta\gamma}\odot\boldsymbol{x}^{(k-m-i)}_{\gamma}\right)=$$
$$=\sum_{m=0}^{k}\binom{k}{m}\boldsymbol{h}^m_{\alpha\beta}\odot\boldsymbol{z}^{(k-m)}_{\beta\gamma}\,. \quad (5.7.18)$$

□

5.8. Développement de Taylor des moyennes de cellule

Les variables élémentaires pour les méthodes de type volumes finis sont les moyennes de cellule des fonctions inconnues. Il est donc intéressant de disposer de développements de Taylor pour les moyennes de cellule de fonctions suffisamment régulières.

PROPOSITION 5.8.1 (Développements de Taylor en maillage non structuré). *Soit v une fonction de classe C^{k+1} sur un ouvert $\mathcal{O}$ telle que la dérivée $D^{(k+1)}v$ soit bornée sur $\mathcal{O}$. Soient $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ deux cellules contenues dans $\mathcal{O}$. On suppose que les segments $[\boldsymbol{x}_\alpha, \boldsymbol{y}]$ sont contenus dans $\mathcal{O}$ pour tout $\boldsymbol{y}\in\mathcal{T}_\beta$ et que la distance $\sup_{\boldsymbol{x}\in\mathcal{T}_\alpha}\sup_{\boldsymbol{y}\in\mathcal{T}_\beta}\|\boldsymbol{x}-\boldsymbol{y}\|_2$ des deux cellules est bornée par h. Les moyennes $\overline{v}_\alpha$ et $\overline{v}_\beta$ de v sur les cellules respectives $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ s'expriment en fonction des moments $\boldsymbol{x}^{(i)}_\alpha$, cf. (5.7.1), et des tenseurs composés $\boldsymbol{z}^{(i)}_{\alpha\beta}$, cf. (5.7.10), par*

$$\overline{v}_\beta = \sum_{i=0}^{k}\frac{1}{i!}\,D^{(i)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\boldsymbol{z}^{(i)}_{\alpha\beta}+\mathrm{O}\left(h^{k+1}\right) \quad (5.8.1)$$

$$\overline{v}_\alpha = \sum_{i=0}^{k}\frac{1}{i!}\,D^{(i)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\boldsymbol{x}^{(i)}_\alpha+\mathrm{O}\left(h^{k+1}\right) \quad (5.8.2)$$

$$\overline{v}_\beta-\overline{v}_\alpha = \sum_{i=1}^{k}\frac{1}{i!}\,D^{(i)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\left[\boldsymbol{z}^{(i)}_{\alpha\beta}-\boldsymbol{x}^{(i)}_\alpha\right]+\mathrm{O}\left(h^{k+1}\right)\,. \quad (5.8.3)$$

En particulier, si p est un polynôme de degré k, la moyenne $\overline{p}_\beta$ de p sur la cellule $\mathcal{T}_\beta$ est

$$\overline{p}_\beta=\sum_{i=0}^{k}\frac{1}{i!}\,D^{(i)}p\Big|_{\boldsymbol{x}_\alpha}\bullet\boldsymbol{z}^{(i)}_{\alpha\beta}\,. \quad (5.8.4)$$

DÉMONSTRATION. La fonction v satisfait la formule de Taylor avec reste intégral

$$v(\boldsymbol{x})=\sum_{j=0}^{k}\frac{1}{j!}\,D^{(j)}v\Big|_{\boldsymbol{x}_\alpha}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^j+r_{v,\boldsymbol{x}_\alpha}(\boldsymbol{x}) \quad (5.8.5)$$

pour $\boldsymbol{x}\in\mathcal{T}_\alpha$ et $\boldsymbol{x}\in\mathcal{T}_\beta$. Le reste intégral est donné par

$$r_{v,\boldsymbol{x}_\alpha}(\boldsymbol{x})=\int_0^1\frac{(1-s)^k}{k!}\,D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha+s(\boldsymbol{x}-\boldsymbol{x}_\alpha)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^{k+1}\,ds\,.$$

et satisfait selon l'hypothèse sur v une estimation

$$r_{v,\boldsymbol{x}_\alpha}(\boldsymbol{x})\le C\,\|\boldsymbol{x}-\boldsymbol{x}_\alpha\|^{k+1}=\mathrm{O}\left(h^{k+1}\right)$$

pour $\boldsymbol{x} \in \mathcal{T}_\alpha$ et $\boldsymbol{x} \in \mathcal{T}_\beta$. L'intégration de (5.8.5) sur la cellule $\mathcal{T}_\beta$ et l'application du lemme 5.7.2 démontre la formule (5.8.1)

$$\overline{v}_\beta = \frac{1}{|\mathcal{T}_\beta|} \int \left[\sum_{j=0}^{k} \frac{1}{j!} \, D^{(j)} v \Big|_{\boldsymbol{x}_\alpha} \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)^j + r_{v,\boldsymbol{x}_\alpha}(\boldsymbol{x}) \right] dx = \\ = \sum_{j=0}^{k} \frac{1}{j!} \, D^{(j)} v \Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(j)} + \mathrm{O}\left(h^{k+1}\right). \quad (5.8.6)$$

L'identité (5.7.12) $\boldsymbol{z}_{\alpha\alpha}^{(j)} = \boldsymbol{x}_\alpha^{(j)}$ prouve la formule (5.8.2). La différence des formules (5.8.1) et (5.8.2) donne (5.8.3) car le premier terme dans la somme est nul en raison de $\boldsymbol{z}_{\alpha\beta}^{(0)} - \boldsymbol{x}_\alpha^{(0)} = 1 - 1 = 0$. Finalement, (5.8.4) est un cas particulier de (5.8.1). □

Si le volume de la cellule $\mathcal{T}_\beta$ tend vers zéro, ses moments $\boldsymbol{x}_\beta^{(k)}$ d'ordre $k \geq 1$ tendent vers zéro et pour cette raison

$$\boldsymbol{z}_{\alpha\beta}^{(j)} \longrightarrow \boldsymbol{h}_{\alpha\beta}^{j} \text{ si } |\mathcal{T}_\beta| \to 0\,. \quad (5.8.7)$$

La limite (5.8.7) montre que le développement de Taylor (5.8.1) devient, dans cette limite, le développement classique de la valeur $v(\boldsymbol{x}_\beta)$ de v au point $\boldsymbol{x}_\beta$ défini par (5.2.30)

$$\overline{v}_\beta \longrightarrow v(\boldsymbol{x}_\beta) = \sum_{j=0}^{k} \frac{1}{j!} \, D^{(j)} v \Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{h}_{\alpha\beta}^{j} + \mathrm{O}\left(h^{k+1}\right)$$

5.9. Identités géométriques en maillage non structuré

Les vecteurs et tenseurs définis à partir des maillages non structurés dans les sections 5.3 et 5.7 sont liés par des identités géométriques. L'objectif de cette section est d'établir un ensemble d'identités qui seront utilisées dans les chapitres suivants.

(1) La première identité (5.5.2) est

$$\sum_\beta \boldsymbol{a}_{\alpha\beta} = 0\,.$$

Elle provient de l'application du théorème de Green à une fonction constante

$$0 = \int_{\mathcal{T}_\alpha} \boldsymbol{\nabla}(1) \, dx = \sum_{\beta \in \mathbb{V}_\alpha} \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x}) \, d\sigma = \sum_{\beta \in \mathbb{V}_\alpha} \boldsymbol{a}_{\alpha\beta}\,.$$

Si l'on interprète les fonctions constantes comme des polynômes de degré zéro, on peut considérer (5.5.2) comme la première d'une suite d'identités obtenues par l'application du théorème de Green à des polynômes de degré croissant.

(2) La deuxième identité s'écrit de façon explicite pour le degré un par

$$|\mathcal{T}_\alpha| \, \delta_{ij} = \int_{\mathcal{T}_\alpha} \frac{\partial}{\partial x_i} (x_j - x_{\alpha,j}) \, dx = \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \nu_i(\boldsymbol{x}) \, (x_j - x_{\alpha,j}) \, d\sigma\,,\ 1 \leq i,j \leq d\,. \quad (5.9.1)$$

La suite nécessite la définition de tenseurs *non symétriques* d'ordre $j+1$

$$\boldsymbol{\kappa}_{\alpha\beta}^{(j)} \triangleq \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x}) \otimes (\boldsymbol{x} - \boldsymbol{x}_\alpha)^j \, d\sigma \quad (5.9.2)$$

de composantes

$$\kappa_{i_0 i_1 \cdots i_j}^{\alpha\beta} \triangleq \int_{\mathcal{A}_{\alpha\beta}} \nu_{i_0}(\boldsymbol{x}) \, (x - x_\alpha)_{i_1} \cdots (x - x_\alpha)_{i_j} \, d\sigma\,. \quad (5.9.3)$$

Si la face $\mathcal{A}_{\alpha\beta}$ est plane, le vecteur normal unitaire $\boldsymbol{\nu}(\boldsymbol{x})$ dans (5.9.2) est constant. Il est alors possible de sortir $\boldsymbol{\nu}(\boldsymbol{x})$ de l'intégrale dans (5.9.2) qui devient, grâce aux définitions (5.3.5) et (5.7.7), l'identité

$$\boldsymbol{\kappa}_{\alpha\beta}^{(j)} = \boldsymbol{\nu}_{\alpha\beta} \otimes \int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{x} - \boldsymbol{x}_\alpha)^j \, d\sigma = \boldsymbol{a}_{\alpha\beta} \otimes \boldsymbol{k}_{\alpha\beta}^{(j)} . \tag{5.9.4}$$

On définit $\boldsymbol{\kappa}_{\alpha\beta}^{(j)} \triangleq 0$ si les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ n'ont pas de face commune. La définition (5.9.3) permet d'écrire l'identité (5.9.1) comme

$$|\mathcal{T}_\alpha| \, \delta_{ij} \quad = \quad \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \nu_i(\boldsymbol{x}) \, (x_j - x_{\alpha,j}) \, d\sigma = \sum_\beta \kappa_{ij}^{\alpha\beta} \tag{5.9.5}$$

qui se simplifie pour une face $\mathcal{A}_{\alpha\beta}$ plane en

$$|\mathcal{T}_\alpha| \, \delta_{ij} = \sum_\beta a_i^{\alpha\beta} k_j^{\alpha\beta} . \tag{5.9.6}$$

En notation tensorielle, les identités (5.9.5) et (5.9.6) s'écrivent respectivement

$$|\mathcal{T}_\alpha| \, \boldsymbol{\delta}^{(2)} = \sum_\beta \boldsymbol{\kappa}_{\alpha\beta}^{(1)} \tag{5.9.7}$$

et

$$|\mathcal{T}_\alpha| \, \boldsymbol{\delta}^{(2)} = \sum_\beta \boldsymbol{a}_{\alpha\beta} \otimes \boldsymbol{k}_{\alpha\beta} . \tag{5.9.8}$$

(3) La troisième identité s'obtient pour le degré deux par la formule

$$\int_{\mathcal{T}_\alpha} \frac{\partial}{\partial x_i} \left[(x_j - x_{\alpha,j}) \, (x_l - x_{\alpha,l}) \right] dx = \int_{\mathcal{T}_\alpha} \left[\delta_{ij} \, (x_l - x_{\alpha,l}) + \delta_{il} \, (x_j - x_{\alpha,j}) \right] dx = 0 \tag{5.9.9}$$

car $\boldsymbol{x}_\alpha$ est le barycentre de la cellule $\mathcal{T}_\alpha$. L'application du théorème de Green à (5.9.9) donne

$$\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \nu_i(\boldsymbol{x}) \, (x_j - x_{\alpha,j}) \, (x_l - x_{\alpha,l}) = 0 . \tag{5.9.10}$$

La définition (5.9.3) permet d'écrire (5.9.10) comme

$$\sum_\beta \kappa_{ijl}^{\alpha\beta} = 0 \, , \; 1 \leq i, j, l \leq d \, .$$

Si la face est plane, (5.9.4) permet de simplifier (5.9.10) en

$$\sum_\beta a_i^{\alpha\beta} k_{jl}^{\alpha\beta} = 0 \, , \; 1 \leq i, j, l \leq d \, ,$$

ce qui s'écrit en notation tensorielle comme

$$\sum_\beta \boldsymbol{a}_{\alpha\beta} \otimes \boldsymbol{k}_{\alpha\beta}^{(2)} = 0 . \tag{5.9.11}$$

(4) Pour le degré trois, on obtient

$$\begin{aligned} \int_{\mathcal{T}_\alpha} \frac{\partial}{\partial x_i} \left[(x_j - x_{\alpha,j}) \, (x_k - x_{\alpha,k}) \, (x_l - x_{\alpha,l}) \right] dx = \\ = \int_{\mathcal{T}_\alpha} \left[\delta_{ij} \, (x_k - x_{\alpha,k}) \, (x_l - x_{\alpha,l}) + \delta_{ik} \, (x_j - x_{\alpha,j}) \, (x_l - x_{\alpha,l}) + \right. \\ \left. + \delta_{il} \, (x_j - x_{\alpha,j}) \, (x_k - x_{\alpha,k}) \right] dx = |\mathcal{T}_\alpha| \left[\delta_{ij} x_{\alpha,kl} + \delta_{ik} x_{\alpha,jl} + \delta_{il} x_{\alpha,jk} \right] . \end{aligned} \tag{5.9.12}$$

La forme de l'identité (5.9.12) suggère la définition des tenseurs *non symétriques* d'ordre $j+1$

$$\boldsymbol{\zeta}_\alpha^{(j)} \triangleq \int_{\mathcal{T}_\alpha} \boldsymbol{\nabla} \otimes (\boldsymbol{x} - \boldsymbol{x}_\alpha)^j \, dx \,, \; 1 \leq j \,. \tag{5.9.13}$$

Dans (5.9.13), les composantes du tenseur de différentiation $\boldsymbol{\nabla}$ sont données par $\frac{\partial}{\partial x_i}$, $1 \leq i \leq d$.

Les composantes de $\boldsymbol{\zeta}_\alpha^{(j)}$ sont

$$\begin{aligned}\zeta_{i_0 i_1 \ldots i_j} \triangleq \int_{\mathcal{T}_\alpha} \big[& \delta_{i_0 i_1} \left(x_{i_2} - x_{\alpha,i_2}\right) \cdots \left(x_{i_j} - x_{\alpha,i_j}\right) + \ldots \\ & \ldots + \left(x_{i_1} - x_{\alpha,i_1}\right) \cdots \delta_{i_0 i_k} \cdots \left(x_{i_j} - x_{\alpha,i_j}\right) + \ldots \\ & \ldots + \left(x_{i_1} - x_{\alpha,i_1}\right) \cdots \left(x_{i_{j-1}} - x_{\alpha,i_{j-1}}\right) \delta_{i_0 i_j} \big] \, dx \,. \end{aligned} \tag{5.9.14}$$

Pour $j = 1$, la définition (5.9.14) implique l'identité

$$\boldsymbol{\zeta}_\alpha^{(1)} = |\mathcal{T}_\alpha| \, \boldsymbol{\delta}^{(2)} \tag{5.9.15}$$

et pour $j = 0$ on définit

$$\boldsymbol{\zeta}_\alpha^{(0)} \triangleq 0 \,. \tag{5.9.16}$$

L'application du théorème de Green à (5.9.12) donne pour une face $\mathcal{A}_{\alpha\beta}$ gauche l'identité

$$\zeta_{\alpha,ijkl} = |\mathcal{T}_\alpha| \left(\delta_{ij} x_{\alpha,kl} + \delta_{ik} x_{\alpha,jl} + \delta_{il} x_{\alpha,jk}\right) = \sum_\beta \kappa_{ijkl}^{\alpha\beta} \tag{5.9.17}$$

et pour une face $\mathcal{A}_{\alpha\beta}$ plane l'identité

$$\zeta_{\alpha,ijkl} = |\mathcal{T}_\alpha| \left(\delta_{ij} x_{\alpha,kl} + \delta_{ik} x_{\alpha,jl} + \delta_{il} x_{\alpha,jk}\right) = \sum_\beta a_i^{\alpha\beta} k_{jkl}^{\alpha\beta} \,. \tag{5.9.18}$$

En notation tensorielle, (5.9.17) s'écrit

$$\boldsymbol{\zeta}_\alpha^{(3)} \;\; = \;\; \sum_\beta \boldsymbol{\kappa}_{\alpha\beta}^{(3)}$$

et (5.9.18) devient

$$\boldsymbol{\zeta}_\alpha^{(3)} = \sum_\beta \boldsymbol{a}_{\alpha\beta} \otimes \boldsymbol{k}_{\alpha\beta}^{(3)} \,.$$

La généralisation de cette procédure conduit au

LEMME 5.9.1 (Identités fondamentales en maillage non structuré). *Pour toute cellule $\mathcal{T}_\alpha$ du maillage, les tenseurs $\boldsymbol{\zeta}_\alpha^{(j)}$ définis par (5.9.13) et les tenseurs $\boldsymbol{\kappa}_{\alpha\beta}^{(j)}$ définis par (5.9.2) satisfont l'identité*

$$\boldsymbol{\zeta}_\alpha^{(j)} = \sum_\beta \boldsymbol{\kappa}_{\alpha\beta}^{(j)} \,, \; 0 \leq j \,. \tag{5.9.19}$$

Si toutes les faces $\mathcal{A}_{\alpha\beta}$ de $\mathcal{T}_\alpha$ sont planes, les tenseurs $\boldsymbol{\zeta}_\alpha^{(j)}$ définis par (5.9.13), les vecteurs surface (5.3.3) et les tenseurs (5.7.7) satisfont

$$\boldsymbol{\zeta}_\alpha^{(j)} = \sum_\beta \boldsymbol{a}_{\alpha\beta} \otimes \boldsymbol{k}_{\alpha\beta}^{(j)} \,, \; 0 \leq j \,. \tag{5.9.20}$$

DÉMONSTRATION. L'application du théorème de Green à la définition (5.9.13) prouve (5.9.19). Si toutes les faces $\mathcal{A}_{\alpha\beta}$ de $\mathcal{T}_\alpha$ sont planes, l'identité (5.9.4) et (5.9.19) démontrent (5.9.20). □

Le lemme 5.9.1 permet de prouver la

PROPOSITION 5.9.2 (Relations entre les moments des cellules et des faces). *Si toutes les faces d'une cellule $\mathcal{T}_\alpha$ sont planes, l'identité suivante est valable pour $j \geq 1$ et tout $\boldsymbol{c} \in \mathbb{R}^d$*

$$j \, |\mathcal{T}_\alpha| \left(\boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(j-1)}\right) = \boldsymbol{c} \cdot \boldsymbol{\zeta}_\alpha^{(j)} = \sum_\beta \left(\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta}\right) \boldsymbol{k}_{\alpha\beta}^{(j)} \,. \tag{5.9.21}$$

DÉMONSTRATION. Pour $j \geq 2$, le tenseur $\boldsymbol{c} \cdot \boldsymbol{\zeta}_\alpha^{(j)}$ a, d'après la définition (5.9.14), les composantes

$$\sum_{i_0=1}^{d} c_{i_0} \zeta_{i_0 i_1 \dots i_j} = \int_{\mathcal{T}_\alpha} \big[c_{i_1} \left(x_{i_2} - x_{\alpha,i_2}\right) \cdots \left(x_{i_j} - x_{\alpha,i_j}\right) + \dots \\ \dots + \left(x_{i_1} - x_{\alpha,i_1}\right) \cdots c_{i_k} \cdots \left(x_{i_j} - x_{\alpha,i_j}\right) + \dots \\ \dots + \left(x_{i_1} - x_{\alpha,i_1}\right) \cdots \left(x_{i_{j-1}} - x_{\alpha,i_{j-1}}\right) c_{i_j} \big] \, dx \,. \quad (5.9.22)$$

Le produit symétrique $\boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(j-1)}$ a, par la définition (5.2.19), les composantes

$$\left(\boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(j-1)}\right)_{i_1 \dots i_j} = \frac{1}{j!} \sum_{\pi \in \mathfrak{S}_j} c_{i_{\pi(1)}} x_{\alpha, i_{\pi(2)} \cdots i_{\pi(j)}} \,. \quad (5.9.23)$$

Pour une valeur $k = \pi(1)$ fixe, il y a $(j-1)!$ permutations de l'ensemble restant $\{2, \dots, j\}$. La symétrie de $\boldsymbol{x}_\alpha^{(j-1)}$ impose

$$x_{\alpha, i_2 \cdots i_j} = x_{\alpha, i_{\pi(2)} \cdots i_{\pi(j)}}$$

pour toute permutation $\pi \in \mathfrak{S}_{j-1}$ de l'ensemble $\{2, \dots, j\}$. Pour cette raison, on peut réarranger les indices dans (5.9.23) et écrire (5.9.23) comme

$$\left(\boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(j-1)}\right)_{i_1 \dots i_j} = \frac{(j-1)!}{j!} \big\{ c_{i_1} x_{\alpha, i_2 i_3 \cdots i_j} + \cdots + \\ + c_{i_k} x_{\alpha, i_1 \cdots i_{k-1} i_{k+1} \cdots i_j} + \cdots + c_{i_j} x_{\alpha, i_1 i_2 \cdots i_{j-1}} \big\} \,. \quad (5.9.24)$$

La définition (5.7.2) des composantes de $\boldsymbol{x}_\alpha^{(j-1)}$ et la comparaison des identités (5.9.22) et (5.9.24) permettent de conclure que

$$j \, |\mathcal{T}_\alpha| \, \boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(j-1)} = \boldsymbol{c} \cdot \boldsymbol{\zeta}_\alpha^{(j)}$$

pour $j \geq 2$. Dans le cas $j = 1$, la combinaison de l'identité (5.7.3) $\boldsymbol{x}_\alpha^{(0)} = 1$, de l'identité (5.9.15)

$$|\mathcal{T}_\alpha| \, \boldsymbol{c} \cdot \boldsymbol{\delta}^{(2)} = \boldsymbol{\zeta}_\alpha^{(1)}$$

et de la définition (5.2.20)

$$\boldsymbol{c} \odot 1 \triangleq \boldsymbol{c}$$

entraîne

$$|\mathcal{T}_\alpha| \, \boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(0)} = |\mathcal{T}_\alpha| \, \boldsymbol{c} \odot 1 = |\mathcal{T}_\alpha| \, \boldsymbol{c} = |\mathcal{T}_\alpha| \, \boldsymbol{c} \cdot \boldsymbol{\delta}^{(2)} = \boldsymbol{c} \cdot \boldsymbol{\zeta}_\alpha^{(1)} \,.$$

Finalement, le lemme 5.9.1 permet de prouver la deuxième identité dans (5.9.21)

$$j \, |\mathcal{T}_\alpha| \, \boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(j-1)} = \sum_\beta \left(\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta}\right) \boldsymbol{k}_{\alpha\beta}^{(j)} \,.$$

□

5.10. Calcul avec des tenseurs symétriques

Le développement de schémas numériques d'ordre élevé nécessite la manipulation de tenseurs symétriques d'ordre deux, trois et plus. Les composantes de tels tenseurs ne sont pas mutuellement indépendantes. Cela pose deux problèmes distincts, l'un d'ordre théorique, l'autre d'ordre pratique.

(1) Si l'on travaille avec des systèmes d'équations linéaires dont les inconnues et les coefficients sont des tenseurs symétriques, il faut éliminer les inconnues et les équations dépendantes. En particulier, il faut disposer d'une méthode pour transcrire ce type de système en une équation matricielle.

(2) Pour une implémentation concrète, *il est inutile de stocker toutes les composantes d'un tenseur symétrique* $\boldsymbol{a}^{(k)}$ en mémoire.

Ce constat entraîne deux questions.

– Quel est le nombre de composantes indépendantes du tenseur $\boldsymbol{a}^{(k)}$?
– Comment peut-on les stocker dans un tableau ?

Pour répondre à ces questions, on rappelle d'abord que deux k-tuples d'indices $(i_1, \cdots, i_k)$ et $(j_1, \cdots, j_k)$ désignent d'après (5.2.6) la même composante du tenseur symétrique $\boldsymbol{a}^{(k)}$

$$a_{i_1,\cdots,i_k} = a_{j_1,\cdots,j_k}$$

s'il existe une permutation $\pi \in \mathfrak{S}_k$ telle que $i_l = j_{\pi(l)}$ pour $1 \leq l \leq k$. Pour les besoins d'indexation des composantes $a_{i_1,\cdots,i_k}$ d'un tenseur symétrique $\boldsymbol{a}^{(k)}$, on peut donc identifier de tels k-tuples d'indices en introduisant une relation d'équivalence sur l'ensemble

$$\mathbb{J}_k^d \triangleq \{\, (i_1, \ldots, i_k) |\, 1 \leq i_1, \ldots, i_k \leq d\} = \{1, \ldots, d\}^k \tag{5.10.1}$$

par la

Définition 5.10.1 (Indices équivalentes). Deux k-tuples d'indices

$$(i_1, \cdots, i_k) \in \mathbb{J}_k^d \text{ et } (j_1, \cdots, j_k) \in \mathbb{J}_k^d$$

sont dits *équivalents* et on écrit

$$(i_1, \cdots, i_k) \sim (j_1, \cdots, j_k) \tag{5.10.2}$$

s'il existe une permutation $\pi \in \mathfrak{S}_k$ telle que $i_l = j_{\pi(l)}$ pour $1 \leq l \leq k$.

La définition 5.10.1 entraîne le

Lemme 5.10.2 (Classes d'indices équivalentes). *La relation définie par (5.10.2) est réflexive, symétrique et transitive et définit par conséquent une partition de $\mathbb{J}_k^d$ en classes d'équivalence. La classe d'équivalence de $(i_1, \cdots, i_k) \in \mathbb{J}_k^d$ est notée $[i_1, \cdots, i_k]$ et l'ensemble de ces classes est noté $\mathbb{J}_k^d/\sim$. Chaque classe a un représentant unique $(i_1, \cdots, i_k)$ pour lequel $i_1 \leq \cdots \leq i_k$.*

Démonstration. Chaque $\pi \in \mathfrak{S}_k$ définit une application de l'ensemble $\mathbb{J}_k^d$ sur lui-même

$$\mathbb{J}_k^d \ni (i_1, \cdots, i_k) \longmapsto \left(i_{\pi(1)}, \cdots, i_{\pi(k)}\right) . \tag{5.10.3}$$

Les applications (5.10.3) définissent une action du groupe des permutations $\mathfrak{S}_k$ sur $\mathbb{J}_k^d$ et les ensembles d'indices équivalentes sont les orbites de cette action. Cela prouve immédiatement que la relation (5.10.2) est une relation d'équivalence, voir par exemple [**35**]. Pour la commodité du lecteur, on donne la démonstration détaillée :

(1) Réflexivité : l'existence de la permutation identique $\pi(l) = l$ pour $1 \leq l \leq k$ montre que la relation (5.10.2) est réflexive : $(i_1, \cdots, i_k) \sim (i_1, \cdots, i_k)$.

(2) Symétrie : pour toute permutation $\pi \in \mathfrak{S}_k$ il existe une permutation réciproque π^{-1}. Si $i_l = j_{\pi(l)}$ pour $1 \leq l \leq k$, la permutation réciproque donne $i_{\pi^{-1}(l)} = j_l$ pour $1 \leq l \leq k$. Cela prouve que la relation (5.10.2) est symétrique car $(i_1, \cdots, i_k) \sim (j_1, \cdots, j_k)$ implique $(j_1, \cdots, j_k) \sim (i_1, \cdots, i_k)$.

(3) Transitivité : La composition de permutations permet de démontrer la transitivité de la relation (5.10.2). Soient deux permutations $\pi_1 \in \mathfrak{S}_k$ et $\pi_2 \in \mathfrak{S}_k$ telles que $i_l = j_{\pi_1(l)}$ pour $1 \leq l \leq k$ et $j_l = n_{\pi_2(l)}$ pour $1 \leq l \leq k$, alors il vient $i_l = n_{\pi_2(\pi_1(l))}$ pour $1 \leq l \leq k$. Cela montre que $(i_1, \cdots, i_k) \sim (j_1, \cdots, j_k)$ et $(j_1, \cdots, j_k) \sim (n_1, \cdots, n_k)$ entraînent $(i_1, \cdots, i_k) \sim (n_1, \cdots, n_k)$.

Finalement, puisque les entiers sont un ensemble totalement ordonné, on peut toujours arranger les k entiers $(i_1, \cdots, i_k)$ dans leur ordre naturel par une permutation et cet ordre est unique. Il existe donc pour chaque classe d'équivalence un représentant unique $(i_1, \cdots, i_k)$ pour lequel $i_1 \leq \cdots \leq i_k$. □

Les définitions 5.10.1 et (5.2.6) ainsi que le lemme 5.10.2 montrent que le nombre de composantes indépendantes d'un tenseur symétrique $\boldsymbol{a}^{(k)}$ en $\mathbb{R}^d$ est égal au nombre de classes d'équivalence dans $\mathbb{J}_k^d/\sim$. À titre d'exemple, pour $k = 6$ et $d = 3$ on a

$$a_{112333} = a_{132133} = a_{312313} = \cdots \tag{5.10.4}$$

et la classe d'équivalence des 6-tuples $(1, 3, 2, 1, 3, 3)$, $(3, 1, 2, 3, 1, 3)$, etc., est caractérisée par le représentant unique $(1, 1, 2, 3, 3, 3)$.

L'exemple (5.10.4) ci-dessus montre que le nombre de composantes indépendantes de $\boldsymbol{a}^{(k)}$ est égal au nombre des mots de longueur k qu'on peut former avec les caractères distincts $l \in \{1, \ldots, d\}$ si l'ordre des caractères dans le mot ne joue pas de rôle et si le même caractère peut apparaître plusieurs fois dans un mot. Pour trouver le nombre de ces mots, on peut s'imaginer d boîtes distinctes dans lesquelles on place un nombre total de k objets identiques. Le nombre d'objets dans la boîte numéro l, $1 \leq l \leq d$, correspond à la multiplicité du caractère l dans le mot. Il suffit de $d-1$ parois de séparation pour séparer les d boîtes. Si l'on désigne les objets par le symbole $\star$ et la paroi de séparation par le symbole $|$, on peut par exemple représenter le mot 132133 par la chaîne

$$\underbrace{\overbrace{\star\star}^{\text{boîte 1}} \mid \overbrace{\star}^{\text{boîte 2}} \mid \overbrace{\star\star\star}^{\text{boîte 3}}}_{k=6 \text{ fois le symbole } \star} .$$

Cela montre que le nombre cherché correspond au nombre des chaînes qu'on peut former avec les k caractères $\star$ et les $d-1$ caractères $|$ et ce nombre est donné par

$$\frac{(k+d-1)!}{k!\,(d-1)!} .$$

La discussion ci-dessus prouve la

PROPOSITION 5.10.3 (Nombre de composantes d'un tenseur symétrique). *Le nombre de composantes indépendantes d'un tenseur symétrique $\boldsymbol{a}^{(k)}$ dans $\mathbb{R}^d$ est donné par*

$$\frac{(k+d-1)!}{k!\,(d-1)!} = \binom{k+d-1}{k} . \tag{5.10.5}$$

Il reste la question du rangement des composantes indépendantes d'un tenseur *symétrique* $\boldsymbol{a}^{(k)}$ dans un tableau de longueur $\binom{k+d-1}{k}$. Pour cela, il faut définir une relation d'ordre pour les composantes indépendantes de $\boldsymbol{a}^{(k)}$. Les définitions 5.10.1 et (5.2.6) ainsi que le lemme 5.10.2 montrent qu'il suffit de définir cet ordre pour le représentant unique $(i_1, \cdots, i_k)$ de chaque classe d'équivalence $[i_1, \cdots, i_k]$ pour lequel $i_1 \leq \cdots \leq i_k$.

DÉFINITION 5.10.4 (Ordre lexicographique pour les composantes de tenseurs symétriques). On définit une relation d'ordre sur les classes d'équivalence que la relation (5.10.2) induit sur l'ensemble $\mathbb{J}_k^d$:

(i) Pour $k = 1$, la relation d'ordre est définie par

$$[i_1] \preceq [j_1] \Leftrightarrow i_1 \leq j_1$$

(ii) Pour $k > 1$, soient $(i_1, \cdots, i_k)$ et $(j_1, \cdots, j_k)$ deux représentants uniques de leurs classes respectives tels que $i_1 \leq \cdots \leq i_k$ et $j_1 \leq \cdots \leq j_k$. La relation d'ordre sur les classes est définie de façon récursive par

$$[i_1, \cdots, i_k] \preceq [j_1, \cdots, j_k] \text{ si et seulement si : } i_1 < j_1 \text{ ou } (i_1 = j_1 \text{ et } [i_2, \cdots, i_k] \preceq [j_2, \cdots, j_k]) .$$

On a le

LEMME 5.10.5. *La relation $\preceq$ introduite par la définition 5.10.4 est une relation d'ordre totale.*

DÉMONSTRATION. Pour $k = 1$, la relation $\preceq$ de la définition 5.10.4 coïncide avec l'ordre naturel des entiers qui est une relation d'ordre totale. Pour $k > 1$, la définition est récursive. Supposons que la relation $\preceq$ soit une relation d'ordre totale pour $k-1$. L'unicité des représentants $(i_1, \cdots, i_k)$ et $(j_1, \cdots, j_k)$ signifie que $(i_2, \cdots, i_k)$ et $(j_2, \cdots, j_k)$ sont définis de manière unique et qu'on a $i_2 \leq \cdots \leq i_k$ et $j_2 \leq \cdots \leq j_k$. Si $i_1 = j_1$, on a, d'après l'hypothèse, soit $[i_1, \cdots, i_k] \preceq [j_1, \cdots, j_k]$ si $[i_2, \cdots, i_k] \preceq [j_2, \cdots, j_k]$, soit $[j_1, \cdots, j_k] \preceq [i_1, \cdots, i_k]$ si $[j_2, \cdots, j_k] \preceq [i_2, \cdots, i_k]$. Si $i_1 \neq j_1$, on a $[i_1, \cdots, i_k] \preceq [j_1, \cdots, j_k]$ si $i_1 < j_1$ et $[j_1, \cdots, j_k] \preceq [i_1, \cdots, i_k]$ si $i_1 > j_1$. Cela montre que la relation $\preceq$ est bien définie pour $k \geq 1$ pour tout couple de classes de $\mathbb{J}_k^d / \sim$. La relation $\sim$ est réflexive pour $k > 1$, c'est-à-dire $[i_1, \cdots, i_k] \preceq [i_1, \cdots, i_k]$, si elle est réflexive pour $k - 1$. Elle est antisymétrique parce que la combinaison de $[i_1, \cdots, i_k] \preceq [j_1, \cdots, j_k]$ et

de $[j_1, \cdots, j_k] \preceq [i_1, \cdots, i_k]$ entraîne les relations $i_1 = j_1$, ainsi que $[i_2, \cdots, i_k] \preceq [j_2, \cdots, j_k]$ et $[j_2, \cdots, j_k] \preceq [i_2, \cdots, i_k]$. En raison de l'hypothèse sur le cas $k-1$, cela permet de conclure $[i_1, \cdots, i_k] = [j_1, \cdots, j_k]$. On suppose maintenant $[i_1, \cdots, i_k] \preceq [j_1, \cdots, j_k]$ et $[j_1, \cdots, j_k] \preceq [n_1, \cdots, n_k]$. Cela entraîne $i_1 \leq j_1 \leq n_1$. Si l'on a $i_1 < j_1 \leq n_1$ ou $i_1 \leq j_1 < n_1$, il vient immédiatement par la définition 5.10.4 $[i_1, \cdots, i_k] \preceq [n_1, \cdots, n_k]$. Sinon, on a $i_1 = j_1 = n_1$ et la transitivité de la relation $\preceq$ pour $k-1$ prouve que la relation $\preceq$ est transitive pour k. □

Considérons un tenseur symétrique $\boldsymbol{a}^{(k)}$. Tous les l-tuples $(i_1, \cdots, i_k) \in \mathbb{J}_k^d$ d'une même classe $[i_1, \cdots, i_k] \in \mathbb{J}_k^d/\sim$ désignent la même composante $a_{i_1,\cdots,i_k}$ de $\boldsymbol{a}^{(k)}$. Pour une classe $[i_1, \cdots, i_k] \in \mathbb{J}_k^d/\sim$, il existe un représentant unique $(i'_1, \cdots, i'_k)$ tel que $i'_1 \leq \cdots \leq i'_k$. Il est donc suffisant de stocker les composantes $a_{i'_1,\cdots,i'_k}$ pour lesquelles $i'_1 \leq \cdots \leq i'_k$. L'intérêt de la relation d'ordre 5.10.4 est de fournir un ordre naturel pour ranger ces composantes de $\boldsymbol{a}^{(k)}$ dans un tableau. Cela est exprimé par la

Définition 5.10.6 (Fonction de rangement). On définit la fonction de rangement d'ordre k en $\mathbb{R}^d$

$$\begin{aligned} \varpi_k^d : \mathbb{J}_k^d/\sim &\rightarrow \left\{1, \ldots, \binom{k+d-1}{k}\right\} \\ [i_1, \ldots, i_k] &\mapsto \varpi_k^d([i_1, \ldots, i_k]) \end{aligned} \tag{5.10.6}$$

de la façon suivante : La valeur $\varpi_k^d([i_1, \ldots, i_k])$ est le numéro de la classe d'équivalence de $(i_1, \ldots, i_k)$ dans la relation d'ordre 5.10.4.□

Dans le cadre de cette thèse, la fonction (5.10.6) a été implémentée par un algorithme récursif. Si k est connu d'avance, il est néanmoins préférable de programmer l'ordre donné par (5.10.6) de façon fixe dans un tableau auxiliaire car un algorithme récursif est consommateur en temps de calcul.

Exemple 5.10.7. En deux dimensions, la définition 5.10.6 permet de ranger les composantes d'un tenseur symétrique $\boldsymbol{a}^{(2)}$ d'ordre deux comme

$$(a_{11}, a_{12}, a_{22})$$

et ceux d'un tenseur symétrique $\boldsymbol{b}^{(3)}$ d'ordre trois comme

$$(b_{111}, b_{112}, b_{122}, b_{222}) \ .$$

En trois dimensions, le rangement est

$$(a_{11}, a_{12}, a_{13}, a_{22}, a_{23}, a_{33})$$

pour un tenseur symétrique $\boldsymbol{a}^{(2)}$ et

$$(b_{111}, b_{112}, b_{113}, b_{122}, b_{123}, b_{133}, b_{222}, b_{223}, b_{233}, b_{333})$$

pour un tenseur symétrique $\boldsymbol{b}^{(3)}$.

5.11. Bilan du chapitre

Le chapitre a permis de mettre en place les outils pour réaliser l'étude.

(1) Les sections 5.3 et 5.4 définissent la notion de maillage non structuré utilisée dans la suite du document. La section 5.5 établit une notation géométrique simplifiée qui permet d'éviter la mention des voisinages dans les sommes sur les cellules du maillage.

(2) La section 5.6 définit le produit tensoriel symétrique $\odot$ qui est particulièrement important pour le chapitre 8.

(3) La section 5.7 définit des tenseurs géométriques associés au maillage non structuré. Ces tenseurs sont importants pour les chapitres 7, 8, 10 et 11.

(4) La section 5.8 définit un développement de Taylor pour les moyennes de maille qui est important pour les chapitres 7, 8 et 11.

(5) La section 5.9 démontre des identités géométriques nécessaire à l'étude de la précision dans le chapitre 11.

(6) Finalement, la section 5.10 est dédiée au calcul avec des tenseurs symétriques. Ces notions sont nécessaires pour l'étude de la reconstruction dans les chapitres 7 et 8.

CHAPITRE 6

Discrétisation spatiale par la méthode des volumes finis

6.1. Objectif du chapitre

L'objectif de ce chapitre est de présenter le cadre général de la discrétisation spatiale par la méthode des volumes finis. Le chapitre sert surtout à clarifier les notions nécessaires à la compréhension des chapitres suivants. Il regroupe entre autres du matériel connu qui est présenté sous une forme qui facilite la lecture du reste du texte.

Une loi de conservation repose sur le principe général de conservation d'une quantité $u(\boldsymbol{x}, t)$: Sur tout volume suffisamment régulier $\mathcal{T} \subseteq \Omega$ du domaine physique $\Omega \subseteq \mathbb{R}^d$, la quantité $u(\boldsymbol{x}, t)$ vérifie une équation de bilan de la forme suivante : la dérivée en temps de la somme de $u(\boldsymbol{x}, t)$ sur le volume $\mathcal{T} \subseteq \Omega$ est égale à la somme des flux à travers la surface $\partial\mathcal{T}$ du volume. Les flux dépendent de la quantité $u(\boldsymbol{x}, t)$ elle-même et leur forme détermine la dynamique du système. Pour des solutions suffisamment régulières, ce principe peut s'exprimer par une équation aux dérivées partielles pour $u(\boldsymbol{x}, t)$ en fonction de l'espace et du temps.

Ni les équations de bilan ni les équations aux dérivées partielles ne sont directement accessibles à la résolution par ordinateur. L'objectif de la discrétisation spatiale est de remplacer ces équations de bilan par une équation approchée qui est une équation différentielle ordinaire en temps. Ce type d'équation se résout sur ordinateur à l'aide des schémas classiques en temps. L'établissement de cette équation différentielle soulève deux questions :

(1) Il faut d'abord identifier un ensemble fini de variables, appelées variables discrètes, qui représentent la solution approchée.

(2) Ensuite, il est nécessaire de déterminer l'équation différentielle proprement dite qui décrit l'évolution temporelle de ces variables discrètes.

Le choix de cette équation différentielle repose sur deux critères :

(1) La solution de l'équation approchée doit être une bonne approximation de la solution de l'équation exacte dans un sens qui sera précisé plus loin.

(2) La solution de l'équation approchée doit respecter une équation de bilan sur l'ensemble du domaine physique. C'est-à-dire que l'accroissement en temps de la quantité $u(\boldsymbol{x}, t)$ contenue dans le domaine global Ω doit être égal à la différence entre les flux entrants et sortants à travers les frontières extérieures $\partial\Omega$ du domaine.

La discrétisation spatiale par la méthode des volumes finis part du principe de conservation défini ci-dessus. Le découpage du domaine en un maillage polyédrique fournit de façon naturelle une famille finie de variables discrètes pour l'équation approchée. Il suffit de prendre pour chaque cellule du maillage la moyenne de $u(\boldsymbol{x}, t)$ sur la cellule. Le produit de cette moyenne et du volume de la cellule est égal à la quantité $u(\boldsymbol{x}, t)$ contenue dans la cellule. On obtient ainsi un ensemble fini de variables, une pour chaque cellule, dont on peut espérer qu'elles approchent la solution exacte si les diamètres des cellules tendent vers zéro. De plus, la somme des produits des moyennes et des volumes sur toutes les cellules est justement égale à la quantité globale contenue dans le domaine physique. Le deuxième critère est donc naturellement rempli si la somme des produits des moyennes et volumes reste constante dans le temps.

Il reste à établir l'équation différentielle qui gouverne l'évolution temporelle des variables discrètes. Pour les moyennes de cellule, une équation peut être déduite directement des équations de bilan. La solution exacte vérifie une équation de bilan sur chacune des cellules car la dérivée en temps de la quantité $u(\boldsymbol{x}, t)$ contenue dans une cellule est égale à la somme des flux à travers les faces de la cellule. Il suffit de remplacer les flux exacts par des flux approchés qui dépendent

seulement des moyennes de $u(\boldsymbol{x}, t)$ sur les cellules afin d'obtenir une équation différentielle qui ne contient plus que les moyennes de $u(\boldsymbol{x}, t)$. Ces flux sont appelés flux numériques et ils dépendent en général de deux états de chaque côté de la face. Dans le plus simple des schémas, ces deux états sont tout simplement les moyennes de cellule dans les deux cellules adjacentes.

L'objectif suivant consiste à spécifier des critères pour que les flux numériques fournissent une dynamique

(1) qui approche suffisamment bien la dynamique exacte de la loi de conservation par rapport à des critères de précision à définir et

(2) qui respecte la conservation de la quantité globale contenue dans le domaine physique.

Le deuxième point est automatiquement satisfait si pour chaque interface entre deux cellules, le flux sortant de la cellule $\mathcal{T}_\alpha$ vers la cellule $\mathcal{T}_\beta$ est égal au flux entrant dans la cellule $\mathcal{T}_\beta$ en provenance de la cellule $\mathcal{T}_\alpha$. Même si les flux ne sont pas exacts, la quantité globale du domaine physique reste conservée.

Avant d'aborder le premier point, il est nécessaire de fixer un critère de précision. Une analyse de l'erreur sur la dérivée en temps des moyennes de $u(\boldsymbol{x}, t)$ à un temps t_0 fixé peut fournir un critère qualitatif. L'objectif est de voir de quelle façon l'erreur sur la dérivée diminue avec le diamètre des cellules. L'approche choisie est la suivante : dans l'équation approchée, la dérivée en temps des moyennes est elle-même une fonction des moyennes. Dans l'équation exacte, cette dérivée en temps est une fonction des flux évalués à l'aide de la solution exacte. L'idée est de supposer l'existence d'une solution exacte et d'insérer les moyennes de cette solution dans l'équation approchée avec le but de trouver une borne pour l'écart entre la dérivée exacte et la dérivée approchée. L'objectif est de montrer que cette borne est proportionnelle à une certaine puissance du diamètre des cellules, ce qui permettra de déduire que l'erreur tend vers zéro lorsqu'on raffine le maillage de plus en plus. On appelle l'exposant de cette puissance *l'ordre de troncature du schéma*. Ce nombre caractérise en effet l'erreur spatiale du schéma de façon qualitative car plus il est élevé plus l'erreur diminue rapidement avec le raffinement du maillage.

L'équation approchée décrit l'évolution des moyennes de cellule de $u(\boldsymbol{x}, t)$ dans le temps. Si les flux numériques aux interfaces entre les cellules dépendent seulement des moyennes de cellule dans les deux cellules adjacentes, il faut s'attendre à ce que l'évaluation de ces flux ne soit pas très précise. Ceci conduirait automatiquement à une erreur élevée dans l'évolution de la solution approchée qui dépend de l'évaluation des flux numériques. La reconstruction d'une fonction régulière, à partir de moyennes de cellule au voisinage de la cellule, peut remédier à cette situation. La finalité de cette fonction est de fournir des valeurs plus précises aux interfaces entre les cellules et de permettre, par conséquent, une évaluation plus précise des flux. Dans la suite, ce type de méthode s'appellera une *reconstruction locale* parce qu'il s'agit de reconstruire localement des fonctions dans chaque cellule. Ce problème ressemble à un problème d'interpolation, mais on préfère ici le terme *reconstruction* au terme *interpolation* pour des raisons expliquées dans la remarque 2.8.1 de la section 2.8.

Avec la reconstruction locale, les flux approchés dépendent de fonctions régulières qui varient le long des interfaces entre les cellules. Pour cette raison il sera également nécessaire de définir comment intégrer ces fonctions sur les faces pour obtenir l'erreur de troncature souhaitée.

Finalement, il faut aborder un aspect des flux numériques qui ne fait pas l'objet du travail de cette thèse mais qui est indispensable pour la définition de schémas numériques pour les équations hyperboliques. Il s'avère en général nécessaire d'introduire des mécanismes qui atténuent suffisamment la solution numérique pour contrer les instabilités et pour masquer les erreurs numériques sur la vitesse. Au niveau des flux numériques, il existe principalement deux techniques pour atteindre ce but. La première consiste à évaluer des flux numériques dit « centrés » et de rajouter des flux de dissipation artificielle. L'expression « flux centré » vient du fait que le flux s'évalue de façon symétrique à partir des états dans les deux cellules adjacentes à la face. L'autre méthode se base sur des flux dits « décentrés » et qui sont évalués en privilégiant l'état dans l'une des cellules adjacentes à la face. Comme les équations hyperboliques décrivent en général des phénomènes de propagation, il est possible de distinguer les cellules en amont et en aval de la propagation. Le présent travail de recherche se restreint à l'utilisation de flux

décentrés qui sont évalués à l'aide de l'état en amont. Pour la convection linéaire, il s'agit du flux décentré classique calculé à partir de l'état en amont de la propagation. Pour les équations de Navier-Stokes, les calculs se font avec un flux numérique de type Roe sur la base des vitesses de propagation perpendiculaires à la face (vitesses du fluide et des ondes sonores).

6.2. Formulation générale de la méthode des volumes finis

L'objectif de cette section est de poser les bases de la discrétisation spatiale par la méthode des volumes finis. Comme expliqué dans la section 6.1, cette approche permet de transformer une loi de conservation hyperbolique en une équation différentielle ordinaire en temps. Des méthodes classiques de discrétisation en temps [**69**, chap. 2,3,4] permettent ensuite de transformer cette dernière en une équation discrète que l'ordinateur peut résoudre. Le point de vue adopté dans cette étude est de procéder à une discrétisation en deux étapes, d'abord à une discrétisation spatiale et ensuite à une discrétisation temporelle. Ceci permet de mieux isoler les effets de la discrétisation spatiale qui est le sujet principal de cette étude. Comme la discrétisation en temps ne fait pas réellement l'objet du présent travail de recherche, elle ne sera mentionnée que si cela s'avère nécessaire. Rappelons que ce point de vue, discrétisation spatiale puis schéma en temps, est connu dans la littérature sous le nom de *méthode de lignes*. Dans la suite du document il est utile de fixer une terminologie en appelant l'équation différentielle ordinaire issue de la discrétisation spatiale un *schéma semi-discret*. Par conséquent, l'équation discrète obtenue par la discrétisation en temps s'appelle un *schéma discret en espace et en temps* ou simplement un *schéma discret.*

On se donne donc une loi de conservation sur un domaine spatial $\Omega \subset \mathbb{R}^d$ où $d = 1, 2, 3$ dans les applications concrètes. Dans cette section, la loi de conservation est une loi scalaire mais le développement se généralise au cas des systèmes de lois de conservation tels que les équations de Navier-Stokes. Dans la suite, on suppose que Ω est un domaine avec des conditions périodiques aux limites.

Dans sa forme la plus générale, une loi de conservation scalaire est donnée sous forme d'équation de bilan [**52, 73, 48**]. Une fonction $u : \Omega \times [0, T] \to \mathbb{R}$, qui est une solution de cette loi, satisfait sur tout sous-volume fixe $\mathcal{T} \subseteq \Omega$ suffisamment régulier la relation

$$\frac{d}{dt} \int_{\mathcal{T}} u(\boldsymbol{x}, t)\, dx = - \int_{\partial \mathcal{T}} \boldsymbol{\nu}(\boldsymbol{x}) \cdot \boldsymbol{f}(u(\boldsymbol{x}, t))\, d\sigma . \tag{6.2.1}$$

Dans l'équation (6.2.1), $\boldsymbol{f}$ est la fonction de flux associée à la loi de conservation, $\boldsymbol{\nu}(\boldsymbol{x})$ est le vecteur normal unitaire au point $\boldsymbol{x} \in \partial\mathcal{T}$ et $d\sigma$ est l'élément de surface de $\partial\mathcal{T}$. L'équation (6.2.1) signifie que la dérivée en temps de la quantité u contenue dans le sous-volume $\mathcal{T}$ est égale au bilan des flux entrants et sortants à travers la surface $\partial\mathcal{T}$ de $\mathcal{T}$. On peut montrer [**52, 73, 48**] qu'une solution suffisamment régulière u de (6.2.1) satisfait une équation aux dérivées partielles de la forme

$$\partial_t u(\boldsymbol{x}, t) + \boldsymbol{\nabla} \cdot \boldsymbol{f}(u(\boldsymbol{x}, t)) = 0\,,\ \boldsymbol{x} \in \Omega\,,\ 0 \leq t \leq T\,. \tag{6.2.2}$$

La première étape de la discrétisation spatiale consiste à découper le domaine Ω en cellules, comme décrit dans la section 5.3. Ceci permet d'appliquer l'équation (6.2.1) à chaque cellule $\mathcal{T}_\alpha$ du maillage non structuré, ce qui donne

$$\frac{d}{dt} \int_{\mathcal{T}_\alpha} u(\boldsymbol{x}, t)\, dx = - \int_{\partial \mathcal{T}_\alpha} \boldsymbol{\nu} \cdot \boldsymbol{f}(u(\boldsymbol{x}, t))\, d\sigma . \tag{6.2.3}$$

La définition de la moyenne de u sur $\mathcal{T}_\alpha$

$$\overline{u}_\alpha(t) = \frac{1}{|\mathcal{T}_\alpha|} \int_{\mathcal{T}_\alpha} u(\boldsymbol{x}, t)\, dx \tag{6.2.4}$$

et la décomposition du bord de la cellule $\mathcal{T}_\alpha$ en faces $\mathcal{A}_{\alpha\beta}$ permettent d'écrire (6.2.3) comme une somme

$$\frac{d\overline{u}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|} \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x}) \cdot \boldsymbol{f}(u(\boldsymbol{x}, t))\, d\sigma . \tag{6.2.5}$$

L'écriture de (6.2.5) fait usage de la convention de la section 5.5 de définir la face $\mathcal{A}_{\alpha\beta}$ comme l'ensemble vide si les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ n'ont pas d'interface commune. Ceci allège la notation en permettant d'omettre la mention du voisinage dans la somme.

La deuxième étape consiste à remplacer dans (6.2.5) la solution exacte $u(\boldsymbol{x},t)$ par ses moyennes de cellule $\overline{u}_\alpha(t)$ avec l'objectif de formuler une équation différentielle ordinaire pour les moyennes de cellule $\overline{u}_\alpha(t)$. Cette nouvelle équation doit être telle que ses solutions approchent les moyennes de cellule de la solution exacte de (6.2.1). Le membre de gauche de (6.2.5) contient seulement la dérivée en temps des moyennes $\overline{u}_\alpha$ et peut donc rester inchangé. Il suffit alors de remplacer le flux exact $\boldsymbol{f}(u(\boldsymbol{x},t))$ dans le membre de droite de (6.2.5) par une expression qui ne dépend que des moyennes de cellule $\overline{u}_\alpha(t)$. Cette substitution nécessite l'introduction d'une nouvelle notion, celle de flux numérique.

Pour motiver l'introduction des flux numériques, il est utile de passer en revue l'approche initiale de S.K. Godounov dans [**53**]. Le résumé donné ici suit les lignes de [**79**, chapitre 4]. Cette approche a initialement été développée pour des lois de conservation en une seule dimension d'espace et formulée par une discrétisation complète en espace et en temps. On revient temporairement à une seule dimension spatiale, où chaque cellule $\mathcal{T}_\alpha$ a exactement deux voisins, $\mathcal{T}_{\alpha-1}$ à gauche et $\mathcal{T}_{\alpha+1}$ à droite. Les intégrales de surface sur les interfaces des cellules dans (6.2.5) deviennent de simples évaluations de la fonction u aux points $x_{\alpha,\alpha-1}$ et $x_{\alpha,\alpha+1}$. Dans ce cas spécifique, l'intégration de (6.2.5) sur un intervalle de temps $[t_n,t_{n+1}]$ fournit la formule

$$\overline{u}_\alpha(t_{n+1})-\overline{u}_\alpha(t_n)=-\frac{1}{|\mathcal{T}_\alpha|}\int_{t_n}^{t_{n+1}}\boldsymbol{f}(u(x_{\alpha,\alpha+1},t))\,dt-\frac{1}{|\mathcal{T}_\alpha|}\int_{t_n}^{t_{n+1}}\boldsymbol{f}(u(x_{\alpha,\alpha-1},t))\,dt \quad (6.2.6)$$

Il faut souligner que la solution exacte satisfait encore la formule (6.2.6).

La première étape de l'approche de Godounov consiste à remplacer la solution exacte u par une solution approchée $\mathfrak{u}$ constante par cellule et qui peut donc être décrite par un vecteur de valeurs $\mathfrak{u}=(u_1,\ldots,u_N)$. La valeur constante u_α de $\mathfrak{u}$ dans la cellule $\mathcal{T}_\alpha$ est censée approcher la moyenne $\overline{u}_\alpha$ de la solution exacte. Avant de continuer, il est important de fixer une notation pour la suite du document. Les moyennes des solutions exactes sont indiquées par une barre supérieure et l'indice de la cellule par α comme $\overline{u}_\alpha$ dans (6.2.4). Les valeurs discrètes censées approcher les moyennes exactes dans une cellule $\mathcal{T}_\alpha$ s'écrivent sans barre transversale comme la valeur u_α de la solution approchée de Godounov. Les valeurs ponctuelles de fonctions sont toujours notées en spécifiant leur argument comme dans $u(\boldsymbol{x},t)$.

La deuxième étape de l'approche de Godounov est de remplacer la solution exacte $u(x_{\alpha,\alpha+1},t)$ dans le terme de droite de (6.2.6) par la solution du problème de Riemann au point $x_{\alpha,\alpha+1}$ avec la donnée initiale u_α pour $x<x_{\alpha,\alpha+1}$ et $u_{\alpha+1}$ pour $x>x_{\alpha,\alpha+1}$. Dans le cadre de cette étude, il suffit de savoir que le problème de Riemann consiste à résoudre la loi de conservation (6.2.1), au sens faible, avec une condition initiale discontinue. Une description complète et détaillée de cette analyse se trouve par exemple dans [**44, 52, 73, 79, 12**]. Dans la suite, la solution du problème de Riemann sera notée $\widetilde{u}(x,t\,;\,u_\alpha,u_{\alpha+1},x_{\alpha,\alpha+1})$. Cette notation indique que la solution $\tilde{u}$ du problème de Riemann dépend également des valeurs initiales $u_{\alpha+1}$ à droite et u_α à gauche ainsi que du point $x_{\alpha,\alpha+1}$ où se situe la discontinuité dans la solution initiale.

Le point-clé de la méthode de Godounov est l'observation que la valeur de $\widetilde{u}(x,t\,;\,u_\alpha,u_{\alpha+1},x_{\alpha,\alpha+1})$ en $x_{\alpha,\alpha+1}$ ne dépend pas du temps. Notons cette valeur

$$\widetilde{u}_R(x_{\alpha,\alpha+1},u_\alpha,u_{\alpha+1})\triangleq\widetilde{u}(x_{\alpha,\alpha+1},t\,;\,u_\alpha,u_{\alpha+1},x_{\alpha,\alpha+1}) \quad (6.2.7)$$

Cela permet de formuler la

Définition 6.2.1 (Flux de Godounov). Soit $\widetilde{u}(x,t\,;\,u_\alpha,u_{\alpha+1},x_{\alpha,\alpha+1})$ la solution du problème de Riemann pour la condition initiale u_α pour $x<x_{\alpha,\alpha+1}$ et $u_{\alpha+1}$ pour $x>x_{\alpha,\alpha+1}$. Le flux numérique de Godounov est défini comme l'évaluation du flux exact $\boldsymbol{f}$ de (6.2.1) à la solution du problème de Riemann au point $x_{\alpha,\alpha+1}$

$$\widetilde{f}_{\alpha,\alpha+1}(u_\alpha,u_{\alpha+1})\triangleq\boldsymbol{f}(\widetilde{u}_R(x_{\alpha,\alpha+1},u_\alpha,u_{\alpha+1})) \quad (6.2.8)$$

où $\widetilde{u}_R(x_{\alpha,\alpha+1},u_\alpha,u_{\alpha+1})$ est défini par (6.2.7).

Par conséquent, la substitution de $\widetilde{u}_R(x_{\alpha,\alpha+1}, u_\alpha, u_{\alpha+1})$ pour $u(x_{\alpha,\alpha+1}, t)$ permet d'évaluer les intégrales dans le membre de droite de (6.2.6) de façon exacte car *les intégrandes ne dépendent plus du temps*. Ceci permet finalement de transformer la relation (6.2.6) en l'équation suivante

$$u_\alpha(t_{n+1}) - u_\alpha(t_n) = -\frac{t_{n+1}-t_n}{|\mathcal{T}_\alpha|}\left\{\widetilde{f}_{\alpha,\alpha+1}(u_\alpha(t_n), u_{\alpha+1}(t_n)) - \widetilde{f}_{\alpha-1,\alpha}(u_{\alpha-1}(t_n), u_\alpha(t_n))\right\}. \tag{6.2.9}$$

La formule (6.2.9) est un schéma discret décrivant l'évolution en temps des variables discrètes.

On peut noter que le schéma (6.2.9) ne comporte pas de paramètres. C'est en fait le seul schéma entièrement défini par le fait qu'il donne au temps t_{n+1} la moyenne de la solution exacte de (6.2.1) avec la donnée initiale $u(\boldsymbol{x}, t_n) = \overline{u}_\alpha$ pour $\boldsymbol{x} \in \mathcal{T}_\alpha$.

Comme rappelé dans le chapitre 2, le schéma de Godounov a été le point de départ de nombreuses variantes de la méthode MUSCL [**109, 108, 122, 12**]. Dans toutes ces approches, la notion de problème de Riemann local est au coeur de la conception globale du schéma.

Une vision alternative, en un sens plus empirique, consiste à considérer (6.2.9) comme un schéma discret obtenu par l'intégration en temps de l'équation semi-discrète

$$\frac{du_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\left\{\widetilde{f}_{\alpha,\alpha+1}(u_\alpha(t), u_{\alpha+1}(t)) - \widetilde{f}_{\alpha-1,\alpha}(u_{\alpha-1}(t), u_\alpha(t))\right\} \tag{6.2.10}$$

à l'aide du schéma d'Euler explicite.

Le point de vue adopté ici est de définir, par la méthode des lignes, une classe de schémas volumes finis construits par une généralisation de la formule (6.2.10) au cas multidimensionnel : on considère a priori le schéma semi-discret

$$\frac{du_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \widetilde{f}_{\alpha\beta}(u_\alpha(t), u_\beta(t))\, d\sigma. \tag{6.2.11}$$

Autrement dit, (6.2.11) s'obtient formellement à partir de (6.2.5) en remplaçant la moyenne exacte $\overline{u}_\alpha(t)$ par l'inconnue discrète $u_\alpha(t)$ et l'intégrale du second membre de (6.2.5) par des flux numériques.

Comme le sujet principal de l'étude est le cas multidimensionnel, il est nécessaire de préciser le traitement de l'intégrale de surface dans le membre de droite de (6.2.5). La convention choisie ici est de remplacer le flux normal $\boldsymbol{\nu}(\boldsymbol{x})\cdot\boldsymbol{f}(u(\boldsymbol{x}, t))$ par $\widetilde{f}_{\alpha\beta}(u_\alpha(t), u_\beta(t))$. Dans le cas d'une face plane $\mathcal{A}_{\alpha\beta}$, le vecteur surface $\boldsymbol{\nu}(\boldsymbol{x})$ est constant et égal à $\boldsymbol{\nu}_{\alpha\beta}$. Par conséquent, $\widetilde{f}_{\alpha\beta}(u_\alpha(t), u_\beta(t))$ est une approximation de $\boldsymbol{\nu}_{\alpha\beta}\cdot\boldsymbol{f}(u(\boldsymbol{x}, t))$. L'intégrale de surface occasionne alors un facteur égal à $|\mathcal{A}_{\alpha\beta}|$ dans (6.2.11). Si la face $\mathcal{A}_{\alpha\beta}$ n'est pas plane, elle est définie comme une réunion de triangles. On peut choisir $\widetilde{f}_{\alpha\beta}(u_\alpha(t), u_\beta(t))$ comme une approximation de $\boldsymbol{\nu}_{\alpha\beta}\cdot\boldsymbol{f}(u(\boldsymbol{x}, t))$ où $\boldsymbol{\nu}_{\alpha\beta}$ est le vecteur normal unitaire (5.3.5) défini dans la section 5.3. Une alternative est de décomposer la face en ses triangles et d'approcher sur le i-ème triangle, avec le vecteur normal unitaire $\boldsymbol{\nu}_{\alpha\beta}^{(i)}$, l'expression $\boldsymbol{\nu}_{\alpha\beta}^{(i)}\cdot\boldsymbol{f}(u(\boldsymbol{x}, t))$ par un flux numérique $\widetilde{f}_{\alpha\beta}^{(i)}(u_\alpha(t), u_\beta(t))$.

Dans le cas de l'utilisation des flux numériques (6.2.8) de Godounov, l'évolution en temps de (6.2.11) est donc déterminée par la résolution de problèmes de Riemann unidimensionnels sur chaque face du maillage. Il faut noter qu'il s'agit d'une approximation supplémentaire car le solveur de Riemann unidimensionnel ne prend pas en compte la direction multidimensionnelle de propagation. De plus, dans une cellule bidimensionnelle ou tridimensionnelle, les solveurs de Riemann de deux faces partageant un sommet peuvent interagir. Un certain nombre de travaux a été mené afin de trouver des solveurs de Riemann plus adaptés au cas multidimensionnel [**52**]. Il existe également des solveurs de Riemann basées sur des conditions initiales qui ne sont pas constantes par morceaux [**12**]. L'étude des flux numériques et des solveurs de Riemann ne fait cependant pas partie du présent document. C'est pourquoi, dans la suite du chapitre, et pour la plupart du document, il n'est pas nécessaire de connaître la forme exacte des flux numériques. Il suffit de définir formellement la notion de "fonction flux numérique".

DÉFINITION 6.2.2 (Flux numériques). Un *flux numérique conservatif et consistant* est une fonction $\widetilde{f}_{\alpha\beta} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ attachée à la face $\mathcal{A}_{\alpha\beta}$ et qui satisfait les critères suivants, cf. [**73**, déf. 3.2.6, p. 159] :

(i) La première propriété est la *consistance*. Elle est en général une condition nécessaire pour la convergence du schéma (6.2.11) et prend la forme

$$\widetilde{f}_{\alpha\beta}(u,u) = \boldsymbol{f}(u) \cdot \boldsymbol{\nu}_{\alpha\beta} \quad \text{pour tout } u \in \mathbb{R}. \tag{6.2.12}$$

(ii) Pour des raisons expliquées ultérieurement, il est souhaitable que le flux numérique soit une fonction lipschitzienne. Si c'est le cas, il existe une constante $C_{\widetilde{f}}$ telle que

$$\left|\widetilde{f}_{\alpha\beta}(w_{\text{int}}, w_{\text{ext}}) - \widetilde{f}_{\alpha\beta}(v_{\text{int}}, v_{\text{ext}})\right| \leq C_{\widetilde{f}} \left[|w_{\text{ext}} - v_{\text{ext}}| + |w_{\text{int}} - v_{\text{int}}|\right] . \tag{6.2.13}$$

(iii) La troisième propriété est la *conservativité*

$$\widetilde{f}_{\alpha\beta}(w_{\text{int}}, w_{\text{ext}}) = -\widetilde{f}_{\beta\alpha}(w_{\text{ext}}, w_{\text{int}}) . \tag{6.2.14}$$

□

Les propriétés (6.2.12) et (6.2.13) sont essentielles pour la précision du schéma semi-discret. La relation (6.2.14) garantit, dans le cas d'un domaine Ω périodique, la conservation de la quantité totale de la solution u de (6.2.11) sur Ω

$$\frac{d}{dt}\left\{\sum_{\alpha=1}^{N} |\mathcal{T}_\alpha|\, u_\alpha(t)\right\} = 0 . \tag{6.2.15}$$

Une autre propriété importante des flux numériques est donnée par la

DÉFINITION 6.2.3 (Flux numériques monotones). Un flux numérique $\widetilde{f}_{\alpha\beta}$ au sens de la définition 6.2.2 est monotone si $\widetilde{f}_{\alpha\beta}$ est une fonction différentiable et

$$\left.\begin{aligned} \frac{\partial}{\partial w_{\text{int}}}\left[\widetilde{f}_{\alpha\beta}(w_{\text{int}}, w_{\text{ext}})\right] &\geq 0 \\ \frac{\partial}{\partial w_{\text{ext}}}\left[\widetilde{f}_{\alpha\beta}(w_{\text{int}}, w_{\text{ext}})\right] &\leq 0 . \end{aligned}\right\} \tag{6.2.16}$$

□

La propriété (6.2.16) s'avère importante pour la stabilité linéaire et non-linéaire, cf. les chapitres 10 et 12.

Dans la littérature, les preuves de convergence concernent en grande partie la version discrète en temps du schéma semi-discret (6.2.11). L'ouvrage [**73**, ch. 3.3] contient une preuve de la convergence du schéma volumes finis discret appliqué aux lois de conservation scalaires dans $\mathbb{R}^2$. Une preuve de la convergence en norme du schéma (6.2.11) se trouve dans [**46**] pour le cas très spécifique de l'équation de convection linéaire. La norme de l'erreur du schéma est dans ce dernier cas $\mathrm{O}\left(h^{1/2}\right)$ ce qui est plus faible que l'erreur $\mathrm{O}(h)$ qu'on peut obtenir pour la même équation sur un maillage régulier [**46, 73**].

6.3. Amélioration de la précision des méthodes des volumes finis

Grâce à l'introduction de flux numériques, il a été possible de définir le schéma numérique semi-discret (6.2.11) à partir de l'équation de bilan (6.2.1). Le prochain objectif est de développer un critère pratique de précision permettant d'estimer l'erreur numérique entre la solution exacte de (6.2.1) et la solution de (6.2.11). Il s'agit d'abord d'obtenir un critère qualitatif précisant comment cette erreur dépend du diamètre h du maillage et non pas un critère rigoureux de convergence. Notons que les résultats mentionnés à la fin de la section 6.2 sont limités aux lois scalaires et que certains nécessitent des outils d'analyse fonctionnelle assez sophistiqués, comme les mesures de Young et le théorème de Tartar, cf. [**73**, chap. 3.3., p. 164].

Pour estimer l'erreur du schéma (6.2.11), on part d'une solution exacte u de la loi de conservation (6.2.1) et on insère les moyennes de cette solution dans le terme de droite du schéma

(6.2.11). Sous l'hypothèse de la régularité suffisante de u, un développement de Taylor autour d'un point $\boldsymbol{x} \in \mathcal{T}_\alpha$ permet de montrer que

$$|\overline{u}_\alpha(t_0) - u(\boldsymbol{x}, t_0)| = \mathrm{O}(h)$$

où $\overline{u}_\alpha(t_0)$ est la moyenne de u à un temps fixe t_0. Cette relation vient du fait que

$$\overline{u}_\alpha(t_0) = \frac{1}{|\mathcal{T}_\alpha|} \int_{\mathcal{T}_\alpha} \left[u(\boldsymbol{x}, t_0) + (\boldsymbol{y} - \boldsymbol{x}) \cdot \boldsymbol{\nabla} u|_{\boldsymbol{x}, t_0} \right] dy + O\left(h^2\right)$$

ce qui entraîne

$$\overline{u}_\alpha(t_0) - u(\boldsymbol{x}, t_0) = \frac{1}{|\mathcal{T}_\alpha|} \int_{\mathcal{T}_\alpha} (\boldsymbol{y} - \boldsymbol{x}) \cdot \boldsymbol{\nabla} u|_{\boldsymbol{x}, t_0} \, dy + O\left(h^2\right) = O(h) \ .$$

Si le flux numérique satisfait les points (i) et (ii) de la définition 6.2.2, on obtient pour chaque face $\mathcal{A}_{\alpha\beta}$ une inégalité de la forme

$$\left| \widetilde{f}_{\alpha\beta}(\overline{u}_\alpha, \overline{u}_\beta) - \boldsymbol{\nu}_{\alpha\beta} \cdot \boldsymbol{f}(u(\boldsymbol{x}, t)) \right| = \left| \widetilde{f}_{\alpha\beta}(\overline{u}_\alpha, \overline{u}_\beta) - \widetilde{f}_{\alpha\beta}(u(\boldsymbol{x}, t), u(\boldsymbol{x}, t)) \right| \leq$$
$$\leq C_{\widetilde{f}} |\overline{u}_\alpha - u(\boldsymbol{x}, t)| + C_{\widetilde{f}} |\overline{u}_\beta - u(\boldsymbol{x}, t)| = \mathrm{O}(h) \quad (6.3.1)$$

qui est valable pour tout $\boldsymbol{x} \in \mathcal{A}_{\alpha\beta}$.

En insérant cette inégalité dans l'équation de bilan exacte (6.2.3), on obtient

$$\frac{d\overline{u}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|} \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta} \cdot \boldsymbol{f}(u(\boldsymbol{x}, t)) \, d\sigma =$$
$$= -\frac{1}{|\mathcal{T}_\alpha|} \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \widetilde{f}_{\alpha\beta}(\overline{u}_\alpha(t), \overline{u}_\beta(t)) \, d\sigma + \mathrm{O}(1) \ . \quad (6.3.2)$$

L'intégration sur la face dans (6.3.2) donne un facteur h^{d-1} alors que la division par le volume de la cellule apporte un facteur h^{-d}. La combinaison de ces deux facteurs compense le facteur h obtenu par l'estimation (6.3.1). L'équation (6.3.2) montre par conséquent que cette approche n'est pas suffisante pour prouver que l'erreur sur la dérivée en temps des moyennes de u diminue avec le diamètre du maillage. Dans le cas de l'équation de convection linéaire, il est même possible de montrer de façon explicite que l'erreur dans (6.3.2) ne peut pas être d'ordre $\mathrm{O}(h)$ [**73**, lm. 3.2.8, p. 161]. Ceci n'empêche pas la convergence de (6.3.2) dans des cas plus ou moins généraux [**46, 73**] où la vitesse de convergence démontrée est en général $\mathrm{O}\left(h^{1/2}\right)$ ou encore plus lente.

Ce constat indique clairement que le schéma (6.2.11) n'est pas suffisamment précis pour des application pratiques. Pour surmonter cela, il est possible d'adapter l'approche suivante. Pour réduire l'erreur exprimée par (6.3.2), on peut remplacer les arguments $u_\alpha(t)$ et $u_\beta(t)$ dans (6.2.11) par des valeurs plus précises. Pour un temps *fixe* t_0, on reconstruit une fonction w à partir des moyennes approchées $\mathfrak{u} = (u_1, \ldots, u_N)$ sur un voisinage de la cellule $\mathcal{T}_\alpha$. La dépendance de w des moyennes de cellule $\mathfrak{u}(t_0)$ est indiquée par des crochets $w[\mathfrak{u}(t_0)]$ et la dépendance de $\boldsymbol{x} \in \Omega$ par les parenthèses $w[\mathfrak{u}(t_0)](\boldsymbol{x})$. La reconstruction se fait cellule par cellule, de façon à ce que seules les moyennes dans un certain voisinage de la cellule $\mathcal{T}_\alpha$ déterminent la restriction de w à $\mathcal{T}_\alpha$. Dans la suite, cette restriction de w à $\mathcal{T}_\alpha$ est notée w_α.

Pour améliorer la précision, on demande que la reconstruction satisfasse une propriété d'approximation de la forme suivante : pour toute fonction suffisamment régulière u, dont les moyennes de cellule sont $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$, il doit exister une constante C_u, indépendante de h et de la cellule $\mathcal{T}_\alpha$, telle que l'estimation

$$|w_\alpha[\mathfrak{u}(t_0)](\boldsymbol{x}) - u(\boldsymbol{x}, t_0)| \leq \mathrm{C}_u \, h^{k+1} = \mathrm{O}\left(h^{k+1}\right) \text{ pour tout } \boldsymbol{x} \in \mathcal{T}_\alpha \quad (6.3.3)$$

soit vraie dans toutes les cellules $\mathcal{T}_\alpha$ du maillage. Le nombre entier $k+1$ est appelé *l'ordre de la reconstruction*.

Un schéma plus précis s'obtient à présent à partir de (6.2.11) en remplaçant les moyennes u_α et u_β par les fonctions reconstruites $w_\alpha[\mathfrak{u}(t)](\boldsymbol{x})$ et $w_\beta[\mathfrak{u}(t)](\boldsymbol{x})$ sur les interfaces des cellules,

ce qui donne

$$\frac{du_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \widetilde{f}_{\alpha\beta}\left(w_\alpha\left[\mathfrak{u}(t)\right](\boldsymbol{x}), w_\beta\left[\mathfrak{u}(t)\right](\boldsymbol{x})\right)\, d\sigma\,. \tag{6.3.4}$$

La propriété de Lipschitz (6.2.13) du flux numérique entraîne pour chaque face $\mathcal{A}_{\alpha\beta}$ l'estimation suivante qui est valable pour tout $\boldsymbol{x} \in \mathcal{A}_{\alpha\beta}$

$$\begin{aligned}
&\left|\widetilde{f}_{\alpha\beta}\left(w_\alpha\left[\mathfrak{u}(t)\right](\boldsymbol{x},t), w_\beta\left[\mathfrak{u}(t)\right](\boldsymbol{x},t)\right) - \boldsymbol{\nu}_{\alpha\beta}\cdot\boldsymbol{f}\left(u(\boldsymbol{x},t)\right)\right| = \\
&\qquad = \left|\widetilde{f}_{\alpha\beta}\left(w_\alpha\left[\mathfrak{u}(t)\right](\boldsymbol{x},t), w_\beta\left[\mathfrak{u}(t)\right](\boldsymbol{x},t)\right) - \widetilde{f}_{\alpha\beta}\left(u(\boldsymbol{x},t), u(\boldsymbol{x},t)\right)\right| \leq \\
&\leq \mathrm{C}_{\widetilde{f}}\left|w_\alpha\left[\mathfrak{u}(t)\right](\boldsymbol{x},t) - u(\boldsymbol{x},t)\right| + \mathrm{C}_{\widetilde{f}}\left|w_\beta\left[\mathfrak{u}(t)\right](\boldsymbol{x},t) - u(\boldsymbol{x},t)\right| \leq 2\,\mathrm{C}_{\widetilde{f}}\mathrm{C}_u\, h^{k+1} = \mathrm{O}\left(h^{k+1}\right).
\end{aligned} \tag{6.3.5}$$

Afin de montrer comment l'estimation (6.3.5) permet d'obtenir une erreur de troncature d'ordre élevé pour le schéma (6.3.4), il est nécessaire de régler une question technique liée aux intégrales de surface dans le membre de droite de (6.3.4).

Une différence importante entre les schémas (6.2.11) et (6.3.4) est que le terme sous l'intégrale dans le membre de droite de (6.3.4) dépend maintenant de $\boldsymbol{x} \in \mathcal{A}_{\alpha\beta}$. L'intégrale de surface ne peut donc plus être évaluée de façon simple. Il faut également prévoir un traitement spécifique des faces non planes. On peut par exemple les décomposer en triangles et effectuer ensuite l'estimation (6.3.5) triangle par triangle. On peut aussi utiliser l'estimation (6.3.5) pour les faces non planes avec le vecteur surface unique $\boldsymbol{a}_{\alpha\beta}$ défini par (5.3.3). Ceci introduit cependant une erreur de discrétisation supplémentaire.

Après le calcul des intégrales de surface, il est possible de procéder à l'estimation principale de cette section. Grâce à (6.3.5), l'estimation d'erreur (6.3.2) peut être améliorée afin de donner

$$\begin{aligned}
\frac{d\overline{u}_\alpha(t)}{dt} &= -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}\cdot\boldsymbol{f}\left(u(\boldsymbol{x},t)\right)\, d\sigma = \\
&= -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \widetilde{f}_{\alpha\beta}\left(w_\alpha\left[\mathfrak{u}(t)\right](\boldsymbol{x}), w_\beta\left[\mathfrak{u}(t)\right](\boldsymbol{x})\right)\, d\sigma + \mathrm{O}\left(h^k\right).
\end{aligned} \tag{6.3.6}$$

De nouveau, la combinaison de l'intégrale de surface et de la division par le volume apporte un facteur h^{-1} dans l'estimation (6.3.6), ce qui réduit l'ordre de l'erreur d'un degré par rapport à l'ordre de (6.3.5). Dans la suite de ce travail, l'exposant de h dans (6.3.6) s'appelle *l'ordre de troncature* du schéma (6.3.4) et l'erreur s'appelle *l'erreur de troncature*. Par conséquent, une reconstruction d'ordre $k+1$ entraîne que l'erreur de troncature du schéma (6.3.4) est d'ordre au moins $\mathrm{O}\left(h^k\right)$. Par contre, il n'est pas exclu que l'erreur de troncature soit d'ordre $\mathrm{O}\left(h^{k'}\right)$ où $k' > k$, ce qui est en général le cas en maillage structuré, par exemple. Pour le moment, l'erreur de troncature sert à motiver l'introduction de la notion de reconstruction.

L'étape finale de la discrétisation spatiale par la méthode des volumes finis consiste à approcher l'intégrale dans (6.3.4) par une formule appropriée d'intégration numérique. Ceci donne le système dynamique

$$\frac{du_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \sum_q \omega_q\, \widetilde{f}_{\alpha\beta}\left(w_\alpha\left[\mathfrak{u}(t)\right](\boldsymbol{x}_{\alpha\beta;q}), w_\beta\left[\mathfrak{u}(t)\right](\boldsymbol{x}_{\alpha\beta;q})\right)\,. \tag{6.3.7}$$

Dans l'équation (6.3.7), les $\boldsymbol{x}_{\alpha\beta;q}$ sont les points de quadrature ou points de Gauss sur $\mathcal{A}_{\alpha\beta}$ et les nombres réels ω_q sont les poids de quadrature. Cette nouvelle étape n'introduit aucune nouvelle erreur de discrétisation si l'expression composée sous l'intégrale dans le membre de droite de (6.3.4) est un polynôme de degré k' en $\boldsymbol{x}$ et si la formule de quadrature intègre exactement les polynômes de degré k'. Dans les autres cas, l'introduction de la quadrature dans (6.3.4) ajoute une nouvelle erreur de discrétisation spatiale. Pour cette raison il est nécessaire que la précision de la méthode de quadrature (intégration numérique) soit suffisamment élevée. En particulier, l'erreur d'approximation de la quadrature doit être du même ordre que l'ordre de reconstruction.

Les méthodes de type volumes finis ont été conçues pour la discrétisation de lois de conservation. Dans chaque étape de la discrétisation, il est donc important de respecter le plus possible le principe de conservativité. Cela justifie d'exiger que la fonction $w_\alpha[\mathfrak{u}]$ reconstruite à partir des moyennes $\mathfrak{u} = (u_1, \ldots, u_N)$ respecte la moyenne u_α sur la cellule $\mathcal{T}_\alpha$ de façon à ce que

$$\frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha} w_\alpha[\mathfrak{u}](\boldsymbol{x})\, dx = u_\alpha\,;\, 1 \leq \alpha \leq N\,. \tag{6.3.8}$$

L'objectif de la condition (6.3.8) est d'augmenter encore la précision bien que cela ne soit pas nécessaire pour l'estimation de l'erreur de troncature (6.3.6). Il faut garder à l'esprit que la fonction $w_\alpha[\mathfrak{u}]$ n'est utilisée qu'à l'intérieur de la cellule $\mathcal{T}_\alpha$. Ceci justifie de renforcer la précision de la reconstruction, particulièrement dans cette cellule, ce que fait la condition (6.3.8). D'ailleurs, la section 7.5 montre que le respect de (6.3.8) simplifie la forme de la reconstruction.

6.4. Bilan du chapitre

Ce chapitre rappelle les principes de la méthode des volumes finis. La section 6.2 présente un résumé de la méthode introduite par Godounov en mettant l'accent sur les points suivants :

(1) Les variables élémentaires sont les moyennes de cellule.

(2) L'introduction des flux numériques permet de déduire les équations du schéma volumes finis directement des équations de bilan de la loi de conservation.

(3) Les propriétés élémentaires des flux numériques sont la consistance (6.2.12), la conservativité (6.2.14) et la continuité au sens de Lipschitz (6.2.13). Une autre propriété importante est la monotonie, introduite par la définition 6.2.3, qui joue un rôle important pour la stabilité et la suppression des oscillations.

La section 6.3 introduit la notion de reconstruction locale qui permet d'augmenter la précision des schémas volumes finis. Cette étape est indispensable car la précision du schéma volumes finis sans reconstruction est trop faible pour les applications réalistes. Le point le plus important est que la reconstruction locale doit satisfaire une propriété d'approximation du type (6.3.3)

$$|w_\alpha[\mathfrak{u}(t_0)](\boldsymbol{x}) - u(\boldsymbol{x}, t_0)| = \mathrm{O}\left(h^{k+1}\right)\,,\, \boldsymbol{x} \in \mathcal{T}_\alpha$$

qui permet d'obtenir une erreur de troncature d'ordre $\mathrm{O}\left(h^k\right)$.

Pour tenir compte de la conservativité, on introduit par ailleurs la notion de reconstruction conservative (6.3.8) qui impose que la fonction reconstruite doit préserver la moyenne sur la cellule, c'est-à-dire

$$\frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha} w_\alpha[\mathfrak{u}](\boldsymbol{x})\, dx = u_\alpha\,;\, 1 \leq \alpha \leq N\,.$$

Il s'agit d'une restriction supplémentaire qui n'est pas nécessaire, dans un sens strict, pour augmenter l'ordre de l'erreur de troncature.

CHAPITRE 7

Étude de la reconstruction locale en maillage non structuré

7.1. Objectif du chapitre

Le chapitre 6 sur la discrétisation spatiale introduit la notion de reconstruction et met en évidence la façon dont celle-ci influe sur l'erreur de troncature du schéma volumes finis (6.3.4). L'objectif du présent chapitre est d'explorer de manière générale des méthodes de reconstruction permettant d'atteindre des erreurs de troncature d'ordre élevé. La question importante de l'implémentation est réservée au chapitre 8 qui est dédié à la recherche d'algorithmes particulièrement adaptés aux grands calculs sur des ordinateurs parallèles et vectoriels.

Une méthode de reconstruction locale détermine dans chaque cellule une fonction à partir des moyennes de cellule au voisinage. Cette fonction sert à interpoler des valeurs aux interfaces entre la cellule et les cellules adjacentes qui sont plus précises que les simples moyennes de cellule. La discussion du chapitre 6 a montré comment les flux numériques calculés à l'aide de ces interpolations permettent d'obtenir une erreur de troncature d'ordre plus élevé, c'est-à-dire une erreur qui diminue plus rapidement avec le diamètre du maillage.

L'opération de reconstruction étudiée ici ressemble à un problème d'interpolation car elle consiste à déterminer une fonction à partir d'un nombre fini de valeurs ponctuelles ou de moyennes de cellule. Cependant, dans le cadre de cette thèse, on préfère le terme reconstruction pour des raisons données dans la remarque 2.8.1 de la section 2.8.

La discussion du chapitre 6 précise la condition pour qu'une reconstruction augmente l'ordre du schéma volumes finis : la reconstruction doit reproduire, à partir des moyennes de n'importe quelle fonction suffisamment régulière, une nouvelle fonction qui approche la fonction initiale avec une erreur bornée par une certaine puissance du diamètre maximal du maillage. L'exposant de cette puissance est appelé *l'ordre de la reconstruction*. Le chapitre 6 introduit également la notion de *reconstruction conservative* : la fonction reconstruite doit avoir la même moyenne que la fonction initiale qu'on veut approcher.

Ces rappels permettent de formuler l'objectif principal du présent chapitre : il s'agit de définir et de caractériser la famille la plus large de méthodes de reconstruction satisfaisant les critères cités ci-dessus. Les méthodes de reconstruction se caractérisent par trois aspects :

(1) Il est d'abord nécessaire de déterminer la façon générale dont la fonction reconstruite dépend des moyennes de cellule. La reconstruction locale tente de reconstruire une fonction à partir de la seule connaissance de ses moyennes. Elle est donc une opération de type inverse approchée de l'opération de moyenne. Du fait que l'opération de moyenne sur les cellules est linéaire, il est naturel de choisir pour l'opération inverse de reconstruction locale une dépendance linéaire.

(2) Deuxièmement, les fonctions reconstruites doivent faire partie d'un espace de fonctions de dimension fini afin qu'un algorithme de reconstruction locale puisse être programmé sur ordinateur. Des espaces de fonctions naturels sont les espaces de polynômes de degré un, deux et trois. Les polynômes ont des propriétés d'approximation bien connues et une erreur d'approximation qu'on peut analyser par un développement de Taylor. De plus, chaque fonction dans un tel espace est identifiée par le même nombre fini de coefficients réels. Dans chaque cellule, il y a donc le même nombre de paramètres réels à calculer et à stocker ce qui est indispensable pour la vectorisation du traitement.

(3) Le dernier aspect concerne le choix d'un voisinage de reconstruction pour chaque cellule, c'est-à-dire la sélection de l'ensemble des cellules voisines qui sont prises en compte pour la reconstruction. Concernant ce point, il est nécessaire de faire le choix entre des

méthodes utilisant des voisinages fixes et des méthodes à voisinage variable. Pour ces dernières, le voisinage de reconstruction d'une cellule peut changer en fonction de la solution au cours du calcul. Pour des raisons précisées plus loin, cette étude est restreinte à l'utilisation de voisinages fixes qui sont indépendants de la solution. Par contre, ceci n'exclut pas des voisinages qui évoluent en raison d'un changement de topologie du maillage.

Les choix ci-dessus ont pour conséquence qu'une reconstruction locale est représentée par un ensemble de matrices. Plus précisément, à chaque cellule est attachée une matrice qui envoie linéairement les moyennes de cellule sur les coefficients du polynôme reconstruit. L'étape suivante est d'identifier les matrices qui engendrent des reconstructions locales menant à des erreurs de troncature d'ordre deux, trois et quatre. Pour cela, il sera nécessaire de traduire le critère d'approximation présenté au chapitre 6 en une équation algébrique pour les matrices de reconstruction.

Un moyen pour atteindre cet objectif est de prouver que la reconstruction satisfait le critère d'approximation du chapitre 6 si elle reproduit tout polynôme d'un certain degré sur un voisinage donné. Une telle condition est appelée *condition de consistance*. L'avantage de cette approche est qu'une condition de consistance s'exprime de façon naturelle par une équation algébrique pour les éléments de la matrice. L'équation algébrique de reconstruction peut ensuite être résolue cellule par cellule sur ordinateur et ainsi donner la matrice de reconstruction dans chaque cellule.

Dans la suite, la condition de consistance sera aussi appelée *critère de consistance* et l'équation algébrique qui en découle une *équation de consistance*. Le terme de *reproduction polynomiale* n'est pas utilisé ici car il est déjà employé dans les références [**121**] et [**120**] pour désigner une notion différente. Il est important de noter que le critère de consistance correspond à la notion de reconstruction k-exacte, introduite par Barth et Frederickson dans [**9**].

L'étape suivante est d'examiner en détail les propriétés de l'équation de consistance afin de caractériser les reconstructions polynomiales de façon générale. Il faut en particulier analyser les conditions dans lesquelles cette équation a une solution et déterminer le nombre minimal de cellules voisines nécessaires à la reconstruction d'un polynôme d'un certain degré. L'influence de la taille du voisinage sur la stabilité et la précision sera traitée dans les chapitres 10 et 11. Un autre but de cette étude est de comparer les aspects de la reconstruction consistante à partir de moyennes de cellule et à partir de valeurs ponctuelles.

Finalement, il est nécessaire de tester sur ordinateur les méthodes numériques développées dans ce chapitre. La reconstruction, telle qu'elle est étudiée ici, est une opération linéaire qui nécessite d'inverser des matrices dépendant d'un maillage non structuré. Le objectif principal des tests est donc de vérifier l'inversibilité et le conditionnement des matrices de reconstruction sur des maillages non structurés.

7.2. Définition de reconstructions précises par une reproduction de polynômes

7.2.1. Définition et propriétés élémentaires de la reconstruction. Comme expliqué dans la section 7.1, l'objectif de ce chapitre est d'explorer des familles générales de reconstructions reproduisant les polynômes de degré k. L'introduction de la notion de reconstruction dans la section 6.3 est motivée par le besoin d'augmenter la précision des méthodes de type volumes finis. Pour cela, la reconstruction doit satisfaire la propriété d'approximation suivante : pour toute fonction suffisamment régulière u, il existe une constante C_u, indépendante de h et de la cellule $\mathcal{T}_\alpha$, telle que l'estimation (6.3.3)

$$|w_\alpha\left[\mathfrak{u}\left(t_0\right)\right]\left(\boldsymbol{x}\right) - u\left(\boldsymbol{x}, t_0\right)| \leq \mathrm{C}_u\, h^{k+1} = \mathrm{O}\left(h^{k+1}\right) \text{ pour tout } \boldsymbol{x} \in \mathcal{T}_\alpha \text{ et } 1 \leq \alpha \leq N\,,$$

soit vraie.

Il est cependant évident que le critère de précision dans sa forme (6.3.3) ne se prête pas à la conception d'algorithmes informatiques. C'est pourquoi l'objectif de ce chapitre est de transformer le critère (6.3.3) en une relation algébrique pour les coefficients de la reconstruction locale.

La discussion préliminaire de la section 7.1 justifie certaines restrictions sur la forme de la reconstruction.

(1) D'abord, il n'est pas nécessaire de reconstruire une fonction globale sur tout le domaine de calcul Ω. Au contraire, il suffit de reconstruire localement dans chaque cellule $\mathcal{T}_\alpha$ une fonction w_α. La section 6.3 montre que les valeurs de cette fonction sont exclusivement utilisées sur les faces de la cellule $\mathcal{T}_\alpha$. Pour ces raisons, ce type de reconstruction est appelé *reconstruction locale* dans la suite du document.

(2) Puisqu'il est suffisant de respecter le critère (6.3.3) localement sur les faces de la cellule $\mathcal{T}_\alpha$ et pour des raisons d'efficacité de calcul, la reconstruction d'une fonction w_α dans la cellule $\mathcal{T}_\alpha$ n'utilise que les informations d'un nombre restreint de cellules au voisinage de $\mathcal{T}_\alpha$. L'ensemble des cellules qui contribuent à la reconstruction dans la cellule $\mathcal{T}_\alpha$ s'appelle *voisinage de reconstruction* de $\mathcal{T}_\alpha$. Dans la littérature, les voisinages de reconstruction s'appellent aussi *molécules de reconstruction* ou *stencils* en anglais. L'ensemble des numéros des cellules dans le voisinage de reconstruction de $\mathcal{T}_\alpha$ est désigné par

$$\mathbb{W}_\alpha \triangleq \{\, \beta |\, \mathcal{T}_\beta \text{ contribue à la reconstruction dans } \mathcal{T}_\alpha \,,\, \beta \neq \alpha \} \,. \tag{7.2.1}$$

Le nombre de cellules dans $\mathbb{W}_\alpha$ est noté

$$m_\alpha \triangleq |\mathbb{W}_\alpha| \,. \tag{7.2.2}$$

Comme la cellule $\mathcal{T}_\alpha$ contribue en général elle-même à la reconstruction, on définit le *voisinage de reconstruction augmenté* $\widehat{\mathbb{W}}_\alpha$ comme

$$\widehat{\mathbb{W}}_\alpha \triangleq \mathbb{W}_\alpha \cup \{\alpha\} \,. \tag{7.2.3}$$

(3) Les fonctions reconstruites dépendent linéairement des moyennes de cellule dans le voisinage de reconstruction. Le même principe vaut pour la reconstruction à partir de valeurs ponctuelles.

(4) Les fonctions reconstruites font partie d'un espace de polynômes de dimension finie. En général, il s'agit de l'espace $\mathbb{P}_k\left(\mathbb{R}^d\right)$ des polynômes réels de degré k.

Les quatre points ci-dessus permettent de considérer la reconstruction dans la cellule $\mathcal{T}_\alpha$ comme une application linéaire $\mathfrak{R}_\alpha$

$$\mathfrak{R}_\alpha : \left\{ \begin{array}{ccc} \mathbb{R}^N & \to & \mathbb{P}_k\left(\mathbb{R}^d\right) \\ \mathfrak{u} & \mapsto & w_\alpha\left[\mathfrak{u}\right] \end{array} \right. \,. \tag{7.2.4}$$

qui associe au vecteur $\mathfrak{u} \triangleq (\overline{u}_1, \ldots, \overline{u}_N)$ des moyennes de cellule d'une fonction u le polynôme $w_\alpha\left[\mathfrak{u}\right] \in \mathbb{P}_k\left(\mathbb{R}^d\right)$. Ce polynôme $w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x})$ doit être une approximation de la fonction $u\,(\boldsymbol{x})$. Comme expliqué dans le point 2 ci-dessus, les valeurs de l'opérateur $\mathfrak{R}_\alpha$ ne dépendent que des moyennes de cellule dans le voisinage $\widehat{\mathbb{W}}_\alpha$ de la cellule $\mathcal{T}_\alpha$. Cependant, pour simplifier la notation, l'opérateur (7.2.4) est défini sur $\mathbb{R}^N$ entier avec la convention que $\mathfrak{R}_\alpha\,(\overline{u}_1, \ldots, \overline{u}_N) = \mathfrak{R}_\alpha\,(\overline{v}_1, \ldots, \overline{v}_N)$ si $\overline{u}_\gamma = \overline{v}_\gamma$ pour tout $\gamma \in \widehat{\mathbb{W}}_\alpha$. Cette convention évite la mention explicite des voisinages $\widehat{\mathbb{W}}_\alpha$.

La linéarité de la reconstruction $\mathfrak{R}_\alpha$ et la forme polynomiale des fonctions reconstruites entraînent que la fonction $w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x})$ a la forme

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}\,(\boldsymbol{x})\ \overline{u}_\beta \tag{7.2.5}$$

où les fonctions $w^{\alpha\beta}$ sont des polynômes réels de degré k en $\boldsymbol{x}$. La convention de notation pour $\mathfrak{R}_\alpha$ entraîne que, par définition,

$$w^{\alpha\beta} \triangleq 0 \text{ si } \beta \notin \widehat{\mathbb{W}}_\alpha \,. \tag{7.2.6}$$

La définition (7.2.6) permet d'omettre la mention des voisinages $\widehat{\mathbb{W}}_\alpha$ dans la somme (7.2.5).

Si la reconstruction se fait à partir de valeurs ponctuelles en $\boldsymbol{x}_\beta$, $\beta \in \widehat{\mathbb{W}}_\alpha$, la fonction $w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x})$ a la forme générale

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}\,(\boldsymbol{x})\ u\,(\boldsymbol{x}_\beta) \tag{7.2.7}$$

où les fonctions $w^{\alpha\beta}$ sont des polynômes réels de degré k en $\boldsymbol{x}$. Dans le cadre de cette thèse, on privilégie cependant la reconstruction à partir de valeurs moyennes.

7.2.2. Définition et existence de reconstructions consistantes. L'objectif de cette section est d'examiner sous quelles conditions un opérateur de reconstruction (7.2.4) de la forme (7.2.5) ou (7.2.7) reproduit tous les polynômes de degré k donné. On commence par introduire la notion de reconstruction consistante de degré k qui coïncide avec la notion d'opérateur k-exact utilisée par Barth et Frederickson dans [**9**].

Définition 7.2.1 (Reconstruction consistante de degré k). Soit $\mathfrak{R}_\alpha$ un opérateur de reconstruction du type (7.2.4) qui reconstruit les polynômes de degré k à partir de moyennes de cellule. L'application $\mathfrak{R}_\alpha$ et la fonction reconstruite

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}(\boldsymbol{x})\ \overline{u}_\beta$$

s'appellent *consistantes de degré k* si la condition suivante est satisfaite : pour tout polynôme p de degré k sur le voisinage de reconstruction, on a l'identité

$$w_\alpha\left[\mathfrak{p}\right](\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}(\boldsymbol{x})\ \overline{p}_\beta = p(\boldsymbol{x})$$

où $\mathfrak{p} \triangleq (\overline{p}_1, \ldots, \overline{p}_N)$ sont les moyennes de cellule de p.

Pour la reconstruction à partir de valeurs ponctuelles, la définition de consistance est analogue.□

Pour la reconstruction à partir de moyennes de cellule, il est utile de définir dans chaque cellule $\mathcal{T}_\alpha$ un opérateur de moyenne $\mathfrak{A}_\alpha$ qui calcule les moyennes d'un polynôme p sur les cellules voisines de $\mathcal{T}_\alpha$

$$\mathfrak{A}_\alpha : \mathbb{P}_k\left(\mathbb{R}^d\right) \to \mathbb{R}^N .$$

Pour éviter la mention explicite des voisinages, on définit les valeurs de $\mathfrak{A}_\alpha$ par

$$\left(\mathfrak{A}_\alpha p\right)_\beta \triangleq \begin{cases} \overline{p}_\beta & \text{si } \beta \in \widehat{\mathbb{W}}_\alpha \\ 0 & \text{sinon.} \end{cases} \tag{7.2.8}$$

Dans (7.2.8), $\overline{p}_\beta$ est la moyenne du polynôme p sur la cellule $\mathcal{T}_\beta$. Pour la reconstruction à partir de valeurs ponctuelles aux points $\boldsymbol{x}_\beta$ pour $\beta \in \widehat{\mathbb{W}}_\alpha$, on définit de la même façon un opérateur d'évaluation

$$\mathfrak{D}_\alpha : \mathbb{P}_k\left(\mathbb{R}^d\right) \to \mathbb{R}^N$$

avec les valeurs

$$\left(\mathfrak{D}_\alpha p\right)_\beta \triangleq \begin{cases} p(\boldsymbol{x}_\beta) & \text{si } \beta \in \widehat{\mathbb{W}}_\alpha \\ 0 & \text{sinon.} \end{cases} \tag{7.2.9}$$

Avec la définition (7.2.8), la reproduction des polynômes de degré k s'exprime par la condition que $\mathfrak{R}_\alpha$ est un inverse à gauche de $\mathfrak{A}_\alpha$ sur l'espace des polynômes $\mathbb{P}_k\left(\mathbb{R}^d\right)$

$$\mathfrak{R}_\alpha \circ \mathfrak{A}_\alpha = \mathrm{id}_{\mathbb{P}_k\left(\mathbb{R}^d\right)} . \tag{7.2.10}$$

La définition (7.2.9) permet de formuler une condition analogue pour la reconstruction à partir de valeurs ponctuelles

$$\mathfrak{R}_\alpha \circ \mathfrak{D}_\alpha = \mathrm{id}_{\mathbb{P}_k\left(\mathbb{R}^d\right)} . \tag{7.2.11}$$

Il suffit maintenant de choisir une base de l'espace $\mathbb{P}_k\left(\mathbb{R}^d\right)$ pour transcrire les conditions (7.2.10) et (7.2.11) en équations matricielles. Cela fera l'objet de la section 7.3. Avant d'aborder ces questions, il faut étudier l'existence de reconstructions consistantes, c'est-à-dire l'existence de solutions $\mathfrak{R}_\alpha$ de (7.2.10) et de (7.2.11).

Autrement dit, est-il possible de reconstruire de façon unique tout polynôme de degré k à partir de ses moyennes de cellule ou de ses valeurs ponctuelles sur un voisinage de $\mathcal{T}_\alpha$?

Une condition nécessaire pour l'existence de solutions $\mathfrak{R}_\alpha$ de (7.2.10) est que les moyennes de cellule dans le voisinage $\widehat{\mathbb{W}}_\alpha$ de $\mathcal{T}_\alpha$ permettent de distinguer les polynômes réels de degré k. Ceci revient à exiger que pour deux polynômes p et q de degré k on ait

$$p \neq q \text{ entraîne } \mathfrak{A}_\alpha p \neq \mathfrak{A}_\alpha q \tag{7.2.12}$$

ou de façon équivalente

$$\mathfrak{A}_\alpha p = \mathfrak{A}_\alpha q \text{ entraîne } p = q\,, \tag{7.2.13}$$

c'est-à-dire que l'opérateur de moyenne $\mathfrak{A}_\alpha$ soit injectif sur $\mathbb{P}_k\left(\mathbb{R}^d\right)$. On dit aussi que $\mathfrak{A}_\alpha$ est unisolvant. Dans le cas contraire, il y aurait deux polynômes distincts ayant les mêmes moyennes sur le voisinage de $\mathcal{T}_\alpha$ et la reconstruction consistante deviendrait impossible. En raison de la linéarité de $\mathfrak{A}_\alpha$, $\mathfrak{A}_\alpha$ est injectif si le seul polynôme p de degré k qui satisfait $\overline{p}_\beta = 0$ pour tout $\beta \in \widehat{\mathbb{W}}_\alpha$ est le polynôme nul $p = 0$. L'injectivité de $\mathfrak{A}_\alpha$ est également une *condition suffisante* pour l'existence de $\mathfrak{R}_\alpha$. Cela est démontré par le résultat suivant.

Proposition 7.2.2 (Existence d'un inverse à gauche). *Une application linéaire $f : \mathbb{R}^n \to \mathbb{R}^m$ admet un inverse à gauche, c'est-à-dire une application linéaire $g : \mathbb{R}^m \to \mathbb{R}^n$ telle que*

$$g \circ f = \mathrm{id}_{\mathbb{R}^n}\,,$$

si et seulement si f est injective.

Démonstration.

(i) Supposons f injective. Un inverse à gauche g se définit sans ambiguïté sur le sous-espace $\mathrm{Im}\,(f) \subset \mathbb{R}^n$ par $g\,(f\,(\boldsymbol{x})) = \boldsymbol{x}$ car l'injectivité de f entraîne que $f\,(\boldsymbol{x}) = f\,(\boldsymbol{y})$ si et seulement si $\boldsymbol{x} = \boldsymbol{y}$. Il suffit alors de choisir un sous-espace $\mathcal{W}$ supplémentaire de $\mathrm{Im}\,(f)$ dans $\mathbb{R}^m$ et de définir $g\,(\boldsymbol{u}) = 0$ si $\boldsymbol{u} \in \mathcal{W}$. Tout $\boldsymbol{v} \in \mathbb{R}^m$ possède une décomposition unique $\boldsymbol{v} = f\,(\boldsymbol{x}) + \boldsymbol{u}$ avec $\boldsymbol{x} \in \mathbb{R}^n$ et $\boldsymbol{u} \in \mathcal{W}$. Pour cette raison, la valeur de g sur $\boldsymbol{v}$ peut être définie de façon univoque par

$$g\,(\boldsymbol{v}) \triangleq g\,(f\,(\boldsymbol{x})) + g\,(\boldsymbol{u}) = \boldsymbol{x}$$

pour tout $\boldsymbol{v} \in \mathbb{R}^m$. L'application g ainsi définie satisfait $g \circ f = \mathrm{id}_{\mathbb{R}^n}$. La linéarité de f entraîne la linéarité de g car pour $\boldsymbol{v}' = f\,(\boldsymbol{x}') + \boldsymbol{u}'$ et $\boldsymbol{v}'' = f\,(\boldsymbol{x}'') + \boldsymbol{u}''$ il vient

$$\boldsymbol{v}' + \boldsymbol{v}'' = f\left(\boldsymbol{x}'\right) + \boldsymbol{u}' + f\left(\boldsymbol{x}''\right) + \boldsymbol{u}'' = f\left(\boldsymbol{x}' + \boldsymbol{x}''\right) + \boldsymbol{u}' + \boldsymbol{u}''$$

ce qui donne

$$g\left(\boldsymbol{v}' + \boldsymbol{v}''\right) = g\left(f\left(\boldsymbol{x}' + \boldsymbol{x}''\right)\right) + g\left(\boldsymbol{u}' + \boldsymbol{u}''\right) = \boldsymbol{x}' + \boldsymbol{x}''\,.$$

Pour $\lambda \in \mathbb{R}$ et $\boldsymbol{v} = f\,(\boldsymbol{x}) + \boldsymbol{u}$ on a $\lambda\boldsymbol{v} = f\,(\lambda\boldsymbol{x}) + \lambda\boldsymbol{u}$ et par conséquent

$$g\,(\lambda\boldsymbol{v}) = g\,(f\,(\lambda\boldsymbol{x})) + g\,(\lambda\boldsymbol{u}) = \lambda\boldsymbol{x}\,.$$

(ii) Supposons que f admet un inverse à gauche g. Dans ce cas,

$$f\,(\boldsymbol{x}) = f\,(\boldsymbol{y}) \text{ entraîne } \boldsymbol{x} = g\,(f\,(\boldsymbol{x})) = g\,(f\,(\boldsymbol{y})) = \boldsymbol{y}\,,$$

ce qui prouve l'injectivité de f.

□

La proposition 7.2.2 suggère de chercher des conditions pour l'injectivité de $\mathfrak{A}_\alpha$. D'après (7.2.8), $\mathfrak{A}_\alpha$ est une application linéaire définie sur $\mathbb{P}_k\left(\mathbb{R}^d\right)$ qui prend ses valeurs dans un sous-espace de dimension $m_\alpha + 1$ de $\mathbb{R}^N$. Pour que $\mathfrak{A}_\alpha$ puisse être injectif, il faut que la dimension de $\mathbb{P}_k\left(\mathbb{R}^d\right)$ soit plus petite que la dimension de l'espace image, c'est-à-dire

$$m_\alpha + 1 \geq= \dim\left(\mathbb{P}_k\left(\mathbb{R}^d\right)\right) = \binom{k+d}{d}\,. \tag{7.2.14}$$

La condition (7.2.14) se réduit pour $d = 2$ à

$$m_\alpha + 1 \geq \frac{1}{2}\,(k+2)\,(k+1) \tag{7.2.15}$$

et pour $d = 3$, à

$$m_\alpha + 1 \geq \frac{1}{6}\,(k+3)\,(k+2)\,(k+1) \tag{7.2.16}$$

ce qui montre qu'en dimension trois, la taille minimale du voisinage croît de façon cubique avec le degré des polynômes. Cette augmentation du nombre des voisins constitue un handicap bien connu des méthodes classiques des volumes finis. Le présent travail tente de contourner ce problème par le développement de méthodes itératives de reconstruction dans le chapitre 8.

La condition (7.2.14) n'est en général pas suffisante pour que $\mathfrak{A}_\alpha$ soit injectif. Cependant, dans un certain nombre de cas, on dispose de conditions suffisantes simples pour la reconstruction à partir de valeurs ponctuelles. Par exemple, la reconstruction d'un polynôme p de degré un dans $\mathbb{R}^2$ nécessite les valeurs de p en trois points non colinéaires. De façon analogue, dans $\mathbb{R}^3$ il faut les valeurs de p en quatre points non coplanaires. Des généralisations de ces règles aux polynômes de degré plus élevé existent, voir par exemple [**120**, chapitre 1] pour $\mathbb{R}^2$. Les calculs numériques effectués dans le cadre de cette étude montrent que les maillages non structurés utilisés admettent une reconstruction polynomiale jusqu'au degré trois si la condition nécessaire (7.2.14) est satisfaite. Dans la pratique, l'existence de solutions $\mathfrak{R}_\alpha$ pour (7.2.10) et (7.2.11) ne pose donc pas de problème majeur. Dans la section 7.2.3, on montre que la réelle difficulté consiste à contrôler la norme de $\mathfrak{R}_\alpha$ lorsque h tend vers zéro.

7.2.3. Influence de la condition de consistance sur la précision de la reconstruction. Dans cette section, on étudie l'influence de $\mathfrak{R}_\alpha$ sur la précision de la reconstruction. On commence par la

Définition 7.2.3 (Ordre de la reconstruction). La reconstruction $\mathfrak{R}_\alpha$ définie par (7.2.4) est appelée une *reconstruction d'ordre* $k+1$ si elle satisfait la condition suivante : pour toute fonction suffisamment régulière u, il doit exister une constante C_u, indépendante de h et de la cellule $\mathcal{T}_\alpha$, telle que l'estimation (6.3.3)

$$|w_\alpha[\mathfrak{u}](\boldsymbol{x}) - u(\boldsymbol{x})| \leq \mathrm{C}_u\, h^{k+1} = \mathrm{O}\left(h^{k+1}\right) \text{ pour tout } \boldsymbol{x} \in \mathcal{T}_\alpha$$

soit vraie dans toutes les cellules $\mathcal{T}_\alpha$ du maillage. Le vecteur $\mathfrak{u}$ est soit le vecteur des moyennes de cellule $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$ soit le vecteur des valeurs ponctuelles $\mathfrak{u} = (u(\boldsymbol{x}_1), \ldots, u(\boldsymbol{x}_N))$.□

L'objectif de cette section est d'analyser les reconstructions qui reproduisent les polynômes de degré k et de voir sous quelles conditions elles satisfont une estimation de type (6.3.3). Cette étude nécessite la formule de Taylor avec reste intégral pour les fonctions réelles.

Soit $u : \mathcal{O} \to \mathbb{R}$ une fonction de classe $C^{k+1}(\mathcal{O})$ sur un voisinage ouvert $\mathcal{O} \subseteq \mathbb{R}^d$ de $\boldsymbol{x}_\alpha$. Soit $\boldsymbol{x} \in \mathbb{R}^d$ un point tel que $[\boldsymbol{x}_\alpha, \boldsymbol{x}] \subset \mathcal{O}$, alors u se décompose en un polynôme $q_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})$ de degré k et un reste $r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})$

$$u(\boldsymbol{x}) = q_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x}) + r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})\ . \tag{7.2.17}$$

Dans la suite, on suppose toujours que le voisinage de reconstruction de $\mathcal{T}_\alpha$ est contenu dans $\mathcal{O}$ et que la dérivée d'ordre $k+1$ de u est bornée sur $\mathcal{O}$. Les conventions de notation de la section 5.2 permettent d'écrire le polynôme $q_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})$ comme

$$q_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x}) = \sum_{j=0}^{k} \frac{1}{j!}\, D^{(j)}u\Big|_{\boldsymbol{x}_\alpha} \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)^j\ .$$

L'expression $r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})$ est le reste intégral

$$r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x}) = \int_0^1 \frac{(1-s)^k}{k!}\, D^{(k+1)}u\Big|_{\boldsymbol{x}_\alpha + s(\boldsymbol{x}-\boldsymbol{x}_\alpha)} \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)^{k+1}\, ds\,.$$

Dans les applications de calcul, le domaine Ω et, par conséquent, le vecteur $\boldsymbol{x} - \boldsymbol{x}_\alpha$ sont toujours bornés. De ce fait et en raison de l'hypothèse sur u, il existe une constante C_{r_u} telle que

$$|r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})| \leq \mathrm{C}_{r_u} \|\boldsymbol{x} - \boldsymbol{x}_\alpha\|^{k+1}\ .$$

Pour le besoin de l'estimation, le point $\boldsymbol{x}$ fait toujours partie du voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$. On suppose que le nombre de cellules dans le voisinage de reconstruction reste

borné lorsque h tend vers zéro. Il existe donc une constante $m_{\max}$ qui ne dépend pas de h et telle que

$$m_\alpha = \left|\widehat{\mathbb{W}}_\alpha\right| \leq m_{\max}$$

pour *toutes les cellules* $\mathcal{T}_\alpha$ du maillage lorsque $h \longrightarrow 0$. Cela implique que

$$\|\boldsymbol{x} - \boldsymbol{x}_\alpha\|^{k+1} \leq (m_{\max})^{k+1} h^{k+1} = \mathrm{O}\left(h^{k+1}\right) \tag{7.2.18}$$

pour $\boldsymbol{x}$ dans le voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$. L'estimation (7.2.18) permet de conclure que

$$|r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})| \leq \mathrm{C}_{r_u} (m_{\max})^{k+1} h^{k+1} = \mathrm{O}\left(h^{k+1}\right) \tag{7.2.19}$$

pour $\boldsymbol{x}$ dans le voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$. La même estimation demeure valable pour la moyenne de $r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})$ sur toute cellule $\mathcal{T}_\beta$ dans le voisinage de reconstruction de $\mathcal{T}_\alpha$ car

$$\left|\frac{1}{|\mathcal{T}_\beta|}\int_{\mathcal{T}_\beta} r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})\, dx\right| \leq \frac{1}{|\mathcal{T}_\beta|}\int_{\mathcal{T}_\beta} |r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x})|\, dx \leq$$

$$\leq \mathrm{C}_{r_u} (m_{\max})^{k+1} h^{k+1} \frac{1}{|\mathcal{T}_\beta|}\int_{\mathcal{T}_\beta} dx = \mathrm{O}\left(h^{k+1}\right). \tag{7.2.20}$$

Les moyennes de u se décomposent de façon linéaire

$$\overline{u}_\beta = \overline{(q_{u,\boldsymbol{x}_\alpha})}_\beta + \overline{(r_{u,\boldsymbol{x}_\alpha})}_\beta .$$

La reconstruction à partir de ces valeurs est

$$w_\alpha[\mathrm{u}](\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}(\boldsymbol{x}) \left[\overline{(q_{u,\boldsymbol{x}_\alpha})}_\beta + \overline{(r_{u,\boldsymbol{x}_\alpha})}_\beta\right]. \tag{7.2.21}$$

Le premier terme dans (7.2.21) est égal à $q_{u,\boldsymbol{x}_\alpha}$ en raison de la condition de consistance de degré k

$$q_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}(\boldsymbol{x})\, \overline{(q_{u,\boldsymbol{x}_\alpha})}_\beta .$$

Pour cette raison, la différence entre (7.2.21) et (7.2.17) est

$$w_\alpha[\mathrm{u}](\boldsymbol{x}) - u(\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}(\boldsymbol{x})\, \overline{(r_{u,\boldsymbol{x}_\alpha})}_\beta - r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x}) . \tag{7.2.22}$$

Pour la reconstruction à partir de valeurs ponctuelles, on obtient de façon analogue

$$w_\alpha[\mathrm{u}](\boldsymbol{x}) - u(\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}(\boldsymbol{x})\, r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x}_\beta) - r_{u,\boldsymbol{x}_\alpha}(\boldsymbol{x}) . \tag{7.2.23}$$

La forme des équations (7.2.22) et (7.2.23) ne permet pas encore de déterminer l'ordre de la reconstruction. On introduit d'abord la

Définition 7.2.4 (Reconstructions régulières). Une reconstruction $\mathfrak{R}_\alpha$ de la forme (7.2.5) ou (7.2.7) est dite *régulière* s'il existe une constante $\mathrm{C}_{\mathrm{reg}}$, indépendante de h et de la cellule $\mathcal{T}_\alpha$, telle que les fonctions coefficients $w^{\alpha\beta}$ satisfont l'estimation

$$\left|w^{\alpha\beta}(\boldsymbol{x})\right| \leq \mathrm{C}_{\mathrm{reg}} \text{ pour tout } \boldsymbol{x} \in \mathcal{T}_\alpha \tag{7.2.24}$$

dans toutes les cellules $\mathcal{T}_\alpha$ du maillage lorsque h tend vers zéro.□

Les définitions 7.2.3 et 7.2.4 permettent de formuler le

Théorème 7.2.5 (Condition de consistance et précision). *Une reconstruction $\mathfrak{R}_\alpha$ de la forme (7.2.5) ou (7.2.7) est d'ordre $k+1$ au sens de la définition 7.2.3 si elle satisfait les deux conditions suivantes :*

(i) *Elle est régulière au sens de la définition 7.2.4.*

(ii) *Elle est consistante de degré k au sens de la définition 7.2.1, c'est-à-dire elle reproduit les polynômes de degré k.*

DÉMONSTRATION. Les estimations (7.2.19) et (7.2.20) montrent que les moyennes et valeurs ponctuelles de $r_{u,\boldsymbol{x}_\alpha}$ dans (7.2.22) sont $\mathrm{O}\left(h^{k+1}\right)$. La condition (7.2.24) permet donc de conclure que l'expression (7.2.22) est $\mathrm{O}\left(h^{k+1}\right)$. La même démonstration est valable pour la reconstruction à partir de valeurs ponctuelles. □

Il est important de voir que la condition (7.2.24) est en effet une condition pour la norme de l'opérateur $\mathfrak{R}_\alpha$ qui envoie le vecteur $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N) \in \mathbb{R}^N$ sur le polynôme (7.2.5)

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}(\boldsymbol{x})\ \overline{u}_\beta\,.$$

Pour cela, il est utile de munir l'espace $\mathbb{R}^N$ de la norme

$$\|\mathfrak{u}\|_\infty \triangleq \sup_{1\leq\alpha\leq N} |u_\alpha|\ ,\ \mathfrak{u}\in\mathbb{R}^N \tag{7.2.25}$$

et de considérer le polynôme reconstruit (7.2.5) comme un élément de l'espace $\mathbb{P}_k(\mathcal{T}_\alpha)$ des polynômes de degré k sur $\mathcal{T}_\alpha$, muni de la norme

$$\|p\|_{\mathbb{P}_k(\mathcal{T}_\alpha)} = \sup_{\boldsymbol{x}\in\mathcal{T}_\alpha} |p(\boldsymbol{x})|\ . \tag{7.2.26}$$

Les deux normes (7.2.25) et (7.2.26) induisent une norme de l'opérateur $\mathfrak{R}_\alpha$ par la formule

$$\|\mathfrak{R}_\alpha\|_{\mathcal{L}(\mathbb{R}^N,\mathbb{P}_k(\mathcal{T}_\alpha))} = \sup_{\|\mathfrak{u}\|_\infty\leq 1}\sup_{\boldsymbol{x}\in\mathcal{T}_\alpha}\left|\sum_\beta w^{\alpha\beta}(\boldsymbol{x})\ \overline{u}_\beta\right|\ . \tag{7.2.27}$$

La forme de la norme (7.2.27) permet de voir que la condition (7.2.24) est satisfaite dans toutes les cellules $\mathcal{T}_\alpha$ si et seulement si la norme (7.2.27) de $\mathfrak{R}_\alpha$ reste bornée uniformément dans toutes les cellules $\mathcal{T}_\alpha$ lorsque $h \longrightarrow 0$. Ceci vaut aussi bien pour la reconstruction à partir de moyennes de cellule qu'à partir de valeurs ponctuelles. Il faut préciser que ce résultat ne dépend pas des normes spécifiques (7.2.25), (7.2.26) et (7.2.27) puisque $\mathbb{R}^N$ et $\mathbb{P}_k(\mathcal{T}_\alpha)$ sont des espaces de dimension finie sur lesquels toutes les normes sont équivalentes.

D'une manière générale, sur des maillages non structurés généraux, il est difficile de trouver des conditions suffisantes pour l'existence de $\mathfrak{R}_\alpha$ qui fournissent également une borne pour la norme de $\mathfrak{R}_\alpha$. Des résultats existent dans [**120**] pour la reconstruction à partir de valeurs ponctuelles et dans [**121**] pour la reconstruction à partir de moyennes de cellule. Ces résultats prouvent que des solutions $\mathfrak{R}_\alpha$ pour (7.2.10) et (7.2.11) existent et que la norme de $\mathfrak{R}_\alpha$ reste bornée lorsque les voisinages de reconstruction contiennent suffisamment de cellules. Par contre, ils ne permettent pas de borner la taille des voisinages de façon explicite, ce qui limite leur utilité pour la présente étude. Pour cette raison, on fait l'hypothèse suivante :

HYPOTHÈSE 7.2.6 (Régularité des reconstructions). *On fait, dans la suite, l'hypothèse que les reconstructions $\mathfrak{R}_\alpha$ considérées satisfont la condition (7.2.24) de la définition 7.2.4.*

La question de la régularité reviendra dans la section 7.3.4 lorsque que la base polynomiale pour la reconstruction aura été fixée.

7.3. Formulation explicite des conditions de consistance

Jusqu'à présent, les conditions de consistance ont été analysées de façon abstraite. L'objectif de cette section est de transformer (7.2.10) et (7.2.11) en équations matricielles afin d'implémenter concrètement des méthodes reproduisant les polynômes de degré k. Cela nécessite le choix d'une base de l'espace des polynômes $\mathbb{P}_k\left(\mathbb{R}^d\right)$. C'est l'objectif de la section suivante.

7.3.1. Choix d'une base polynomiale. Pour la reconstruction dans la cellule $\mathcal{T}_\alpha$, il est intéressant d'utiliser une base de l'espace des polynômes $\mathbb{P}_k\left(\mathbb{R}^d\right)$ centrée sur le barycentre $\boldsymbol{x}_\alpha$ de $\mathcal{T}_\alpha$. Une telle base peut être construite à partir de l'ensemble des polynômes

$$\widehat{p}_{i_1\cdots i_l}(\boldsymbol{x}) = (x_{i_1} - x_{\alpha,i_1})\cdots(x_{i_l} - x_{\alpha,i_l})\ ,\ 1\leq i_1,\ldots,i_l\leq d\,,\ 0\leq l\leq k\,. \tag{7.3.1}$$

Ces polynômes satisfont par définition

$$\widehat{p}_{i_1\cdots i_l}(\boldsymbol{x}_\alpha) = 0\,,\ 1 \le i_1, \ldots, i_l \le d\,,\ 1 \le l \le k.$$

Les fonctions (7.3.1) ne sont pas indépendantes en raison de leur symétrie par rapport aux permutations des indices $(i_1, \ldots, i_l)$. Soit π une permutation de l'ensemble $\{1, \ldots, l\}$, alors on a

$$\widehat{p}_{i_{\pi(1)}\cdots i_{\pi(l)}}(\boldsymbol{x}) = \widehat{p}_{i_1\cdots i_l}(\boldsymbol{x})\ .$$

Pour cette raison, il suffit de choisir parmi les polynômes (7.3.1) ceux pour lesquels les indices se trouvent dans l'ordre croissant $1 \le i_1 \le i_2 \le \ldots \le i_l \le d$. Une base de l'espace des polynômes $\mathbb{P}_k\left(\mathbb{R}^d\right)$ est alors donnée par l'ensemble des polynômes

$$\mathbb{B}_\alpha^{(k)} \triangleq \{\widehat{p}_{i_1\cdots i_l}(\boldsymbol{x}) |\, 1 \le i_1 \le i_2 \le \ldots \le i_l \le d\,,\ 0 \le l \le k\}\ . \tag{7.3.2}$$

Lorsqu'on développe les fonctions $w^{\alpha\beta}(\boldsymbol{x})$ de (7.2.5) dans la base (7.3.2), la somme doit être restreinte aux indices dans l'ordre croissant $1 \le i_1 \le i_2 \le \ldots \le i_l \le d$. Cela signifie que le développement de $w^{\alpha\beta}(\boldsymbol{x})$ s'écrit sous la forme

$$w^{\alpha\beta}(\boldsymbol{x}) = \sum_{l=0}^{k} \sum_{i_1=1}^{d} \sum_{i_2=i_1}^{d} \sum_{i_3=i_2}^{d} \cdots \sum_{i_l=i_{l-1}}^{d} \widehat{w}_{i_1\ldots i_l}^{\alpha\beta} \widehat{p}_{i_1\cdots i_l}(\boldsymbol{x}) \tag{7.3.3}$$

où les $\widehat{w}_{i_1\ldots i_l}^{\alpha\beta}$, $1 \le i_1 \le i_2 \le \ldots \le i_l \le d$, sont les coefficients de $w^{\alpha\beta}(\boldsymbol{x})$ dans la base $\mathbb{B}_\alpha^{(k)}$.

Pour simplifier l'écriture du développement de $w^{\alpha\beta}(\boldsymbol{x})$, on élargit la définition des coefficients $\widehat{w}_{i_1\ldots i_l}^{\alpha\beta}$ par symétrie : soit $(i_1, \ldots, i_l)$ un l-tuple d'indices et soit $\pi \in \mathfrak{S}_l$ une permutation telle que

$$i_{\pi(1)} \le \cdots \le i_{\pi(l)}\,,$$

alors on définit

$$\widehat{w}_{i_1\ldots i_l}^{\alpha\beta} \triangleq \widehat{w}_{i_{\pi(1)}\cdots i_{\pi(l)}}^{\alpha\beta}\ .$$

On introduit alors de nouveau coefficients symétriques

$$w_{i_1\ldots i_l}^{\alpha\beta} \triangleq \frac{l!}{\widehat{C}_{i_1\ldots i_l}} \widehat{w}_{i_1\ldots i_l}^{\alpha\beta}\,, \tag{7.3.4}$$

où les nombres $\widehat{C}_{i_1\ldots i_l}$ expriment le nombre des permutations distinctes du l-tuple $(i_1, \ldots, i_l)$.

Les coefficients (7.3.4) permettent de réécrire le développement (7.3.3) de $w^{\alpha\beta}(\boldsymbol{x})$ sous la forme plus commode

$$\begin{aligned} w^{\alpha\beta}(\boldsymbol{x}) = \sum_{l=0}^{k} \frac{1}{l!} \sum_{i_1=1}^{d} \cdots \sum_{i_l=1}^{d} w_{i_1\ldots i_l}^{\alpha\beta} \widehat{p}_{i_1\cdots i_l}(\boldsymbol{x}) = \\ = \sum_{l=0}^{k} \frac{1}{l!} \sum_{i_1=1}^{d} \cdots \sum_{i_l=1}^{d} w_{i_1\ldots i_l}^{\alpha\beta} (x_{i_1} - x_{\alpha,i_1}) \cdots (x_{i_l} - x_{\alpha,i_l})\ . \end{aligned} \tag{7.3.5}$$

La forme (7.3.5) du développement de $w^{\alpha\beta}(\boldsymbol{x})$ a l'avantage de ressembler à un polynôme de Taylor, ce qui simplifie, dans la suite, l'analyse des conditions de consistance.

Les coefficients $w_{i_1\ldots i_l}^{\alpha\beta}$ dans (7.3.5) forment les composantes de tenseurs symétriques $\boldsymbol{w}_{\alpha\beta}^{(l)}$ d'ordre l pour $0 \le l \le k$. La notation tensorielle de la section 5.2 permet d'écrire (7.3.5) sous la forme compacte

$$w^{\alpha\beta}(\boldsymbol{x}) = \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{w}_{\alpha\beta}^{(l)} \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)^l\ . \tag{7.3.6}$$

L'équation (7.3.6) montre que le tenseur $\boldsymbol{w}_{\alpha\beta}^{(l)}$ coïncide avec la dérivée d'ordre l de $w^{\alpha\beta}$ en $\boldsymbol{x}_\alpha$

$$\boldsymbol{w}_{\alpha\beta}^{(l)} = D^{(l)} w^{\alpha\beta}\Big|_{\boldsymbol{x}_\alpha}\ .$$

L'insertion de (7.3.6) dans (7.2.5) démontre la

PROPOSITION 7.3.1 (Reconstruction de polynômes dans la base (7.3.2)). *La forme générale des polynômes reconstruits dans la base (7.3.2) est*

$$w_\alpha\left[\mathrm{u}\right](\boldsymbol{x}) = \sum_\beta \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{w}_{\alpha\beta}^{(l)} \bullet (\boldsymbol{x}-\boldsymbol{x}_\alpha)^l \, \overline{u}_\beta \tag{7.3.7}$$

où les coefficients $\boldsymbol{w}_{\alpha\beta}^{(l)}$ sont des tenseurs symétriques d'ordre l dans $\mathbb{R}^d$.

7.3.2. Formulation des conditions de consistance pour la reconstruction à partir de moyennes de cellule. Les coefficients $\boldsymbol{w}_{\alpha\beta}^{(l)}$ doivent satisfaire la condition de consistance de degré k (7.2.10). La prochaine étape consiste donc à exprimer cette condition par des équations pour les inconnues $\boldsymbol{w}_{\alpha\beta}^{(l)}$. Tout polynôme p de degré k s'écrit comme

$$p(\boldsymbol{x}) = \sum_{i=0}^{k} \frac{1}{i!} D^{(i)}p\Big|_{\boldsymbol{x}_\alpha} \bullet (\boldsymbol{x}-\boldsymbol{x}_\alpha)^i \, .$$

La formule (5.8.4) permet d'exprimer la moyenne $\overline{p}_\beta$ de p sur la cellule $\mathcal{T}_\beta$ par

$$\overline{p}_\beta = \sum_{j=0}^{k} \frac{1}{j!} D^{(j)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(j)} \, .$$

L'insertion du polynôme p et de la fonction reconstruite (7.3.7) dans la condition de consistance (7.2.10) conduit à la relation

$$p(\boldsymbol{x}) = \sum_{i=0}^{k} \frac{1}{i!} D^{(i)}p\Big|_{\boldsymbol{x}_\alpha} \bullet (\boldsymbol{x}-\boldsymbol{x}_\alpha)^i = \left[\sum_\beta \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{w}_{\alpha\beta}^{(l)} \bullet (\boldsymbol{x}-\boldsymbol{x}_\alpha)^l\right] \left[\sum_{j=0}^{k} \frac{1}{j!} D^{(j)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(j)}\right] \tag{7.3.8}$$

que les coefficients inconnus $\boldsymbol{w}_{\alpha\beta}^{(l)}$ doivent satisfaire quel que soit le polynôme p de degré k. En raison de l'indépendance des fonctions de base, la condition (7.3.8) est équivalente au système d'équations

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \otimes \boldsymbol{z}_{\alpha\beta}^{(j)} = \begin{cases} j!\boldsymbol{\delta}^{(2j)} & \text{si } l=j \\ 0 & \text{si } l \neq j \end{cases} \quad 0 \leq j,l \leq k \tag{7.3.9}$$

pour les inconnues $\boldsymbol{w}_{\alpha\beta}^{(l)}$. Le système de conditions (7.3.9) assure avec l'identité tensorielle (5.2.26) que

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \left(\frac{1}{j!} D^{(j)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(j)}\right) = \begin{cases} \boldsymbol{\delta}^{(2j)} \cdot D^{(j)}p\big|_{\boldsymbol{x}_\alpha} = D^{(j)}p\big|_{\boldsymbol{x}_\alpha} & \text{si } l=j \\ 0 & \text{si } l \neq j \end{cases}$$

pour $0 \leq j,l \leq k$.

En explicitant et en utilisant la définition (5.2.22) de $\boldsymbol{\delta}^{(2j)}$, les équations (7.3.9) s'écrivent

$$\sum_\beta w_{i_1\cdots i_l}^{\alpha\beta} z_{n_1\cdots n_j}^{\alpha\beta} = \begin{cases} j! \left(\boldsymbol{\delta}^{(2j)}\right)_{i_1\cdots i_j \, n_1\cdots n_j} & \text{si } l=j \\ 0 & \text{si } l \neq j \end{cases} \quad 0 \leq j,l \leq k \, . \tag{7.3.10}$$

Avant de discuter l'existence de solutions $\boldsymbol{w}_{\alpha\beta}^{(l)}$ de l'équation (7.3.9) , il faut d'abord compter le nombre d'inconnues et d'équations indépendantes. Fixons d'abord l. D'après la proposition 5.10.3, le tenseur symétrique $\boldsymbol{w}_{\alpha\beta}^{(l)}$ a

$$(m_\alpha+1)\binom{l+d-1}{l} \tag{7.3.11}$$

composantes indépendantes et pour l et j fixés, il existe

$$\binom{l+d-1}{l}\binom{j+d-1}{j} \tag{7.3.12}$$

équations indépendantes dans (7.3.9). Pour l fixé, cela fait

$$\sum_{j=0}^{k} \binom{l+d-1}{l}\binom{j+d-1}{j} = \binom{l+d-1}{l}\binom{k+d}{d} \tag{7.3.13}$$

équations indépendantes pour les composantes de $\boldsymbol{w}_{\alpha\beta}^{(l)}$. Puisque le nombre d'inconnues doit être supérieur ou égal au nombre d'équations, il faut que

$$(m_\alpha+1)\binom{l+d-1}{l} \geq \binom{l+d-1}{l}\binom{k+d}{d} \tag{7.3.14}$$

pour que (7.3.9) puisse admettre une solution pour $\boldsymbol{w}_{\alpha\beta}^{(l)}$. La division de (7.3.14) par $\binom{l+d-1}{l}$ montre que la condition (7.3.14) est indépendante de l et coïncide avec la condition (7.2.14)

$$m_\alpha + 1 \geq \binom{k+d}{d}$$

de la section 7.2.2.

Finalement, la somme de (7.3.11) de $l=0$ à $l=k$ donne le nombre total d'inconnues

$$\sum_{l=0}^{k} (m_\alpha+1)\binom{l+d-1}{l} = (m_\alpha+1)\binom{k+d}{d} \tag{7.3.15}$$

et la somme de (7.3.13) de $l=0$ à $l=k$ donne le nombre total d'équations indépendantes

$$\sum_{l=0}^{k} \binom{l+d-1}{l}\binom{k+d}{d} = \binom{k+d}{d}^2 . \tag{7.3.16}$$

La comparaison de (7.3.15) et (7.3.16) conduit de nouveau à la condition (7.2.14).

Dans les sections suivantes, il est intéressant de remplacer le système d'équations (7.3.9) par un système équivalent. Cela est exprimé par la

Proposition 7.3.2 (Conditions alternatives de consistance). *Le système d'équations (7.3.9) est équivalent au système*

$$\sum_{\beta} w_{\alpha\beta}^{(0)} \otimes \boldsymbol{z}_{\alpha\beta}^{(j)} = \begin{cases} 1 & si\ j=0 \\ 0 & si\ j\neq 0 \end{cases}, \ 0 \leq j \leq k \tag{7.3.17}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(l)} \otimes \left[\boldsymbol{z}_{\alpha\beta}^{(j)} - \boldsymbol{x}_{\alpha}^{(j)}\right] = \begin{cases} j!\boldsymbol{\delta}^{(2j)} & si\ l=j \\ 0 & si\ l\neq j \end{cases}, \ 1 \leq j,l \leq k \tag{7.3.18}$$

$$\boldsymbol{w}_{\alpha\alpha}^{(l)} = -\sum_{\beta\neq\alpha} \boldsymbol{w}_{\alpha\beta}^{(l)} = 0\,, \ 1 \leq l \leq k \tag{7.3.19}$$

où les coefficients $\boldsymbol{w}_{\alpha\alpha}^{(l)}$ pour $l \geq 1$ apparaissent uniquement dans l'équation (7.3.19).

Démonstration. Les équations (7.3.17) pour les coefficients $w_{\alpha\beta}^{(0)}$ restent inchangées. Les équations (7.3.9) pour $l \geq 1$ et $j=0$ s'écrivent

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(l)} = 0\,, \ 1 \leq l \leq k \tag{7.3.20}$$

car $\boldsymbol{z}_{\alpha\beta}^{(0)} = 1$. Les équations (7.3.20) servent uniquement à déterminer les tenseurs $\boldsymbol{w}_{\alpha\alpha}^{(l)}$ en fonction des $\boldsymbol{w}_{\alpha\beta}^{(l)}$ pour $\beta \neq \alpha$ et donnent les équations (7.3.19). Pour $l \geq 1$ et $j \geq 1$, il est possible d'éliminer les inconnues $\boldsymbol{w}_{\alpha\alpha}^{(l)}$ des équations (7.3.9). Pour cela, on forme les produits tensoriels de (7.3.20) et des tenseurs $\boldsymbol{x}_{\alpha}^{(j)}$ pour $1 \leq j \leq k$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(l)} \otimes \boldsymbol{x}_{\alpha}^{(j)} = 0\,, \ 1 \leq j,l \leq k\,. \tag{7.3.21}$$

La soustraction de (7.3.21) de (7.3.9) donne les équations (7.3.18) qui ne contiennent plus les inconnues $\boldsymbol{w}_{\alpha\alpha}^{(l)}$ car $\boldsymbol{z}_{\alpha\alpha}^{(l)} = \boldsymbol{x}_{\alpha}^{(l)}$. Cela prouve que le système d'équations (7.3.9) est équivalent au système formé par les équations (7.3.17), (7.3.18) et (7.3.19). □

7.3.3. Formulation des conditions de consistance pour la reconstruction à partir de valeurs ponctuelles. La reconstruction à partir de valeurs ponctuelles se fait de façon analogue à la reconstruction à partir de moyennes de cellule. Il suffit de remplacer le développement de la moyenne (5.8.4) des polynômes par un développement des valeurs de p aux points $\boldsymbol{x}_{\beta}$, $\beta \in \widehat{\mathbb{W}}_{\alpha}$. Soit p un polynôme de degré k. La valeur du polynôme p en $\boldsymbol{x}_{\beta}$ est

$$p\left(\boldsymbol{x}_{\beta}\right) = \sum_{j=0}^{k} \frac{1}{j!}\, D^{(j)}p\Big|_{\boldsymbol{x}_{\alpha}} \bullet \boldsymbol{h}_{\alpha\beta}^{j}\,.$$

L'insertion du polynôme p dans la condition de consistance (7.2.11) conduit à la relation

$$p\left(\boldsymbol{x}\right) = \sum_{i=0}^{k} \frac{1}{i!}\, D^{(i)}p\Big|_{\boldsymbol{x}_{\alpha}} \bullet \left(\boldsymbol{x}-\boldsymbol{x}_{\alpha}\right)^{i} = \left[\sum_{\beta}\sum_{l=0}^{k} \frac{1}{l!}\boldsymbol{w}_{\alpha\beta}^{(l)} \bullet \left(\boldsymbol{x}-\boldsymbol{x}_{\alpha}\right)^{l}\right]\left[\sum_{j=0}^{k} \frac{1}{j!}\, D^{(j)}p\Big|_{\boldsymbol{x}_{\alpha}} \bullet \boldsymbol{h}_{\alpha\beta}^{j}\right] \tag{7.3.22}$$

que les coefficients inconnus $\boldsymbol{w}_{\alpha\beta}^{(l)}$ doivent satisfaire, quel que soit le polynôme p de degré k.

En raison de l'indépendance des fonctions de base (7.3.2), la condition (7.3.22) est équivalente au système d'équations

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(l)} \otimes \boldsymbol{h}_{\alpha\beta}^{j} = \begin{cases} j!\boldsymbol{\delta}^{(2j)} & \text{si } l = j \\ 0 & \text{si } l \neq j \end{cases} \quad 0 \leq j,l \leq k \tag{7.3.23}$$

qui est de la même forme que (7.3.9). Pour cette raison, la discussion de la section 7.3.2 s'applique également à la reconstruction à partir de valeurs ponctuelles.

Remarque 7.3.3. Le cas $k = 1$ mérite une attention particulière car les identités $\boldsymbol{h}_{\alpha\beta} = \boldsymbol{z}_{\alpha\beta}^{(1)}$ et $\boldsymbol{h}_{\alpha\beta}^{0} = \boldsymbol{z}_{\alpha\beta}^{(0)} = 1$ impliquent que les équations (7.3.9) et (7.3.23) sont identiques pour $k = 1$. Les reconstructions à partir de valeurs ponctuelles et à partir de moyennes coïncident donc pour $k = 1$. □

7.3.4. Régularité de la reconstruction. L'objectif de cette section est de présenter quelques observations sur la relation entre les conditions de reconstruction (7.3.9) et (7.3.23) et la condition de régularité de la reconstruction exprimée par la définition 7.2.4.

Les équations (7.3.9) et (7.3.23) ont une influence sur le comportement des inconnues $\boldsymbol{w}_{\alpha\beta}^{(j)}$ lorsque h tend vers zéro car les tenseurs $\boldsymbol{z}_{\alpha\beta}^{(j)}$ et $\boldsymbol{h}_{\alpha\beta}^{j}$ sont $\mathrm{O}\left(h^{j}\right)$. Il suffit de considérer la reconstruction à partir de moyennes de cellule car la situation est analogue pour les valeurs ponctuelles. Pour des indices $i_1 \cdots i_j$ fixes, la relation

$$\sum_{\beta} w_{i_1\cdots i_j}^{\alpha\beta} z_{i_1\cdots i_j}^{\alpha\beta} = j!\left(\boldsymbol{\delta}^{(2j)}\right)_{i_1\cdots i_j\, i_1\cdots i_j} \geq 1$$

entraîne en raison de l'inégalité de Cauchy-Schwarz

$$1 \leq \left[\sum_{\beta}\left(w_{i_1\cdots i_j}^{\alpha\beta}\right)^{2}\right]^{\frac{1}{2}}\left[\sum_{\beta}\left(z_{i_1\cdots i_j}^{\alpha\beta}\right)^{2}\right]^{\frac{1}{2}}. \tag{7.3.24}$$

Puisque les tenseurs $\boldsymbol{z}_{\alpha\beta}^{(j)}$ sont $\mathrm{O}\left(h^{j}\right)$, l'inégalité (7.3.24) montre que la moyenne des carrés des composantes $w_{i_1\cdots i_j}^{\alpha\beta}$ *augmente au moins comme* h^{-j} pour $j \geq 1$.

Si l'hypothèse 7.2.6 sur la condition de régularité (7.2.24) est satisfaite, il existe une constante $\mathrm{C}_{\mathrm{reg}}$, indépendante de h et de la cellule $\mathcal{T}_\alpha$, telle que

$$\left|w^{\alpha\beta}\left(\boldsymbol{x}\right)\right| = \left|\sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{w}_{\alpha\beta}^{(l)} \bullet \left(\boldsymbol{x} - \boldsymbol{x}_\alpha\right)^l\right| \leq \mathrm{C}_{\mathrm{reg}} \tag{7.3.25}$$

pour $\boldsymbol{x} \in \mathcal{T}_\alpha$ et $h \longrightarrow 0$. Pour $\boldsymbol{x} \in \mathcal{T}_\alpha$, les expressions $\left(\boldsymbol{x} - \boldsymbol{x}_\alpha\right)^j$ sont $\mathrm{O}\left(h^j\right)$ car la définition de h entraîne l'inégalité

$$\left|\left(x_{i_1} - x_{\alpha,i_1}\right) \cdots \left(x_{i_j} - x_{\alpha,i_j}\right)\right| \leq h^j \text{ pour } 1 \leq i_1, \cdots, i_j \leq d \text{ si } \boldsymbol{x} \in \mathcal{T}_\alpha\,. \tag{7.3.26}$$

La condition (7.3.25) permet de formuler une condition suffisante pour la régularité de la reconstruction.

Proposition 7.3.4 (Condition suffisante pour la régularité de la reconstruction). *La reconstruction $\mathfrak{R}_\alpha$ est régulière au sens de la définition 7.2.4 s'il existe une constante $\widetilde{\mathrm{C}}_{\mathrm{reg}}$, indépendante de h et de la cellule $\mathcal{T}_\alpha$, telle que les composantes de $\boldsymbol{w}_{\alpha\beta}^{(j)}$, pour $0 \leq j \leq k$, satisfont des estimations*

$$\left|w_{i_1\cdots i_j}^{\alpha\beta}\right| \leq \frac{\widetilde{\mathrm{C}}_{\mathrm{reg}}}{h^j} \tag{7.3.27}$$

Démonstration. Les estimations (7.3.26) et (7.3.27) entraînent immédiatement l'inégalité (7.3.25). □

La proposition 7.3.4 assure donc que l'hypothèse 7.2.6 est satisfaite. Malheureusement, il paraît difficile de prouver qu'une méthode de reconstruction satisfait (7.3.27), ce qui montre que la proposition 7.3.4 n'est pas très utile en pratique. Une analyse plus détaillée de cette question dépasserait le cadre de cette thèse. Pour cette raison, la présente étude adopte l'hypothèse 7.2.6 sur la régularité des reconstructions.

7.4. Interprétation de la consistance comme approximation des dérivées

Les équations (7.3.9) et (7.3.23) permettent de donner une interprétation simple des $\boldsymbol{w}_{\alpha\beta}^{(l)}$. Soit u une fonction suffisamment régulière. La moyenne de u possède un développement à l'ordre $k+1$ défini par (5.8.1)

$$\overline{u}_\beta = \sum_{j=0}^{k+1} \frac{1}{j!}\, D^{(j)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(j)} + \mathrm{O}\left(h^{k+2}\right)\,.$$

Puisque les tenseurs $\boldsymbol{w}_{\alpha\beta}^{(l)}$ satisfont les conditions (7.3.9), il vient

$$\begin{aligned}\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \overline{u}_\beta &= \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \left[\sum_{j=0}^{k+1} \frac{1}{j!}\, D^{(j)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(j)} + \mathrm{O}\left(h^{k+2}\right)\right] = \\ &= D^{(l)}u\Big|_{\boldsymbol{x}_\alpha} + \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \left[\frac{1}{(k+1)!}\, D^{(k+1)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(k+1)} + \mathrm{O}\left(h^{k+2}\right)\right]\,. \end{aligned} \tag{7.4.1}$$

La discussion à la fin de la section 7.3.4 montre que l'hypothèse 7.2.6 est satisfaite si les composantes $\boldsymbol{w}_{\alpha\beta}^{(l)}$ satisfont une estimation du type (7.3.27). Si l'on suppose (7.3.27) vraie, alors l'estimation $\boldsymbol{z}_{\alpha\beta}^{(k+1)} = \mathrm{O}\left(h^{k+1}\right)$ entraîne que

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \overline{u}_\beta = D^{(l)}u\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k-l+1}\right)\,,\ 0 \leq l \leq k \tag{7.4.2}$$

est une approximation de la dérivée d'ordre l de u en $\boldsymbol{x}_\alpha$. On obtient en particulier pour $l = 0$

$$w_\alpha\left[\mathfrak{u}\right]\left(\boldsymbol{x}_\alpha\right) = \sum_\beta w_{\alpha\beta}^{(0)} \overline{u}_\beta = u\left(\boldsymbol{x}_\alpha\right) + \mathrm{O}\left(h^{k+1}\right)\,. \tag{7.4.3}$$

La reconstruction à partir de valeurs ponctuelles permet de retrouver des résultats analogues.

Les résultats ci-dessus sont une motivation pour la

DÉFINITION 7.4.1 (Dérivée consistante). On appelle *dérivée consistante d'ordre* $l \in \mathbb{N}$ *de précision à l'ordre* $j \in \mathbb{N}$ dans la cellule $\mathcal{T}_\alpha$ tout opérateur linéaire de la forme

$$\boldsymbol{w}_\alpha^{(l)}\left[\mathfrak{u}\right] \triangleq \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \overline{u}_\beta$$

tel que pour tout polynôme de degré $l+j-1$

$$\boldsymbol{w}_\alpha^{(l)}\left[\mathfrak{p}\right] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \overline{p}_\beta = D^{(l)}p\Big|_{\boldsymbol{x}_\alpha}$$

et pour toute fonction u de classe C^{l+j} sur le voisinage de $\mathcal{T}_\alpha$

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \overline{u}_\beta = D^{(l)}u\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^j\right) .$$

Pour $l=1$, on parle de *gradient consistant de précision à l'ordre* j, pour $l=2$ de *dérivée seconde consistante* de *précision à l'ordre* j et pour $l=3$ de *dérivée troisième consistante* de *précision à l'ordre* j.

La définition 7.4.1 permet d'exprimer (7.4.2) sous la forme suivante.

PROPOSITION 7.4.2 (Critère de consistance et dérivées consistantes). *Les tenseurs* $\boldsymbol{w}_{\alpha\beta}^{(l)}$ *qui satisfont l'hypothèse 7.2.6 et les conditions (7.3.9) ou (7.3.23) forment les coefficients d'une dérivée consistante d'ordre* l *de précision à l'ordre* $k-l+1$ *dans la cellule* $\mathcal{T}_\alpha$.

REMARQUE 7.4.3. La forme générale (7.3.7) de la reconstruction

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \sum_\beta \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{w}_{\alpha\beta}^{(l)} \bullet (\boldsymbol{x}-\boldsymbol{x}_\alpha)^l \, \overline{u}_\beta$$

peut donc être interprétée comme un polynôme de Taylor approché dont les coefficients sont des approximations des dérivées exactes. Ce point de vue rejoint la notion de "Truncated Taylor Series Expansion Reconstruction" de Delanaye [**44**].

La définition 7.4.1 permet également de formuler le

LEMME 7.4.4. *Toute dérivée consistante d'ordre* $l \geq 1$ *de précision à l'ordre* $j \geq 1$ *s'exprime sous les deux formes équivalentes*

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \overline{u}_\beta = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \left(\overline{u}_\beta - \overline{u}_\alpha\right) \tag{7.4.4}$$

DÉMONSTRATION. La dérivée d'ordre $l \geq 1$ d'une fonction constante est nulle. Comme la dérivée consistante d'ordre $l \geq 1$ doit également reproduire la dérivée d'ordre l de tout polynôme de degré 0, il vient

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \overline{u}_\alpha = 0$$

ce qui démontre (7.4.4). □

Le lemme 7.4.4 montre que l'on n'a pas besoin du coefficient $\boldsymbol{w}_{\alpha\alpha}^{(l)}$ pour former l'expression (7.4.4) si $l \geq 1$.

7.5. Définition des reconstructions conservatives

Les reconstructions générales définies par les conditions (7.3.9) reproduisent les polynômes de degré k mais elles ne satisfont pas encore la contrainte de conservation (6.3.8) introduite à la fin de la section 6.3. La prochaine étape est d'imposer (6.3.8)

$$\frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha} w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x})\; dx = \overline{u}_\alpha$$

qui stipule que la moyenne du polynôme $w_\alpha\left[\mathfrak{u}\right]$ reconstruit à partir des moyennes $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$ d'une fonction u doit toujours être égale à $\overline{u}_\alpha$, même si u *n'est pas un polynôme*. Cette condition supplémentaire renforce la précision de la reconstruction dans la cellule $\mathcal{T}_\alpha$, là où les valeurs de $w_\alpha\left[\mathfrak{u}\right]$ sont utilisées.

Définition 7.5.1 (Reconstructions conservatives). Une reconstruction $\mathfrak{R}_\alpha$ de type (7.2.5) qui satisfait (6.3.8) est appelée une *reconstruction conservative*. □

Remarque 7.5.2. Il n'est à priori pas clair de demander la condition (6.3.8) à une reconstruction à partir de valeurs ponctuelles car la connaissance des valeurs de u aux points $\boldsymbol{x}_\beta$ ne permet pas a priori de déterminer sa moyenne sur la cellule $\mathcal{T}_\alpha$. Il faudrait une formule d'intégration numérique permettant d'approcher l'intégrale de u sur $\mathcal{T}_\alpha$ par des valeurs aux points $\boldsymbol{x}_\beta$. Cela est possible en théorie mais paraît compliqué en pratique. Pour cette raison, cette section concerne seulement la reconstruction à partir de moyennes de cellule. □

L'insertion de la reconstruction (7.3.7)

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \sum_\beta \sum_{l=0}^{k} \frac{1}{l!}\boldsymbol{w}_{\alpha\beta}^{(l)} \bullet (\boldsymbol{x}-\boldsymbol{x}_\alpha)^l\; \overline{u}_\beta\,.$$

dans (6.3.8) donne l'équation

$$\sum_\beta \sum_{l=0}^{k} \frac{1}{l!}\boldsymbol{w}_{\alpha\beta}^{(l)} \bullet \boldsymbol{x}_\alpha^{(l)}\, \overline{u}_\beta = \overline{u}_\alpha\,. \tag{7.5.1}$$

que les $\boldsymbol{w}_{\alpha\beta}^{(l)}$ doivent satisfaire quelles que soient les valeurs $(\overline{u}_1, \ldots, \overline{u}_N)$. L'identité $x_\alpha^{(0)} = 1$ permet d'écrire (7.5.1) comme

$$\sum_\beta \left\{ w_{\alpha\beta}^{(0)} + \sum_{l=1}^{k} \frac{1}{l!}\boldsymbol{w}_{\alpha\beta}^{(l)} \bullet \boldsymbol{x}_\alpha^{(l)} \right\} \overline{u}_\beta = \overline{u}_\alpha\,. \tag{7.5.2}$$

Comme (7.5.2) doit être vrai pour tout $(\overline{u}_1, \ldots, \overline{u}_N)$, (7.5.2) s'exprime par la *condition de conservativité*

$$w_{\alpha\beta}^{(0)} = \delta_{\alpha\beta} - \sum_{l=1}^{k} \frac{1}{l!}\boldsymbol{w}_{\alpha\beta}^{(l)} \bullet \boldsymbol{x}_\alpha^{(l)}\,. \tag{7.5.3}$$

Une reconstruction conservative d'ordre k est donc une reconstruction définie par des coefficients $\boldsymbol{w}_{\alpha\beta}^{(j)}$, $0 \le j \le k$, qui satisfont la condition de conservativité (7.5.3).

Pour définir une reconstruction consistante de degré k, les $\boldsymbol{w}_{\alpha\beta}^{(j)}$, $0 \le j \le k$, doivent être solutions de (7.3.9) ou, de façon équivalente, solutions du système formé par (7.3.17), (7.3.18) et (7.3.19), comme démontré par la proposition 7.3.2. On peut se demander si les solutions de (7.5.3) peuvent être des solutions du système (7.3.17-7.3.19) ou s'il y a une contradiction. La réponse est fournie par la

Proposition 7.5.3 (Reconstruction consistante et reconstruction conservative). *Si les tenseurs $\boldsymbol{w}_{\alpha\beta}^{(l)}$, $1 \le l \le k$, satisfont les équations (7.3.18)*

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \otimes \left[\boldsymbol{z}_{\alpha\beta}^{(j)} - \boldsymbol{x}_\alpha^{(j)}\right] = \begin{cases} j!\boldsymbol{\delta}^{(2j)} & \text{si } l = j \\ 0 & \text{si } l \neq j \end{cases},\ 1 \le j, l \le k$$

et (7.3.19)

$$\boldsymbol{w}_{\alpha\alpha}^{(l)} \;=\; -\sum_{\beta\neq\alpha}\boldsymbol{w}_{\alpha\beta}^{(l)}=0\,,\;1\leq l\leq k\,,$$

alors les coefficients $w_{\alpha\beta}^{(0)}$ *donnés par la condition de conservativité (7.5.3)*

$$w_{\alpha\beta}^{(0)}=\delta_{\alpha\beta}-\sum_{l=1}^{k}\frac{1}{l!}\boldsymbol{w}_{\alpha\beta}^{(l)}\bullet\boldsymbol{x}_{\alpha}^{(l)}$$

satisfont automatiquement la condition (7.3.17)

$$\sum_{\beta}w_{\alpha\beta}^{(0)}\otimes\boldsymbol{z}_{\alpha\beta}^{(j)}=\begin{cases}1 & si\ j=0\\ 0 & si\ j\neq 0\end{cases},\;0\leq j\leq k\,.$$

Il est donc possible d'imposer la condition de conservativité (7.5.3) à une reconstruction consistante de degré k qui satisfait le système d'équations composé de (7.3.18) et de (7.3.19).

Démonstration. Il est suffisant de vérifier que les coefficients (7.5.3) satisfont les équations (7.3.17)

$$\sum_{\beta}w_{\alpha\beta}^{(0)}\otimes\boldsymbol{z}_{\alpha\beta}^{(j)}=\begin{cases}1 & \text{si } j=1\\ 0 & \text{si } j\neq 1\end{cases}\;0\leq j\leq k\,.$$

L'insertion de (7.5.3) dans (7.3.17) donne pour $j=0$

$$\sum_{\beta}w_{\alpha\beta}^{(0)}\otimes\boldsymbol{z}_{\alpha\beta}^{(0)}=\boldsymbol{z}_{\alpha\alpha}^{(0)}-\sum_{\beta}\sum_{l=1}^{k}\frac{1}{l!}\left(\boldsymbol{w}_{\alpha\beta}^{(l)}\bullet\boldsymbol{x}_{\alpha}^{(l)}\right)\boldsymbol{z}_{\alpha\beta}^{(0)}=1$$

car $\boldsymbol{z}_{\alpha\alpha}^{(0)}=1$ par la définition (5.7.10) et $\boldsymbol{w}_{\alpha\beta}^{(l)}\otimes\boldsymbol{z}_{\alpha\beta}^{(0)}=0$ pour $l\geq 1$ en raison de $\boldsymbol{z}_{\alpha\beta}^{(0)}=1$ et (7.3.19). Pour $j\geq 1$ il vient en raison des équations (7.3.18) et $\boldsymbol{z}_{\alpha\alpha}^{(j)}=\boldsymbol{x}_{\alpha}^{(j)}$

$$\sum_{\beta}w_{\alpha\beta}^{(0)}\otimes\boldsymbol{z}_{\alpha\beta}^{(j)}=\boldsymbol{z}_{\alpha\alpha}^{(j)}-\sum_{\beta}\sum_{l=1}^{k}\frac{1}{l!}\left(\boldsymbol{w}_{\alpha\beta}^{(l)}\bullet\boldsymbol{x}_{\alpha}^{(l)}\right)\boldsymbol{z}_{\alpha\beta}^{(j)}=\boldsymbol{x}_{\alpha}^{(j)}-\boldsymbol{x}_{\alpha}^{(j)}=0\,.$$

□

Sous la contrainte (6.3.8), la forme générale des polynômes $w^{\alpha\beta}(\boldsymbol{x})$ de (7.2.5) est

$$w^{\alpha\beta}(\boldsymbol{x})=\delta_{\alpha\beta}+\sum_{l=1}^{k}\frac{1}{l!}\boldsymbol{w}_{\alpha\beta}^{(l)}\bullet\left[(\boldsymbol{x}-\boldsymbol{x}_{\alpha})^{l}-\boldsymbol{x}_{\alpha}^{(l)}\right]\tag{7.5.4}$$

et la forme générale de la reconstruction (7.2.5) est

$$w_{\alpha}[\mathfrak{u}](\boldsymbol{x})=\overline{u}_{\alpha}+\sum_{\beta}\sum_{l=1}^{k}\frac{1}{l!}\boldsymbol{w}_{\alpha\beta}^{(l)}\bullet\left[(\boldsymbol{x}-\boldsymbol{x}_{\alpha})^{l}-\boldsymbol{x}_{\alpha}^{(l)}\right]\overline{u}_{\beta}\,.\tag{7.5.5}$$

Puisque la contrainte (6.3.8) a permis d'exprimer les inconnues $w_{\alpha\beta}^{(0)}$ en fonction des coefficients $w_{\alpha\beta}^{(l)}$ pour $1\leq l\leq k$, elle a éliminé les conditions (7.3.17) du système des équations (7.3.17), (7.3.18) et (7.3.19). Il faut cependant faire la

Remarque 7.5.4. La condition (6.3.8) de la reconstruction conservative ne modifie pas la condition nécessaire (7.2.14) sur le nombre de voisins

$$m_{\alpha}+1\geq\binom{k+d}{d}\,.$$

Cela vient du fait que la condition (7.3.14) doit être satisfaite séparément pour chaque $0 \leq l \leq k$. Une autre façon de formuler cette observation est de dire que la condition (6.3.8) élimine $(m_\alpha + 1)$ inconnues et $\binom{k+d}{d}$ équations indépendantes. Le nouveau nombre total d'inconnues est donc

$$N_{\text{inc}} = (m_\alpha + 1)\left[\binom{k+d}{d} - 1\right]$$

et le nouveau nombre d'équations indépendantes

$$N_{\text{eq}} = \binom{k+d}{d}\left[\binom{k+d}{d} - 1\right].$$

La condition $N_{\text{inc}} \geq N_{\text{eq}}$ est donc équivalente à (7.2.14).

7.6. Formulation matricielle des conditions de consistance

Les sections 7.3.2 et 7.5 ont permis d'établir un système d'équations pour les paramètres $\boldsymbol{w}_{\alpha\beta}^{(l)}$, $0 \leq l \leq k$, de la reconstruction. Il s'agit des conditions de consistance de degré k, (7.3.17), (7.3.18), (7.3.19), et de la condition de conservativité (7.5.3). Pour une analyse théorique et pour des calculs pratiques, ce système de conditions présente les inconvénients suivants dans sa forme actuelle :

(1) Les inconnues et les coefficients de ces équations sont des tenseurs symétriques, ce qui engendre des relations de dépendance, aussi bien entre les inconnues qu'entre les équations. Il est par conséquent nécessaire de déterminer un sous-ensemble d'inconnues indépendantes et un sous-ensemble d'équations indépendantes.

(2) Il est difficile d'étudier la structure générale des solutions de ce système.

(3) Il n'est pas évident d'implémenter des algorithmes de résolution pour ce système.

L'objectif de cette section est de transcrire le système formé par (7.3.17), (7.3.18), (7.3.19) et (7.5.3) en une équation matricielle car les équations matricielles peuvent être analysées et résolues par des méthodes bien établies.

On commence par éliminer les équations et les inconnues dépendantes. L'équation (7.5.3)

$$w_{\alpha\beta}^{(0)} = \delta_{\alpha\beta} - \sum_{l=1}^{k} \frac{1}{l!} \boldsymbol{w}_{\alpha\beta}^{(l)} \bullet \boldsymbol{x}_\alpha^{(l)}$$

détermine uniquement les inconnues $w_{\alpha\beta}^{(0)}$ en fonction des inconnues $\boldsymbol{w}_{\alpha\beta}^{(l)}$, pour $1 \leq l \leq k$. De façon analogue, les équations (7.3.19)

$$\boldsymbol{w}_{\alpha\alpha}^{(l)} = -\sum_{\beta \neq \alpha} \boldsymbol{w}_{\alpha\beta}^{(l)} = 0\,,\ 1 \leq l \leq k\,,$$

servent uniquement à déterminer les inconnues $\boldsymbol{w}_{\alpha\alpha}^{(l)}$ en fonction des inconnues $\boldsymbol{w}_{\alpha\beta}^{(l)}$, pour $1 \leq l \leq k$ et $\beta \neq \alpha$. La proposition 7.5.3 montre que si les inconnues $\boldsymbol{w}_{\alpha\beta}^{(l)}$, pour $1 \leq l \leq k$, satisfont (7.3.18) et (7.3.19), alors les inconnues $w_{\alpha\beta}^{(0)}$ déterminées par (7.5.3) satisfont automatiquement (7.3.17). Cela démontre la

Proposition 7.6.1. *Afin de résoudre le système formé par les conditions (7.3.17), (7.3.18), (7.3.19) et (7.5.3) pour les inconnues $\boldsymbol{w}_{\alpha\beta}^{(l)}$, $0 \leq l \leq k$, il suffit de résoudre les équations (7.3.18)*

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(l)} \otimes \left[\boldsymbol{z}_{\alpha\beta}^{(j)} - \boldsymbol{x}_\alpha^{(j)}\right] = \begin{cases} j!\boldsymbol{\delta}^{(2j)} & \text{si } l = j \\ 0 & \text{si } l \neq j \end{cases} \quad 1 \leq j, l \leq k$$

pour les inconnues $\boldsymbol{w}_{\alpha\beta}^{(l)}$, pour $1 \leq l \leq k$ et $\beta \neq \alpha$. Les inconnues $\boldsymbol{w}_{\alpha\alpha}^{(l)}$, $1 \leq l \leq k$ et $w_{\alpha\beta}^{(0)}$, $\beta \in \widehat{\mathbb{W}}_\alpha$, sont ensuite déterminées par les conditions respectives (7.3.19) et (7.5.3).

Le système (7.3.18) s'écrit de façon explicite

$$\sum_{\beta} w^{\alpha\beta}_{i_1\cdots i_l}\left[z^{\alpha\beta}_{n_1\cdots n_j} - x_{\alpha,n_1\cdots n_j}\right] = \begin{cases} j!\,\left(\boldsymbol{\delta}^{(2j)}\right)_{i_1\cdots i_j\, n_1\cdots n_j} & \text{si } l = j \\ 0 & \text{si } l \neq j \end{cases} \tag{7.6.1}$$

où $1 \leq j, l \leq k$.

Les inconnues $w^{\alpha\beta}_{i_1\cdots i_l}$ et les coefficients $z^{\alpha\beta}_{n_1\cdots n_j} - x_{\alpha,n_1\cdots n_j}$ du système (7.6.1) sont des composantes de tenseurs symétriques. La proposition 5.10.3 permet de compter le nombre d'inconnues et d'équations indépendantes du système (7.6.1). Pour β fixé, il existe

$$\sum_{l=1}^{k} \binom{l+d-1}{l} = \binom{k+d}{d} - 1$$

inconnues $w^{\alpha\beta}_{i_1\cdots i_l}$ indépendantes et autant de coefficients $z^{\alpha\beta}_{n_1\cdots n_j} - x_{\alpha,n_1\cdots n_j}$ indépendants. Si m_α cellules voisines de $\mathcal{T}_\alpha$ contribuent à la reconstruction, il existe

$$m_\alpha\left[\binom{k+d}{d} - 1\right]$$

inconnues indépendantes $w^{\alpha\beta}_{i_1\cdots i_l}$, pour $1 \leq l \leq k$ et $\beta \neq \alpha$. Le nombre d'équations indépendantes du système (7.6.1) est

$$\left[\binom{k+d}{d} - 1\right]^2.$$

La comparaison du nombre des inconnues avec le nombre d'équations donne à nouveau la condition (7.2.14).

Pour rappel, d'après la définition 5.10.1 de la relation (5.10.2), deux l-tuples d'indices $(i_1, \cdots, i_l)$ et $(i'_1, \cdots, i'_l)$ sont dits *équivalents* s'il existe une permutation $\pi \in \mathfrak{S}_l$ de l'ensemble $\{1, \ldots, l\}$ telle que $i_m = i'_{\pi(m)}$, $1 \leq m \leq l$. La relation (5.10.2) définit une partition de l'ensemble (5.10.1)

$$\mathbb{J}^d_k \triangleq \{\,(i_1, \ldots, i_k) |\, 1 \leq i_1, \ldots, i_k \leq d\} = \{1, \ldots, d\}^k$$

en classes d'équivalence. Par définition, tous les l-tuples d'indices $(i_1, \cdots, i_l)$ d'une classe d'équivalence désignent la même inconnue $w^{\alpha\beta}_{i_1\cdots i_l}$. Il suffit par conséquent de résoudre le système (7.6.1) pour le représentant $(i_1, \cdots, i_l)$ de chaque classe pour lequel $i_1 \leq \cdots \leq i_l$. Ce représentant est unique d'après le lemme 5.10.2. Le même raisonnement s'applique aux coefficients $z^{\alpha\beta}_{n_1\cdots n_j} - x_{\alpha,n_1\cdots n_j}$ des équations. Il suffit de résoudre les équations du système (7.6.1) pour le représentant $(n_1, \ldots, n_j)$ de chaque classe pour lequel $n_1 \leq \cdots \leq n_j$.

Par définition des tenseurs $\boldsymbol{\delta}^{(2j)}$, le membre de droite de (7.6.1) est non nul si et seulement si $(i_1, \cdots, i_l)$ est équivalent à $(n_1, \ldots, n_l)$ au sens de la relation (5.10.2). Pour la transcription du système (7.6.1) en une équation matricielle, il est préférable de normaliser chaque équation pour que le membre de droite prenne uniquement les valeurs 0 ou 1. Pour cela, on divise chaque équation du système (7.6.1) par un facteur adapté, ce qui donne le système

$$\sum_{\beta} w^{\alpha\beta}_{i_1\cdots i_l} \frac{z^{\alpha\beta}_{n_1\cdots n_j} - x_{\alpha,n_1\cdots n_j}}{j!\left(\boldsymbol{\delta}^{(2j)}\right)_{n_1\cdots n_j\, n_1\cdots n_j}} = \begin{cases} \dfrac{\left(\boldsymbol{\delta}^{(2j)}\right)_{i_1\cdots i_j\, n_1\cdots n_j}}{\left(\boldsymbol{\delta}^{(2j)}\right)_{n_1\cdots n_j\, n_1\cdots n_j}} & \text{si } l = j \\ 0 & \text{si } l \neq j \end{cases} \tag{7.6.2}$$

où $1 \leq j, l \leq k$ et $i_1 \leq \cdots \leq i_l$ ainsi que $n_1 \leq \cdots \leq n_j$.

Le système (7.6.2) est prêt pour être transcrit en une équation matricielle.

Définition 7.6.2 (Matrices des inconnues indépendantes). On définit la matrice des inconnues S_α de la façon suivante : Soit S_α une matrice avec $\binom{k+d}{d} - 1$ lignes et m_α colonnes. On classe les indices des cellules voisines de $\mathcal{T}_\alpha$ dans l'ordre naturel $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_{m_\alpha}$, ce qui permet d'associer la colonne numéro j de S_α à la cellule numéro β_j. Ensuite, on classe les l-tuples d'indices $(i_1, \cdots, i_l)$ qui satisfont $i_1 \leq \cdots \leq i_l$ pour $1 \leq l \leq k$ dans l'ordre suivant. Pour

$l' < l$, les l-tuples d'indices $(i_1, \cdots, i_{l'})$ précèdent les l-tuples d'indices $(i_1, \cdots, i_l)$. Pour l fixé, la fonction de rangement ϖ_k^d, définie par (5.10.6), permet de ranger les l-tuples d'indices dans l'ordre donné par la définition 5.10.4. L'élément de S_α dans la colonne numéro $j \in \{1, \ldots, m_\alpha\}$ et dans la ligne associée à $(i_1, \cdots, i_l)$ est alors, par définition, la composante $w^{\alpha\beta_j}_{i_1 \cdots i_l}$. □

Définition 7.6.3 (Matrices des coefficients indépendants). La définition de la matrice des coefficients H_α procède de la même façon que la définition 7.6.2. Soit H_α une matrice avec m_α lignes et $\binom{k+d}{d} - 1$ colonnes. L'élément de H_α dans la colonne numéro $j \in \{1, \ldots, m_\alpha\}$ et dans la ligne associée à $(n_1, \cdots, n_l)$ est par définition

$$\frac{z^{\alpha\beta_j}_{n_1 \cdots n_j} - x_{\alpha, n_1 \cdots n_j}}{j! \left(\boldsymbol{\delta}^{(2j)}\right)_{n_1 \cdots n_j \, n_1 \cdots n_j}}.$$

□

Ces définitions permettent de formuler la

Proposition 7.6.4 (Formulation matricielle des conditions de consistance). *La définition 7.6.2 de la matrice S_α et la définition 7.6.3 de la matrice H_α permettent d'écrire le système d'équations (7.6.2) comme l'équation matricielle*

$$S_\alpha H_\alpha = \mathrm{I}_{\binom{k+d}{d}-1} \tag{7.6.3}$$

pour la matrice inconnue S_α.

La proposition 7.6.4 justifie d'étendre la définition 7.2.1 de la reconstruction consistante aux solutions de (7.6.3).

Définition 7.6.5. On appelle *reconstruction consistante* de degré k une solution S_α de la condition (7.6.3).□

L'intérêt de la proposition 7.6.4 est le suivant :

(1) Par construction, les inconnues et les équations de la condition (7.6.3) sont *indépendantes*.
(2) Il est facile d'analyser la structure générale et le comportement des solutions de (7.6.3).
(3) La forme de l'équation (7.6.3) et la proposition 7.2.2 entraînent que des solutions S_α existent si et seulement si H_α est une matrice injective.
(4) Si H_α est une matrice injective, une solution de (7.6.3) est immédiatement donnée par la pseudo-inverse ou inverse de Moore-Penrose

$$H_\alpha^\dagger \triangleq \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t \tag{7.6.4}$$

de H_α, voir par exemple [**54**].

La reconstruction définie par (7.6.4) occupe une place particulière parmi les solutions de (7.6.3). Cet aspect est détaillé dans la section suivante.

7.7. Minimisation de la norme de la reconstruction

On introduit le nom suivant pour la solution (7.6.4).

Définition 7.7.1 (Méthode des moindres carrés ou méthode de la pseudo-inverse). On appelle la solution (7.6.4) de (7.6.3) la *reconstruction par la méthode des moindres carrés* ou la *reconstruction par la méthode de la pseudo-inverse*. □

La définition 7.7.1 se justifie par le fait que la solution (7.6.4) coïncide avec la méthode classique des moindres carrés dans le cas $k = 1$. Cela est démontré dans la section 7.9.1.

La définition 7.6.2 de S_α entraîne que la norme de Frobenius (5.2.5) de S_α est égale à la moyenne des carrés des composantes indépendantes $w^{\alpha\beta}_{i_1 \cdots i_j}$ des tenseurs $\boldsymbol{w}^{(l)}_{\alpha\beta}$, pour $1 \leq l \leq k$,

$$\|S_\alpha\|_F = \sqrt{\operatorname{trace}\left(S_\alpha^t S_\alpha\right)} = \sqrt{\sum_{\beta \neq \alpha} \sum_{l=1}^{k} \sum_{i_1=1}^{d} \sum_{i_2=i_1}^{d} \sum_{i_3=i_2}^{d} \cdots \sum_{i_l=i_{l-1}}^{d} \left(w^{\alpha\beta}_{i_1 \ldots i_l}\right)^2}. \tag{7.7.1}$$

Il est maintenant possible de prouver la proposition suivante.

PROPOSITION 7.7.2 (Minimisation de la norme de Frobenius). *Supposons la matrice H_α injective. Dans ce cas, la reconstruction (7.6.4)*

$$S_\alpha = H_\alpha^\dagger = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t$$

minimise la norme de Frobenius (7.7.1) parmi les solutions de l'équation (7.6.3).

DÉMONSTRATION. Une matrice S_α minimise (7.7.1) si et seulement si elle minimise la fonction

$$\mathbb{M}_{\binom{k+d}{d}-1,m_\alpha}(\mathbb{R}) \ni S_\alpha \longmapsto \operatorname{trace}\left(S_\alpha^t S_\alpha\right) . \tag{7.7.2}$$

La fonction (7.7.2) est différentiable ce qui permet de minimiser (7.7.2) sous la contrainte (7.6.3)

$$S_\alpha H_\alpha = \mathrm{I}_{\binom{k+d}{d}-1}$$

à l'aide de multiplicateurs de Lagrange.

Pour cela, on introduit une matrice L_α comme multiplicateur de Lagrange pour la contrainte (7.6.3) et on cherche le minimum de la fonction

$$F(S_\alpha, L_\alpha) = \operatorname{trace}\left(S_\alpha^t S_\alpha\right) + \operatorname{trace}\left(L_\alpha^t \left(S_\alpha H_\alpha - \mathrm{I}_{\binom{k+d}{d}-1}\right)\right) . \tag{7.7.3}$$

La dérivation de (7.7.3) par rapport à S_α donne la condition

$$2S_\alpha^t = -H_\alpha L_\alpha^t \tag{7.7.4}$$

et la dérivation de (7.7.3) par rapport à L_α donne la condition (7.6.3). La multiplication de (7.7.4) par H_α^t à gauche et l'utilisation de (7.6.3) permettent de déterminer L_α

$$L_\alpha = -2\left(H_\alpha^t H_\alpha\right)^{-1} . \tag{7.7.5}$$

La combinaison de (7.7.4) et de (7.7.5) donne le résultat du problème de minimisation

$$S_\alpha = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t$$

qui coïncide avec la reconstruction (7.6.4). □

La proposition 7.7.2 montre donc que la reconstruction (7.6.4) coïncide avec la *reconstruction d'énergie minimale* que Barth et Frederickson ont utilisée dans [**9**]. Cependant, le résultat 7.7.2 est en vérité le corollaire d'un résultat plus général, le théorème 10.6.7 présenté dans la section 10.6, voir la remarque 10.6.8.

La proposition 7.7.2 prouve qu'il existe une méthode, la méthode des moindres carrés ou de la pseudo-inverse, qui minimise la somme des carrés des coefficients indépendants $w_{i_1\dots i_l}^{\alpha\beta}$. Les coefficients de cette reconstruction sont complètement déterminés par les tenseurs géométriques $z_{i_1\dots i_l}^{\alpha\beta}$ via l'équation (7.6.4)

$$S_\alpha = H_\alpha^\dagger = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t .$$

REMARQUE 7.7.3 (Régularité et reconstruction des moindres carrés). Une question mathématique intéressante consiste à se demander dans quelles conditions géométriques la reconstruction spécifique des moindres carrés est régulière, c'est-à-dire dans quelles conditions les coefficients indépendants $w_{i_1\dots i_l}^{\alpha\beta}$ définis par l'équation (7.6.4) satisfont une estimation du type (7.3.27)

$$\left|w_{i_1\cdots i_j}^{\alpha\beta}\right| \leq \frac{\widetilde{\mathrm{C}}_{\mathrm{reg}}}{h^j}$$

où la constante $\widetilde{\mathrm{C}}_{\mathrm{reg}}$ est indépendante de h et de la cellule $\mathcal{T}_\alpha$. Une analyse de cette question pour les maillages non structurés serait trop difficile dans le cadre de cette étude.□

7.8. Interprétation algébrique des reconstructions consistantes

L'objectif de cette section est d'exprimer les solutions de l'équation matricielle de consistance (7.6.3) sous deux formes spécifiques qui serviront dans la suite à analyser le comportement des schémas numériques.

On rappelle la définition (7.2.1) du voisinage de reconstruction

$$\mathbb{W}_\alpha \triangleq \{\beta |\, \mathcal{T}_\beta \text{ contribue à la reconstruction dans } \mathcal{T}_\alpha\,,\ \beta \neq \alpha\}$$

et la définition (7.2.2)

$$m_\alpha \triangleq |\mathbb{W}_\alpha|$$

du nombre des cellules qui contribuent à la reconstruction dans la cellule $\mathcal{T}_\alpha$, à l'exclusion de la cellule $\mathcal{T}_\alpha$ elle-même.

Comme la forme générale de l'équation (7.6.3) est indépendante du degré k des polynômes reconstruits, on peut, pour les besoins de cette section, supposer que $k = 1$, ce qui entraîne

$$\binom{k+d}{d} - 1 = d\,.$$

La matrice géométrique H_α a m_α lignes qui sont, dans le cas $k = 1$, les m_α vecteurs géométriques $\boldsymbol{h}^t_{\alpha\beta}$, pour $\beta \in \mathbb{W}_\alpha = \{\beta_1, \ldots, \beta_{m_\alpha}\}$

$$H_\alpha = \begin{bmatrix} \boldsymbol{h}^t_{\alpha\beta_1} \\ \vdots \\ \boldsymbol{h}^t_{\alpha\beta_{m_\alpha}} \end{bmatrix} \in \mathbb{M}_{m_\alpha,d}\,(\mathbb{R})\ . \tag{7.8.1}$$

Dans le cas $k = 1$, la forme générale des fonctions reconstruites (7.5.5) est

$$w_\alpha\,[\mathfrak{u}]\,(\boldsymbol{x}) = \overline{u}_\alpha + \boldsymbol{w}^{(1)}_{\alpha\beta}\,[\mathfrak{u}] \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)$$

où

$$\boldsymbol{w}^{(1)}_\alpha\,[\mathfrak{u}] = \sum_\beta \boldsymbol{w}^{(1)}_{\alpha\beta}\,\overline{u}_\beta$$

est un gradient consistant au sens de la définition 7.4.1 et $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$ est le vecteur des moyennes de cellule. Pour les besoins de la reconstruction du gradient, le gradient $\boldsymbol{w}^{(1)}_\alpha\,[\mathfrak{u}]$ est désormais noté $\boldsymbol{\sigma}_\alpha\,[\mathfrak{u}]$ et les coefficients $\boldsymbol{w}^{(1)}_{\alpha\beta}$ s'appellent $\boldsymbol{\sigma}_{\alpha\beta} \triangleq \boldsymbol{w}^{(1)}_{\alpha\beta}$

$$\boldsymbol{\sigma}_\alpha\,[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta\,. \tag{7.8.2}$$

Rappelons que $\boldsymbol{\sigma}_{\alpha\beta} \triangleq 0$ par définition dans (7.8.2) si la cellule $\mathcal{T}_\beta$ n'est pas dans le voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$. La matrice S_α a d lignes et m_α colonnes qui sont les m_α vecteurs $\boldsymbol{\sigma}_{\alpha\beta}$, pour $\beta \in \mathbb{W}_\alpha = \{\beta_1, \ldots, \beta_{m_\alpha}\}$

$$S_\alpha = \left[\boldsymbol{\sigma}_{\alpha\beta_1}, \ldots, \boldsymbol{\sigma}_{\alpha\beta_{m_\alpha}}\right] \in \mathbb{M}_{d,m_\alpha}\,(\mathbb{R})\ . \tag{7.8.3}$$

Avec ces définitions, l'équation (7.6.3) est donnée par

$$S_\alpha H_\alpha = \mathrm{I}_d \tag{7.8.4}$$

On peut alors énoncer le

Théorème 7.8.1 (Caractérisation des reconstructions consistantes comme espace affine de matrices). *Si la matrice H_α dans (7.8.4) est injective, la solution générale de l'équation (7.8.4) pour l'inconnue S_α est de la forme*

$$S_\alpha = \widetilde{S}_\alpha + \Lambda_\alpha B_\alpha\,. \tag{7.8.5}$$

Dans l'équation (7.8.5), $\widetilde{S}_\alpha \in \mathbb{M}_{d,m_\alpha}\,(\mathbb{R})$ est une solution particulière de (7.8.4). La matrice $B_\alpha \in \mathbb{M}_{m_\alpha-d,m_\alpha}\,(\mathbb{R})$ est une solution de l'équation homogène associée à (7.8.4), donnée par

$$B_\alpha H_\alpha = 0_{m_\alpha-d,d} \tag{7.8.6}$$

où $0_{m_\alpha-d,d} \in \mathbb{M}_{m_\alpha-d,d}(\mathbb{R})$ *est la matrice nulle. Le rang de* B_α *est égal à* $m_\alpha - d$ *et* $\Lambda_\alpha \in \mathbb{M}_{d,m_\alpha-d}(\mathbb{R})$ *est une matrice réelle arbitraire.*

Démonstration. Si la matrice $H_\alpha \in \mathbb{M}_{m_\alpha,d}(\mathbb{R})$ est injective, la proposition 7.2.2 montre que l'équation (7.8.4) admet au moins une solution particulière $\widetilde{S}_\alpha$. De plus, on a nécessairement $m_\alpha \geq d$. Dans le cas $m_\alpha = d$, l'injectivité de H_α entraîne la bijectivité de H_α et la solution de (7.8.4) admet l'unique solution $\widetilde{S}_\alpha = H_\alpha^{-1}$.

Si $m_\alpha > d$, la solution de (7.8.4) n'est pas unique car dans ce cas la matrice $H_\alpha^t \in \mathbb{M}_{d,m_\alpha}(\mathbb{R})$ a un noyau de dimension $m_\alpha - d > 0$. Il existe donc des vecteurs $\mathfrak{b} \in \mathbb{R}^{m_\alpha}$ tels que

$$H_\alpha^t \mathfrak{b}^t = 0$$

ou de façon équivalente

$$\mathfrak{b} H_\alpha = 0\,. \tag{7.8.7}$$

Puisque la dimension du noyau de H_α^t est $m_\alpha - d$, il existe une famille de $m_\alpha - d$ vecteurs libres $\{\mathfrak{b}_1, \ldots, \mathfrak{b}_{m_\alpha-d}\}$ qui satisfont (7.8.7). Ces vecteurs forment les lignes d'une matrice $B_\alpha \in \mathbb{M}_{m_\alpha-d,m_\alpha}(\mathbb{R})$ qui satisfait l'équation homogène (7.8.6). Comme les $m_\alpha - d$ vecteurs lignes de B_α sont libres, la matrice B_α a un rang égal à $m_\alpha - d$. Cela est le rang maximal qu'elle peut avoir car le nombre m_α de ses colonnes est plus grand que le nombre de ses lignes, c'est-à-dire $m_\alpha > m_\alpha - d$. Si $\widetilde{S}_\alpha$ est une solution particulière de (7.8.4) et $\Lambda_\alpha \in \mathbb{M}_{d,m_\alpha-d}(\mathbb{R})$ est une matrice arbitraire, la matrice

$$S_\alpha = \widetilde{S}_\alpha + \Lambda_\alpha B_\alpha$$

est également solution de (7.8.4) car

$$S_\alpha H_\alpha = \left(\widetilde{S}_\alpha + \Lambda_\alpha B_\alpha\right) H_\alpha = \widetilde{S}_\alpha H_\alpha = \mathrm{I}_d\,.$$

Réciproquement, si $\widetilde{S}_\alpha$ et S_α sont deux solutions de (7.8.4), leur différence est une solution de l'équation

$$\left(S_\alpha - \widetilde{S}_\alpha\right) H_\alpha = 0_{d,d}$$

où $0_{d,d} \in \mathbb{M}_{d,d}(\mathbb{R})$ est la matrice nulle. Les lignes de la matrice $S_\alpha - \widetilde{S}_\alpha$ sont donc des éléments du noyau de H_α^t, ce qui signifie qu'elles peuvent être écrites sous la forme de combinaisons linéaires des vecteurs $\{\mathfrak{b}_1, \ldots, \mathfrak{b}_{m_\alpha-d}\}$. Par conséquent, la matrice $S_\alpha - \widetilde{S}_\alpha$ peut être représentée comme

$$S_\alpha - \widetilde{S}_\alpha = \Lambda_\alpha B_\alpha\,.$$

Cela prouve que l'ensemble des solutions de l'équation (7.8.4) forme un sous-espace affine de l'espace des $d \times m_\alpha$ matrices. □

L'espace des solutions de (7.8.4) possède une autre interprétation mathématique basée sur une analyse algébrique de (7.8.4). Dans cette section, l'indice α est omis afin de faciliter la notation. Les matrices (7.8.1) et (7.8.3) s'interprètent comme des applications linéaires

$$\begin{aligned} H : \mathbb{R}^d &\to \mathbb{R}^m \\ S : \mathbb{R}^m &\to \mathbb{R}^d\,. \end{aligned}$$

Cette notation permet de formuler le

Théorème 7.8.2 (Caractérisation des reconstructions consistantes par des projecteurs). *Soit* $H \in \mathbb{M}_{m,d}(\mathbb{R})$ *avec* $d \leq m$. *Supposons que* H *soit de rang* d *et considérons l'équation (7.8.4)*

$$SH = \mathrm{I}_d$$

pour l'inconnue matricielle $S \in \mathbb{M}_{d,m}(\mathbb{R})$.

(i) *L'équation (7.8.4) a des solutions si et seulement si* $\operatorname{rang}(H) = d$.

(ii) *Le choix d'une solution* S *de (7.8.4) est équivalent au choix d'un sous-espace* $\mathcal{W} = Ker(S)$ *supplémentaire de* $Im(H)$ *dans* $\mathbb{R}^m$.

(iii) *L'application* $HS : \mathbb{R}^m \longrightarrow \mathbb{R}^m$ *est un projecteur qui projette les vecteurs de* $\mathbb{R}^m$ *sur* $Im(H)$ *parallèlement à* $\mathcal{W} = Ker(S)$.

(iv) *La forme générale de la solution est*

$$S_{\mathcal{W}} = \left(W^t H\right)^{-1} W^t \tag{7.8.8}$$

où $W \in \mathbb{M}_{m,d}(\mathbb{R})$ est une matrice dont les lignes forment une base du complément orthogonal $\mathcal{W}^\perp$ de $\mathcal{W}$. La solution $S_{\mathcal{W}}$ ne dépend que de la matrice H et du sous-espace $\mathcal{W}$ et non pas de la matrice particulière W.

DÉMONSTRATION. Si le rang de H est inférieur à d, il existe un vecteur $\boldsymbol{\sigma} \in \mathbb{R}^d$, $\boldsymbol{\sigma} \neq 0$, tel que $H\boldsymbol{\sigma} = 0$. Par conséquent, (7.8.4) ne peut avoir de solutions. Si le rang de H est égal à d, H est injective et la proposition 7.2.2 prouve l'existence de solutions S de (7.8.4). Cela prouve le point (i).

Dans la suite de la démonstration, on suppose que des solutions S de (7.8.4) existent, c'est-à-dire que H est injective, ce qui entraîne que le sous-espace $\mathrm{Im}\,(H) \subseteq \mathbb{R}^m$ est de dimension d. Si une matrice S est solution de l'équation (7.8.4), S est nécessairement surjective. Le noyau $\mathrm{Ker}\,(S) \subseteq \mathbb{R}^m$ de l'application S est par conséquent de dimension $m - d$. De plus, l'équation (7.8.4) implique que l'intersection de $\mathrm{Ker}\,(S)$ et $\mathrm{Im}\,(H)$ est $\{0\}$. La somme de $\mathrm{Im}\,(H)$ et de $\mathrm{Ker}\,(S)$ est par conséquent une somme directe, c'est-à-dire que tout vecteur dans cette somme se décompose de façon unique sous la forme

$$\mathfrak{u} = \mathfrak{u}_H + \mathfrak{u}_S\,,\ \mathfrak{u}_H \in \mathrm{Im}\,(H)\ ,\ \mathfrak{u}_S \in \mathrm{Ker}\,(S)\ . \tag{7.8.9}$$

La prochaine étape est de prouver que la somme de $\mathrm{Im}\,(H)$ et de $\mathrm{Ker}\,(S)$ est égale à $\mathbb{R}^m$. On rappelle le théorème suivant, cf. [**35**, section 4.4, p.79]. Soient $\mathcal{U}'$ et $\mathcal{U}''$ deux sous-espaces arbitraires d'un espace vectoriel de dimension finie, alors la dimension de leur somme est donnée par

$$\dim\left(\mathcal{U}' + \mathcal{U}''\right) = \dim\left(\mathcal{U}'\right) + \dim\left(\mathcal{U}''\right) - \dim\left(\mathcal{U}' \cap \mathcal{U}''\right)\ .$$

L'application de cette relation aux sous-espaces $\mathrm{Im}\,(H)$ et $\mathrm{Ker}\,(S)$ donne

$$\begin{aligned}\dim\left(\mathrm{Im}\,(H) + \mathrm{Ker}\,(S)\right) &= \dim\left(\mathrm{Im}\,(H)\right) + \dim\left(\mathrm{Ker}\,(S)\right) - \dim\left(\mathrm{Im}\,(H) \cap \mathrm{Ker}\,(S)\right) \\ &= d + (m-d) + 0 = m\,.\end{aligned}$$

Les sous-espaces $\mathrm{Im}\,(H)$ et $\mathrm{Ker}\,(S)$ sont donc en somme directe et

$$\mathbb{R}^m = \mathrm{Im}\,(H) \oplus \mathrm{Ker}\,(S)\ . \tag{7.8.10}$$

Cette relation signifie que *tout vecteur* $\mathfrak{u} \in \mathbb{R}^m$ possède une *décomposition unique* de la forme (7.8.9).

Ce résultat permet maintenant de voir les solutions de (7.8.4) sous un autre angle. La relation (7.8.10) montre en effet que le choix de S est équivalent au choix d'un sous-espace $\mathcal{W}$ supplémentaire de $\mathrm{Im}\,(H)$ dans $\mathbb{R}^m$.

(1) D'une part, le choix d'une matrice S satisfaisant (7.8.4) détermine un sous-espace $W = \mathrm{Ker}\,(S)$ qui est un supplémentaire de $\mathrm{Im}\,(H)$ d'après la discussion ci-dessus.

(2) D'autre part, si $\mathcal{W}$ est un sous-espace de $\mathbb{R}^m$ tel que $\mathbb{R}^m = \mathrm{Im}\,(H) \oplus \mathcal{W}$, alors tout vecteur $\mathfrak{u} \in \mathbb{R}^m$ possède une *décomposition unique* de la forme $\mathfrak{u} = \mathfrak{u}_H + \mathfrak{u}_W$, $\mathfrak{u}_H \in \mathrm{Im}\,(H)$, $\mathfrak{u}_W \in \mathcal{W}$. D'ailleurs, $\mathfrak{u}_H \in \mathrm{Im}\,(H)$ signifie qu'il existe un $\boldsymbol{\sigma} \in \mathbb{R}^d$ tel que $\mathfrak{u}_H = H\boldsymbol{\sigma}$ et ce $\boldsymbol{\sigma}$ est unique en raison du rang de H. Il suffit maintenant de poser $S\mathfrak{u}_W = 0$ pour tout $\mathfrak{u}_W \in \mathcal{W}$ pour déterminer complètement les valeurs de S. En effet, il vient d'abord

$$S\mathfrak{u} = S\mathfrak{u}_H + S\mathfrak{u}_W = S\mathfrak{u}_H = SH\boldsymbol{\sigma} = \boldsymbol{\sigma}\,. \tag{7.8.11}$$

Ceci montre que le choix de $\mathcal{W} = \mathrm{Ker}\,(S)$ détermine complètement les valeurs de S et démontre le point (ii).

La multiplication de (7.8.11) par H à gauche

$$HS\mathfrak{u} = HS\left(\mathfrak{u}_H + \mathfrak{u}_W\right) = HS\mathfrak{u}_H = H\sigma = \mathfrak{u}_H$$

montre que HS est un projecteur sur le sous-espace $\mathrm{Im}\,(H)$. La matrice HS est effectivement idempotente en vertu de (7.8.4)

$$(HS)(HS) = H(SH)S = HS\,.$$

On peut donc interpréter HS comme un projecteur qui projette les vecteurs de $\mathbb{R}^m$ sur $\mathrm{Im}\,(H)$ parallèlement à $\mathcal{W} = \mathrm{Ker}\,(S)$.

Quelle est la forme générale de S dans cette approche ? Il suffit de déterminer d vecteurs libres qui forment une base du complément orthogonal $\mathcal{W}^\perp$ de $\mathcal{W}$. Le sous-espace $\mathcal{W}^\perp$ est déterminé de façon unique par $\mathcal{W}$. Ces d vecteurs forment les lignes d'une $d \times m$ matrice W et les colonnes de sa transposée W^t. La transposée W^t a des propriétés importantes. On a par définition $\mathcal{W} \subseteq \mathrm{Ker}\left(W^t\right)$. Le rang de W^t est égal à d, donc la dimension de $\mathrm{Ker}\left(W^t\right)$ est $m-d$, ce qui est la même dimension que celle de $\mathcal{W}$. Ceci prouve que $\mathrm{Ker}\left(W^t\right) = \mathcal{W}$. Ensuite, $\mathrm{Im}\,(H) \cap \mathcal{W} = (\mathbf{0})$ signifie que $\mathrm{Im}\,(H) \cap \mathrm{Ker}\left(W^t\right) = (\mathbf{0})$. Cela entraîne que la matrice W^tH est inversible car pour tout vecteur $\boldsymbol{\sigma} \in \mathbb{R}^d$, $\boldsymbol{\sigma} \neq \mathbf{0}$ il vient $W^tH\sigma \neq \mathbf{0}$. Ceci permet de former la matrice (7.8.8)

$$S_{\mathcal{W}} = \left(W^tH\right)^{-1} W^t$$

qui est visiblement une solution de (7.8.4) et dont le noyau est par définition égal à $\mathcal{W}$. Même si la matrice W^t n'est pas unique car le sous-espace $\mathcal{W}^\perp$ a une infinité de bases, la matrice S_W est unique car uniquement déterminée par le sous-espace $\mathcal{W}$ et la matrice H dans la relation (7.8.11). □

La forme générale (7.8.8) suggère plusieurs remarques. D'un côté, cette formule donne une méthode générale pour fabriquer des reconstructions consistantes de gradients. De l'autre côté, elle montre un problème potentiel de la reconstruction qui se manifeste parfois dans les calculs. Le sous-espace $\mathcal{W}$ peut être tel que la matrice W^tH soit presque singulière. Dans ce cas, la norme de son inverse $\left(W^tH\right)^{-1}$ peut devenir très grande. Ceci signifie que pour de "petits" vecteurs $\mathfrak{u}$ l'image $\left(W^tH\right)^{-1} W^t\mathfrak{u}$ peut devenir très grande. *La conséquence est que de petites fluctuations autour d'une cellule peuvent provoquer des gradients forts, ce qui n'est pas souhaitable dans une simulation numérique.*

Finalement, le fait que la reconstruction consistante du gradient soit identifiable à une projection sur un sous-espace pose une question supplémentaire. Quelle est la méthode associée à la projection orthogonale, c'est-à-dire à la projection parallèle à $(\mathrm{Im}\,(H))^\perp$? Dans ce cas, $\mathcal{W} = (\mathrm{Im}\,(H))^\perp$, donc $\mathcal{W}^\perp = \mathrm{Im}\,(H)$ et le choix $W = H$ est possible car les colonnes de H forment une base de $\mathcal{W}^\perp = \mathrm{Im}\,(H)$. La matrice d'interpolation devient la matrice

$$S = \left(H^tH\right)^{-1} H^t$$

dans laquelle on peut reconnaître la pseudo-inverse ou inverse de Moore-Penrose de H. La section 7.9.1 apporte pour $k = 1$ la preuve que cette solution particulière de (7.8.4) coïncide avec la reconstruction classique des moindres carrés.

7.9. Analyse de deux méthodes particulières pour la reconstruction linéaire par morceaux

Cette section présente deux méthodes particulières pour la reconstruction des polynômes de degré $k = 1$, c'est-à-dire pour la reconstruction linéaire par morceaux :

(1) La méthode des moindres carrés, présentée dans la section 7.9.1.

(2) Une méthode particulière, présentée dans la section 7.9.2, qui repose sur le théorème de Green.

Pour rappel, dans le cas $k = 1$, la forme générale des fonctions reconstruites (7.5.5) est

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \overline{u}_\alpha + \boldsymbol{\sigma}_\alpha[\mathfrak{u}] \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)$$

où $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$, donné par (7.8.2)

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta\,,$$

est un gradient consistant au sens de la définition 7.4.1 et $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$ est le vecteur des moyennes de cellule. Rappelons que $\boldsymbol{\sigma}_{\alpha\beta} \triangleq 0$ par définition dans (7.8.2) si la cellule $\mathcal{T}_\beta$ n'est pas dans le voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$.

7.9.1. Étude de la méthode des moindres carrés. Cette section présente un résultat sur la reconstruction du gradient par la méthode des moindres carrés et le lien de celle-ci avec l'interprétation algébrique de la reconstruction dans la section 7.8.

PROPOSITION 7.9.1 (Méthode des moindres carrés). *On suppose que la matrice H_α de (7.8.1) est de rang d et $\boldsymbol{\sigma}_\alpha \in \mathbb{R}^d$ est la solution du problème des moindres carrés donné par*

$$\min_{\boldsymbol{\sigma}\in\mathbb{R}^d} \left\{ \sum_{\beta\in\mathbb{W}_\alpha} \left(u_\beta - u_\alpha - \boldsymbol{h}_{\alpha\beta}\cdot\boldsymbol{\sigma}\right)^2 \right\} \tag{7.9.1}$$

où $\mathbb{W}_\alpha$ est un voisinage de la cellule $\mathcal{T}_\alpha$. Dans ces conditions $\boldsymbol{\sigma}_\alpha$ est unique et les coefficients $\boldsymbol{\sigma}_{\alpha\beta}$ forment les colonnes d'une matrice $\widetilde{S}_\alpha$ qui minimise la norme de Frobenius parmi les solutions de l'équation (7.8.4).

DÉMONSTRATION. La solution $\boldsymbol{\sigma}_\alpha$ du problème (7.9.1) satisfait

$$\sum_{\beta\in\mathbb{W}_\alpha} \boldsymbol{h}_{\alpha\beta}\left(\boldsymbol{h}_{\alpha\beta}\cdot\boldsymbol{\sigma}\right) = \sum_{\beta\in\mathbb{W}_\alpha} \boldsymbol{h}_{\alpha\beta}\left(u_\beta - u_\alpha\right) . \tag{7.9.2}$$

La définition $\delta\mathfrak{u} \triangleq \left(u_{\beta_1} - u_\alpha, \ldots, u_{\beta_{m_\alpha}} - u_\alpha\right)$ permet d'écrire la solution unique de (7.9.2) comme

$$\boldsymbol{\sigma} = S_\alpha \delta\mathfrak{u} \tag{7.9.3}$$

où la matrice des moindres carrés est définie par

$$S_\alpha = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t . \tag{7.9.4}$$

Les paramètres non nuls de la reconstruction du gradient sont donnés par les colonnes de la matrice S_α

$$\boldsymbol{\sigma}_{\alpha\beta} = \begin{cases} \left(H_\alpha^t H_\alpha\right)^{-1} \boldsymbol{h}_{\alpha\beta} & \beta \in \mathbb{W}_\alpha \\ 0 & \beta \notin \mathbb{W}_\alpha \end{cases} . \tag{7.9.5}$$

L'inverse de la matrice $H_\alpha^t H_\alpha$ dans (7.9.3) existe en raison du rang de H_α. La forme de la matrice S_α dans (7.9.3) montre qu'elle est la pseudo-inverse $H_\alpha^\dagger$ de H_α. La pseudo-inverse $H_\alpha^\dagger$, qui porte également le nom d'inverse de *Moore-Penrose,* est la matrice qui minimise la norme de Frobenius

$$\left\|S_\alpha\right\|_F \triangleq \sqrt{\operatorname{trace}\left(S_\alpha^t S_\alpha\right)}$$

parmi les solutions de (7.8.4), d'après la proposition 7.7.2. □

7.9.2. Étude de la méthode de Green. Une autre méthode pour reconstruire des gradients consistants au sens de la définition 7.4.1 s'obtient par le théorème de Green, cf. [**78**]. Elle utilise la définition des vecteurs $\boldsymbol{j}_{\alpha\beta}$ et $\boldsymbol{b}_{\alpha\beta}$ de la section 5.3 et un point auxiliaire $\boldsymbol{y}_{\alpha\beta} \triangleq \boldsymbol{x}_\alpha + \boldsymbol{j}_{\alpha\beta}$. L'idée est de considérer un polynôme de degré un $\boldsymbol{x} \mapsto u_\alpha + \boldsymbol{\sigma}\cdot(\boldsymbol{x} - \boldsymbol{x}_\alpha)$ et de dériver une formule pour son gradient $\boldsymbol{\sigma}$ par le théorème de Green

$$\int_{\mathcal{T}_\alpha} \boldsymbol{\nabla} u\left(\boldsymbol{x}\right)\, dx = \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}\left(\boldsymbol{x}\right) u\left(\boldsymbol{x}\right)\, d\sigma \tag{7.9.6}$$

que u doit satisfaire. De cette façon, la formule pour le gradient est consistante par définition. La forme de u entraîne $\boldsymbol{\nabla} u\left(\boldsymbol{x}\right) = \boldsymbol{\sigma}$, $u\left(\boldsymbol{x}\right) = u\left(\boldsymbol{y}_{\alpha\beta}\right) + \boldsymbol{\sigma}\cdot\left(\boldsymbol{x} - \boldsymbol{y}_{\alpha\beta}\right)$ et $\overline{u}_\alpha = u\left(\boldsymbol{x}_\alpha\right) = u_\alpha$ où $\overline{u}_\alpha$ est la moyenne de u sur la cellule $\mathcal{T}_\alpha$. La valeur $u\left(\boldsymbol{y}_{\alpha\beta}\right)$ satisfait

$$u\left(\boldsymbol{y}_{\alpha\beta}\right) = u\left(\boldsymbol{x}_\alpha + \boldsymbol{j}_{\alpha\beta}\right) = u\left(\boldsymbol{x}_\alpha\right) + \frac{\left\|\boldsymbol{j}_{\alpha\beta}\right\|}{\left\|\boldsymbol{h}_{\alpha\beta}\right\|}\left(u\left(\boldsymbol{x}_\beta\right) - u\left(\boldsymbol{x}_\alpha\right)\right) .$$

L'équation (7.9.6) devient

$$|\mathcal{T}_\alpha|\,\boldsymbol{\sigma} = \sum_\beta \boldsymbol{a}_{\alpha\beta} u_\alpha + \sum_\beta \boldsymbol{a}_{\alpha\beta} \frac{\|\boldsymbol{j}_{\alpha\beta}\|}{\|\boldsymbol{h}_{\alpha\beta}\|} (u_\beta - u_\alpha) + \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}(\boldsymbol{x}) \left[\boldsymbol{\sigma}\cdot\left(\boldsymbol{x}-\boldsymbol{y}_{\alpha\beta}\right)\right] d\sigma\,. \quad (7.9.7)$$

Le premier terme dans le membre de droite de (7.9.7) est nul en raison de (5.5.2). La notation $\boldsymbol{y}_{\alpha\beta} - \boldsymbol{x}_\alpha = \boldsymbol{j}_{\alpha\beta}$ permet d'écrire l'identité

$$\begin{aligned}|\mathcal{T}_\alpha|\,\boldsymbol{\sigma} = \int_{\mathcal{T}_\alpha} \boldsymbol{\nabla}\left[\boldsymbol{\sigma}\cdot(\boldsymbol{x}-\boldsymbol{x}_\alpha)\right] dx = \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}(\boldsymbol{x}) \left[\boldsymbol{\sigma}\cdot\left(\boldsymbol{x}-\boldsymbol{y}_{\alpha\beta}+\boldsymbol{y}_{\alpha\beta}-\boldsymbol{x}_\alpha\right)\right] d\sigma = \\ = \sum_\beta \boldsymbol{a}_{\alpha\beta}\left(\boldsymbol{\sigma}\cdot\boldsymbol{j}_{\alpha\beta}\right) + \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}(\boldsymbol{x}) \left[\boldsymbol{\sigma}\cdot\left(\boldsymbol{x}-\boldsymbol{y}_{\alpha\beta}\right)\right] d\sigma\,. \quad (7.9.8)\end{aligned}$$

L'insertion de (7.9.8) dans (7.9.7) et la colinéarité de $\boldsymbol{j}_{\alpha\beta}$ et $\boldsymbol{h}_{\alpha\beta}$ donnent une relation entre $\boldsymbol{\sigma}$ et le vecteur $\delta\mathfrak{u}$ de composantes $u_\beta - u_\alpha$

$$\sum_\beta \frac{\|\boldsymbol{j}_{\alpha\beta}\|}{\|\boldsymbol{h}_{\alpha\beta}\|} (\boldsymbol{a}_{\alpha\beta}\otimes\boldsymbol{h}_{\alpha\beta})\,\boldsymbol{\sigma} \;=\; \sum_\beta \frac{\|\boldsymbol{j}_{\alpha\beta}\|}{\|\boldsymbol{h}_{\alpha\beta}\|} \boldsymbol{a}_{\alpha\beta}(u_\beta - u_\alpha)\,. \quad (7.9.9)$$

La définition

$$\boldsymbol{a}'_{\alpha\beta} \triangleq \frac{\|\boldsymbol{j}_{\alpha\beta}\|}{\|\boldsymbol{h}_{\alpha\beta}\|} \boldsymbol{a}_{\alpha\beta}$$

simplifie la forme de (7.9.9)

$$\sum_\beta (\boldsymbol{a}'_{\alpha\beta}\otimes\boldsymbol{h}_{\alpha\beta})\,\boldsymbol{\sigma} = \sum_\beta \boldsymbol{a}'_{\alpha\beta}(u_\beta - u_\alpha)\,. \quad (7.9.10)$$

Finalement, la définition de la matrice N_α dont les lignes sont les vecteurs $\boldsymbol{a}'_{\alpha\beta}$ permet d'écrire la formule de Green (7.9.10) sous la forme

$$\boldsymbol{\sigma} = S_\alpha \delta\mathfrak{u} = \left(N_\alpha^t H_\alpha\right)^{-1} N_\alpha^t \delta\mathfrak{u}\,. \quad (7.9.11)$$

La matrice S_α dans (7.9.11) est bien de la forme (7.8.8) avec $W = N_\alpha$.

7.10. Étude numérique

L'étude théorique de la reconstruction en maillage non structuré général laisse certaines questions ouvertes. Pour les élucider, il faut mener des expériences numériques. On rappelle que le problème de trouver des reconstructions consistantes de degré k se réduit à la résolution de l'équation matricielle (7.6.3)

$$S_\alpha H_\alpha = \mathrm{I}_{\binom{k+d}{d}-1}$$

pour la matrice inconnue S_α. La matrice géométrique H_α, introduite par la définition 7.6.3, exprime la géométrie de la reconstruction au voisinage de la cellule $\mathcal{T}_\alpha$.

Dans le contexte considéré ici, il faut notamment examiner les points suivants :

(1) Existence de reconstructions consistantes de degré k : On peut se demander si les logiciels utilisés dans les milieux industriel et scientifique génèrent des maillages qui admettent des reconstructions consistantes de degré un, deux et trois sur certains types de voisinages. Pour cela, il suffit de vérifier que la solution des moindres carrés de (7.6.3), donnée par la pseudo-inverse (7.6.4) de H_α

$$S_\alpha = H_\alpha^\dagger \triangleq \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t\,,$$

existe dans chaque cellule du maillage.

(2) Dérivées des polynômes reconstruits : Soit $w_\alpha[\mathfrak{u}](\boldsymbol{x})$ le polynôme reconstruit dans la cellule $\mathcal{T}_\alpha$. D'après la définition 7.6.2, les éléments de la matrice S_α sont les dérivées du polynôme $w_\alpha[\mathfrak{u}](\boldsymbol{x})$ au point $\boldsymbol{x}_\alpha$. Plus les éléments de S_α sont grands, plus les dérivées de $w_\alpha[\mathfrak{u}](\boldsymbol{x})$ en $\boldsymbol{x}_\alpha$ sont grandes. Puisqu'il semble préférable que les dérivées de $w_\alpha[\mathfrak{u}](\boldsymbol{x})$ soient aussi petites que possible, il est intéressant d'analyser si les éléments

de S_α sont plus grandes sur certains types de maillages que sur d'autres. La façon la plus simple consiste à calculer et comparer des normes de S_α, par exemple la norme de Frobenius $\|S_\alpha\|_F$. Le problème de cette approche réside dans le fait que les matrices S_α ne sont pas invariantes par des changements d'échelle du maillage, ce qui rend difficile la comparaison de $\|S_\alpha\|_F$ entre différents maillages. Le chapitre 10 présente un nouveau critère adimensionnel, invariant par des changements d'échelle, qui est un meilleur indicateur pour ce problème spécifique de reconstruction. L'étude de cette question est donc reportée au chapitre 10.

(3) Taux de convergence réel en maillage non structuré général : La discussion de la section 7.2.3 montre que la seule condition de consistance de degré k ne garantit pas que la reconstruction est d'ordre $k+1$ en maillage non structuré. Il faut par conséquent tester si la reconstruction obéit à une estimation (6.3.3)

$$|w_\alpha[\mathfrak{u}](\boldsymbol{x}) - u(\boldsymbol{x})| \leq \mathrm{C}_u\, h^{k+1} = \mathrm{O}\left(h^{k+1}\right) \text{ pour tout } \boldsymbol{x} \in \mathcal{T}_\alpha$$

dans toutes les cellules $\mathcal{T}_\alpha$ du maillage.

Les tests de reconstruction ont été effectués ensemble avec les tests de stabilité et de précision des chapitres 10 et 11. Les tests couvrent la reconstruction des polynômes de degré un, deux et trois, par la méthode des moindres carrés, voir la définition 7.7.1, sur plusieurs types de maillages non structurés et structurés. Les domaines de calcul sont un carré et un cube avec des conditions de périodicité au bord. Les cas tests incluent les maillages suivants :

(1) Des maillages de tétraèdres en dimension trois et des maillages de triangles en dimension deux.

(2) Des maillages mixtes constitués de tétraèdres et de prismes en dimension trois.

(3) Des maillages mixtes constitués de triangles et de quadrilatères en dimension deux.

(4) Des maillages cartésiens et des maillages cartésiens déformés en dimension deux et trois.

Les maillages utilisés ont été générés avec les logiciels Centaur et Gmsh.

Les résultats données dans les sections 10.9 et 11.3 permettent de conclure que les reconstructions suivantes sont possibles sur les maillages testés et cités ci-dessus :

(1) Reconstruction de polynômes de degré trois sur le troisième et quatrième voisinage.

(2) Reconstruction de polynômes de degré deux sur le deuxième et troisième voisinage.

(3) Reconstruction de polynômes de degré un sur le premier et deuxième voisinage.

L'étude numérique des taux de convergence est restreinte à la dimension deux car les tests en maillage de tétraèdres demandent un nombre élevé de cellules pour obtenir une précision suffisante. L'outil de test n'était pas suffisamment performant pour faire des tests en dimension trois dans des conditions satisfaisantes.

On se place sur un carré

$$\Omega = \left\{ (x,y) \in \mathbb{R}^2 \middle| \, 0 \leq x \leq 1\,,\, 0 \leq y \leq 1 \right\}$$

avec des conditions de périodicité au bord. Comme fonction test on choisit la fonction périodique

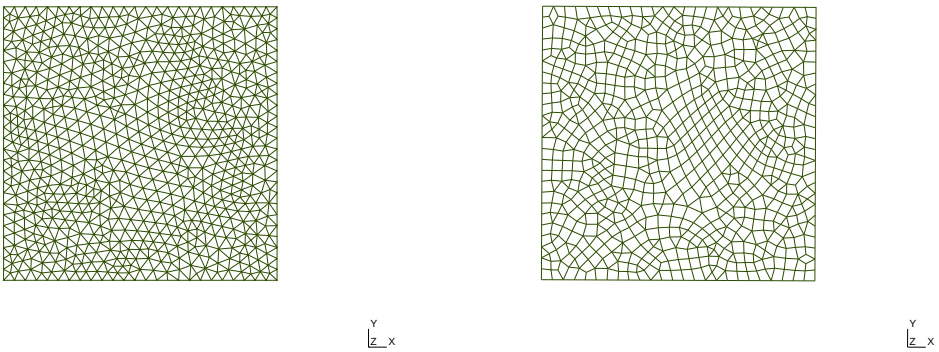
$$u_0(x,y) = \sin(2\pi x)\sin(2\pi y) \tag{7.10.1}$$

de moyennes de cellule $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$. A partir de ces moyennes, on reconstruit dans chaque cellule un polynôme $w_\alpha[\mathfrak{u}](\boldsymbol{x})$ et on calcule l'erreur d'approximation

$$\varepsilon = \frac{\sqrt{\sum_{\alpha=1}^{N} \int_{\mathcal{T}_\alpha} |w_\alpha[\mathfrak{u}](\boldsymbol{x}) - u_0(\boldsymbol{x})|^2 \, dx}}{\sqrt{\sum_{\alpha=1}^{N} \int_{\mathcal{T}_\alpha} |u_0(\boldsymbol{x})|^2 \, dx}} \tag{7.10.2}$$

par des formules d'intégration numérique d'une précision suffisante.

La figure 7.10.1 montre deux exemples de maillages non structurés utilisés pour les tests et générés avec le logiciel Gmsh. Pour le maillage le plus grossier, soit ε_0 l'erreur (7.10.2) et h_0 le diamètre du maillage. Lorsqu'on dessine $\log(\varepsilon/\varepsilon_0)$ en fonction de $\log(h/h_0)$ pour les différents



(a) Maillage de triangles numéro 4 (b) Maillage hybride numéro 4

FIG. 7.10.1: Deux exemples de maillages non structurés en dimension deux.

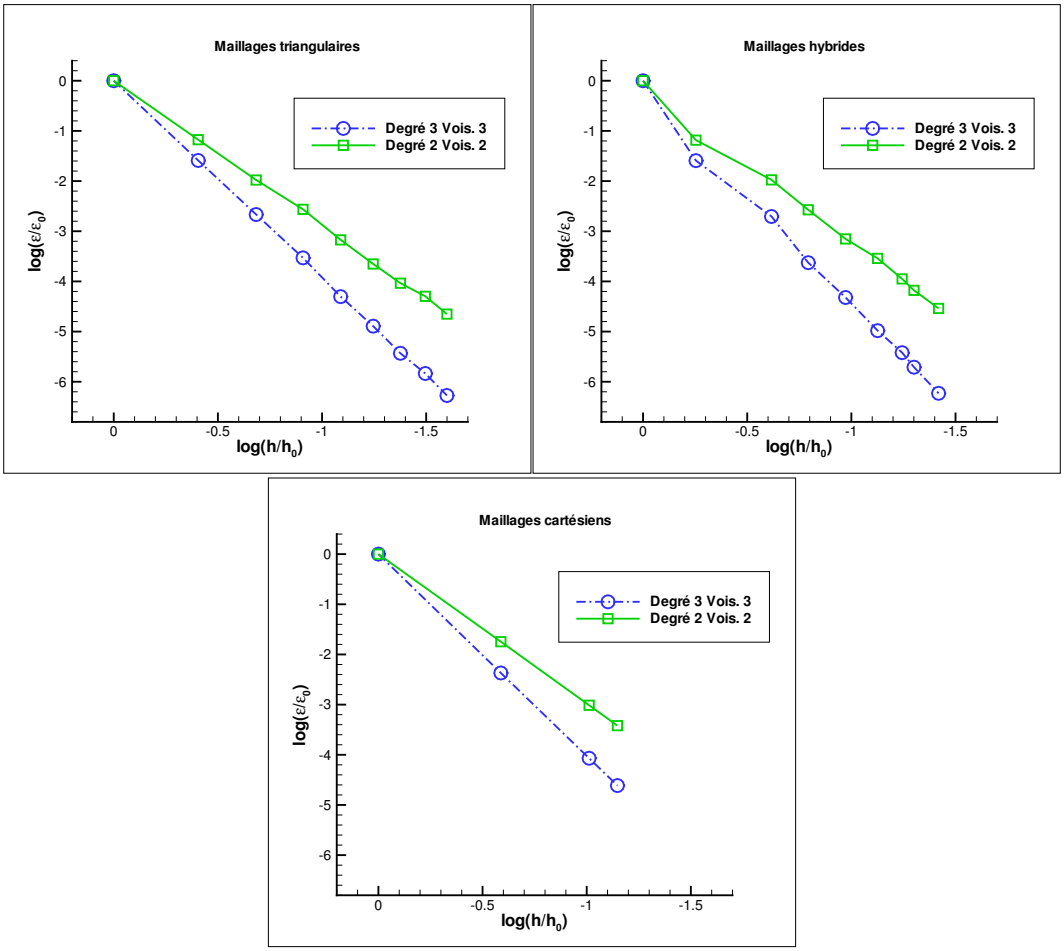


FIG. 7.10.2: Taux de convergence de la reconstruction en dimension deux

maillages testés, on obtient une courbe dont la pente est une mesure pour le taux de convergence de l'erreur d'approximation en fonction du diamètre des mailles. La figure 7.10.2 montre ces courbes pour des maillages triangulaires, des maillages hybrides composés de triangles et de quadrangles et des maillages cartésiens. On constate que les pentes sont effectivement de -4 pour la reconstruction de degré trois et de -3 pour la reconstruction de degré deux, et cela pour les trois types de maillages testés. Dans les cas testés, la reconstruction satisfait par conséquent une estimation du type (6.3.3).

7.11. Bilan du chapitre

L'étude a permis de revoir la théorie de la reconstruction locale en maillage non structuré général et de préciser la forme générale des reconstructions locales consistantes. L'objectif de la reconstruction est de déterminer, dans chaque cellule $\mathcal{T}_\alpha$, à partir de moyennes de maille ou de valeurs ponctuelles au voisinage de la cellule, une fonction qui sert à interpoler des flux précis aux interfaces entre les cellules. L'approche de la reconstruction étudiée ici repose sur trois principes fondamentaux :

(1) La fonction reconstruite dépend linéairement des moyennes de cellule ou des valeurs ponctuelles.

(2) Les fonctions reconstruites sont des polynômes de degré k.

(3) La reconstruction est locale, c'est-à-dire seules les moyennes de cellule dans un voisinage de la cellule contribuent à la reconstruction.

Ces principes permettent de représenter la reconstruction par un opérateur linéaire $\mathfrak{R}_\alpha$, de la forme générale (7.2.4), qui envoie les moyennes de maille ou les valeurs ponctuelles sur le polynôme reconstruit. Dans le cas de la reconstruction à partir des moyennes de cellule, les fonctions reconstruites sont alors de la forme (7.2.5)

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \sum_\beta w^{\alpha\beta}(\boldsymbol{x})\ \overline{u}_\beta$$

où les $w^{\alpha\beta}(\boldsymbol{x})$ sont des polynômes réels et $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$.

La discussion du chapitre 6 montre que la reconstruction améliore la précision du schéma si elle satisfait une propriété d'approximation du type (6.3.3). Cette propriété est formalisée par la définition 7.2.3 : on dit que la reconstruction est d'ordre $k+1$ si la condition suivante est satisfaite : pour toute fonction suffisamment régulière u, dont les moyennes de cellule sont $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$, il doit exister une constante C_u, indépendante de h et de la cellule $\mathcal{T}_\alpha$, telle que l'estimation

$$\left|w_\alpha\left[\mathfrak{u}(t_0)\right](\boldsymbol{x}) - u(\boldsymbol{x}, t_0)\right| \leq \mathrm{C}_u\, h^{k+1} = \mathrm{O}\left(h^{k+1}\right) \text{ pour tout } \boldsymbol{x} \in \mathcal{T}_\alpha$$

soit vraie dans toutes les cellules $\mathcal{T}_\alpha$ du maillage.

Il faut alors trouver un critère pour que la reconstruction soit d'ordre $k+1$. Le critère considéré ici consiste à exiger que la reconstruction reproduise les polynômes de degré k. On appelle une telle reconstruction une *reconstruction consistante de degré k*. Cette notion, formalisée par la définition 7.2.1, coïncide avec la notion de reconstruction k-exacte introduite par Barth et Frederickson dans [**9**]. L'opérateur de moyenne $\mathfrak{A}_\alpha$, défini par (7.2.8) et l'opérateur de l'évaluation ponctuelle $\mathfrak{D}_\alpha$, défini par (7.2.9) ont permis de reformuler la question de l'existence de reconstruction consistantes :

(1) Dans le cas de la reconstruction à partir de valeurs moyennes, le critère de consistance est équivalent à la condition que l'opérateur $\mathfrak{R}_\alpha$ est un inverse à gauche de $\mathfrak{A}_\alpha$ sur l'espace des polynômes $\mathbb{P}_k\left(\mathbb{R}^d\right)$.

(2) Dans le cas de la reconstruction à partir de valeurs ponctuelles, le critère de consistance est équivalent à la condition que l'opérateur $\mathfrak{R}_\alpha$ est un inverse à gauche de $\mathfrak{D}_\alpha$ sur l'espace des polynômes $\mathbb{P}_k\left(\mathbb{R}^d\right)$.

La proposition 7.2.2 montre que l'opérateur $\mathfrak{A}_\alpha$ (ou $\mathfrak{D}_\alpha$) admet des inverses à gauche si et seulement s'il est injectif. Pour que les opérateurs $\mathfrak{A}_\alpha$ et $\mathfrak{D}_\alpha$ puissent être injectifs, il faut que la taille du voisinage de reconstruction satisfasse l'inégalité (7.2.14). Cette inégalité constitue par conséquent une condition nécessaire pour la reproduction des polynômes de degré k. Elle est en même temps un obstacle majeur pour l'implémentation des algorithmes car la reconstruction de polynômes de degré élevé nécessite des voisinages très larges.

La discussion à la fin de la section 7.2.2 aborde la question des conditions suffisantes pour l'injectivité de $\mathfrak{A}_\alpha$ et $\mathfrak{D}_\alpha$. Dans le cas de la reconstruction à partir de valeurs ponctuelles, il existe des conditions géométriques simples pour l'injectivité de $\mathfrak{D}_\alpha$, cf. le premier chapitre de [**120**] pour $\mathbb{R}^2$. Dans le cas de la reconstruction à partir de moyennes de cellule, l'article [**121**] fournit une preuve pour l'injectivité de $\mathfrak{A}_\alpha$. Cette preuve suppose cependant que les voisinages de reconstruction soient suffisamment larges, ce qui rend ce résultat peu intéressant pour les applications considérées. Il faut en effet connaître la taille exacte des voisinages nécessaires à la reconstruction pour implémenter des algorithmes. C'est pourquoi cette étude a recours aux tests numériques pour confirmer l'injectivité de l'opérateur $\mathfrak{A}_\alpha$. Les tests effectués ont permis de confirmer que plusieurs types des maillages, produits par des logiciels Centaur et Gmsh, admettent la reproduction des polynômes de degré un, deux et trois, sur différents types de voisinage, cf. section 7.10.

Il est alors important de noter qu'en maillage non structuré général, la seule condition de consistance de degré k n'est pas suffisante pour que la reconstruction soit d'ordre $k+1$ au sens de la définition 7.2.3. Il faut par conséquent chercher des conditions supplémentaires qui garantissent que la reconstruction soit d'ordre $k+1$, ce qui fait l'objet de la section 7.2.3. Le théorème 7.2.5 établit une condition pour qu'une reconstruction consistante de degré k soit d'ordre $k+1$: il suffit qu'il existe une constante $\mathrm{C}_{\mathrm{reg}}$ telle que les fonctions coefficients $w^{\alpha\beta}(\boldsymbol{x})$ dans (7.2.5) satisfont la condition (7.2.24)

$$\left|w^{\alpha\beta}(\boldsymbol{x})\right| \leq \mathrm{C}_{\mathrm{reg}}, \text{ pour tout } \boldsymbol{x} \in \mathcal{T}_\alpha\,, \text{ pour tout } 1 \leq \alpha \leq N\,, \text{ lorsque } h \longrightarrow 0\,.$$

La définition 7.2.4 introduit le terme de *reconstruction régulière* pour désigner les reconstructions qui satisfont une telle condition.

Cette discussion a permis de dégager deux conditions distinctes dont la combinaison suffit pour que la reconstruction soit d'ordre $k+1$:

(1) La reconstruction doit être consistante de degré k au sens de la définition 7.2.1.

(2) La reconstruction doit être régulière au sens de la définition 7.2.4.

La discussion à la fin de la section 7.2.3 aborde brièvement la question de conditions suffisantes pour la condition (7.2.24) sur maillage non structuré. Cette question touche directement à la théorie de l'approximation de données en maillage irrégulier, voir par exemple [**121, 120**]. Cependant, cette question théorique est très difficile et ne rentre donc pas dans le cadre de cette thèse. Pour cette raison, on suppose ici que les reconstructions considérées sont régulières, ce qui est exprimé par l'hypothèse 7.2.6.

Pour transcrire la condition de consistance en une équation d'algèbre linéaire, il est indispensable de choisir une base de l'espace des polynômes. Le choix considéré ici est la base (7.3.2) de fonctions centrées sur le barycentre de la cellule $\mathcal{T}_\alpha$. Dans cette base, la forme générale des polynômes reconstruits est (7.3.7)

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \sum_\beta \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{w}_{\alpha\beta}^{(l)} \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)^l \; \overline{u}_\beta$$

où les coefficients $\boldsymbol{w}_{\alpha\beta}^{(l)}$ sont des tenseurs symétriques d'ordre l et $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$ est le vecteur des moyennes de cellule.

Dans cette base, la condition de consistance, introduite par la définition 7.2.1, s'exprime, dans le cas de la reconstruction à partir de moyennes de cellule, par les équations d'algèbre linéaire (7.3.9). L'étude de la section 7.4 montre que le coefficient $\boldsymbol{w}_{\alpha\beta}^{(l)}$ peut être interprété comme une approximation de la dérivée d'ordre l en $\boldsymbol{x}_\alpha$. Le polynôme reconstruit (7.3.6) s'interprète

donc comme un polynôme de Taylor approché dont les coefficients sont des approximations des dérivées exactes. Ce point de vue correspond à la notion de "Truncated Taylor Series Expansion Reconstruction" de Delanaye [**44**].

Dans le cas de la reconstruction à partir de valeurs ponctuelles, la condition de consistance s'exprime par les équations (7.3.23) qui ont la même structure mathématique que les conditions (7.3.9). Il n'y a donc pas de différence mathématique fondamentale entre les conditions de consistance pour la reconstruction à partir de moyennes de cellule et pour la reconstruction à partir de valeurs ponctuelles.

La prochaine étape consiste à imposer la condition de conservativité (6.3.8) à la reconstruction. L'analyse de la section 7.5 montre que cette condition supplémentaire est compatible avec la condition de consistance et conduit même à une simplification de la forme des fonctions reconstruites, dans le cas de la reconstruction à partir de moyennes de cellule. Il n'est pas évident d'imposer cette condition à une reconstruction à partir de valeurs ponctuelles, ce qui justifie de poursuivre l'analyse uniquement pour la reconstruction à partir de moyennes.

Pour continuer l'étude, il a été nécessaire de transcrire les équations (7.3.9) en une équation matricielle. La matrice des inconnues S_α, introduite par la définition 7.6.2, et la matrice des coefficients H_α, introduite par la définition 7.6.3 permettent d'écrire les équations (7.3.9) sous la forme de l'équation matricielle (7.6.3)

$$S_\alpha H_\alpha = \mathrm{I}_{\binom{k+d}{d}-1}\,.$$

La forme de la condition (7.6.3) permet de définir de façon naturelle la reconstruction par la méthode des moindres carrés, donnée par la solution particulière

$$H_\alpha^\dagger \triangleq \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t$$

de (7.6.3), cf. la définition 7.7.1. Elle est également appelée méthode de la pseudo-inverse. Cette solution particulière minimise la norme de Frobenius parmi les solutions de (7.6.3). Elle donne lieu à une remarque supplémentaire sur la question des reconstructions régulières, voir la remarque 7.7.3.

La forme de (7.6.3) conduit à deux interprétations distinctes de la condition de consistance. Le théorème 7.8.1 permet de voir l'ensemble des solutions de (7.6.3) comme un sous-espace affine d'un espace de matrices. Cela permet d'établir la forme (7.8.5) de la solution générale

$$S_\alpha = \widetilde{S}_\alpha + \Lambda_\alpha B_\alpha\,.$$

Dans (7.8.5), $\widetilde{S}_\alpha$ est une solution particulière de (7.6.3), la matrice B_α est une solution de l'équation homogène (7.8.6)

$$B_\alpha H_\alpha = 0\,,$$

telle que le rang de B_α soit maximal et Λ_α est une matrice arbitraire.

Le théorème 7.8.2 permet de donner une autre interprétation géométrique de la solution générale de (7.6.3) :

(1) Le choix d'une solution S_α de (7.6.3) est équivalent au choix d'un sous-espace $\mathcal{W}_\alpha = \mathrm{Ker}\,(S_\alpha)$ supplémentaire de $\mathrm{Im}\,(H_\alpha)$ dans $\mathbb{R}^{m_\alpha}$.

(2) L'application $H_\alpha S_\alpha : \mathbb{R}^{m_\alpha} \longrightarrow \mathbb{R}^{m_\alpha}$ est un projecteur qui projette les vecteurs de $\mathbb{R}^{m_\alpha}$ parallèlement à $\mathcal{W}_\alpha = \mathrm{Ker}\,(S_\alpha)$ sur $\mathrm{Im}\,(H_\alpha)$.

(3) La forme générale (7.8.8) de la solution est donnée par

$$S_{\mathcal{W}_\alpha} = \left(W_\alpha^t H_\alpha\right)^{-1} W_\alpha^t$$

où $W_\alpha \in \mathbb{M}_{m,d}\,(\mathbb{R})$ est une matrice dont les lignes forment une base du complément orthogonal $\mathcal{W}_\alpha^\perp$ de $\mathcal{W}_\alpha$. La solution $S_{\mathcal{W}_\alpha}$ ne dépend que de la matrice H_α et du sous-espace $\mathcal{W}_\alpha$ et non pas de la matrice particulière W_α.

La reconstruction des moindres carrés s'identifie alors au cas particulier $W_\alpha = H_\alpha$.

Pour le cas $k = 1$, la section 7.9.2 présente une méthode de reconstruction, basée sur le théorème de Green, qui s'écrit sous la forme

$$S_\alpha = \left(N_\alpha^t H_\alpha\right)^{-1} N_\alpha^t$$

où la matrice N_α dépend des vecteurs $\boldsymbol{a}_{\alpha\beta}$, $\boldsymbol{h}_{\alpha\beta}$ et $\boldsymbol{j}_{\alpha\beta}$. Il s'agit donc d'une méthode particulière de reconstruction pour laquelle $W = N_\alpha$.

La forme (7.8.8) de la solution met en évidence un problème potentiel : si la matrice $W_\alpha^t H_\alpha$ est presque singulière, la matrice inverse $\left(W_\alpha^t H_\alpha\right)^{-1}$ peut devenir très grande. De petites fluctuations dans la solution peuvent alors générer des reconstructions ayant des dérivées trop fortes. Il y a également la possibilité que la matrice $W_\alpha^t H_\alpha$ soit singulière, ce qui a été constaté dans CEDRE pour la méthode de Green mentionnée ci-dessus.

Le chapitre a permis de formuler la reconstruction de polynômes de degré k pour des schémas volumes finis en maillage non structuré. Quelques principes fondamentaux ont permis de transcrire les conditions de reconstruction en l'équation matricielle (7.6.3) et d'établir les formes (7.8.5) et (7.8.8) de la solution générale de (7.6.3).

Cette étude est centrée sur l'aspect algébrique des conditions de reconstruction. Par conséquent, il faut la compléter par rapport aux points suivants :

(1) Pour prouver que la condition de consistance conduit à une estimation du type (6.3.3)

$$|w_\alpha \left[\mathfrak{u}\left(t_0\right)\right]\left(\boldsymbol{x}\right) - u\left(\boldsymbol{x}, t_0\right)| = \mathrm{O}\left(h^{k+1}\right)\ , \boldsymbol{x} \in \mathcal{T}_\alpha\ ,$$

il a fallu faire l'hypothèse 7.2.6 sur les propriétés d'approximation des polynômes en maillage non structuré. Une analyse approfondie de cette hypothèse dans le cadre de la théorie d'approximation en maillage non structuré représente un travail trop important pour être traitée dans le cadre de cette thèse.

(2) L'étude de ce chapitre a fait abstraction des aspects d'implémentation des méthodes de reconstruction. Cette question fait l'objet du chapitre 8 qui est dédié à le recherche de méthodes de reconstruction spécialement conçues pour une implémentation rapide et efficace.

CHAPITRE 8

Étude de méthodes de reconstruction compactes

8.1. Objectif du chapitre

Les sections 6.3 et 7.2 montrent que des reconstructions basées sur la reproduction des polynômes de degré k donnent des schémas volumes finis (6.3.4) dont l'erreur de troncature est au moins d'ordre k. Il paraît donc à première vue souhaitable d'utiliser des polynômes du degré k le plus élevé possible. L'objectif de cette section est de mettre en évidence les problèmes que les reconstructions de degré élevé créent pour l'implémentation des algorithmes sur ordinateur. Cela donne lieu à de nouveaux types d'algorithmes de reconstruction particulièrement faciles à implémenter. Ces algorithmes font l'objet des sections 8.2 et 8.3.

Les conditions (7.2.15) et (7.2.16) montrent que la taille minimale des voisinages de reconstruction augmente en dimensions deux et trois de façon quadratique et cubique avec le degré des polynômes. La grande taille des voisinages est un problème pour l'implémentation d'algorithmes informatiques efficaces car il faut calculer, trier, stocker et lire les données de connectivité entre les cellules. Cette situation constitue un obstacle sérieux à l'augmentation de la précision des méthodes de type volumes finis. Un problème supplémentaire est le découpage des maillages en sous-domaines pour le calcul parallèle, ce qui nécessite de transmettre les informations des cellules voisines à travers les frontières entre les domaines. L'implémentation informatique est plus simple si l'échange d'informations se fait uniquement entre cellules adjacentes, c'est-à-dire entre cellules voisines du premier voisinage.

Les schémas numériques de type volumes finis se trouvent par conséquent face à deux objectifs contradictoires.

(1) D'une part, il faut augmenter la taille des voisinages pour augmenter la robustesse et la précision.

(2) D'autre part, il faut diminuer la taille des voisinages pour rendre la réalisation des algorithmes sur ordinateur plus rapide, plus souple et aussi plus évolutive.

Une solution consiste à concevoir des algorithmes capables d'obtenir l'information de cellules éloignées tout en échangeant des données entre premiers voisins. Une possibilité pour réaliser un tel algorithme est de travailler par itérations.

(1) Lors d'une première itération, chaque cellule reçoit les informations de ses premiers voisins.

(2) Une deuxième itération permet ensuite d'accéder à l'information des premiers voisins des premiers voisins, donc des deuxièmes voisins de la cellule.

Ce processus peut continuer itérativement. La difficulté est d'assembler correctement cette information pour réaliser des reconstructions précises.

Dans les sections 8.2 et 8.3, on explore deux méthodes différentes. La première est appelée *méthode des moindres carrés couplés* et a été proposée dans [**40**]. Elle est présentée dans la section 8.2. La deuxième méthode est appelée *méthode des corrections successives*. Elle fait l'objet de la section 8.3 et a été introduite dans [**20**].

8.2. Analyse de la méthode des moindres carrés couplés

Cette méthode [**40**] repose sur le fait que les dérivées successives d'un polynôme p de degré k satisfont les relations

$$\overline{p}_\beta - \overline{p}_\alpha = \sum_{i=1}^{k} \frac{1}{i!}\, D^{(i)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \left[\boldsymbol{z}_{\alpha\beta}^{(i)} - \boldsymbol{x}_\alpha^{(i)}\right] ,\ \beta \in \mathbb{W}_\alpha \tag{8.2.1}$$

$$D^{(j)}p\Big|_{\boldsymbol{x}_\beta} = \sum_{l=j}^{k} \frac{1}{(l-j)!}\, D^{(l)}p\Big|_{\boldsymbol{x}_\alpha} \cdot \boldsymbol{h}_{\alpha\beta}^{l-j} ,\ 1 \le j \le k-1 ,\ \beta \in \mathbb{W}_\alpha \tag{8.2.2}$$

où $\mathbb{W}_\alpha$ est pour le moment un voisinage non spécifié de la cellule $\mathcal{T}_\alpha$. La relation (8.2.1) détermine les différences des moyennes de p sur le voisinage de la cellule $\mathcal{T}_\alpha$ en fonction des dérivées de p au barycentre de $\mathcal{T}_\alpha$. La condition (8.2.2) relie les dérivées de p au barycentre de $\mathcal{T}_\alpha$ avec les dérivées de p aux barycentres des cellules voisines $\mathcal{T}_\beta$.

Considérons une fonction u, suffisamment régulière, dont les moyennes de cellule sont $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$. On souhaite approcher les dérivées de u en $\boldsymbol{x}_\alpha$ par des dérivées consistantes au sens de la définition 7.4.1. Cela signifie qu'on cherche des tenseurs symétriques $\boldsymbol{w}_\alpha^{(l)}$ tels que

$$\boldsymbol{w}_\alpha^{(l)} = D^{(l)}u\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k-l+1}\right)$$

pour $1 \le l \le k$.

Si u est une fonction suffisamment régulière, ses moyennes et ses dérivées sont des solutions approchées de (8.2.1) puisque selon la formule (5.8.3)

$$\overline{u}_\beta - \overline{u}_\alpha = \sum_{i=1}^{k} \frac{1}{i!}\, D^{(i)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \left[\boldsymbol{z}_{\alpha\beta}^{(i)} - \boldsymbol{x}_\alpha^{(i)}\right] + \mathrm{O}\left(h^{k+1}\right) . \tag{8.2.3}$$

De façon analogue, les dérivées de u sont des solutions approchées de (8.2.2) car

$$D^{(j)}u\Big|_{\boldsymbol{x}_\beta} = \sum_{l=j}^{k} \frac{1}{(l-j)!}\, D^{(l)}u\Big|_{\boldsymbol{x}_\alpha} \cdot \boldsymbol{h}_{\alpha\beta}^{l-j} + \mathrm{O}\left(h^{k-j+1}\right) ,\ 1 \le j \le k-1 . \tag{8.2.4}$$

Considérons maintenant des *dérivées consistantes* au sens de la définition 7.4.1

$$\boldsymbol{w}_\beta^{(l)}[\mathfrak{u}] = \sum_{\gamma} \boldsymbol{w}_{\beta\gamma}^{(l)} \overline{u}_\gamma = D^{(l)}u\Big|_{\boldsymbol{x}_\beta} + \mathrm{O}\left(h^{k+1-l}\right) ,\ 1 \le l \le k . \tag{8.2.5}$$

Les dérivées consistantes $\boldsymbol{w}_\beta^{(l)}[\mathfrak{u}]$, qui sont des tenseurs symétriques, sont par définition des approximations des dérivées de u. L'insertion de (8.2.5) dans (8.2.3) et (8.2.4) montre que les tenseurs $w_\beta^{(l)}[\mathfrak{u}]$ sont également des solutions approchées de (8.2.1) et de (8.2.2) car

$$\overline{u}_\beta - \overline{u}_\alpha = \sum_{i=1}^{k} \frac{1}{i!} \boldsymbol{w}_\alpha^{(i)}[\mathfrak{u}] \bullet \left[\boldsymbol{z}_{\alpha\beta}^{(i)} - \boldsymbol{x}_\alpha^{(i)}\right] + \mathrm{O}\left(h^{k+1}\right) \tag{8.2.6}$$

$$\boldsymbol{w}_\beta^{(j)}[\mathfrak{u}] = \sum_{l=j}^{k} \frac{1}{(l-j)!} \boldsymbol{w}_\alpha^{(l)}[\mathfrak{u}] \cdot \boldsymbol{h}_{\alpha\beta}^{l-j} + \mathrm{O}\left(h^{k-j+1}\right) \tag{8.2.7}$$

pour $1 \le j \le k-1$ et $\beta \in \mathbb{W}_\alpha$.

Le point de départ de la méthode des moindres carrés couplés est de considérer le système (8.2.6-8.2.7) comme un système d'équations approchées pour les inconnues $w_\beta^{(l)}[\mathfrak{u}]$, $1 \le l \le k$

$$\overline{u}_\beta - \overline{u}_\alpha = \sum_{i=1}^{k} \frac{1}{i!} \boldsymbol{w}_\alpha^{(i)}[\mathfrak{u}] \bullet \left[\boldsymbol{z}_{\alpha\beta}^{(i)} - \boldsymbol{x}_\alpha^{(i)}\right] \tag{8.2.8}$$

$$\boldsymbol{w}_\beta^{(j)}[\mathfrak{u}] = \sum_{l=j}^{k} \frac{1}{(l-j)!} \boldsymbol{w}_\alpha^{(l)}[\mathfrak{u}] \cdot \boldsymbol{h}_{\alpha\beta}^{l-j} \tag{8.2.9}$$

où $1 \leq j \leq k-1$ et $\beta \in \mathbb{W}_\alpha$.

Pour trouver des solutions approchées de (8.2.8-8.2.9), il faut surmonter deux problèmes.

(1) Si u n'est pas un polynôme, le système (8.2.8-8.2.9) n'admet en général pas de solution exacte. Il est par contre possible de trouver des solutions approchées de (8.2.8-8.2.9) en recourant à la méthode des moindres carrés.

(2) Les équations (8.2.9) couplent les inconnues dans différentes cellules. Pour éviter la résolution d'un système global, il faut décomposer la résolution approchée de (8.2.8-8.2.9) en une suite de problèmes locaux.

Pour résoudre ces deux problèmes, la méthode des moindres carrés couplés procède de la manière suivante : on fait usage de l'identité $\boldsymbol{z}_{\alpha\beta}^{(1)} - \boldsymbol{x}_\alpha^{(1)} = \boldsymbol{h}_{\alpha\beta}$ dans (8.2.8) pour réécrire le système (8.2.8-8.2.9) sous la forme

$$\boldsymbol{h}_{\alpha\beta} \cdot \boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] = \overline{u}_\beta - \overline{u}_\alpha - \sum_{i=2}^{k} \frac{1}{i!} \boldsymbol{w}_\alpha^{(i)}[\mathfrak{u}] \bullet \left[\boldsymbol{z}_{\alpha\beta}^{(i)} - \boldsymbol{x}_\alpha^{(i)}\right], \ \beta \in \mathbb{W}_\alpha \tag{8.2.10}$$

$$\begin{aligned} \boldsymbol{h}_{\alpha\beta} \cdot \boldsymbol{w}_\alpha^{(j+1)}[\mathfrak{u}] &= \boldsymbol{w}_\beta^{(j)}[\mathfrak{u}] - \boldsymbol{w}_\alpha^{(j)}[\mathfrak{u}] - \\ &\quad - \sum_{l=j+2}^{k} \frac{1}{(l-j)!} \boldsymbol{w}_\alpha^{(l)}[\mathfrak{u}] \cdot \boldsymbol{h}_{\alpha\beta}^{l-j}, \ 1 \leq j \leq k-1, \ \beta \in \mathbb{W}_\alpha. \end{aligned} \tag{8.2.11}$$

Ensuite, on détermine pour chaque équation (8.2.10-8.2.11) l'inconnue dans le membre de gauche par la méthode des moindres carrés *en considérant les inconnues* $\boldsymbol{w}_\alpha^{(l)}[\mathfrak{u}]$ *dans le membre de droite comme des paramètres fixes*. Ce procédé donne un système d'équations linéaires pour les inconnues $\boldsymbol{w}_\alpha^{(l)}[\mathfrak{u}]$ qu'on se propose d'expliciter dans la suite.

On introduit d'abord les définitions

$$\widetilde{\delta u}_{\alpha\beta}^{(0)} \triangleq \overline{u}_\beta - \overline{u}_\alpha - \sum_{i=2}^{k} \frac{1}{i!} \boldsymbol{w}_\alpha^{(i)}[\mathfrak{u}] \bullet \left[\boldsymbol{z}_{\alpha\beta}^{(i)} - \boldsymbol{x}_\alpha^{(i)}\right] \tag{8.2.12}$$

$$\widetilde{\boldsymbol{\delta u}}_{\alpha\beta}^{(j)} \triangleq \boldsymbol{w}_\beta^{(j)}[\mathfrak{u}] - \boldsymbol{w}_\alpha^{(j)}[\mathfrak{u}] - \sum_{l=j+2}^{k} \frac{1}{(l-j)!} \boldsymbol{w}_\alpha^{(l)}[\mathfrak{u}] \cdot \boldsymbol{h}_{\alpha\beta}^{l-j} \tag{8.2.13}$$

où $1 \leq j \leq k-1$, ce qui permet d'écrire le système (8.2.10-8.2.11) comme

$$\boldsymbol{h}_{\alpha\beta} \cdot \boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] = \widetilde{\delta u}_{\alpha\beta}^{(0)}, \ \beta \in \mathbb{W}_\alpha \tag{8.2.14}$$

$$\boldsymbol{h}_{\alpha\beta} \cdot \boldsymbol{w}_\alpha^{(j+1)}[\mathfrak{u}] = \widetilde{\boldsymbol{\delta u}}_{\alpha\beta}^{(j)}, \ \beta \in \mathbb{W}_\alpha, \ 1 \leq j \leq k-1. \tag{8.2.15}$$

La solution au sens des moindres carrés de (8.2.14) s'obtient par la formule des moindres carrés (7.9.2) pour la reconstruction du gradient. Plus précisément, on a

$$\boldsymbol{w}_\alpha^{(1)} = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \widetilde{\delta u}_{\alpha\beta}^{(0)} \tag{8.2.16}$$

où les $\boldsymbol{\sigma}_{\alpha\beta}$ sont les paramètres (7.9.5)

$$\boldsymbol{\sigma}_{\alpha\beta} = \begin{cases} \left(H_\alpha^t H_\alpha\right)^{-1} \boldsymbol{h}_{\alpha\beta} & \beta \in \mathbb{W}_\alpha \\ 0 & \beta \notin \mathbb{W}_\alpha \end{cases}$$

de la reconstruction du gradient par la méthode des moindres carrés.

La solution au sens des moindres carrés de (8.2.15) demande plus d'effort car il s'agit d'équations tensorielles dont les inconnues $w_{\alpha,i_1\cdots i_j}$ ne sont pas mutuellement indépendantes. La transcription de l'équation (8.2.15) en une équation matricielle procède de façon similaire aux définitions 7.6.2 et 7.6.3. L'équation (8.2.15) s'écrit de façon explicite

$$\sum_{i_0=1}^{d} h_{i_0}^{\alpha\beta} w_{\alpha,i_0 i_1\cdots i_j} = \delta u_{i_1\cdots i_j}^{\alpha\beta}, \ \beta \in \mathbb{V}_\alpha, \ 1 \leq i_1, \ldots, i_j \leq d \tag{8.2.17}$$

où $1 \leq j \leq k-1$.

Gardons j fixe. Pour chaque $(j+1)$-tuple d'indices $(i_0, i_1, \ldots, i_j) \in \mathbb{J}^d_{j+1}$, il existe, d'après le lemme 5.10.2, un unique $(j+1)$-tuple d'indices $\left(i'_0, i'_1, \ldots, i'_j\right) \in \mathbb{J}^d_{j+1}$ qui désigne la même composante $w_{\alpha,i_0 i_1 \cdots i_j}$ que $(i_0, i_1, \ldots, i_j)$ et tel que $i'_0 \leq i'_1 \leq \ldots \leq i'_j$. Cela signifie qu'il suffit d'écrire l'équation (8.2.17) pour les composantes $w_{\alpha,i_0 i_1 \cdots i_j}$ dont les indices satisfont $i_0 \leq i_1 \leq \ldots \leq i_j$. La fonction de rangement ϖ^d_{j+1} définie par (5.10.6) établit un ordre sur les $(j+1)$-tuples $(i_0, i_1, \ldots, i_j)$ tels que $i_0 \leq i_1 \leq \ldots \leq i_j$. Cela permet de classer les composantes $w_{\alpha,i_0 i_1 \cdots i_j}$, $i_0 \leq i_1 \leq \ldots \leq i_j$, dans un vecteur de $\binom{j+d}{j+1}$ éléments noté $\widetilde{\boldsymbol{w}}^{(j+1)}_\alpha$.

On arrange ensuite les indices des cellules voisines $\beta \in \mathbb{V}_\alpha$ dans leur ordre naturel $\beta_1 \leq \ldots \leq \beta_{l_\alpha}$. Cela permet de classer les composantes $\delta u^{\alpha\beta}_{i_1 \cdots i_j}$, $\beta \in \mathbb{V}_\alpha$, $1 \leq i_1 \leq \ldots \leq i_j \leq d$, dans un vecteur de $l_\alpha \binom{j+d-1}{j}$ éléments où l_α est le nombre des premiers voisins de $\mathcal{T}_\alpha$. Ce vecteur est noté $\widetilde{\boldsymbol{\delta u}}^{(j)}_\alpha$.

La relation (8.2.17) devient alors une relation linéaire entre le vecteur $\widetilde{\boldsymbol{w}}^{(j+1)}_\alpha$ de $\binom{j+d}{j+1}$ éléments et le vecteur $\widetilde{\boldsymbol{\delta u}}^{(j)}_\alpha$ de $l_\alpha \binom{j+d-1}{j}$ éléments. Cette relation s'exprime par une matrice $\widetilde{H}^{(j+1)}_\alpha$

$$\widetilde{H}^{(j+1)}_\alpha \widetilde{\boldsymbol{w}}^{(j+1)}_\alpha = \widetilde{\boldsymbol{\delta u}}^{(j)}_\alpha \,. \tag{8.2.18}$$

La solution au sens des moindres carrés de (8.2.18) est donnée par la pseudo-inverse ou inverse de Moore-Penrose de $\widetilde{H}^{(j+1)}_\alpha$

$$\widetilde{\boldsymbol{w}}^{(j+1)}_\alpha = \left(\widetilde{H}^{(j+1)\,t}_\alpha \widetilde{H}^{(j+1)}_\alpha \right)^{-1} \widetilde{H}^{(j+1)\,t}_\alpha \widetilde{\boldsymbol{\delta u}}^{(j)}_\alpha \,. \tag{8.2.19}$$

Le système résultant est une relation linéaire entre les $\boldsymbol{w}^{(j+1)}_\alpha$ dans le membre de gauche et les $\boldsymbol{w}^{(j)}_\alpha$ ainsi que les $\boldsymbol{w}^{(l)}_\alpha$, $j+2 \leq l \leq k$, dans le membre de droite.

Le système complet d'équations de la méthode des moindres carrés couplés est donc composé des équations (8.2.16) et (8.2.19) qu'on réécrit pour la commodité du lecteur

$$\begin{aligned} \boldsymbol{w}^{(1)}_\alpha &= \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \widetilde{\delta u}^{(0)}_{\alpha\beta} \\ \widetilde{\boldsymbol{w}}^{(j+1)}_\alpha &= \left(\widetilde{H}^{(j+1)\,t}_\alpha \widetilde{H}^{(j+1)}_\alpha \right)^{-1} \widetilde{H}^{(j+1)\,t}_\alpha \widetilde{\boldsymbol{\delta u}}^{(j)}_\alpha \end{aligned}$$

où $1 \leq j \leq k-1$ et $1 \leq \alpha \leq N$.

Si p est un polynôme de degré k sur le voisinage de la cellule $\mathcal{T}_\alpha$, ses moyennes et ses dérivées sont une solution exacte du système (8.2.8-8.2.9). Puisque la solution exacte d'une équation linéaire est aussi une solution au sens des moindres carrés, les moyennes et les dérivées de p sont une solution des équations (8.2.16) et (8.2.19) qui représentent justement les solutions au sens des moindres carrés de (8.2.8-8.2.9). *Le système formé par (8.2.16) et (8.2.19) est donc consistant dans le sens suivant : si p est un polynôme de degré k, les moyennes et les dérivées de p sont solutions de (8.2.16) et (8.2.19).*

L'avantage principal de cette approche est la possibilité d'utiliser le premier voisinage $\mathbb{V}_\alpha$ de la cellule $\mathcal{T}_\alpha$ pour $\mathbb{W}_\alpha$. Cela vient du fait que la solution au sens des moindres carrés de (8.2.10) et de (8.2.11) existe même si les équations sont surdéterminées. *Il est alors possible de développer des algorithmes de résolution du système formé par (8.2.16) et (8.2.19) qui échangent des informations uniquement entre premiers voisins.*

Dans la suite, on explicite le système composé de (8.2.16) et (8.2.19) pour le premier voisinage $\mathbb{W}_\alpha = \mathbb{V}_\alpha$ et le degré $k = 2$ en introduisant les notations spécifiques

$$\begin{aligned} \boldsymbol{\sigma}_\alpha &= \boldsymbol{w}^{(1)}_\alpha [\mathfrak{u}] \\ \boldsymbol{\theta}_\alpha &= \boldsymbol{w}^{(2)}_\alpha [\mathfrak{u}] \,. \end{aligned}$$

Le système (8.2.10-8.2.11) devient

$$\boldsymbol{h}_{\alpha\beta}\cdot\boldsymbol{\sigma}_\alpha = \overline{u}_\beta - \overline{u}_\alpha - \frac{1}{2!}\left[\boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta - \boldsymbol{x}^{(2)}_\alpha\right]\bullet\boldsymbol{\theta}_\alpha\,,\ \beta\in\mathbb{V}_\alpha \tag{8.2.20}$$

$$\boldsymbol{h}_{\alpha\beta}\cdot\boldsymbol{\theta}_\alpha = \boldsymbol{\sigma}_\beta - \boldsymbol{\sigma}_\alpha\,,\ \beta\in\mathbb{V}_\alpha \tag{8.2.21}$$

où $\mathbb{V}_\alpha$ est le premier voisinage et $1\leq\alpha\leq N$. La solution au sens des moindres carrés de (8.2.20) s'écrit avec les paramètres $\boldsymbol{\sigma}_{\alpha\beta}$ de (7.9.5) comme

$$\boldsymbol{\sigma}_\alpha = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\left\{\overline{u}_\beta - \overline{u}_\alpha - \frac{1}{2!}\left[\boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta - \boldsymbol{x}^{(2)}_\alpha\right]\bullet\boldsymbol{\theta}_\alpha\right\}\,. \tag{8.2.22}$$

En dimension deux, le système (8.2.21) peut s'écrire sous forme matricielle comme

$$\begin{pmatrix} h_1^{\alpha\beta_1} & h_2^{\alpha\beta_1} & 0\\ 0 & h_1^{\alpha\beta_1} & h_2^{\alpha\beta_1}\\ h_1^{\alpha\beta_2} & h_2^{\alpha\beta_2} & 0\\ 0 & h_1^{\alpha\beta_2} & h_2^{\alpha\beta_2}\\ \vdots & \vdots & \vdots \end{pmatrix}\begin{pmatrix}\theta_{\alpha,11}\\ \theta_{\alpha,12}\\ \theta_{\alpha,22}\end{pmatrix} = \begin{pmatrix}\sigma_{\beta_1,1}-\sigma_{\alpha,1}\\ \sigma_{\beta_1,2}-\sigma_{\alpha,2}\\ \sigma_{\beta_2,1}-\sigma_{\alpha,1}\\ \sigma_{\beta_2,2}-\sigma_{\alpha,2}\\ \vdots\end{pmatrix}\,. \tag{8.2.23}$$

En dimension trois, la forme matricielle du système (8.2.21) est

$$\begin{pmatrix} h_1^{\alpha\beta_1} & h_2^{\alpha\beta_1} & h_3^{\alpha\beta_1} & 0 & 0 & 0\\ 0 & h_1^{\alpha\beta_1} & 0 & h_2^{\alpha\beta_1} & h_3^{\alpha\beta_1} & 0\\ 0 & 0 & h_1^{\alpha\beta_1} & 0 & h_2^{\alpha\beta_1} & h_3^{\alpha\beta_1}\\ h_1^{\alpha\beta_2} & h_2^{\alpha\beta_2} & h_2^{\alpha\beta_2} & 0 & 0 & 0\\ 0 & h_1^{\alpha\beta_2} & 0 & h_2^{\alpha\beta_2} & h_3^{\alpha\beta_2} & 0\\ 0 & 0 & h_1^{\alpha\beta_2} & 0 & h_2^{\alpha\beta_2} & h_3^{\alpha\beta_2}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}\begin{pmatrix}\theta_{\alpha,11}\\ \theta_{\alpha,12}\\ \theta_{\alpha,13}\\ \theta_{\alpha,22}\\ \theta_{\alpha,23}\\ \theta_{\alpha,33}\end{pmatrix} = \begin{pmatrix}\sigma_{\beta_1,1}-\sigma_{\alpha,1}\\ \sigma_{\beta_1,2}-\sigma_{\alpha,2}\\ \sigma_{\beta_1,3}-\sigma_{\alpha,3}\\ \sigma_{\beta_2,1}-\sigma_{\alpha,1}\\ \sigma_{\beta_2,2}-\sigma_{\alpha,2}\\ \sigma_{\beta_2,3}-\sigma_{\alpha,3}\\ \vdots\end{pmatrix}\,. \tag{8.2.24}$$

Introduisons la notation

$$\widetilde{H}^{(2)}_\alpha\widetilde{\boldsymbol{\theta}}_\alpha = \widetilde{\boldsymbol{\delta\sigma}}_\alpha$$

pour écrire les équations (8.2.23) et (8.2.24). La solution au sens des moindres carrés des équations (8.2.23) et (8.2.24) est alors donnée par l'inverse de Moore-Penrose ou pseudo-inverse de $\widetilde{H}^{(2)}_\alpha$

$$\widetilde{\boldsymbol{\theta}}_\alpha = \left(\widetilde{H}^{(2)\,t}_\alpha\,\widetilde{H}^{(2)}_\alpha\right)^{-1}\,\widetilde{H}^{(2)\,t}_\alpha\,\widetilde{\boldsymbol{\delta\sigma}}_\alpha\,. \tag{8.2.25}$$

Les solutions au sens des moindres carrés des équations (8.2.20) et (8.2.21) s'écrivent alors

$$\boldsymbol{\sigma}_\alpha = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\left\{\overline{u}_\beta - \overline{u}_\alpha - \frac{1}{2!}\left[\boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta - \boldsymbol{x}^{(2)}_\alpha\right]\bullet\boldsymbol{\theta}_\alpha\right\} \tag{8.2.26}$$

$$\widetilde{\boldsymbol{\theta}}_\alpha = \left(\widetilde{H}^{(2)\,t}_\alpha\,\widetilde{H}^{(2)}_\alpha\right)^{-1}\,\widetilde{H}^{(2)\,t}_\alpha\,\widetilde{\boldsymbol{\delta\sigma}}_\alpha \tag{8.2.27}$$

où $1\leq\alpha\leq N$.

On observe que l'équation (8.2.26) contient uniquement les inconnues $\boldsymbol{\sigma}_\alpha$ et $\boldsymbol{\theta}_\alpha$ dans la cellule $\mathcal{T}_\alpha$. L'équation (8.2.27) contient les inconnues $\boldsymbol{\sigma}_\alpha$ et $\boldsymbol{\theta}_\alpha$ ainsi que les gradients $\boldsymbol{\sigma}_\beta$ des cellules voisines $\beta\in\mathbb{V}_\alpha$. Si l'on considère le système formé par (8.2.26) et (8.2.27) comme un système global où α parcourt toutes les cellules, les coefficients de l'équation (8.2.26) sont tous situés sur la diagonale. Par contre, l'équation (8.2.27) occasionne des coefficients non nuls pour $\beta\neq\alpha$. Ces observations justifient l'introduction de la

Définition 8.2.1. On appelle le système formé par les équations (8.2.26) et (8.2.27) le *système des moindres carrés couplés (pour la reconstruction de degré deux).*

La forme des équations (8.2.26) et (8.2.27) suggère d'utiliser des algorithmes itératifs de type *Jacobi* ou *Gauss-Seidel* pour résoudre le système des moindres carrés couplés. Il faut cependant veiller à ce que ces algorithmes travaillent de façon locale, c'est-à-dire qu'ils ne fassent intervenir

que le premier voisinage. Une variante est donnée par l'algorithme suivant, appelé méthode MCCI.

ALGORITHME 8.2.2 (Reconstruction des polynômes de degré deux par la méthode des moindres carrés couplés par itération ou méthode MCCI). *Une façon itérative de résoudre (8.2.26-8.2.27) est de calculer pour $\boldsymbol{\theta}_\alpha$ fixe un gradient $\boldsymbol{\sigma}_\alpha$ par la formule (8.2.26). L'insertion de ce gradient calculé $\boldsymbol{\sigma}_\alpha$ dans (8.2.27) permet de déterminer une nouvelle valeur pour $\boldsymbol{\theta}_\alpha$. Ce $\boldsymbol{\theta}_\alpha$ peut ensuite être inséré itérativement dans (8.2.26) pour ajuster la valeur de $\boldsymbol{\sigma}_\alpha$. L'algorithme peut continuer de cette façon. Il peut être initialisé par $\boldsymbol{\theta}_\alpha = 0$.*

Il n'est pas évident de démontrer que l'algorithme 8.2.2 converge en maillage non structuré général. Il est donc nécessaire de faire des essais numériques pour s'assurer de la convergence. Il est par contre possible de démontrer que l'algorithme 8.2.2 converge en maillage cartésien au bout d'une seule itération.

8.3. Analyse de la méthode des corrections successives

Cette méthode, proposée initialement dans [**20**], permet de réaliser des reconstructions polynomiales de degré élevé à partir d'une reconstruction de polynômes de degré un. *Le grand avantage de cette approche réside dans le fait que la reconstruction de polynômes de degré un, aussi appelée "reconstruction linéaire" ou "reconstruction d'ordre deux", peut s'effectuer sur de petits voisinages et même le premier voisinage de chaque cellule.* La section 8.3.1 présente la reconstruction de polynômes de degré deux et la section 8.3.2 présente la reconstruction de polynômes de degré plus élevé.

Pour les besoins de cette section, il est utile d'expliciter les conditions de consistance (7.3.18) et (7.3.19) dans le cas de la reconstruction des polynômes de degré $k = 1$ et $k = 2$.

Dans le cas $k = 1$, la forme générale des fonctions reconstruites (7.5.5) est

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \overline{u}_\alpha + \boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)$$

où

$$\boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(1)} \overline{u}_\beta$$

est un gradient consistant au sens de la définition 7.4.1 et $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$ est le vecteur des moyennes de cellule. Dans la suite, le gradient $\boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}]$ est noté $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$ et les coefficients $\boldsymbol{w}_{\alpha\beta}^{(1)}$ s'appellent $\boldsymbol{\sigma}_{\alpha\beta} \triangleq \boldsymbol{w}_{\alpha\beta}^{(1)}$

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \overline{u}_\beta \,. \tag{8.3.1}$$

Par convention, on définit $\boldsymbol{\sigma}_{\alpha\beta} \triangleq 0$ dans (8.3.1) si la cellule $\mathcal{T}_\beta$ n'est pas dans le voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$.

Le gradient (8.3.1) permet de reconstruire les polynômes de degré un si les coefficients satisfont les conditions de consistance (7.3.19) et (7.3.18), données dans ce cas particulier par

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta} = 0 \tag{8.3.2}$$

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} = \boldsymbol{\delta}^{(2)} \,. \tag{8.3.3}$$

Pour une fonction suffisamment régulière u, l'erreur d'approximation du gradient

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] - D^{(1)}u\Big|_{\boldsymbol{x}_\alpha} = \frac{1}{2!} \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \left[D^{(2)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(2)} + \mathrm{O}\left(h^3\right) \right]$$

est $\mathrm{O}(h)$ d'après la définition 7.4.1.

Dans le cas $k = 2$, c'est-à-dire dans le cas d'une reconstruction quadratique par cellule, la forme générale des fonctions reconstruites (7.5.5) est

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \overline{u}_\alpha + \boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha) + \frac{1}{2!}\boldsymbol{w}_\alpha^{(2)}[\mathfrak{u}] \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)^2$$

où

$$\boldsymbol{w}_\alpha^{(2)}[\mathfrak{u}] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(2)}\,\overline{u}_\beta$$

est une dérivée seconde consistante au sens de la définition 7.4.1 et

$$\boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(1)}\,\overline{u}_\beta$$

est un *gradient consistant de précision à l'ordre deux* au sens de la définition 7.4.1.

Dans la suite, la dérivée seconde $\boldsymbol{w}_\alpha^{(2)}[\mathfrak{u}]$ est notée $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$ et les coefficients $\boldsymbol{w}_{\alpha\beta}^{(2)}$ s'appellent $\boldsymbol{\theta}_{\alpha\beta} \triangleq \boldsymbol{w}_{\alpha\beta}^{(2)}$

$$\boldsymbol{\theta}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}\overline{u}_\beta\,. \tag{8.3.4}$$

Pour reconstruire la dérivée seconde de tout polynôme de degré deux, les coefficients $\boldsymbol{\theta}_{\alpha\beta}$ doivent satisfaire les conditions de consistance (7.3.19) et (7.3.18), qui s'écrivent de façon explicite

$$\sum_\beta \boldsymbol{\theta}_{\alpha\beta} = 0 \tag{8.3.5}$$

$$\sum_\beta \boldsymbol{\theta}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} = 0 \tag{8.3.6}$$

$$\sum_\beta \frac{1}{2}\boldsymbol{\theta}_{\alpha\beta} \otimes \boldsymbol{z}_{\alpha\beta}^{(2)} = \boldsymbol{\delta}^{(4)}\,. \tag{8.3.7}$$

On rappelle que le tenseur $\boldsymbol{\delta}^{(4)}$ est défini par (5.2.22). Le gradient consistant de précision à l'ordre deux $\boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}]$ est noté $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$, avec le même symbole que le gradient consistant (8.3.1). Pour reconstruire le gradient de tout polynôme de degré deux, les coefficients $\boldsymbol{\sigma}_{\alpha\beta}$ doivent satisfaire les conditions (8.3.2), (8.3.3) et la condition de précision à l'ordre deux

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \otimes \boldsymbol{z}_{\alpha\beta}^{(2)} = 0\,. \tag{8.3.8}$$

On rappelle que la reconstruction des polynômes de degré un, deux et trois est détaillée de façon explicite dans l'annexe A.

8.3.1. Analyse de la méthode des corrections successives dans le cas quadratique. Le point de départ de la méthode est une reconstruction consistante des polynômes de degré $k = 1$, définie par un gradient consistant (8.3.1).

Le gradient consistant (8.3.1) permet de définir un opérateur par

$$\widehat{\boldsymbol{\theta}}_\alpha[\mathfrak{u}] \triangleq \sum_\beta \sum_\gamma \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}\overline{u}_\gamma = \frac{1}{2}\sum_\beta \sum_\gamma \left[\boldsymbol{\sigma}_{\alpha\beta} \otimes \boldsymbol{\sigma}_{\beta\gamma} + \boldsymbol{\sigma}_{\beta\gamma} \otimes \boldsymbol{\sigma}_{\alpha\beta}\right] \overline{u}_\gamma\,. \tag{8.3.9}$$

L'objectif de l'opérateur (8.3.9) est d'approcher la dérivée seconde de u. L'utilisation du produit tensoriel symétrique (5.2.19) rend (8.3.9) symétrique, comme la dérivée seconde d'une fonction de classe C^2. Avec la définition des tenseurs $\widehat{\boldsymbol{\theta}}_{\alpha\gamma}$ d'ordre deux

$$\widehat{\boldsymbol{\theta}}_{\alpha\gamma} \triangleq \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}\,, \tag{8.3.10}$$

(8.3.9) devient

$$\widehat{\boldsymbol{\theta}}_\alpha[\mathfrak{u}] = \sum_\gamma \widehat{\boldsymbol{\theta}}_{\alpha\gamma}\,\overline{u}_\gamma = \sum_\beta \sum_\gamma \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}\,\overline{u}_\gamma\,.$$

De façon explicite, (8.3.9) s'écrit

$$\widehat{\theta}_{\alpha,ij}\left[\mathfrak{u}\right] = \frac{1}{2}\sum_{\beta}\sum_{\gamma}\left[\sigma_i^{\alpha\beta}\sigma_j^{\beta\gamma} + \sigma_j^{\alpha\beta}\sigma_i^{\beta\gamma}\right]\overline{u}_\gamma\,,\ 1 \leq i,j \leq d.$$

Le problème est que (8.3.9) *n'est pas une dérivée seconde consistante* au sens de la définition 7.4.1. En général, si p est un polynôme de degré deux avec les moyennes $\mathfrak{p} = (\overline{p}_1, \ldots, \overline{p}_N)$, alors

$$\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{p}\right] \neq D^{(2)}p\Big|_{\boldsymbol{x}_\alpha}\,.$$

Soit $\mathbb{S}_2^d$ l'espace des tenseurs symétriques d'ordre deux dans $\mathbb{R}^d$. On peut se demander s'il existe une transformation linéaire $\mathfrak{C}_\alpha^{(2)} : \mathbb{S}_2^d \to \mathbb{S}_2^d$ qui ne dépend que de la géométrie du maillage et tel que

$$\mathfrak{C}_\alpha^{(2)} \cdot \widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{p}\right] = D^{(2)}p\Big|_{\boldsymbol{x}_\alpha} \tag{8.3.11}$$

pour tout polynôme p de degré deux. *L'objectif de cette section est de trouver une telle transformation* $\mathfrak{C}_\alpha$.

Pour cela, on prouve d'abord le

LEMME 8.3.1 (Noyau de l'opérateur approché $\widehat{\boldsymbol{\theta}}_\alpha$). *Si*

$$\sigma_\alpha\left[\mathfrak{u}\right] = \sum_{\beta}\boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta\,.$$

est une reconstruction consistante du gradient au sens de la définition 7.4.1, les coefficients $\boldsymbol{\sigma}_{\alpha\beta}$ *satisfont les équations*

$$\sum_{\beta}\sum_{\gamma}\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma} = 0 \tag{8.3.12}$$

$$\sum_{\beta}\sum_{\gamma}\left(\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma}\right)\otimes\boldsymbol{h}_{\alpha\gamma} = 0\,. \tag{8.3.13}$$

DÉMONSTRATION. L'identité (8.3.12) découle directement de (8.3.2) car

$$\sum_{\beta}\sum_{\gamma}\sigma_i^{\alpha\beta}\sigma_j^{\beta\gamma} = \sum_{\beta}\sigma_i^{\alpha\beta}\left(\sum_{\gamma}\sigma_j^{\beta\gamma}\right) = 0$$

pour $1 \leq i,j \leq d$. L'équation (8.3.13) vient de l'identité $\boldsymbol{h}_{\alpha\gamma} = \boldsymbol{h}_{\alpha\beta} + \boldsymbol{h}_{\beta\gamma}$ et du fait que

$$\sum_{\beta}\sum_{\gamma}\sigma_i^{\alpha\beta}\sigma_j^{\beta\gamma}\left(h_l^{\alpha\beta} + h_l^{\beta\gamma}\right) = \sum_{\beta}\left[\sigma_i^{\alpha\beta}h_l^{\alpha\beta}\left(\sum_{\gamma}\sigma_j^{\beta\gamma}\right) + \sigma_i^{\alpha\beta}\left(\sum_{\gamma}\sigma_j^{\beta\gamma}h_l^{\beta\gamma}\right)\right] =$$

$$= \left(\sum_{\beta}\sigma_i^{\alpha\beta}\right)\delta_{jl} = 0$$

pour $1 \leq i,j,l \leq d$, en raison de (8.3.2) et de (8.3.3). □

Le lemme 8.3.1 permet de démontrer le

COROLLAIRE 8.3.2. *L'opérateur (8.3.9) s'annule sur les polynômes de degré un.*

DÉMONSTRATION. Si q est un polynôme de degré un, la moyenne de q sur une cellule $\mathcal{T}_\gamma$ est

$$\overline{q}_\gamma = q\left(\boldsymbol{x}_\alpha\right) + D^{(1)}q\Big|_{\boldsymbol{x}_\alpha}\cdot\boldsymbol{h}_{\alpha\gamma}\,. \tag{8.3.14}$$

L'insertion de (8.3.14) dans (8.3.9) et l'application des identités (8.3.12) et (8.3.13) montrent que

$$\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{q}\right] = \frac{1}{2}\sum_{\beta}\sum_{\gamma}\left[\boldsymbol{\sigma}_{\alpha\beta}\otimes\boldsymbol{\sigma}_{\beta\gamma} + \boldsymbol{\sigma}_{\beta\gamma}\otimes\boldsymbol{\sigma}_{\alpha\beta}\right]\overline{q}_\gamma = 0\,.$$

□

Le corollaire 8.3.2 permet de calculer (8.3.9) pour un polynôme p de degré deux avec les moyennes $\mathfrak{p} = (\overline{p}_1, \ldots, \overline{p}_N)$. La moyenne de p sur une cellule $\mathcal{T}_\gamma$ s'écrit selon la formule (5.8.4)

$$\overline{p}_\gamma \;=\; p\left(\boldsymbol{x}_\alpha\right) + D^{(1)}p\Big|_{\boldsymbol{x}_\alpha} \cdot \boldsymbol{h}_{\alpha\gamma} + \frac{1}{2!}\, D^{(2)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}^{(2)}_{\alpha\gamma}\,. \tag{8.3.15}$$

L'insertion de (8.3.15) dans (8.3.9) et le corollaire 8.3.2 entraînent

$$\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{p}\right] = \sum_\beta \sum_\gamma \widehat{\boldsymbol{\theta}}_{\alpha\gamma}\overline{p}_\gamma = \sum_\beta \sum_\gamma \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}\,\overline{p}_\gamma = \\ = \sum_\beta \sum_\gamma \frac{1}{2!}\left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}\right)\left(\boldsymbol{z}^{(2)}_{\alpha\gamma} \bullet D^{(2)}p\Big|_{\boldsymbol{x}_\alpha}\right)\,. \tag{8.3.16}$$

On définit un nouveau tenseur $\boldsymbol{\tau}^{(4)}_\alpha$ d'ordre quatre par

$$\boldsymbol{\tau}^{(4)}_\alpha \triangleq \frac{1}{2}\sum_{\beta,\gamma}\left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}\right) \otimes \boldsymbol{z}^{(2)}_{\alpha\gamma}\,. \tag{8.3.17}$$

En utilisant la notation (5.2.11), on peut réécrire la relation (8.3.16) sous la forme

$$\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{p}\right] = \boldsymbol{\tau}^{(4)}_\alpha \cdot D^{(2)}p\Big|_{\boldsymbol{x}_\alpha}\,. \tag{8.3.18}$$

De façon explicite, (8.3.18) s'écrit en utilisant l'identité $\boldsymbol{z}^{(2)}_{\alpha\gamma} = \boldsymbol{h}^2_{\alpha\gamma} + \boldsymbol{x}^{(2)}_\gamma$ comme

$$\widehat{\theta}_{\alpha,ij} = \sum_{\beta,\gamma}\sum_{k,l=1}^{d} \frac{1}{4}\left[\sigma_i^{\alpha\beta}\sigma_j^{\beta\gamma} + \sigma_i^{\beta\gamma}\sigma_j^{\alpha\beta}\right]\left[h_k^{\alpha\gamma}h_l^{\alpha\gamma} + x_{\gamma,kl}\right]\frac{\partial^2 p\left(\boldsymbol{x}_\alpha\right)}{\partial x_k \partial x_l}$$

où $1 \leq i, j \leq d$ et les composantes de $\boldsymbol{\tau}^{(4)}_\alpha$ sont données par

$$\tau_{\alpha,ijkl} = \frac{1}{4}\sum_{\beta,\gamma}\left[\sigma_i^{\alpha\beta}\sigma_j^{\beta\gamma} + \sigma_j^{\alpha\beta}\sigma_i^{\beta\gamma}\right]\left[h_k^{\alpha\gamma}h_l^{\alpha\gamma} + x_{\gamma,kl}\right] \tag{8.3.19}$$

pour $1 \leq i, j, k, l \leq d$.

La relation (8.3.18) justifie la

Définition 8.3.3 (*Application de Brenner d'ordre deux*). Soit $\mathbb{S}^d_2$ l'espace des tenseurs symétriques d'ordre deux dans $\mathbb{R}^d$. La relation (8.3.18) définit une application linéaire de l'espace $\mathbb{S}^d_2$ dans lui-même. Cette application, dite *application de Brenner d'ordre deux*, est définie par

$$\mathfrak{T}^{(2)}_\alpha : \left\{ \begin{array}{ccc} \mathbb{S}^d_2 & \longrightarrow & \mathbb{S}^d_2 \\ \boldsymbol{a}^{(2)} & \longmapsto & \boldsymbol{\tau}^{(4)}_\alpha \cdot \boldsymbol{a}^{(2)} \end{array} \right. . \tag{8.3.20}$$

□

Supposons que l'application de Brenner *(8.3.20)* puisse être inversée. *Dans ce cas, son inverse est justement la transformation linéaire* $\mathfrak{C}^{(2)}_\alpha$ *cherchée. L'inverse de* $\mathfrak{T}^{(2)}_\alpha$ *permet par conséquent de déterminer la dérivée seconde exacte* $D^{(2)}p\big|_{\boldsymbol{x}_\alpha}$ *de* p *à partir de l'opérateur* $\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{p}\right]$. Puisque le calcul de $\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{p}\right]$ ne fait intervenir que la reconstruction des polynômes de degré un, il devient possible de calculer des dérivées secondes consistantes par deux itérations sur de petits voisinages.

Une application linéaire d'un espace vectoriel de dimension finie dans lui-même peut être inversée si et seulement si elle est injective. En raison de la linéarité, l'application de Brenner (8.3.20) est injective si et seulement si

$$\boldsymbol{\tau}^{(4)}_\alpha \cdot \boldsymbol{a}^{(2)} = 0 \text{ entraîne } \boldsymbol{a}^{(2)} = 0 \tag{8.3.21}$$

pour tout tenseur symétrique $\boldsymbol{a}^{(2)}$ d'ordre deux. Il faut donc chercher des conditions pour que la condition (8.3.21) soit vraie.

Il est bien connu qu'une application linéaire d'un espace vectoriel dans lui-même est inversible si elle est suffisamment proche de l'application identité dans une norme quelconque. L'application identité est dans le cas présent donnée par le tenseur $\boldsymbol{\delta}^{(4)}$, défini par (5.2.22), car

$$\boldsymbol{\delta}^{(4)} \cdot \boldsymbol{a}^{(2)} = \boldsymbol{a}^{(2)}$$

pour tout tenseur symétrique $\boldsymbol{a}^{(2)}$ d'ordre deux. Il est donc intéressant de pouvoir représenter le tenseur $\boldsymbol{\tau}_\alpha^{(4)}$ comme la somme de l'identité $\boldsymbol{\delta}^{(4)}$ et d'une perturbation. Ceci fait l'objet du

LEMME 8.3.4 (Forme alternative du tenseur $\boldsymbol{\tau}_\alpha^{(4)}$). *Le tenseur $\boldsymbol{\tau}_\alpha^{(4)}$ défini par (8.3.17) peut être mis sous la forme*

$$\boldsymbol{\tau}_\alpha^{(4)} \;=\; \boldsymbol{\delta}^{(4)} + \frac{1}{2}\sum_{\beta,\gamma} (\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}) \otimes \boldsymbol{z}_{\beta\gamma}^{(2)}$$

où $\boldsymbol{\delta}^{(4)}$ est le tenseur défini par (5.2.22).

DÉMONSTRATION. Pour $k = 1$ et $k = 2$, la définition (5.7.10) est

$$\begin{aligned} \boldsymbol{z}_{\alpha\gamma}^{(1)} &\triangleq \boldsymbol{h}_{\alpha\gamma} \\ \boldsymbol{z}_{\alpha\gamma}^{(2)} &\triangleq \boldsymbol{h}_{\alpha\gamma}^2 + \boldsymbol{x}_\gamma^{(2)} \,. \end{aligned}$$

La formule (5.7.18) devient pour $k = 2$ avec la définition (5.2.19) du produit tensoriel symétrique

$$\boldsymbol{z}_{\alpha\gamma}^{(2)} = \sum_{m=0}^{2} \binom{2}{m} \boldsymbol{h}_{\alpha\beta}^m \odot \boldsymbol{z}_{\beta\gamma}^{(2-m)} = \boldsymbol{h}_{\alpha\beta}^2 + \boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\beta\gamma} + \boldsymbol{h}_{\beta\gamma} \otimes \boldsymbol{h}_{\alpha\beta} + \boldsymbol{z}_{\beta\gamma}^{(2)} \,. \tag{8.3.22}$$

La preuve consiste à insérer (8.3.22) dans (8.3.17)

$$\boldsymbol{\tau}_\alpha^{(4)} \;\triangleq\; \frac{1}{2}\sum_{\beta,\gamma} (\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}) \otimes \boldsymbol{z}_{\alpha\gamma}^{(2)} \,.$$

Pour simplifier les formules, il est nécessaire de calculer quelques identités de façon explicite. Il vient d'abord en raison de (8.3.2)

$$\sum_\beta \sum_\gamma \sigma_i^{\alpha\beta} \sigma_j^{\beta\gamma} h_k^{\alpha\beta} h_l^{\alpha\beta} = \sum_\beta \sigma_i^{\alpha\beta} h_k^{\alpha\beta} h_l^{\alpha\beta} \left(\sum_\gamma \sigma_j^{\beta\gamma} \right) = 0$$

pour $1 \le i, j, k, l \le d$, ce qui implique en notation tensorielle

$$\sum_{\beta,\gamma} (\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}) \otimes \boldsymbol{h}_{\alpha\beta}^2 = 0 \,. \tag{8.3.23}$$

Ensuite, (8.3.3) entraîne l'identité

$$\sum_\beta \sum_\gamma \sigma_i^{\alpha\beta} \sigma_j^{\beta\gamma} h_k^{\alpha\beta} h_l^{\beta\gamma} = \sum_\beta \sigma_i^{\alpha\beta} h_k^{\alpha\beta} \left(\sum_\gamma \sigma_j^{\beta\gamma} h_l^{\beta\gamma} \right) = \left(\sum_\beta \sigma_i^{\alpha\beta} h_k^{\alpha\beta} \right) \delta_{jl} = \delta_{ik}\delta_{jl}$$

pour $1 \le i, j, k, l \le d$, ce qui implique

$$\frac{1}{2}\sum_\beta \sum_\gamma \left(\sigma_i^{\alpha\beta} \sigma_j^{\beta\gamma} + \sigma_j^{\alpha\beta} \sigma_i^{\beta\gamma} \right) \left(h_k^{\alpha\beta} h_l^{\beta\gamma} + h_l^{\alpha\beta} h_k^{\beta\gamma} \right) = \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} \,,\; 1 \le i, j, k, l \le d \,. \tag{8.3.24}$$

L'équation (8.3.24) s'écrit en notation tensorielle

$$\sum_{\beta,\gamma} (\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_{\beta\gamma}) \otimes (\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\beta\gamma} + \boldsymbol{h}_{\beta\gamma} \otimes \boldsymbol{h}_{\alpha\beta}) = 2\boldsymbol{\delta}^{(4)} \,. \tag{8.3.25}$$

La combinaison de (8.3.22) avec (8.3.23) et (8.3.25) donne le résultat

$$\frac{1}{2}\sum_{\beta,\gamma}(\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma})\otimes\left[\boldsymbol{h}^2_{\alpha\gamma}+\boldsymbol{x}^{(2)}_{\gamma}\right]=$$

$$=\frac{1}{2}\sum_{\beta,\gamma}(\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma})\otimes\left[\boldsymbol{h}^2_{\alpha\beta}+\boldsymbol{h}_{\alpha\beta}\otimes\boldsymbol{h}_{\beta\gamma}+\boldsymbol{h}_{\beta\gamma}\otimes\boldsymbol{h}_{\alpha\beta}+\boldsymbol{z}^{(2)}_{\beta\gamma}\right]=$$

$$=\boldsymbol{\delta}^{(4)}+\frac{1}{2}\sum_{\beta,\gamma}(\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma})\otimes\boldsymbol{z}^{(2)}_{\beta\gamma}$$

ce qui prouve le lemme 8.3.4. □

Le lemme 8.3.4 montre que la condition

$$\sum_{\beta,\gamma}(\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma})\otimes\boldsymbol{z}^{(2)}_{\beta\gamma}=0 \tag{8.3.26}$$

suffit pour que $\boldsymbol{\tau}^{(4)}_{\alpha}$ soit égal à $\boldsymbol{\delta}^{(4)}$. Dans ce cas particulier, la relation (8.3.18) devient l'identité

$$\widehat{\boldsymbol{\theta}}_{\alpha}[\mathfrak{p}]=D^{(2)}p\Big|_{\boldsymbol{x}_{\alpha}}$$

car la définition (5.2.22) de $\boldsymbol{\delta}^{(4)}$ entraîne, par la relation (5.2.26),

$$\boldsymbol{a}^{(2)}=\operatorname{sym}\left(\boldsymbol{a}^{(2)}\right)=\boldsymbol{\delta}^{(4)}\cdot\boldsymbol{a}^{(2)}$$

pour tout tenseur symétrique $\boldsymbol{a}^{(2)}$ d'ordre deux. Cela implique que l'opérateur (8.3.9) reproduit dans ce cas la dérivée seconde de tout polynôme de degré deux. On peut faire la

Remarque 8.3.5. Le lemme 8.3.4 permet d'interpréter l'application de Brenner (8.3.20) induite par $\boldsymbol{\tau}^{(4)}_{\alpha}$ comme une perturbation de l'application identité induite par $\boldsymbol{\delta}^{(4)}$. L'espace $\mathbb{S}^d_2$ est un espace vectoriel réel de dimension finie. Pour cette raison, une application linéaire de $\mathbb{S}^d_2$ en lui-même, qui est une perturbation de l'identité, peut être inversée s'il existe une norme telle que la norme de la perturbation soit plus petite que un.

Dans la suite, on identifie un cas particulier où cette perturbation est nulle et dans lequel l'application de Brenner devient par conséquent l'identité. Dans le cas général, seuls des essais numériques montreront si l'application de Brenner (8.3.20) peut être inversée. □

La proposition suivante donne une condition suffisante pour que la relation (8.3.26) soit satisfaite.

Proposition 8.3.6 (Condition suffisante pour la consistance). *Si le gradient consistant (8.3.1)*

$$\sigma_{\alpha}[\mathfrak{u}]=\sum_{\beta}\boldsymbol{\sigma}_{\alpha\beta}\overline{u}_{\beta}$$

satisfait la condition (8.3.8) de précision à l'ordre deux

$$\sum_{\gamma}\boldsymbol{\sigma}_{\beta\gamma}\otimes\boldsymbol{z}^{(2)}_{\beta\gamma}=0\,,$$

l'opérateur approché (8.3.9)

$$\widehat{\boldsymbol{\theta}}_{\alpha}[\mathfrak{u}]\triangleq\sum_{\beta}\sum_{\gamma}\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma}\,\overline{u}_{\gamma}$$

est une dérivée seconde consistante au sens de la définition 7.4.1.

Démonstration. Un gradient consistant d'ordre deux satisfait (8.3.2) et (8.3.3) ainsi que la condition de précision à l'ordre deux (8.3.8) qui s'écrit composante par composante

$$\sum_{\gamma}\sigma_i^{\beta\gamma}\left[h_k^{\beta\gamma}h_l^{\beta\gamma}+x_{\gamma,kl}\right]=0 \tag{8.3.27}$$

pour $1 \leq i,k,l \leq d$. L'équation (8.3.27) montre que

$$\frac{1}{2}\sum_{\beta}\sum_{\gamma}\left(\sigma_i^{\alpha\beta}\sigma_j^{\beta\gamma}+\sigma_j^{\alpha\beta}\sigma_i^{\beta\gamma}\right)\left(h_k^{\beta\gamma}h_l^{\beta\gamma}+x_{\gamma,kl}\right)=0\,,$$

ce qui s'écrit en notation tensorielle

$$\frac{1}{2}\sum_{\beta,\gamma}\left(\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma}\right)\otimes\left[\boldsymbol{h}_{\beta\gamma}^{2}+\boldsymbol{x}_{\gamma}^{(2)}\right]=0\,. \tag{8.3.28}$$

La combinaison du lemme 8.3.4 et de (8.3.28) prouve alors que $\boldsymbol{\tau}_{\alpha}^{(4)}=\boldsymbol{\delta}^{(4)}$. Cela signifie que l'opérateur (8.3.9) reproduit la dérivée seconde de tout polynôme p de degré deux

$$\widehat{\boldsymbol{\theta}}_{\alpha}\left[\mathfrak{p}\right]=\boldsymbol{\tau}_{\alpha}^{(4)}\cdot D^{(2)}p\Big|_{\boldsymbol{x}_{\alpha}}=\boldsymbol{\delta}^{(4)}\cdot D^{(2)}p\Big|_{\boldsymbol{x}_{\alpha}}=D^{(2)}p\Big|_{\boldsymbol{x}_{\alpha}}\,.$$

Il reste à calculer l'erreur d'approximation pour la dérivée seconde d'une fonction u suffisamment régulière. La fonction u a les moyennes $\mathfrak{u}=(\overline{u}_1,\ldots,\overline{u}_N)$ qui se développent selon la formule (5.8.1) comme

$$\overline{u}_{\gamma}\;=\;\sum_{j=0}^{3}\frac{1}{j!}\,D^{(j)}u\Big|_{\boldsymbol{x}_{\alpha}}\bullet\boldsymbol{z}_{\alpha\gamma}^{(j)}+\mathrm{O}\left(h^{4}\right)\,. \tag{8.3.29}$$

L'insertion de (8.3.29) dans (8.3.9) donne l'erreur d'approximation

$$\widehat{\boldsymbol{\theta}}_{\alpha}\left[\mathfrak{u}\right]-D^{(2)}u\Big|_{\boldsymbol{x}_{\alpha}}=\sum_{\beta}\sum_{\gamma}\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma}\left\{\frac{1}{3!}\,D^{(3)}u\Big|_{\boldsymbol{x}_{\alpha}}\bullet\boldsymbol{z}_{\alpha\gamma}^{(3)}+\mathrm{O}\left(h^{4}\right)\right\}$$

qui est $\mathrm{O}\left(h\right)$ puisque $\boldsymbol{z}_{\alpha\beta}^{(3)}$ est $\mathrm{O}\left(h^{3}\right)$ et la définition 7.4.1 implique que $\|\boldsymbol{\sigma}_{\alpha\beta}\|=\mathrm{O}\left(h^{-1}\right)$. □

Grâce à la relation (8.3.18), l'opérateur approché $\widehat{\boldsymbol{\theta}}_{\alpha}$ permet de reproduire la dérivée seconde de tout polynôme de degré deux si l'application de Brenner $\mathfrak{T}_{\alpha}^{(2)}$ est inversible. Pour reproduire les polynômes de degré deux, il faut également un gradient consistant de précision à l'ordre deux. Le gradient (8.3.1) est a priori seulement de précision à l'ordre un, ce qui signifie qu'on a en général, pour un polynôme p de degré deux,

$$\boldsymbol{\sigma}_{\alpha}\left[\mathfrak{p}\right]\neq D^{(1)}p\Big|_{\boldsymbol{x}_{\alpha}}\,.$$

La solution à ce problème est fournie par le

Lemme 8.3.7 (Gradient corrigé de précision à l'ordre deux). *Supposons l'application de Brenner (8.3.20) inversible d'inverse*

$$\mathfrak{C}_{\alpha}^{(2)}\cdot\widehat{\boldsymbol{\theta}}_{\alpha}\left[\mathfrak{p}\right]=D^{(2)}p\Big|_{\boldsymbol{x}_{\alpha}} \tag{8.3.30}$$

et soit u une fonction suffisamment régulière. L'application $\mathfrak{C}_{\alpha}^{(2)}$ et l'opérateur approché $\widehat{\boldsymbol{\theta}}_{\alpha}$ permettent de définir un opérateur de gradient corrigé

$$\check{\boldsymbol{\sigma}}_{\alpha}\left[\mathfrak{u}\right]\triangleq\sum_{\beta}\boldsymbol{\sigma}_{\alpha\beta}\overline{u}_{\beta}-\frac{1}{2!}\sum_{\beta}\boldsymbol{\sigma}_{\alpha\beta}\left(\left(\mathfrak{C}_{\alpha}^{(2)}\cdot\widehat{\boldsymbol{\theta}}_{\alpha}\left[\mathfrak{u}\right]\right)\bullet\boldsymbol{z}_{\alpha\beta}^{(2)}\right) \tag{8.3.31}$$

qui reproduit le gradient de tout polynôme de degré deux. Sous la condition

$$\mathfrak{C}_{\alpha}^{(2)}\cdot\widehat{\boldsymbol{\theta}}_{\alpha\gamma}=\mathrm{O}\left(h^{-2}\right)\,,$$

l'erreur d'approximation est

$$\begin{aligned}\check{\boldsymbol{\sigma}}_{\alpha}\left[\mathfrak{u}\right]-D^{(1)}u\Big|_{\boldsymbol{x}_{\alpha}}&=\sum_{\beta}\boldsymbol{\sigma}_{\alpha\beta}\left\{\frac{1}{3!}\,D^{(3)}u\Big|_{\boldsymbol{x}_{\alpha}}\bullet\boldsymbol{z}_{\alpha\beta}^{(3)}+\mathrm{O}\left(h^{4}\right)\right\}-\\&-\frac{1}{2!}\sum_{\beta}\boldsymbol{\sigma}_{\alpha\beta}\sum_{\gamma}\left(\boldsymbol{z}_{\alpha\beta}^{(2)}\bullet\left(\mathfrak{C}_{\alpha}^{(2)}\cdot\widehat{\boldsymbol{\theta}}_{\alpha\gamma}\right)\right)\left\{\frac{1}{3!}\,D^{(3)}u\Big|_{\boldsymbol{x}_{\alpha}}\bullet\boldsymbol{z}_{\alpha\gamma}^{(3)}+\mathrm{O}\left(h^{4}\right)\right\}=\mathrm{O}\left(h^{2}\right)\,.\end{aligned} \tag{8.3.32}$$

Démonstration. Soit p un polynôme de degré deux. Les moyennes de p satisfont

$$\overline{p}_\beta = p\left(\boldsymbol{x}_\alpha\right) + D^{(1)}p\Big|_{\boldsymbol{x}_\alpha} \cdot \boldsymbol{h}_{\alpha\beta} + \frac{1}{2!}\, D^{(2)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}^{(2)}_{\alpha\beta}\,.$$

L'application du gradient consistant de précision à l'ordre un à p donne

$$\boldsymbol{\sigma}_\alpha\left[\mathfrak{p}\right] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{p}_\beta = D^{(1)}p\Big|_{\boldsymbol{x}_\alpha} + \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\frac{1}{2!}\, D^{(2)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}^{(2)}_{\alpha\beta}\,.$$

L'identité (8.3.30) prouve alors

$$\check{\boldsymbol{\sigma}}_\alpha\left[\mathfrak{p}\right] \;=\; \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{p}_\beta - \frac{1}{2!}\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\left(\mathfrak{C}^{(2)}_\alpha \cdot \widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{p}\right]\right) \bullet \boldsymbol{z}^{(2)}_{\alpha\beta} = D^{(1)}p\Big|_{\boldsymbol{x}_\alpha}\,.$$

Pour une fonction u suffisamment régulière, il faut insérer son développement de Taylor (5.8.1) dans la formule (8.3.31) pour obtenir (8.3.32). □

L'inverse $\mathfrak{C}^{(2)}_\alpha$ de l'application de Brenner et l'opérateur approché $\widehat{\boldsymbol{\theta}}_\alpha$ permettent de reconstruire le gradient et la dérivée seconde de tout polynôme de degré deux. Cela permet d'implémenter l'algorithme suivant, appelé méthode CS.

Algorithme 8.3.8 (Reconstruction quadratique par la méthode des corrections successives ou méthode CS). *La méthode des corrections successives permet de reconstruire les polynômes de degré deux en plusieurs étapes.*

(1) *La première étape consiste à calculer dans chaque cellule un gradient consistant* $\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right]$ *de précision à l'ordre un par la formule*

$$\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta\,.$$

Le calcul de $\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right]$ *peut s'effectuer sur le premier voisinage. Les gradients* $\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right]$ *sont stockés dans un tableau indexé par le numéro de cellule.*

(2) *Dans une seconde étape, l'algorithme calcule l'opérateur approché* $\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{u}\right]$ *par la formule*

$$\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{\sigma}_\beta\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \odot \left(\sum_\gamma \boldsymbol{\sigma}_{\beta\gamma}\overline{u}_\gamma\right)\,.$$

Ce calcul peut s'effectuer sur le premier voisinage.

(3) *L'application* $\mathfrak{C}^{(2)}_\alpha$ *permet ensuite de déterminer à partir de* $\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{u}\right]$ *une dérivée seconde consistante au sens de la définition 7.4.1 par la formule*

$$\boldsymbol{\theta}_\alpha\left[\mathfrak{u}\right] = \mathfrak{C}^{(2)}_\alpha \cdot \widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{u}\right]\,.$$

(4) *Finalement, l'application* $\mathfrak{C}^{(2)}_\alpha$ *permet de calculer un gradient consistant* $\check{\boldsymbol{\sigma}}_\alpha\left[\mathfrak{u}\right]$ *de précision à l'ordre deux par la formule (8.3.31).*

L'application $\mathfrak{C}^{(2)}_\alpha$ *peut être réalisée par une matrice qui est déterminée avant le début du calcul. Elle corrige l'erreur de l'opérateur approché* $\widehat{\boldsymbol{\theta}}_\alpha\left[\mathfrak{u}\right]$*, d'où le nom de « correction successive ».*

L'algorithme 8.3.8 permet a priori de reproduire les polynômes de degré deux par un processus itératif qui ne nécessite que la connectivité entre cellules adjacentes. Il est nécessaire que l'application de Brenner soit inversible dans toutes les cellules pour que cet algorithme puisse fonctionner. Comme les travaux théoriques n'ont pas pu démontrer l'inversibilité de l'application de Brenner dans le cas général, il est indispensable de conduire une campagne d'expériences numériques pour déterminer si l'application de Brenner est inversible sur des maillages utilisés en pratique.

Le calcul de l'application $\mathfrak{C}^{(2)}_\alpha$ nécessite le calcul du tenseur $\boldsymbol{\tau}^{(4)}_\alpha$. Dans une implémentation concrète, il est intéressant d'exprimer le tenseur $\boldsymbol{\tau}^{(4)}_\alpha$ en fonction des barycentres $\boldsymbol{x}_\gamma$, ce qui évite de calculer les vecteurs $\boldsymbol{h}_{\alpha\gamma}$. La formule est explicitée par le lemme suivant qui est donné sans preuve.

LEMME 8.3.9. *Le tenseur $\boldsymbol{\tau}_\alpha^{(4)}$ défini par (8.3.17) peut être mis sous la forme*

$$\boldsymbol{\tau}_\alpha^{(4)} = \frac{1}{2}\sum_{\beta,\gamma}\left(\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{\sigma}_{\beta\gamma}\right)\otimes\left[\boldsymbol{x}_\gamma\otimes\boldsymbol{x}_\gamma+\boldsymbol{x}_\gamma^{(2)}\right].$$

8.3.2. Analyse de la méthode des corrections successives pour les polynômes de degré élevé. La méthode discutée dans la section 8.3.1 se généralise aux reconstructions de degré supérieur à deux. Dans cette section, on suppose donnés un gradient consistant et une dérivée consistante d'ordre k. L'objectif est de construire une dérivée consistante d'ordre $k+1$ à partir de ces deux opérateurs. Cette approche permet en théorie de calculer des dérivées consistantes d'ordre arbitraire itérativement à partir d'un gradient consistant.

Soit u une fonction suffisamment régulière de moyennes $\mathfrak{u}=(\overline{u}_1,\ldots,\overline{u}_N)$. Soit

$$\boldsymbol{w}_\alpha^{(k)}[\mathfrak{u}]=\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(k)}\overline{u}_\beta \tag{8.3.33}$$

une dérivée consistante au sens de la définition 7.4.1. Par définition de la consistance, les coefficients $\boldsymbol{w}_{\alpha\beta}^{(k)}$ satisfont les conditions (7.3.9)

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(k)}\otimes\boldsymbol{z}_{\alpha\beta}^{(j)}=\begin{cases} j!\boldsymbol{\delta}^{(2j)} & \text{si } j=k\\ 0 & \text{si } j\neq k\end{cases}\quad 0\le j\le k$$

et l'erreur d'approximation

$$\boldsymbol{w}_{\alpha\beta}^{(k)}-D^{(k)}u\Big|_{\boldsymbol{x}_\alpha}=\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(k)}\left[\frac{1}{(k+1)!}\,D^{(k+1)}u\Big|_{\boldsymbol{x}_\alpha}\bullet\boldsymbol{z}_{\alpha\beta}^{(k+1)}+\mathrm{O}\left(h^{k+2}\right)\right]$$

est $\mathrm{O}(h)$. Soit

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}]=\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta \tag{8.3.34}$$

un gradient consistant de précision à l'ordre un au sens de la définition 7.4.1 qui permet de reproduire les polynômes de degré un grâce aux identités (8.3.2) et (8.3.3)

$$\begin{aligned}\sum_\beta \boldsymbol{\sigma}_{\alpha\beta} &= 0\\ \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\otimes\boldsymbol{h}_{\alpha\beta} &= \boldsymbol{\delta}^{(2)}.\end{aligned}$$

On rappelle que $\boldsymbol{w}_{\alpha\beta}^{(k)}\triangleq 0$ et $\boldsymbol{\sigma}_{\alpha\beta}\triangleq 0$ si la cellule $\mathcal{T}_\beta$ ne contribue pas à la reconstruction dans la cellule $\mathcal{T}_\alpha$. D'après la définition 7.4.1, l'erreur d'approximation

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}]-D^{(1)}u\Big|_{\boldsymbol{x}_\alpha}=\frac{1}{2}\sum_\beta\boldsymbol{\sigma}_{\alpha\beta}\left[D^{(2)}u\Big|_{\boldsymbol{x}_\alpha}\bullet\boldsymbol{z}_{\alpha\beta}^{(2)}+\mathrm{O}\left(h^3\right)\right]$$

est $\mathrm{O}(h)$.

Les opérateurs (8.3.33) et (8.3.34) permettent de formuler l'opérateur

$$\widehat{\boldsymbol{w}}_\alpha^{(k+1)}[\mathfrak{u}]\triangleq\sum_\beta\sum_\gamma\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{w}_{\beta\gamma}^{(k)}\,\overline{u}_\gamma\,. \tag{8.3.35}$$

Le tenseur (8.3.35) est symétrique comme la dérivée d'ordre $k+1$ d'une fonction de classe C^{k+1}. On définit les tenseurs $\widehat{\boldsymbol{w}}_{\alpha\gamma}^{(k+1)}$ par

$$\widehat{\boldsymbol{w}}_{\alpha\gamma}^{(k+1)}\triangleq\sum_\beta\boldsymbol{\sigma}_{\alpha\beta}\odot\boldsymbol{w}_{\beta\gamma}^{(k)}\,. \tag{8.3.36}$$

On peut alors énoncer un lemme analogue au lemme 8.3.1.

LEMME 8.3.10 (Noyau de l'opérateur approché $\widehat{\boldsymbol{w}}_\alpha^{(k+1)}$). *Soit $\boldsymbol{w}_\alpha^{(k)}[\mathfrak{u}]$ une dérivée consistante qui satisfait (7.3.9) et $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$ un gradient consistant qui satisfait (8.3.2) et (8.3.3). L'opérateur approché (8.3.35), défini par*

$$\widehat{\boldsymbol{w}}_\alpha^{(k+1)}[\mathfrak{u}] = \sum_\gamma \boldsymbol{w}_{\alpha\gamma}^{(k+1)} \overline{u}_\gamma \triangleq \sum_\beta \sum_\gamma \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)} \overline{u}_\gamma \,,$$

satisfait

$$\sum_\gamma \widehat{\boldsymbol{w}}_{\alpha\gamma}^{(k+1)} \otimes \boldsymbol{z}_{\alpha\gamma}^{(j)} = 0$$

pour $0 \le j \le k$.

DÉMONSTRATION. Les identités (7.3.9) s'écrivent de façon explicite

$$\sum_\gamma w_{i_1\cdots i_k}^{\beta\gamma} z_{n_1\cdots n_j}^{\beta\gamma} = \begin{cases} k! \left(\boldsymbol{\delta}^{(2k)}\right)_{i_1\cdots i_k\, n_1\cdots n_k} & \text{si } j = k \\ 0 & \text{si } j \ne k \end{cases} \quad 0 \le j \le k\,.$$

Soit $0 \le j \le k$. La formule du binôme (5.7.15)

$$\boldsymbol{z}_{\alpha\gamma}^{(j)} = \sum_{m=0}^{j} \binom{j}{m} \boldsymbol{z}_{\beta\gamma}^{(m)} \odot \boldsymbol{h}_{\alpha\beta}^{j-m}$$

permet d'écrire

$$\sum_\gamma \widehat{\boldsymbol{w}}_{\alpha\gamma}^{(k+1)} \otimes \boldsymbol{z}_{\alpha\gamma}^{(j)} = \sum_{m=0}^{j} \binom{j}{m} \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\right) \otimes \left(\boldsymbol{z}_{\beta\gamma}^{(m)} \odot \boldsymbol{h}_{\alpha\beta}^{j-m}\right) . \tag{8.3.37}$$

Le tenseur

$$\sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\right) \otimes \left(\boldsymbol{z}_{\beta\gamma}^{(m)} \odot \boldsymbol{h}_{\alpha\beta}^{j-m}\right) \tag{8.3.38}$$

est de composantes

$$\frac{1}{(k+1)!j!} \sum_\beta \sum_\gamma \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_j} \sigma_{i_{\pi(1)}}^{\alpha\beta} w_{i_{\pi(2)}\cdots i_{\pi(k+1)}}^{\beta\gamma} z_{n_{\varsigma(1)}\cdots n_{\varsigma(m)}}^{\beta\gamma} h_{n_{\varsigma(m+1)}}^{\alpha\beta} \cdots h_{n_{\varsigma(j)}}^{\alpha\beta} \,.$$

La condition (7.3.9) prouve que les composantes de (8.3.38) sont nulles si $m \ne k$. Puisque $m \le j \le k$ dans (8.3.37), il suffit de calculer les composantes de (8.3.38) pour le cas $m = j = k$, ce qui donne

$$\frac{1}{(k+1)!k!} \sum_\beta \sum_\gamma \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_k} \sigma_{i_{\pi(1)}}^{\alpha\beta} w_{i_{\pi(2)}\cdots i_{\pi(k+1)}}^{\beta\gamma} z_{n_{\varsigma(1)}\cdots n_{\varsigma(k)}}^{\beta\gamma} =$$

$$= \frac{1}{(k+1)!k!} \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_k} \left(\sum_\beta \sigma_{i_{\pi(1)}}^{\alpha\beta}\right) k! \left(\boldsymbol{\delta}^{(2k)}\right)_{i_{\pi(2)}\cdots i_{\pi(k+1)}\, n_{\varsigma(1)}\cdots n_{\varsigma(k)}} = 0$$

en raison des identités (7.3.9) et de la propriété (8.3.2)

$$\sum_\beta \sigma_i^{\alpha\beta} = 0\,,\ 1 \le i \le d.$$

Par conséquent, toutes les composantes de (8.3.37) sont nulles, ce qui prouve le lemme. □

Le lemme 8.3.10 entraîne le

COROLLAIRE 8.3.11. *L'opérateur (8.3.35) s'annule sur les polynômes de degré k.*

DÉMONSTRATION. Soit q un polynôme de degré k. La formule (5.8.4) permet de développer les moyennes de q comme

$$\overline{q}_\gamma = \sum_{i=0}^{k} \frac{1}{i!}\, D^{(i)} q\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\gamma}^{(i)}\,.$$

Le lemme 8.3.10 permet de conclure que

$$\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{q}] = \sum_\beta \sum_\gamma \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)} \overline{q}_\gamma = 0\,.$$

□

Le corollaire 8.3.11 permet de calculer (8.3.35) pour un polynôme p de degré $k+1$ avec les moyennes $\mathfrak{p} = (\overline{p}_1, \ldots, \overline{p}_N)$. La moyenne de p sur une cellule $\mathcal{T}_\gamma$ s'écrit selon la formule (5.8.4)

$$\overline{p}_\gamma \;=\; \sum_{i=0}^{k+1} \frac{1}{i!}\, D^{(i)} p\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\gamma}^{(i)}\,. \tag{8.3.39}$$

L'insertion de (8.3.39) dans (8.3.35) et le corollaire 8.3.11 entraînent

$$\begin{aligned}\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{p}] = \sum_\beta \sum_\gamma \widehat{\boldsymbol{w}}_{\alpha\gamma}^{(k+1)} \overline{p}_\gamma = \sum_\beta \sum_\gamma \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\, \overline{p}_\gamma = \\ = \frac{1}{(k+1)!} \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\right) \left(\boldsymbol{z}_{\alpha\gamma}^{(k+1)} \bullet D^{(k+1)} p\Big|_{\boldsymbol{x}_\alpha}\right)\,. \end{aligned} \tag{8.3.40}$$

La définition du tenseur $\boldsymbol{\tau}_\alpha^{(2k+2)}$ d'ordre $2k+2$ par

$$\boldsymbol{\tau}_\alpha^{(2k+2)} \triangleq \frac{1}{(k+1)!} \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\right) \otimes \boldsymbol{z}_{\alpha\gamma}^{(k+1)} \tag{8.3.41}$$

permet d'écrire (8.3.40), à l'aide de la notation (5.2.11), comme la relation linéaire

$$\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{p}] = \boldsymbol{\tau}_\alpha^{(2k+2)} \cdot D^{(k+1)} p\Big|_{\boldsymbol{x}_\alpha}\,. \tag{8.3.42}$$

La relation (8.3.42) justifie la

DÉFINITION 8.3.12 (*Application de Brenner d'ordre $k+1$*). Soit $\mathbb{S}_{k+1}^d$ l'espace des tenseurs symétriques d'ordre $k+1$ dans $\mathbb{R}^d$. La relation (8.3.42) définit une application linéaire de l'espace des tenseurs symétriques $\mathbb{S}_{k+1}^d$ dans lui-même. Cette application, dite *application de Brenner d'ordre $k+1$*, est définie par

$$\mathfrak{T}_\alpha^{(2k+2)} : \left\{ \begin{array}{ccc} \mathbb{S}_{k+1}^d & \longrightarrow & \mathbb{S}_{k+1}^d \\ \boldsymbol{a}^{(k+1)} & \longmapsto & \boldsymbol{\tau}_\alpha^{(2k+2)} \cdot \boldsymbol{a}^{(k+1)} \end{array} \right. . \tag{8.3.43}$$

□

Les valeurs $\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{p}]$ de l'opérateur (8.3.35) se calculent facilement à l'aide des opérateurs $\boldsymbol{w}_\alpha^{(k)}$ et $\boldsymbol{\sigma}_\alpha$. Si l'application de Brenner d'ordre $k+1$ (8.3.43) peut être inversée, son inverse $\mathfrak{C}_\alpha^{(2k+2)}$ permet de calculer la dérivée exacte $D^{(k+1)} p\big|_{\boldsymbol{x}_\alpha}$ comme

$$D^{(k+1)} p\Big|_{\boldsymbol{x}_\alpha} = \mathfrak{C}_\alpha^{(2k+2)} \cdot \widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{p}]\,.$$

Une application linéaire d'un espace vectoriel de dimension finie dans lui-même peut être inversée si et seulement si elle est injective. En raison de la linéarité, l'application de Brenner (8.3.43) est injective si et seulement si

$$\boldsymbol{\tau}_\alpha^{(2k+2)} \cdot \boldsymbol{a}^{(k+1)} = 0 \text{ entraîne } \boldsymbol{a}^{(k+1)} = 0 \tag{8.3.44}$$

pour tout tenseur symétrique $\boldsymbol{a}^{(k+1)}$ d'ordre $k+1$. Il faut donc chercher des conditions pour que la condition (8.3.44) soit vraie.

Il est bien connu qu'une application linéaire d'un espace vectoriel dans lui-même est inversible si elle est suffisamment proche de l'application identité dans une norme quelconque. L'application identité est dans le cas présent donnée par le tenseur $\boldsymbol{\delta}^{(2k+2)}$, défini par (5.2.22), car

$$\boldsymbol{\delta}^{(2k+2)} \cdot \boldsymbol{a}^{(k+1)} = \boldsymbol{a}^{(k+1)}$$

pour tout tenseur symétrique $\boldsymbol{a}^{(k+1)}$ d'ordre $k+1$. Il est donc intéressant de pouvoir représenter le tenseur $\boldsymbol{\tau}_\alpha^{(2k+2)}$ comme la somme de l'identité $\boldsymbol{\delta}^{(2k+2)}$ et d'une perturbation. Ceci fait l'objet du

LEMME 8.3.13 (Forme alternative du tenseur $\boldsymbol{\tau}_\alpha^{(2k+2)}$). *Le tenseur $\boldsymbol{\tau}_\alpha^{(2k+2)}$ défini par (8.3.41) s'écrit sous la forme*

$$\boldsymbol{\tau}_\alpha^{(2k+2)} \;=\; \boldsymbol{\delta}^{(2k+2)} + \frac{1}{(k+1)!} \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)} \right) \otimes \boldsymbol{z}_{\beta\gamma}^{(k+1)} .$$

DÉMONSTRATION. La formule du binôme (5.7.15)

$$\boldsymbol{z}_{\alpha\gamma}^{(k+1)} = \sum_{m=0}^{k+1} \binom{k+1}{m} \boldsymbol{z}_{\beta\gamma}^{(m)} \odot \boldsymbol{h}_{\alpha\beta}^{k+1-m}$$

permet d'écrire

$$\begin{aligned} \boldsymbol{\tau}_\alpha^{(2k+2)} = \frac{1}{(k+1)!} \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)} \right) \otimes \boldsymbol{z}_{\alpha\gamma}^{(k+1)} = \\ = \frac{1}{(k+1)!} \sum_{m=0}^{k+1} \binom{k+1}{m} \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)} \right) \otimes \left(\boldsymbol{z}_{\beta\gamma}^{(m)} \odot \boldsymbol{h}_{\alpha\beta}^{k+1-m} \right) . \end{aligned} \tag{8.3.45}$$

Le tenseur

$$\boldsymbol{\upsilon}_\alpha^{(k|m)} \triangleq \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)} \right) \otimes \left(\boldsymbol{z}_{\beta\gamma}^{(m)} \odot \boldsymbol{h}_{\alpha\beta}^{k+1-m} \right) \tag{8.3.46}$$

est de composantes

$$\begin{aligned} \upsilon_{\alpha,\, i_1 \cdots i_{k+1}\, n_1 \cdots n_{k+1}}^{(k|m)} = \\ = \frac{1}{((k+1)!)^2} \sum_\beta \sum_\gamma \sum_{\pi \in \mathfrak{S}_{k+1}} \sum_{\varsigma \in \mathfrak{S}_{k+1}} \sigma_{i_{\pi(1)}}^{\alpha\beta} w_{i_{\pi(2)} \cdots i_{\pi(k+1)}}^{\beta\gamma} z_{n_{\varsigma(1)} \cdots n_{\varsigma(m)}}^{\beta\gamma} h_{n_{\varsigma(m+1)}}^{\alpha\beta} \cdots h_{n_{\varsigma(k+1)}}^{\alpha\beta} . \end{aligned}$$

La condition (7.3.9) prouve que les composantes de (8.3.46) sont nulles si $m < k$, ce qui permet d'écrire (8.3.45) sous la forme

$$\boldsymbol{\tau}_\alpha^{(2k+2)} = \frac{1}{(k+1)!} \left\{ (k+1)\, \boldsymbol{\upsilon}_\alpha^{(k|k)} + \boldsymbol{\upsilon}_\alpha^{(k|k+1)} \right\} . \tag{8.3.47}$$

Il suffit donc de calculer les composantes de (8.3.46) pour $m = k$ et $m = k+1$. Pour $m = k$, les composantes du tenseur $\boldsymbol{\upsilon}_\alpha^{(k|k)}$ sont

$$\upsilon_{\alpha,\, i_1 \cdots i_{k+1}\, n_1 \cdots n_{k+1}}^{(k|k)} = \frac{1}{((k+1)!)^2} \sum_\beta \sum_\gamma \sum_{\pi \in \mathfrak{S}_{k+1}} \sum_{\varsigma \in \mathfrak{S}_{k+1}} \sigma_{i_{\pi(1)}}^{\alpha\beta} w_{i_{\pi(2)} \cdots i_{\pi(k+1)}}^{\beta\gamma} z_{n_{\varsigma(1)} \cdots n_{\varsigma(k)}}^{\beta\gamma} h_{n_{\varsigma(k+1)}}^{\alpha\beta} . \tag{8.3.48}$$

Comme la somme sur $\pi \in \mathfrak{S}_{k+1}$ dans (8.3.48) parcourt toutes les permutations de $\{1, \ldots, k+1\}$, il est possible de réarranger les indices afin d'écrire (8.3.48) sous la forme

$$\upsilon_{\alpha,\, i_1 \cdots i_{k+1}\, n_1 \cdots n_{k+1}}^{(k|k)} = \frac{1}{((k+1)!)^2} \sum_\beta \sum_\gamma \sum_{\pi \in \mathfrak{S}_{k+1}} \sum_{\varsigma \in \mathfrak{S}_{k+1}} w_{i_{\pi(1)} \cdots i_{\pi(k)}}^{\beta\gamma} z_{n_{\varsigma(1)} \cdots n_{\varsigma(k)}}^{\beta\gamma} \sigma_{i_{\pi(k+1)}}^{\alpha\beta} h_{n_{\varsigma(k+1)}}^{\alpha\beta} . \tag{8.3.49}$$

Les conditions (7.3.9) et (8.3.3) donnent

$$\upsilon^{(k|k)}_{\alpha,\, i_1\cdots i_{k+1}\, n_1\cdots n_{k+1}} = \frac{k!}{\left((k+1)!\right)^2} \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_{k+1}} \left(\boldsymbol{\delta}^{(2k)}\right)_{i_{\pi(1)}\cdots i_{\pi(k)}\, n_{\varsigma(1)}\cdots n_{\varsigma(k)}} \delta_{i_{\pi(k+1)} n_{\varsigma(k+1)}} \,. \tag{8.3.50}$$

La définition (5.2.22)

$$\left(\boldsymbol{\delta}^{(2k)}\right)_{i_1\cdots i_k\, n_1\cdots n_k} \triangleq \frac{1}{k!} \sum_{\varrho\in\mathfrak{S}_k} \delta_{i_1 n_{\varrho(1)}} \cdots \delta_{i_k n_{\varrho(k)}}$$

entraîne pour toute permutation $\varsigma \in \mathfrak{S}_{k+1}$

$$\left(\boldsymbol{\delta}^{(2k)}\right)_{i_1\cdots i_k\, n_{\varsigma(1)}\cdots n_{\varsigma(k)}} \triangleq \frac{1}{k!} \sum_{\varrho\in\mathfrak{S}_k} \delta_{i_1 n_{\varsigma(\varrho(1))}} \cdots \delta_{i_k n_{\varsigma(\varrho(k))}} \,. \tag{8.3.51}$$

L'insertion de (8.3.51) dans (8.3.50) a pour résultat

$$\upsilon^{(k|k)}_{\alpha,\, i_1\cdots i_{k+1}\, n_1\cdots n_{k+1}} = \frac{1}{\left((k+1)!\right)^2} \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_{k+1}} \sum_{\varrho\in\mathfrak{S}_k} \delta_{i_{\pi(1)} n_{\varsigma(\varrho(1))}} \cdots \delta_{i_{\pi(k)} n_{\varsigma(\varrho(k))}} \delta_{i_{\pi(k+1)} n_{\varsigma(k+1)}} \,. \tag{8.3.52}$$

Il est possible d'effectuer les sommes sur les permutations $\pi \in \mathfrak{S}_{k+1}$, $\varsigma \in \mathfrak{S}_{k+1}$ et $\varrho \in \mathfrak{S}_k$ en (8.3.52) dans un ordre arbitraire. Gardons d'abord les permutations $\pi \in \mathfrak{S}_{k+1}$ et $\varrho \in \mathfrak{S}_k$ fixes. Comme $\mathfrak{S}_{k+1}$ est un groupe, la somme sur les permutations $\varsigma \in \mathfrak{S}_{k+1}$ peut parcourir tous les éléments $\varsigma \circ \widetilde{\varsigma} \in \mathfrak{S}_{k+1}$ où $\widetilde{\varsigma}$ est une permutation fixe définie par

$$\widetilde{\varsigma}(j) \triangleq \begin{cases} \varrho^{-1}(j) & \text{si } 1 \leq j \leq k \\ k+1 & \text{si } j = k+1 \end{cases} .$$

Cela permet d'écrire (8.3.52) sous la forme

$$\upsilon^{(k|k)}_{\alpha,\, i_1\cdots i_{k+1}\, n_1\cdots n_{k+1}} = \frac{1}{\left((k+1)!\right)^2} \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_{k+1}} \sum_{\varrho\in\mathfrak{S}_k} \delta_{i_{\pi(1)} n_{\varsigma(1)}} \cdots \delta_{i_{\pi(k)} n_{\varsigma(k)}} \delta_{i_{\pi(k+1)} n_{\varsigma(k+1)}} \tag{8.3.53}$$

où la permutation $\varrho \in \mathfrak{S}_k$ n'apparaît plus dans les indices des termes de la somme. La somme sur $\varrho \in \mathfrak{S}_k$ se calcule donc facilement car elle occasionne uniquement un facteur $k!$, ce qui donne

$$\upsilon^{(k|k)}_{\alpha,\, i_1\cdots i_{k+1}\, n_1\cdots n_{k+1}} = \frac{k!}{\left((k+1)!\right)^2} \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_{k+1}} \delta_{i_{\pi(1)} n_{\varsigma(1)}} \cdots \delta_{i_{\pi(k)} n_{\varsigma(k)}} \delta_{i_{\pi(k+1)} n_{\varsigma(k+1)}} \tag{8.3.54}$$

Gardons ensuite $\pi \in \mathfrak{S}_{k+1}$ fixe. Comme $\mathfrak{S}_{k+1}$ est un groupe, la somme sur les permutations $\varsigma \in \mathfrak{S}_{k+1}$ peut parcourir tous les éléments $\varsigma \circ \pi \in \mathfrak{S}_{k+1}$, ce qui autorise à écrire (8.3.54) sous la forme

$$\upsilon^{(k|k)}_{\alpha,\, i_1\cdots i_{k+1}\, n_1\cdots n_{k+1}} = \frac{k!}{\left((k+1)!\right)^2} \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_{k+1}} \delta_{i_{\pi(1)} n_{\varsigma(\pi(1))}} \cdots \delta_{i_{\pi(k+1)} n_{\varsigma(\pi(k+1))}} \,. \tag{8.3.55}$$

Un simple réarrangement des facteurs $\delta_{i_{\pi(l)} n_{\varsigma(\pi(l))}}$ dans (8.3.55) permet alors d'écrire

$$\upsilon^{(k|k)}_{\alpha,\, i_1\cdots i_{k+1}\, n_1\cdots n_{k+1}} = \frac{k!}{\left((k+1)!\right)^2} \sum_{\pi\in\mathfrak{S}_{k+1}} \sum_{\varsigma\in\mathfrak{S}_{k+1}} \delta_{i_1 n_{\varsigma(1)}} \cdots \delta_{i_{k+1} n_{\varsigma(k+1)}} \,. \tag{8.3.56}$$

La somme sur $\pi \in \mathfrak{S}_{k+1}$ engendre uniquement un facteur $(k+1)!$ et le résultat final est

$$\upsilon^{(k|k)}_{\alpha,\, i_1\cdots i_{k+1}\, n_1\cdots n_{k+1}} = \frac{k!}{(k+1)!} \sum_{\varsigma\in\mathfrak{S}_{k+1}} \delta_{i_1 n_{\varsigma(1)}} \cdots \delta_{i_{k+1} n_{\varsigma(k+1)}} \,, \tag{8.3.57}$$

ce qui s'écrit en notation tensorielle

$$\boldsymbol{\upsilon}^{(k|k)}_{\alpha} = k!\boldsymbol{\delta}^{(2k+2)} \,. \tag{8.3.58}$$

L'insertion de (8.3.58) dans (8.3.47) a pour résultat

$$\boldsymbol{\tau}^{(2k+2)}_{\alpha} = \boldsymbol{\delta}^{(2k+2)} + \frac{1}{(k+1)!} \boldsymbol{\upsilon}^{(k|k+1)}_{\alpha} \,,$$

ce qui prouve avec la définition

$$\boldsymbol{v}_\alpha^{(k|k+1)} \triangleq \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\right) \otimes \boldsymbol{z}_{\beta\gamma}^{(k+1)}$$

le lemme 8.3.13. □

Le lemme 8.3.13 montre que la condition

$$\sum_{\beta,\gamma} \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\right) \otimes \boldsymbol{z}_{\beta\gamma}^{(k+1)} = 0 \tag{8.3.59}$$

suffit pour que $\boldsymbol{\tau}_\alpha^{(2k+2)}$ soit égal à $\boldsymbol{\delta}^{(2k+2)}$. Dans ce cas particulier, l'opérateur (8.3.35) reproduit la dérivée d'ordre $k+1$ de tout polynôme de degré $k+1$. Ce résultat donne lieu à la

Remarque 8.3.14. Le lemme 8.3.13 permet d'interpréter l'application de Brenner (8.3.43), induite par $\boldsymbol{\tau}_\alpha^{(2k+2)}$, comme une perturbation de l'application identité induite par $\boldsymbol{\delta}^{(2k+2)}$. Puisque $\mathbb{S}_{k+1}^d$ est un espace vectoriel réel de dimension finie, $\mathbb{S}_{k+1}^d$ est un espace de Banach pour une norme arbitraire. Par conséquent, une application linéaire de $\mathbb{S}_{k+1}^d$ en lui-même, qui est une perturbation de l'identité, peut être inversée s'il existe une norme telle que la norme de la perturbation soit plus petite que un.

Dans la suite, on identifie un cas particulier où cette perturbation est nulle et dans lequel l'application de Brenner devient par conséquent l'identité. Dans le cas général, seuls des essais numériques montreront si l'application de Brenner (8.3.43) peut être inversée. □

La proposition suivante donne une condition suffisante pour (8.3.59).

Proposition 8.3.15 (Condition suffisante pour la consistance). *Si la dérivée consistante d'ordre k*

$$\boldsymbol{w}_\beta^{(k)}\,[\mathfrak{u}] = \sum_\gamma \boldsymbol{w}_{\beta\gamma}^{(k)} \overline{u}_\gamma$$

satisfait la condition de précision à l'ordre deux

$$\sum_\gamma \boldsymbol{w}_{\beta\gamma}^{(k)} \otimes \boldsymbol{z}_{\beta\gamma}^{(k+1)} = 0\,, \tag{8.3.60}$$

l'opérateur (8.3.35) défini par

$$\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{u}] \triangleq \sum_\beta \sum_\gamma \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\, \overline{u}_\gamma$$

est une dérivée consistante d'ordre $k+1$ au sens de la définition 7.4.1.

Démonstration. Le lemme 8.3.13 permet d'écrire

$$\boldsymbol{\tau}_\alpha^{(2k+2)} = \boldsymbol{\delta}^{(2k+2)} + \frac{1}{(k+1)!} \sum_\beta \sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\right) \otimes \boldsymbol{z}_{\beta\gamma}^{(k+1)}\,.$$

Par conséquent, il suffit de montrer que (8.3.60) implique

$$\sum_\gamma \left(\boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}_{\beta\gamma}^{(k)}\right) \otimes \boldsymbol{z}_{\beta\gamma}^{(k+1)} = 0 \tag{8.3.61}$$

pour prouver la proposition 8.3.15 car $\boldsymbol{\tau}_\alpha^{(2k+2)} = \boldsymbol{\delta}^{(2k+2)}$ signifie que pour tout polynôme p de degré $k+1$

$$\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{p}] = \boldsymbol{\tau}_\alpha^{(2k+2)} \cdot D^{(k+1)}p\Big|_{\boldsymbol{x}_\alpha} = \boldsymbol{\delta}^{(2k+2)} \cdot D^{(k+1)}p\Big|_{\boldsymbol{x}_\alpha} = D^{(k+1)}p\Big|_{\boldsymbol{x}_\alpha}\,.$$

De façon explicite, (8.3.61) s'écrit

$$\sum_\gamma \sum_{\pi\in\mathfrak{S}_{k+1}} \sigma_{i_{\pi(1)}}^{\alpha\beta} w_{i_{\pi(2)}\cdots i_{\pi(k+1)}}^{\beta\gamma} z_{j_1\cdots j_{k+1}}^{\beta\gamma}\,,$$

ce qui est nul en raison de l'identité (8.3.60)

$$\sum_{\gamma} w^{\beta\gamma}_{n_1\cdots n_k} z^{\beta\gamma}_{j_1\cdots j_{k+1}} = 0$$

pour tous les indices $(n_1, \cdots, n_k)$ et $(j_1, \cdots, j_{k+1})$. Il reste à calculer l'erreur d'approximation pour la dérivée d'ordre $k+1$ d'une fonction u suffisamment régulière. Soit $\mathfrak{u} \triangleq (\overline{u}_1, \ldots, \overline{u}_N)$ le vecteur des moyennes de u qui se développent selon la formule (5.8.1) comme

$$\overline{u}_\gamma \;=\; u(\boldsymbol{x}_\alpha) + \sum_{j=1}^{k+2} \frac{1}{j!}\, D^{(j)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}^{(j)}_{\alpha\gamma} + \mathrm{O}\left(h^{k+3}\right). \tag{8.3.62}$$

L'insertion de (8.3.62) dans (8.3.35) donne l'erreur d'approximation

$$\widehat{\boldsymbol{w}}^{(k+1)}_\alpha[\mathfrak{u}] - D^{(k+1)}u\Big|_{\boldsymbol{x}_\alpha} = \sum_\beta \sum_\gamma \boldsymbol{\sigma}_{\alpha\beta} \odot \boldsymbol{w}^{(k)}_{\beta\gamma} \left\{ \frac{1}{(k+2)!}\, D^{(k+2)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}^{(k+2)}_{\alpha\gamma} + \mathrm{O}\left(h^{k+3}\right) \right\}$$

qui est $\mathrm{O}(h)$ puisque $\boldsymbol{z}^{(k+2)}_{\alpha\beta}$ est $\mathrm{O}\left(h^{k+2}\right)$ et la définition 7.4.1 implique $\|\boldsymbol{\sigma}_{\alpha\beta}\| = \mathrm{O}\left(h^{-1}\right)$ et $\left\|\boldsymbol{w}^{(k)}_{\beta\gamma}\right\| = \mathrm{O}\left(h^{-k}\right)$. □

Grâce à la relation (8.3.42), l'opérateur approché $\widehat{\boldsymbol{w}}^{(k+1)}_\alpha$ permet de reproduire la dérivée d'ordre $k+1$ de tout polynôme de degré $k+1$ si l'application de Brenner $\mathfrak{T}^{(2k+2)}_\alpha$ est inversible. Il faut cependant encore corriger la dérivée consistante $\boldsymbol{w}^{(k)}_\alpha$ d'ordre k pour qu'elle devienne de précision à l'ordre deux. Elle est a priori seulement de précision à l'ordre un, ce qui signifie qu'on a en général, pour un polynôme p de degré $k+1$,

$$\boldsymbol{w}^{(k)}_\alpha[\mathfrak{p}] \neq D^{(k)}p\Big|_{\boldsymbol{x}_\alpha}.$$

La solution à ce problème est fournie par le

LEMME 8.3.16 (Dérivée corrigée d'ordre k de précision à l'ordre deux). *Supposons l'application de Brenner (8.3.43) inversible d'inverse*

$$\mathfrak{C}^{(2k+2)}_\alpha \cdot \widehat{\boldsymbol{w}}^{(k+1)}_\alpha[\mathfrak{p}] = D^{(2k+2)}p\Big|_{\boldsymbol{x}_\alpha} \tag{8.3.63}$$

et soit u une fonction suffisamment régulière. L'application $\mathfrak{C}^{(2k+2)}_\alpha$ et l'opérateur approché $\widehat{\boldsymbol{w}}^{(k+1)}_\alpha$ permettent de définir une dérivée consistante d'ordre k de précision à l'ordre deux

$$\check{\boldsymbol{w}}^{(k)}_\alpha[\mathfrak{u}] \triangleq \sum_\beta \boldsymbol{w}^{(k)}_{\alpha\beta}\overline{u}_\beta - \frac{1}{(k+1)!}\sum_\beta \boldsymbol{w}^{(k)}_{\alpha\beta}\left(\left(\mathfrak{C}^{(2k+2)}_\alpha \cdot \widehat{\boldsymbol{w}}^{(k+1)}_\alpha[\mathfrak{u}]\right) \bullet \boldsymbol{z}^{(k+1)}_{\alpha\beta}\right) \tag{8.3.64}$$

qui reproduit la dérivée d'ordre k de tout polynôme de degré $k+1$. Sous la condition

$$\mathfrak{C}^{(2k+2)}_\alpha \cdot \widehat{\boldsymbol{w}}^{(k+1)}_{\alpha\gamma} = \mathrm{O}\left(h^{-k-1}\right),$$

l'erreur d'approximation dans

$$\begin{aligned}\check{\boldsymbol{w}}^{(k)}_\alpha[\mathfrak{u}] = D^{(k)}u\Big|_{\boldsymbol{x}_\alpha} &+ \sum_\beta \boldsymbol{w}^{(k)}_{\alpha\beta}\left\{ D^{(k+2)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \frac{\boldsymbol{z}^{(k+2)}_{\alpha\beta}}{(k+2)!} + \mathrm{O}\left(h^{k+3}\right)\right\} - \\ &- \sum_\beta\sum_\gamma \boldsymbol{w}^{(k)}_{\alpha\beta}\left(\frac{\boldsymbol{z}^{(k+1)}_{\alpha\beta}}{(k+1)!} \bullet \left(\mathfrak{C}^{(2k+2)}_\alpha \cdot \widehat{\boldsymbol{w}}^{(k+1)}_{\alpha\gamma}\right)\right)\left\{ D^{(k+2)}u\Big|_{\boldsymbol{x}_\alpha} \bullet \frac{\boldsymbol{z}^{(k+2)}_{\alpha\gamma}}{(k+2)!} + \mathrm{O}\left(h^{k+3}\right)\right\}\end{aligned} \tag{8.3.65}$$

est $\mathrm{O}\left(h^2\right)$.

DÉMONSTRATION. Soit p un polynôme de degré $k+1$. Les moyennes de p satisfont

$$\overline{p}_\beta = \sum_{j=0}^{k+1} \frac{1}{j!}\, D^{(j)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(j)}\,.$$

L'application de la dérivée $\boldsymbol{w}_\alpha^{(k)}$ de précision à l'ordre un à p donne

$$\boldsymbol{w}_\alpha^{(k)}\,[\mathfrak{p}] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(k)}\overline{p}_\beta = D^{(k)}p\Big|_{\boldsymbol{x}_\alpha} + \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(k)}\, D^{(k+1)}p\Big|_{\boldsymbol{x}_\alpha} \bullet \frac{\boldsymbol{z}_{\alpha\beta}^{(k+1)}}{(k+1)!}\,.$$

L'identité (8.3.63) prouve alors

$$\check{\boldsymbol{w}}_\alpha^{(k)}\,[\mathfrak{p}] \;=\; \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(k)}\overline{p}_\beta - \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(k)}\left(\mathfrak{C}_\alpha^{(2k+2)}\cdot \widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{u}]\right) \bullet \frac{\boldsymbol{z}_{\alpha\beta}^{(k+1)}}{(k+1)!} = D^{(k)}p\Big|_{\boldsymbol{x}_\alpha}\,.$$

Pour une fonction u suffisamment régulière, il faut insérer son développement de Taylor (5.8.1) dans la formule (8.3.64) pour obtenir (8.3.65). □

L'inverse $\mathfrak{C}_\alpha^{(2k+2)}$ de l'application de Brenner et l'opérateur approché $\widehat{\boldsymbol{w}}_\alpha^{(k+1)}$ permettent de reconstruire la dérivée d'ordre $k+1$ de tout polynôme de degré $k+1$. Cela permet d'implémenter l'algorithme suivant.

ALGORITHME 8.3.17 (Reconstruction de degré élevé par la méthode des corrections successives). *Supposons donnés une dérivée consistante d'ordre k et un gradient consistant au sens de la définition 7.4.1. La méthode des corrections successives permet alors de calculer une dérivée consistante d'ordre $k+1$ en suivant les étapes ci-dessous.*

(1) *La première étape consiste à calculer dans chaque cellule la dérivée consistante $\boldsymbol{w}_\alpha^{(k)}\,[\mathfrak{u}]$ d'ordre k par la formule*

$$\boldsymbol{w}_\alpha^{(k)}\,[\mathfrak{u}] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(k)}\overline{u}_\beta\,.$$

(2) *Dans une deuxième étape, l'application du gradient consistant $\boldsymbol{\sigma}_\alpha$ à $\boldsymbol{w}_\alpha^{(k)}\,[\mathfrak{u}]$ permet de déterminer l'opérateur*

$$\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \odot \left(\sum_\gamma \boldsymbol{w}_{\beta\gamma}^{(k)}\overline{u}_\gamma\right)\,.$$

(3) *L'application $\mathfrak{C}_\alpha^{(2k+2)}$ permet ensuite de calculer, à partir de $\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{u}]$, une dérivée d'ordre $k+1$ consistante au sens de la définition 7.4.1. Elle corrige l'erreur de l'opérateur $\widehat{\boldsymbol{w}}_\alpha^{(k+1)}\,[\mathfrak{u}]$.*

(4) *Finalement, le lemme 8.3.16 permet de calculer dans chaque cellule une dérivée consistante $\check{\boldsymbol{w}}_\alpha^{(k)}\,[\mathfrak{u}]$ d'ordre k de précision à l'ordre deux. Les dérivées d'ordre inférieur doivent également être corrigées mais cette étape n'est pas explicitée ici.*

L'application $\mathfrak{C}_\alpha^{(2k+2)}$ peut être réalisée par une matrice qui est déterminée avant le début du calcul.

L'algorithme 8.3.17 permet, à partir d'un gradient consistant, de calculer des dérivées consistantes d'ordre $k > 1$ par un processus itératif. Le gradient consistant peut être utilisé pour calculer d'abord une dérivée seconde consistante. Cette dérivée seconde sert à calculer une dérivée troisième consistante et ainsi de suite jusqu'à la dérivée consistante d'ordre k. Une fois celle-ci calculée, il faut encore corriger les dérivées d'ordre inférieur. Cette dernière étape n'est pas explicitée ici.

L'algorithme 8.3.17 permet donc a priori de reproduire les polynômes de degré k par un processus itératif qui ne nécessite que la connectivité entre cellules adjacentes. Il est nécessaire que l'application de Brenner soit inversible dans toutes les cellules pour que cet algorithme puisse

fonctionner. Comme les travaux théoriques n'ont pas pu démontrer l'inversibilité de l'application de Brenner dans le cas général, il est indispensable de conduire une campagne d'expériences numériques pour déterminer si l'application de Brenner d'ordre $k+1$ est inversible sur des maillages utilisés en pratique.

8.4. Élargissement des voisinages de reconstruction par des méthodes itératives

Les algorithmes présentés dans les sections 8.2 et 8.3 travaillent par itérations dans le but d'augmenter l'ordre de la reconstruction polynomiale. On peut noter que l'approche des itérations peut simplement servir à élargir le voisinage de la reconstruction, sans augmentation de l'ordre de la reconstruction. L'étude du chapitre 10 montrera en effet que l'élargissement des voisinages de reconstruction fournit des schémas numériques plus robustes, ce qui rend cette technique intéressante. L'étude de cette section nécessite de façon explicite les définitions des voisinages de la section 5.4, et en particulier la définition (5.4.2) du premier voisinage augmenté $\widehat{\mathbb{V}}_\alpha$ et la définition (5.4.3) du deuxième voisinage augmenté $\widehat{\mathbb{V}}_\alpha^{(2)}$.

Considérons d'abord le cas le plus simple, celui d'une reconstruction des polynômes de degré un, c'est-à-dire d'une reconstruction linéaire par morceaux. Le point de départ de la méthode est un gradient consistant (8.3.1), de précision à l'ordre un au sens de la définition 7.4.1,

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}(\overline{u}_\beta - \overline{u}_\alpha)\,.$$

On suppose que le gradient $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$ se calcule sur le premier voisinage augmenté $\widehat{\mathbb{V}}_\alpha$ de la cellule, ce qui permet une implémentation rapide et efficace. Par convention, $\boldsymbol{\sigma}_{\alpha\beta} \triangleq 0$ si $\beta \notin \widehat{\mathbb{V}}_\alpha$, c'est-à-dire si la cellule $\mathcal{T}_\beta$ n'est pas dans le voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$.

L'opérateur (8.3.1) permet de reproduire les polynômes de degré un grâce aux identités (8.3.2) et (8.3.3)

$$\begin{aligned}\sum_\beta \boldsymbol{\sigma}_{\alpha\beta} &= 0\\ \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\otimes \boldsymbol{h}_{\alpha\beta} &= \boldsymbol{\delta}^{(2)}\,.\end{aligned}$$

On se donne comme objectif de calculer un gradient sur un voisinage plus grand que le premier voisinage $\widehat{\mathbb{V}}_\alpha$, par un algorithme qui ne fait intervenir que $\widehat{\mathbb{V}}_\alpha$. Une idée simple consiste à utiliser une combinaison linéaire des gradients au premier voisinage de la cellule $\mathcal{T}_\alpha$

$$\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}] \triangleq \sum_\beta \xi_{\alpha\beta}\boldsymbol{\sigma}_\beta[\mathfrak{u}] = \sum_\beta \xi_{\alpha\beta}\left(\sum_\gamma \boldsymbol{\sigma}_{\beta\gamma}\overline{u}_\gamma\right) \tag{8.4.1}$$

où les $\xi_{\alpha\beta}$ sont des coefficients réels. Par convention, on définit $\xi_{\alpha\beta} \triangleq 0$ si $\beta \notin \widehat{\mathbb{V}}_\alpha$, afin de simplifier l'écriture de (8.4.1).

Il résulte directement de la formule (8.4.1) que le gradient $\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}]$ est consistant si

$$\sum_\beta \xi_{\alpha\beta} = 1\,. \tag{8.4.2}$$

Si tous les $\xi_{\alpha\beta}$ sont positifs, le gradient $\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}]$ est une combinaison convexe des gradients $\boldsymbol{\sigma}_\beta[\mathfrak{u}]$, $\beta \in \widehat{\mathbb{V}}_\alpha$. Dans ce cas, il est possible d'interpréter le gradient (8.4.1) comme une moyenne pondérée des gradients sur le premier voisinage de la cellule $\mathcal{T}_\alpha$. Un choix naturel pour les coefficients $\xi_{\alpha\beta}$ est une pondération par les volumes des cellules. Cela donne l'algorithme suivant.

Algorithme 8.4.1 (Reconstruction du gradient sur un voisinage élargi). *Supposons donné un gradient consistant $\boldsymbol{\sigma}_\beta[\mathfrak{u}]$ calculé sur le premier voisinage. Il est possible d'implémenter une reconstruction du gradient sur le deuxième voisinage qui fait intervenir uniquement des sommes sur les premiers voisinages $\widehat{\mathbb{V}}_\alpha$.*

(1) *Dans une première étape, l'algorithme calcule un gradient consistant sur le premier voisinage de chaque cellule $\mathcal{T}_\alpha$ par la formule*

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta\,.$$

Ce gradient est stocké dans un tableau de travail indexé par le numéro de cellule.

(2) *Dans une deuxième étape, l'algorithme calcule le gradient (8.4.1) par la formule*

$$\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}] \triangleq \sum_{\beta\in\widehat{\mathbb{V}}_\alpha} \xi_{\alpha\beta}\boldsymbol{\sigma}_\beta[\mathfrak{u}]$$

où la somme est effectuée sur le premier voisinage. Les coefficients $\xi_{\alpha\beta}$ sont données par

$$\xi_{\alpha\beta} = \frac{|\mathcal{T}_\beta|}{\sum_{\gamma\in\widehat{\mathbb{V}}_\alpha}|\mathcal{T}_\gamma|}\,.$$

Il est possible d'utiliser des formules différentes pour les coefficients $\xi_{\alpha\beta}$, ce qui donne des variantes de l'algorithme 8.4.1.

L'algorithme 8.4.1 s'étend facilement à la reconstruction des polynômes de degré deux. Supposons donnés une dérivée seconde consistante

$$\boldsymbol{\theta}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}\overline{u}_\beta = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}(\overline{u}_\beta - \overline{u}_\alpha)$$

et un gradient consistant de précision à l'ordre deux

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}(\overline{u}_\beta - \overline{u}_\alpha)\,.$$

Il est alors possible de calculer une dérivée seconde moyennée par la formule

$$\breve{\boldsymbol{\theta}}_\alpha[\mathfrak{u}] \triangleq \sum_\beta \xi_{\alpha\beta}\boldsymbol{\theta}_\beta[\mathfrak{u}] = \sum_\beta \xi_{\alpha\beta}\left(\sum_\gamma \boldsymbol{\theta}_{\beta\gamma}\overline{u}_\gamma\right)\,. \tag{8.4.3}$$

Pour moyenner correctement le gradient de précision à l'ordre deux, il faut tenir compte de la dérivée seconde. Le gradient moyenné $\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}]$ doit donc être calculé par la formule

$$\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}] \triangleq \sum_\beta \eta_{\alpha\beta}\left(\boldsymbol{\sigma}_\beta[\mathfrak{u}] - \breve{\boldsymbol{\theta}}_\alpha[\mathfrak{u}]\cdot\boldsymbol{h}_{\alpha\beta}\right) \tag{8.4.4}$$

pour qu'il soit de précision à l'ordre deux. Dans (8.4.4), les $\eta_{\alpha\beta}$ sont des coefficients réels qui satisfont également (8.4.2).

Soit p un polynôme de degré deux dont les moyennes de cellule sont $\mathfrak{p} = (\overline{p}_1,\dots,\overline{p}_N)$. Il vient, par la consistance de $\boldsymbol{\theta}_\beta$,

$$\boldsymbol{\theta}_\alpha[\mathfrak{p}] = D^{(2)}p\Big|_{\boldsymbol{x}_\alpha}$$

et, par la consistance de $\boldsymbol{\sigma}_\beta$,

$$\boldsymbol{\sigma}_\beta[\mathfrak{p}] = D^{(1)}p\Big|_{\boldsymbol{x}_\beta} = D^{(1)}p\Big|_{\boldsymbol{x}_\alpha} + D^{(2)}p\Big|_{\boldsymbol{x}_\alpha}\cdot\boldsymbol{h}_{\alpha\beta}\,.$$

Les formules (8.4.3), (8.4.4) ainsi que la condition (8.4.2) pour les coefficients $\xi_{\alpha\beta}$ et $\eta_{\alpha\beta}$ entraînent alors

$$\breve{\boldsymbol{\theta}}_\alpha[\mathfrak{p}] = D^{(2)}p\Big|_{\boldsymbol{x}_\alpha}$$
$$\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{p}] = D^{(1)}p\Big|_{\boldsymbol{x}_\alpha}\,,$$

ce qui prouve que $\breve{\boldsymbol{\theta}}_\alpha$ et $\breve{\boldsymbol{\sigma}}_\alpha$ reconstruisent correctement la dérivée seconde et le gradient de tout polynôme de degré deux.

L'implémentation de cette méthode donne l'algorithme suivant.

ALGORITHME 8.4.2 (Reconstruction de la dérivée seconde et du gradient sur un voisinage élargi). *Supposons donnés une dérivée seconde consistante $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$ et un gradient consistant $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$ de précision à l'ordre deux, calculés sur un voisinage arbitraire $\widehat{\mathbb{W}}_\alpha$. Il est possible d'implémenter une reconstruction du gradient et de la dérivée seconde sur le voisinage $\bigcup_{\beta\in\widehat{\mathbb{V}}_\alpha}\widehat{\mathbb{W}}_\beta$ qui fait intervenir uniquement des sommes sur les premiers voisinages $\widehat{\mathbb{V}}_\alpha$ et les voisinages $\widehat{\mathbb{W}}_\alpha$.*

(1) *Dans une première étape, l'algorithme calcule la dérivée seconde consistante $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$ et le gradient consistant $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$ sur le voisinage $\widehat{\mathbb{W}}_\alpha$ de chaque cellule $\mathcal{T}_\alpha$ par les formules*

$$\boldsymbol{\theta}_\alpha[\mathfrak{u}] = \sum_{\beta\in\widehat{\mathbb{W}}_\alpha} \boldsymbol{\theta}_{\alpha\beta}\overline{u}_\beta$$
$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_{\beta\in\widehat{\mathbb{W}}_\alpha} \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta\,.$$

Le gradient et la dérivée seconde sont stockés dans un tableau de travail indexé par le numéro de cellule.

(2) *Dans une deuxième étape, l'algorithme calcule la dérivée seconde (8.4.3) et le gradient (8.4.4) par les formules*

$$\breve{\boldsymbol{\theta}}_\alpha[\mathfrak{u}] \triangleq \sum_{\beta\in\widehat{\mathbb{V}}_\alpha} \xi_{\alpha\beta}\boldsymbol{\theta}_\beta[\mathfrak{u}]$$
$$\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}] \triangleq \sum_{\beta\in\widehat{\mathbb{V}}_\alpha} \eta_{\alpha\beta}\left(\boldsymbol{\sigma}_\beta[\mathfrak{u}] - \breve{\boldsymbol{\theta}}_\alpha[\mathfrak{u}]\cdot\boldsymbol{h}_{\alpha\beta}\right)$$

où la somme est effectuée sur le premier voisinage $\widehat{\mathbb{V}}_\alpha$. Les coefficients $\xi_{\alpha\beta}$ et $\eta_{\alpha\beta}$ sont données par

$$\eta_{\alpha\beta} = \xi_{\alpha\beta} = \frac{|\mathcal{T}_\beta|}{\sum_{\gamma\in\widehat{\mathbb{V}}_\alpha}|\mathcal{T}_\gamma|}\,.$$

Il est possible d'utiliser des formules différentes pour les coefficients $\xi_{\alpha\beta}$ et $\eta_{\alpha\beta}$, ce qui donne des variantes de l'algorithme 8.4.2.

Les algorithmes 8.4.1 et 8.4.2 permettent de définir des variantes de la méthode des moindres carrés couplés et de la méthode des corrections successives.

Dans le cas de la méthode des moindres carrés couplés, l'algorithme 8.4.2 peut être utilisé pour moyenner les dérivées reconstruites par l'algorithme 8.2.2. Cela donne l'algorithme suivant, appelé méthode MCCIE.

ALGORITHME 8.4.3 (Reconstruction des polynômes de degré deux par la méthode itérative des moindres carrés couplés sur un voisinage élargi ou méthode MCCIE). *Il est possible de reconstruire les polynômes de degré deux sur un voisinage élargi de la façon suivante :*

(1) *Dans une première étape, l'algorithme 8.2.2 calcule dans chaque cellule une dérivée seconde consistante $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$ et un gradient consistant $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$ de précision à l'ordre deux.*

(2) *Dans une deuxième étape, l'algorithme 8.4.2 calcule une dérivée seconde consistante moyennée $\breve{\boldsymbol{\theta}}_\alpha[\mathfrak{u}]$ et un gradient consistant moyenné $\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}]$ de précision à l'ordre deux.*

Il est également possible d'élargir le voisinage effectif de l'algorithme 8.3.8 de la méthode des corrections successives. Pour cela, on peut utiliser l'opérateur de gradient $\breve{\boldsymbol{\sigma}}_\alpha[\mathfrak{u}]$, défini par (8.4.1) et calculé par l'algorithme 8.4.1, à la place de l'opérateur de gradient $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$ utilisé par la méthode des corrections successives.

Pour expliciter l'algorithme, il est utile d'introduire quelques notations. On définit les vecteurs

$$\breve{\boldsymbol{\sigma}}_{\alpha\beta} \triangleq \begin{cases} \sum_{\gamma\in\widehat{\mathbb{V}}_\alpha}\xi_{\alpha\gamma}\boldsymbol{\sigma}_{\gamma\beta} & \text{si } \beta\in\widehat{\mathbb{V}}_\alpha^{(2)} \\ 0 & \text{sinon} \end{cases},$$

ce qui permet d'exprimer l'opérateur de gradient (8.4.1) sous la forme simple

$$\breve{\boldsymbol{\sigma}}_\alpha [\mathfrak{u}] \triangleq \sum \breve{\boldsymbol{\sigma}}_{\alpha\beta} \overline{u}_\beta = \sum_{\beta \in \widehat{\mathbb{V}}_\gamma} \sum_{\gamma \in \widehat{\mathbb{V}}_\alpha} \xi_{\alpha\gamma} \boldsymbol{\sigma}_{\gamma\beta} \overline{u}_\beta = \sum_{\gamma \in \widehat{\mathbb{V}}_\alpha} \xi_{\alpha\gamma} \boldsymbol{\sigma}_\gamma [\mathfrak{u}] \, .$$

Le gradient $\breve{\boldsymbol{\sigma}}_\alpha [\mathfrak{u}]$ permet de calculer le tenseur $\breve{\boldsymbol{\tau}}_\alpha^{(4)}$, l'équivalent du tenseur (8.3.17), par la formule

$$\breve{\boldsymbol{\tau}}_\alpha^{(4)} \triangleq \frac{1}{2} \sum_{\beta,\gamma} (\breve{\boldsymbol{\sigma}}_{\alpha\beta} \odot \breve{\boldsymbol{\sigma}}_{\beta\gamma}) \otimes \boldsymbol{z}_{\alpha\gamma}^{(2)} \, . \tag{8.4.5}$$

Le tenseur (8.4.5) permet de définir l'application de Brenner

$$\breve{\mathfrak{T}}_\alpha^{(2)} : \left\{ \begin{array}{ccc} \mathbb{S}_2^d & \longrightarrow & \mathbb{S}_2^d \\ \boldsymbol{a}^{(2)} & \longmapsto & \breve{\boldsymbol{\tau}}_\alpha^{(4)} \cdot \boldsymbol{a}^{(2)} \end{array} \right. \tag{8.4.6}$$

dont l'application inverse, si elle existe, sera notée

$$\breve{\mathfrak{C}}_\alpha^{(2)} : \left\{ \begin{array}{ccc} \mathbb{S}_2^d & \longrightarrow & \mathbb{S}_2^d \\ \boldsymbol{a}^{(2)} & \longmapsto & \breve{\mathfrak{C}}_\alpha^{(2)} \left(\boldsymbol{a}^{(2)} \right) \end{array} \right. \, . \tag{8.4.7}$$

Si l'application inverse (8.4.7) existe, il est possible d'implémenter l'algorithme suivant, appelé méthode CSE.

ALGORITHME 8.4.4 (Reconstruction quadratique par la méthode des corrections successives sur un voisinage élargi ou méthode CSE). *L'algorithme travaille de la même manière que l'algorithme 8.3.8 :*

(1) *Dans une première étape, l'algorithme 8.4.1 calcule le gradient (8.4.1) par la formule*

$$\breve{\boldsymbol{\sigma}}_\alpha [\mathfrak{u}] \triangleq \sum_\beta \breve{\boldsymbol{\sigma}}_{\alpha\beta} \overline{u}_\beta \, .$$

Les gradients $\breve{\boldsymbol{\sigma}}_\alpha [\mathfrak{u}]$ sont stockés dans un tableau indexé par le numéro de cellule.

(2) *Dans une deuxième étape, l'algorithme calcule l'opérateur approché $\widehat{\breve{\boldsymbol{\theta}}}_\alpha [\mathfrak{u}]$ par la formule*

$$\widehat{\breve{\boldsymbol{\theta}}}_\alpha [\mathfrak{u}] = \sum_\beta \breve{\boldsymbol{\sigma}}_{\alpha\beta} \odot \breve{\boldsymbol{\sigma}}_\beta [\mathfrak{u}] = \sum_\beta \breve{\boldsymbol{\sigma}}_{\alpha\beta} \odot \left(\sum_\gamma \breve{\boldsymbol{\sigma}}_{\beta\gamma} \overline{u}_\gamma \right) \, .$$

(3) *Dans une troisième étape, l'application $\breve{\mathfrak{C}}_\alpha^{(2)}$ (8.4.7) permet de déterminer à partir de $\widehat{\breve{\boldsymbol{\theta}}}_\alpha [\mathfrak{u}]$ une dérivée seconde consistante au sens de la définition 7.4.1 par la formule*

$$\breve{\boldsymbol{\theta}}_\alpha [\mathfrak{u}] = \breve{\mathfrak{C}}_\alpha^{(2)} \cdot \widehat{\breve{\boldsymbol{\theta}}}_\alpha [\mathfrak{u}] \, .$$

(4) *Finalement, l'application $\breve{\mathfrak{C}}_\alpha^{(2)}$ permet de calculer un gradient consistant $\breve{\breve{\boldsymbol{\sigma}}}_\alpha [\mathfrak{u}]$ de précision à l'ordre deux par la formule (8.3.31). Il suffit pour cela de remplacer les vecteurs $\boldsymbol{\sigma}_{\alpha\beta}$ par les vecteurs $\breve{\boldsymbol{\sigma}}_{\alpha\beta}$, ce qui donne*

$$\breve{\breve{\boldsymbol{\sigma}}}_\alpha [\mathfrak{u}] \triangleq \sum_\beta \breve{\boldsymbol{\sigma}}_{\alpha\beta} \overline{u}_\beta - \frac{1}{2!} \sum_\beta \breve{\boldsymbol{\sigma}}_{\alpha\beta} \left(\left(\breve{\mathfrak{C}}_\alpha^{(2)} \cdot \widehat{\breve{\boldsymbol{\theta}}}_\alpha [\mathfrak{u}] \right) \bullet \boldsymbol{z}_{\alpha\beta}^{(2)} \right)$$

L'application $\breve{\mathfrak{C}}_\alpha^{(2)}$ peut être réalisée par une matrice qui est déterminée avant le début du calcul.

Les algorithmes 8.4.1, 8.4.2, 8.4.3 et 8.4.4 ont été implémentés et testés pour l'équation de la convection linéaire sur des maillages non structurés, cf. section 8.5. L'algorithme 8.4.1 a également été implémenté dans CEDRE et utilisé pour les calculs décrits dans les chapitres 13 et 14.

8.5. Étude numérique

Le chapitre présente deux méthodes itératives de reconstruction de degré élevé dont l'étude théorique n'a pas permis d'éclaircir tous les aspects. Il faut donc recourir à des expériences numériques pour vérifier la faisabilité de ces méthodes. Les points suivants restent ouverts :

(1) Pour la méthode des moindres carrés couplés, introduite dans la section 8.2, il reste à prouver la précision des algorithmes itératifs 8.2.2 (MCCI) et 8.4.3 (MCCIE) en maillage non structuré. En maillage cartésien, l'algorithme calcule la bonne solution à la première itération.

(2) Pour la méthode des corrections successives CS, introduite dans la section 8.3, il reste à démontrer l'inversibilité des applications de Brenner, introduites par les définitions 8.3.3 et 8.3.12. En maillage cartésien, les applications de Brenner sont inversibles car elles sont égales à l'application identité.

Cela signifie que des expériences numériques sont indispensables pour s'assurer du bon fonctionnement des deux méthodes en maillage non structuré. Les tests numériques de cette section sont restreints à la reconstruction des polynômes de degré deux. Pour des raisons de simplicité, ces tests s'effectuent en même temps que les tests de stabilité et précision des chapitres 10 et 11.

Les reconstructions suivantes ont été testées :

(1) La reconstruction par la méthode MCCI, définie par l'algorithme 8.2.2.

(2) La reconstruction par la méthode MCCIE, définie par l'algorithme 8.4.3.

(3) La reconstruction par la méthode CS, définie par l'algorithme 8.3.8.

(4) La reconstruction par la méthode CSE, définie par l'algorithme 8.4.4.

Ces méthodes ont été testées sur les maillages suivants :

(1) Des maillages de tétraèdres en dimension trois et des maillages de triangles en dimension deux.

(2) Des maillages hybrides constitués de tétraèdres et de prismes en dimension trois.

(3) Des maillages hybrides constitués de triangles et de quadrilatères en dimension deux.

(4) Des maillages cartésiens et des maillages cartésiens déformés en dimensions deux et trois.

Pour les méthodes CS et CSE, il faut résoudre le système linéaire

$$\widehat{\boldsymbol{\theta}}_\alpha[\mathfrak{u}] = \boldsymbol{\tau}_\alpha^{(4)} \cdot \boldsymbol{\theta}_\alpha[\mathfrak{u}] \tag{8.5.1}$$

pour l'inconnue $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$. Dans (8.5.1), $\widehat{\boldsymbol{\theta}}_\alpha[\mathfrak{u}]$ est l'opérateur (8.3.9), $\boldsymbol{\tau}_\alpha^{(4)}$ est le tenseur (8.3.17) qui exprime l'application de Brenner et $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$ est la dérivée seconde consistante qu'on souhaite calculer. Il est important d'étudier le nombre de conditionnement des matrices associées au système (8.5.1). Notons $T_\alpha^{(2)}$ la matrice de l'application de Brenner $\mathfrak{T}_\alpha^{(2)}$ dans la base canonique de $\mathbb{S}_2^d$, l'espace des tenseurs symétriques d'ordre deux dans $\mathbb{R}^d$. Le nombre de conditionnement de $T_\alpha^{(2)}$, défini par

$$\kappa\left(T_\alpha^{(2)}\right) \triangleq \left\|T_\alpha^{(2)}\right\|_2 \left\|\left(T_\alpha^{(2)}\right)^{-1}\right\|_2, \tag{8.5.2}$$

mesure la dépendance de la solution $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$ du problème numérique (8.5.1) par rapport à la donnée $\widehat{\boldsymbol{\theta}}_\alpha[\mathfrak{u}]$. Si (8.5.2) est grand, de petites fluctuations de $\widehat{\boldsymbol{\theta}}_\alpha[\mathfrak{u}]$ peuvent entraîner de grandes fluctuations de $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$, ce qui n'est pas souhaitable.

Les tableaux 8.5.1 et 8.5.2 montrent les statistiques du nombre de conditionnement (8.5.2) pour les méthodes CS et CSE sur quatre maillages de tétraèdres. Les données indiquent clairement que le nombre de conditionnement des matrices $T_\alpha^{(2)}$ peut devenir très grand en maillage de tétraèdres dans le cas de la méthode CS, alors que la méthode CSE fournit des nombres de conditionnement beaucoup plus petits.

Pour vérifier l'ordre de la reconstruction, on effectue une étude des taux de convergence de l'erreur numérique comme dans la section 7.10. Cette étude est restreinte à la dimension deux

N° de maillage	Minimum	Moyenne	Écart type	Maximum
1	3.285911833	21.09174648	142.1172384	3821.921630
2	3.431990220	15.12346084	22.07730547	523.5432088
3	2.836825741	18.87527358	82.85973253	2099.992411
4	2.867492427	20.35547288	166.5128024	7960.788417

TAB. 8.5.1: Nombre de conditionnement (8.5.2) de l'application de Brenner pour la méthode CS (algorithme 8.3.8) en maillage de tétraèdres.

N° de maillage	Minimum	Moyenne	Écart type	Maximum
1	1.188013434	1.673177682	0.2991381013	3.570445351
2	1.216102652	1.712074288	0.2845642871	3.400043570
3	1.194822559	1.600451133	0.2567874414	4.761886966
4	1.167254242	1.688491073	0.2710019992	4.158304507

TAB. 8.5.2: Nombre de conditionnement (8.5.2) de l'application de Brenner pour la méthode CSE (algorithme 8.4.4) en maillage de tétraèdres.

car les tests en maillage de tétraèdres demandent un nombre élevé de cellules pour obtenir une précision suffisante. L'outil de test n'était pas suffisamment performant pour faire des tests en dimension trois dans des conditions satisfaisantes.

On se place donc sur un carré

$$\Omega = \left\{ (x, y) \in \mathbb{R}^2 \middle| \, 0 \leq x \leq 1 \, , \, 0 \leq y \leq 1 \right\}$$

avec des conditions de périodicité au bord. La fonction test est la fonction périodique (7.10.1)

$$u_0(x, y) = \sin(2\pi x) \sin(2\pi y)$$

dont les moyennes de cellule sont notées $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$. A partir de ces moyennes, on reconstruit dans chaque cellule un polynôme $w_\alpha[\mathfrak{u}](\boldsymbol{x})$ et on calcule l'erreur d'approximation (7.10.2)

$$\varepsilon = \frac{\sqrt{\sum_{\alpha=1}^{N} \int_{\mathcal{T}_\alpha} |w_\alpha[\mathfrak{u}](\boldsymbol{x}) - u_0(\boldsymbol{x})|^2 \, dx}}{\sqrt{\sum_{\alpha=1}^{N} \int_{\mathcal{T}_\alpha} |u_0(\boldsymbol{x})|^2 \, dx}}$$

par des formules d'intégration numérique d'une précision suffisante.

La figure 7.10.1 sur la page 98 montre deux exemples de maillages non structurés. Ces maillages ont été générés avec le logiciel GMSH. Pour le maillage le plus grossier, soit ε_0 l'erreur (7.10.2) et h_0 le diamètre du maillage. Lorsqu'on dessine $\log(\varepsilon/\varepsilon_0)$ en fonction de $\log(h/h_0)$ pour les différents maillages testés, on obtient une courbe dont la pente est une mesure pour le taux de convergence de l'erreur d'approximation en fonction du diamètre des mailles.

La figure 8.5.1 montre les courbes pour des maillages triangulaires et des maillages hybrides composés de triangles et de quadrangles. La courbe appelée *Degré 2 Vois. 2* correspond à la reconstruction par la méthode des moindres carrés, introduite par la définition 7.7.1, sur le deuxième voisinage. Cette courbe permet de comparer les taux de convergence des méthodes itératives avec celui de la méthode des moindres carrés. On constate que les pentes sont effectivement de -3 pour toutes les méthodes testées, ce qui montre que les reconstructions itératives satisfont une estimation du type (6.3.3). Cependant, lorsque h diminue, la méthode MCCI montre un taux de convergence légèrement plus faible que les autres méthodes en maillage triangulaire. L'étude du chapitre 10 montre que la méthode MCCI s'avère instable sur des maillages de triangles et de tétraèdres et doit donc être remplacée par la méthode MCCIE. Pour cette raison, la question du taux de convergence de la méthode MCCI n'a pas été analysée de façon plus approfondie.

Les tests permettent de tirer les conclusions suivantes :

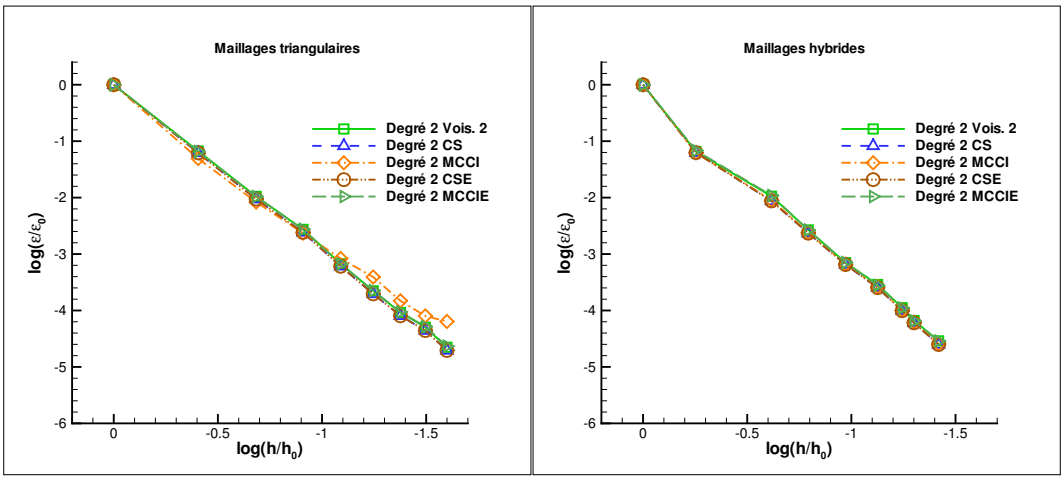


FIG. 8.5.1: Taux de convergence des méthodes de reconstruction itératives en dimension deux. Les méthodes MCCI et MCCIE utilisent ici 3 itérations.

(1) Les algorithmes 8.2.2 (MCCI) et 8.4.3 (MCCIE) de la méthode des moindres carrés couplés permettent de reconstruire les polynômes de degré deux en dimension deux et dimension trois. Comme ces méthodes dépendent de la convergence des itérations, l'erreur sur les dérivées reconstruites est légèrement plus grande que pour la méthode des corrections successives.

(2) Les tests de la méthode des corrections successives montrent que l'application de Brenner est inversible sur tous les maillages testés. En dimension trois, par contre, l'algorithme 8.3.8 (CS) produit des matrices de correction très mal conditionnées, cf. le tableau 8.5.1. Il faut par conséquent remplacer l'algorithme 8.3.8 (CS) par l'algorithme 8.4.4 (CSE) qui n'a pas ce problème.

Il est donc possible d'utiliser les méthodes de reconstruction présentées dans ce chapitre. Cependant, il faut noter que l'étude de stabilité du chapitre 10 montre clairement que les algorithmes 8.2.2 et 8.3.8 conduisent à des schémas instables, en particulier en maillage de tétraèdres mais aussi en maillage de triangles. Il est donc nécessaire de les remplacer par les algorithmes 8.4.3 et 8.4.4.

8.6. Bilan du chapitre

L'étude a permis de concevoir des méthodes de reconstruction particulièrement adaptées à une implémentation rapide et facile dans des logiciels de grand calcul comme CEDRE. L'avantage principal de ces méthodes est d'éviter des voisinages de reconstruction trop larges et d'échanger des données uniquement entre cellules adjacentes. Cela permet de réduire les données de connectivité entre les cellules à un minimum et facilite l'échange des données entre les différentes partitions du maillage dans le cas du calcul parallèle.

L'étude de ce chapitre a d'abord permis d'analyser deux méthodes de reconstruction de polynômes de degré élevé :

(1) La section 8.2 a présenté la méthode des moindres carrés couplés. L'algorithme 8.2.2 (MCCI) permet, à partir d'un gradient consistant, de calculer une dérivée seconde consistante et un gradient consistant de précision à l'ordre deux. Cela permet de reconstruire les polynômes de degré deux. Il manque cependant encore une preuve mathématique rigoureuse de la précision de l'algorithme 8.2.2. L'extension de l'algorithme 8.2.2 aux polynômes de degré $k > 2$ paraît possible mais ce travail supplémentaire est trop important pour rentrer dans le cadre de cette thèse.

(2) La section 8.3 a introduit la méthode des corrections successives. L'algorithme 8.3.8 (CS) permet, à partir d'un gradient consistant, de calculer une dérivée seconde consistante et un gradient de précision à l'ordre deux, ce qui permet de reconstruire les polynômes de degré deux. Il manque encore une preuve mathématique rigoureuse de l'inversibilité de l'application de Brenner (8.3.20). Finalement, l'algorithme 8.3.17 permet en principe de reconstruire les polynômes de degré $k > 2$, mais cette méthode n'a pas été testée dans le cadre de cette thèse car le travail aurait été trop important.

L'étude a également servi à concevoir des méthodes qui permettent d'élargir les voisinages de reconstruction sans engendrer des coûts de calcul informatique exorbitants. La section 8.4 présente les algorithmes suivants :

(1) Les algorithmes 8.4.1 et 8.4.2 permettent tout simplement d'élargir le voisinage de reconstruction sans augmenter l'ordre de la reconstruction.

(2) L'algorithme 8.4.3, appelé méthode MCCIE, est une variante de l'algorithme 8.2.2, basée sur un voisinage élargi.

(3) L'algorithme 8.4.4, appelé méthode CSE, est une variante de l'algorithme 8.3.8, basée sur un voisinage élargi.

Les tests numériques des chapitres 10 et 11 ont permis de vérifier que les algorithmes 8.2.2, 8.4.3, 8.3.8 et 8.4.4 reconstruisent correctement les polynômes de degré deux sur tous les maillages testés en dimension deux et trois, voir les sections 10.9.3 et 11.3. L'étude de stabilité du chapitre 10 montre toutefois que les algorithmes 8.2.2 et 8.3.8 ne donnent pas de schémas stables, en particulier sur des maillages de tétraèdres. Il est donc nécessaire de les remplacer par les algorithmes 8.4.3 et 8.4.4.

Il reste maintenant à vérifier en détail la stabilité et la précision des schémas volumes finis qui utilisent ces méthodes de reconstruction. Ces questions font l'objet des chapitres 10 et 11.

CHAPITRE 9

Étude des intégrales de surface des flux

9.1. Objectif du chapitre

Le chapitre 6 sur la discrétisation spatiale montre qu'il est nécessaire d'intégrer les fonctions reconstruites avec une précision suffisante sur les faces entre les cellules pour atteindre l'erreur de troncature désirée. L'objectif du présent chapitre est de développer des méthodes d'intégration sur les faces qui peuvent être réalisées de façon efficace dans des codes de calcul en maillage non structuré général.

Dans le cas d'une reconstruction locale utilisant un espace de polynômes, on peut envisager deux façons d'intégrer les flux numériques sur les faces.

(1) Une première approche, présentée dans la section 9.2, consiste à déterminer des fonctions de base pour l'espace vectoriel des polynômes en question. En intégrant ces fonctions de base sur chaque face, on obtient des intégrales élémentaires. Dans cette approche, l'intégrale d'un polynôme est simplement une combinaison linéaire des intégrales élémentaires. Cette méthode nécessite le calcul des intégrales élémentaires au début du calcul et leur stockage en mémoire. L'avantage est que les intégrales des flux peuvent être calculées assez rapidement par cette méthode. Un inconvénient de cette méthode constitue sa mise en pratique pour des flux numériques non linéaires. Pour intégrer des flux numériques non linéaires il faut calculer un développement de Taylor du flux et ensuite intégrer le polynôme de Taylor. Il est donc nécessaire d'adapter la méthode à chaque type de flux. Un autre désavantage de cette méthode est l'obligation de recalculer toutes les intégrales élémentaires si la géométrie du maillage change au cours du calcul.

(2) Une autre approche, présentée dans la section 9.3, consiste à utiliser des formules d'intégration numérique. On définit sur chaque face des points de Gauss. L'intégrale d'une fonction sur la face est ensuite approchée par une somme pondérée des évaluations de la fonction aux points de Gauss. Les poids sont choisis de façon à ce que la formule soit exacte pour des polynômes d'un certain degré. La méthode nécessite d'accéder aux sommets des faces afin de calculer les points de Gauss car chaque point de Gauss est une combinaison convexe des sommets. Ce procédé s'intègre facilement avec l'utilisation des flux décentrés car il suffit d'évaluer la fonction de flux numérique aux points de Gauss pour calculer le flux total. Cette méthode est donc totalement transparente par rapport au type de flux numérique. La méthode des points de Gauss est également flexible par rapport à l'ordre de la quadrature. Pour modifier l'ordre de la quadrature, il suffit d'utiliser de nouvelles combinaisons convexes des sommets et de nouveaux poids, ce qui se programme très facilement dans un code de calcul. Le désavantage de cette méthode est le coût de calcul car elle nécessite l'évaluation des flux complets à plusieurs points de la face. Par exemple, pour une intégration exacte de polynômes d'ordre trois, il faut au minimum quatre points de Gauss sur un triangle. Il faut donc évaluer quatre fois les flux de convection, ce qui augmente le temps de calcul de cette partie d'un facteur quatre. Par contre, la méthode d'intégration numérique présente un avantage pour des calculs sur des maillages mobiles. Lorsque la face se déforme au cours du calcul, il suffit de connaître les nouvelles positions des sommets pour déterminer immédiatement les nouveaux points de Gauss par combinaison convexe des sommets. Ceci constitue un avantage par rapport à la méthode des fonctions de base qui nécessite d'intégrer à chaque cycle les fonctions élémentaires sur la face déformée.

9.2. Intégration par des fonctions de base

Dans cette section, on se place dans le contexte de la reconstruction de degré un avec un gradient consistant

$$\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta$$

qui est reconstruit à partir des moyennes de cellule $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$. Dans la suite, on écrit $\boldsymbol{\sigma}_\alpha$ au lieu de $\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right]$ afin de simplifier les formules. Soit

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \overline{u}_\alpha + \boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha)$$

le polynôme de degré un reconstruit dans la cellule $\mathcal{T}_\alpha$ et soit $\boldsymbol{f} : \mathbb{R} \longrightarrow \mathbb{R}^d$ une fonction de flux. Pour simplifier la démonstration, on suppose pour le moment que $\boldsymbol{f}$ ne dépend que d'une seule variable u. L'objectif est d'approcher l'intégrale de surface

$$\int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x}) \cdot \boldsymbol{f}\left(w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x})\right)\, d\sigma = \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x}) \cdot \boldsymbol{f}\left(\overline{u}_\alpha + \boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha)\right)\, d\sigma \tag{9.2.1}$$

sur une face $\mathcal{A}_{\alpha\beta}$ de la cellule $\mathcal{T}_\alpha$ où $\boldsymbol{\nu}(\boldsymbol{x})$ est le vecteur normal unitaire de la face en $\boldsymbol{x} \in \mathcal{A}_{\alpha\beta}$.

Avec l'identité $\boldsymbol{x}_{\alpha\beta} = \boldsymbol{x}_\alpha + \boldsymbol{k}_{\alpha\beta}$ pour le barycentre de la face et la définition

$$u_{\alpha\beta} \triangleq \overline{u}_\alpha + \boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta}$$

pour la valeur reconstruite en $\boldsymbol{x}_{\alpha\beta}$, $w_\alpha\left[\mathfrak{u}\right]$ peut s'écrire sous la forme

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \overline{u}_\alpha + \boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} + \boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta}) = u_{\alpha\beta} + \boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\ .$$

Un développement de la fonction $\boldsymbol{f}$ autour de $u_{\alpha\beta}$ donne

$$\begin{aligned}
&\boldsymbol{f}\left(u_{\alpha\beta} + \boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\right) = \\
&= \boldsymbol{f}\left(u_{\alpha\beta}\right) + \left.\frac{d\boldsymbol{f}}{du}\right|_{u_{\alpha\beta}} \left[\boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\right] + \frac{1}{2!}\left.\frac{d^2\boldsymbol{f}}{du^2}\right|_{u_{\alpha\beta}} \left[\boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\right]^2 + \mathrm{O}\left(\left[\boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\right]^3\right)\ .
\end{aligned} \tag{9.2.2}$$

Dans le cas des équations de Navier-Stokes, la fonction de flux dépend de plusieurs variables et il faut alors remplacer le développement de Taylor (9.2.2) par un développement en plusieurs variables.

Si l'on néglige les termes d'ordre cubique dans (9.2.2), on peut approcher l'intégrale (9.2.1) par

$$\begin{aligned}
&\int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x}) \cdot \boldsymbol{f}\left(u_{\alpha\beta} + \boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\right)\, d\sigma \approx \\
&\qquad \approx \int_{\mathcal{A}_{\alpha\beta}} \left\{ \boldsymbol{\nu}(\boldsymbol{x}) \cdot \boldsymbol{f}\left(u_{\alpha\beta}\right) + \boldsymbol{\nu}(\boldsymbol{x}) \cdot \left.\frac{d\boldsymbol{f}}{du}\right|_{u_{\alpha\beta}} \left[\boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\right] + \right. \\
&\qquad\qquad \left. + \frac{1}{2}\boldsymbol{\nu}(\boldsymbol{x}) \cdot \left.\frac{d^2\boldsymbol{f}}{du^2}\right|_{u_{\alpha\beta}} \left[\boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\right]^2 \right\} d\sigma\ .
\end{aligned} \tag{9.2.3}$$

Supposons la face $\mathcal{A}_{\alpha\beta}$ plane, ce qui signifie que le vecteur normal unitaire $\boldsymbol{\nu}(\boldsymbol{x})$ est constant

$$\boldsymbol{\nu}(\boldsymbol{x}) = \boldsymbol{\nu}_{\alpha\beta}\ .$$

Dans ce cas, le deuxième terme dans le membre de droite de (9.2.3) est nul en raison de la définition du barycentre $\boldsymbol{x}_{\alpha\beta}$ de la face. Avec la définition (5.7.5) des moments $\boldsymbol{l}^{(2)}_{\alpha\beta}$ d'ordre deux de la face $\mathcal{A}_{\alpha\beta}$ et la notation (5.2.10), on obtient alors l'approximation suivante pour l'intégrale (9.2.1)

$$\int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}(\boldsymbol{x}) \cdot \boldsymbol{f}\left(u_{\alpha\beta} + \boldsymbol{\sigma}_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha\beta})\right)\, d\sigma \approx \boldsymbol{a}_{\alpha\beta} \cdot \boldsymbol{f}\left(u_{\alpha\beta}\right) + \frac{1}{2!}\boldsymbol{a}_{\alpha\beta} \cdot \left.\frac{d^2\boldsymbol{f}}{du^2}\right|_{u_{\alpha\beta}} \left(\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{l}^{(2)}_{\alpha\beta} \cdot \boldsymbol{\sigma}_\alpha\right)\ . \tag{9.2.4}$$

Le deuxième terme dans le membre de droite de (9.2.4) est une correction quadratique en $\boldsymbol{\sigma}_\alpha$ pour la valeur numérique de l'intégrale (9.2.1). On rappelle que $\boldsymbol{a}_{\alpha\beta}$ est le vecteur surface de la face $\mathcal{A}_{\alpha\beta}$ qui satisfait

$$\|\boldsymbol{a}_{\alpha\beta}\|_2 = |\mathcal{A}_{\alpha\beta}| \ .$$

Le chapitre 6 montre que la discrétisation spatiale par la méthode des volumes finis remplace la fonction flux $\boldsymbol{f}$ de la loi de conservation par un flux numérique décentré. Pour cela, il suffit de remplacer le premier terme dans le membre de droite de (9.2.4) par un flux décentré. La correction quadratique dans le membre de droite de (9.2.4) peut être évaluée de façon symétrique à partir des deux états de chaque côté de la face. Cette approche a été utilisée pour implémenter l'approximation (9.2.4) dans Cedre.

9.3. Intégration par formule de quadrature

Les méthodes de quadrature sont des approximations de la valeur numérique d'une intégrale. En général, on remplace le calcul de l'intégrale par une somme pondérée prise en un certain nombre de points du domaine d'intégration. Dans le contexte des schémas volumes finis, les domaines d'intégration sont surtout les interfaces entre les cellules, c'est-à-dire les triangles et les quadrangles dans $\mathbb{R}^3$ ainsi que les segments de droite dans $\mathbb{R}^2$. En dimension trois, les maillages non structurés de type général peuvent également contenir des faces non planes. Dans Cedre, ces faces sont représentées par une réunion de triangles, voir la section 5.3 pour les détails. On exige en général qu'une formule de quadrature permette de calculer l'intégrale des polynômes de degré k de façon exacte.

Le rapport [**114**] présente un certain nombre de formules de quadrature pour les simplexes. Soit $\mathcal{A} \subseteq \mathbb{R}^2$ un triangle de sommets $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ et $\boldsymbol{v}_3$ et de surface $|\mathcal{A}|$. Une formule d'intégration pour intégrer exactement les polynômes $p \in \mathbb{P}_3\left(\mathbb{R}^2\right)$ sur $\mathcal{A}$ est donnée par la formule

$$\mathfrak{I}_{\mathcal{A}}^{(3)}(p) = |\mathcal{A}| \left\{ -\frac{9}{16} p(\boldsymbol{s}_0) + \frac{25}{48} \left(p(\boldsymbol{s}_1) + p(\boldsymbol{s}_2) + p(\boldsymbol{s}_3) \right) \right\} \tag{9.3.1}$$

où les points $\boldsymbol{s}_0$ à $\boldsymbol{s}_3$ sont définis par

$$\left.\begin{aligned} \boldsymbol{s}_0 &\triangleq \frac{1}{3}(\boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3) \\ \boldsymbol{s}_1 &\triangleq \frac{1}{5}(3\boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3) \\ \boldsymbol{s}_2 &\triangleq \frac{1}{5}(\boldsymbol{v}_1 + 3\boldsymbol{v}_2 + \boldsymbol{v}_3) \\ \boldsymbol{s}_3 &\triangleq \frac{1}{5}(\boldsymbol{v}_1 + \boldsymbol{v}_2 + 3\boldsymbol{v}_3) \end{aligned}\right\} . \tag{9.3.2}$$

L'avantage des formules de quadrature du type (9.3.1) réside dans le fait qu'elles s'appliquent directement aux faces triangulaires de $\mathbb{R}^3$: si $\mathcal{A} \subseteq \mathbb{R}^3$ est un triangle de sommets $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ et $\boldsymbol{v}_3$ et p un polynôme $p \in \mathbb{P}_3\left(\mathbb{R}^3\right)$, la formule (9.3.1) permet de calculer l'intégrale de surface

$$\int_{\mathcal{A}} p(\boldsymbol{x}) \, d\sigma = |\mathcal{A}| \left\{ -\frac{9}{16} p(\boldsymbol{s}_0) + \frac{25}{48} \left(p(\boldsymbol{s}_1) + p(\boldsymbol{s}_2) + p(\boldsymbol{s}_3) \right) \right\} .$$

La formule (9.3.1) a été implémentée et testée dans Cedre.

Pour les segments de droite, on dispose de la formule de Gauss. Soit $\mathcal{C}$ un intervalle de $\mathbb{R}$, délimité par les points $v_1 \in \mathbb{R}$ et $v_2 \in \mathbb{R}$. Une formule d'intégration numérique exacte pour les polynômes de degré trois est

$$\mathfrak{I}_{\mathcal{C}}^{(2)} = |\mathcal{C}| \left\{ \frac{1}{2} p(s_1) + \frac{1}{2} p(s_2) \right\}$$

où les points s_1 et s_2 sont définis par

$$\left.\begin{aligned} s_1 &\triangleq \frac{1}{2}\left(1+\frac{1}{\sqrt{3}}\right)v_1+\frac{1}{2}\left(1-\frac{1}{\sqrt{3}}\right)v_2 \\ s_2 &\triangleq \frac{1}{2}\left(1-\frac{1}{\sqrt{3}}\right)v_1+\frac{1}{2}\left(1+\frac{1}{\sqrt{3}}\right)v_2 \end{aligned}\right\} \tag{9.3.3}$$

et $|\mathcal{C}| = |v_2 - v_1|$.

Cette formule s'applique au cas des segments de droite $\mathcal{C}$ dans $\mathbb{R}^2$, délimité par les sommets $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$: si l'on calcule les points $\boldsymbol{s}_1 \in \mathbb{R}^2$ et $\boldsymbol{s}_2 \in \mathbb{R}^2$ par

$$\left.\begin{aligned} \boldsymbol{s}_1 &\triangleq \frac{1}{2}\left(1+\frac{1}{\sqrt{3}}\right)\boldsymbol{v}_1+\frac{1}{2}\left(1-\frac{1}{\sqrt{3}}\right)\boldsymbol{v}_2 \\ \boldsymbol{s}_2 &\triangleq \frac{1}{2}\left(1-\frac{1}{\sqrt{3}}\right)\boldsymbol{v}_1+\frac{1}{2}\left(1+\frac{1}{\sqrt{3}}\right)\boldsymbol{v}_2 \end{aligned}\right\}, \tag{9.3.4}$$

la formule

$$\mathfrak{I}_{\mathcal{C}}^{(2)} = |\mathcal{C}|\left\{\frac{1}{2}p(\boldsymbol{s}_1)+\frac{1}{2}p(\boldsymbol{s}_2)\right\} = \int_{\mathcal{C}} p\left(\boldsymbol{x}\right)\, d\sigma \tag{9.3.5}$$

est exacte pour les polynômes $p \in \mathbb{P}_3\left(\mathbb{R}^2\right)$. La formule (9.3.5) a également été implémentée et testée dans Cedre.

La formule (9.3.5) permet par ailleurs de définir une formule de quadrature pour les rectangles dans $\mathbb{R}^2$ et $\mathbb{R}^3$. Un rectangle de $\mathbb{R}^2$ peut être considéré comme le produit de deux intervalles réels $\mathcal{C}_1 \subset \mathbb{R}$ et $\mathcal{C}_2 \subset \mathbb{R}$. Une application successive de la formule (9.3.5) permet d'établir une formule à quatre points qui intègre exactement les polynômes de degré trois sur un rectangle. Cette formule pour les rectangles a été implémentée et testée dans Cedre où un traitement en début de calcul détermine si une face est rectangulaire ou non.

9.4. Bilan du chapitre

Le chapitre a permis de présenter deux méthodes pour approcher la valeur numérique des intégrales sur les interfaces des cellules. Chaque méthode a ses avantages et ses inconvénients. Pour le moment, les deux méthodes ont été implémentées et testées dans Cedre : la méthode des fonctions de base seulement pour les polynômes de degré deux et la méthode des quadratures pour les polynômes de degré trois.

CHAPITRE 10

Étude de la stabilité en maillage non structuré général

10.1. Objectif du chapitre

L'objectif est d'examiner la stabilité des schémas numériques reposant sur les méthodes de reconstruction développées dans les chapitres 7 et 8.

Pour commencer, il faut préciser la notion de stabilité utilisée dans cette étude. La discrétisation spatiale par la méthode des volumes finis transforme une loi de conservation hyperbolique en un système dynamique, c'est-à-dire en une équation différentielle ordinaire en temps. Dans la suite, ce type de système dynamique s'appellera *équation semi-discrète* ou *schéma semi-discret*. Les solutions de cette équation semi-discrète sont censées approcher les solutions exactes de la loi de conservation. Elles doivent donc respecter certaines propriétés des solutions exactes et, en particulier, rester dans des bornes imposées par la théorie et l'expérience.

Dans la pratique des simulations numériques pour l'énergétique, les lois de conservation auxquelles on s'intéresse sont en général les équations de Navier-Stokes pour les fluides compressibles. Les propriétés des solutions exactes de ces équations sont à ce jour mal connues. Ce sont les phénomènes physiques décrits par les équations de Navier-Stokes qui permettent d'établir des bornes que la solution doit respecter. Les grandeurs physiques ne doivent pas devenir infinies et certaines variables comme la pression, la température, la densité de masse et encore l'énergie interne doivent rester positives. Pour la théorie mathématique des équations de Navier-Stokes des fluides incompressibles, on peut par exemple consulter les ouvrages récents [**104**] et [**51**], mais pour le cas des fluides compressibles il existe beaucoup moins de littérature.

En raison de ces difficultés mathématiques, on analyse souvent les schémas numériques à l'aide d'une équation modèle, notamment l'équation de convection linéaire à vitesse constante, dont les solutions exactes sont bien connues. La linéarisation des équations de Navier-Stokes fournit par ailleurs un système d'équations linéaires de convection-diffusion. Ceci implique que le cas linéaire est important pour décrire le comportement des solutions de Navier-Stokes, ce qui justifie de concentrer l'étude de stabilité sur l'équation linéaire de convection à vitesse constante. Le cas des équations de Navier-Stokes sera discuté dans la dernière partie de ce chapitre.

Il est nécessaire de faire deux remarques afin de préciser la portée de l'étude du présent chapitre :

(1) L'application d'une méthode de discrétisation en temps transforme le schéma semi-discret en un schéma discret en espace et en temps qui est accessible à la résolution sur ordinateur. En général, ce schéma discret sera stable si et seulement si le schéma semi-discret est stable et si le domaine de stabilité de la méthode d'intégration en temps contient tout le spectre du système semi-discret. L'étude de ce chapitre est uniquement consacrée à la stabilité de l'équation semi-discrète afin d'isoler les effets de la discrétisation spatiale sur la stabilité. Les questions spécifiques à la stabilité des méthodes d'intégration en temps ne font pas l'objet du présent chapitre.

(2) Pour assurer la stabilité au voisinage de discontinuités et de chocs, les schémas volumes finis utilisent des algorithmes comme les limiteurs ou les reconstructions ENO/WENO. Ces approches sont intrinsèquement non linéaires dans la mesure où elles transforment l'équation de convection linéaire en une équation semi-discrète non-linéaire. L'objectif de la présente étude est d'analyser l'influence de la reconstruction sur la stabilité du schéma numérique. Pour cette raison, il est nécessaire d'étudier la discrétisation spatiale sans limiteur ou autre mécanisme non linéaire. Les mécanismes de limitation et de stabilité non-linéaire font l'objet du chapitre 12.

L'application des méthodes de discrétisation spatiale développées dans les chapitres 6 et 7 à la convection linéaire conduit à une équation semi-discrète qui gouverne l'évolution des moyennes de cellule. Cette équation semi-discrète est une équation différentielle linéaire. Elle est donc entièrement décrite par une matrice carrée qui est appelée *opérateur de discrétisation spatiale*. Il est important de noter que les solutions exactes de l'équation linéaire de convection restent bornées en tout temps et que deux solutions proches à un temps donné le demeurent ensuite. Il est donc naturel d'exiger que les solutions de l'équation semi-discrète montrent le même comportement concernant la stabilité. Selon la théorie des équations différentielles, ce sont les valeurs propres de la matrice de discrétisation spatiale qui déterminent le comportement des solutions en grand temps. S'il existe une valeur propre avec une partie réelle positive, la norme du vecteur propre associé est non bornée en temps. Il est absolument nécessaire que la méthode de discrétisation spatiale évite de créer de tels vecteurs propres. Il faut noter que la discrétisation en temps ne joue aucun rôle dans ce problème car le phénomène se manifeste sur l'équation continue en temps.

Le premier objectif est donc de mener une étude numérique sur des échantillons de maillages afin de confirmer le fait que les instabilités du schéma MUSCL observées avec CEDRE apparaissent déjà dans le cas de la convection linéaire. Le calcul des valeurs propres de la matrice de discrétisation spatiale de l'équation de convection linéaire sur un carré ou un cube avec des conditions de périodicité aux limites permettra de détecter des valeurs propres correspondant à des modes non bornés.

À partir de ce constat, il faut mener deux études complémentaires, l'une théorique et l'autre numérique, afin d'analyser l'influence du type de maillage, de la méthode de reconstruction et du voisinage de reconstruction sur la stabilité du schéma. Comme expliqué ci-dessus, ces études se concentrent sur la stabilité linéaire dans le cas semi-discret.

L'étude numérique a pour but de tester les méthodes de reconstruction sur différents types de maillages avec différents voisinages de reconstruction pour déterminer dans quelle configuration des instabilités apparaissent. L'objectif principal de l'étude numérique est de trouver une propriété de la reconstruction locale corrélée à l'apparition de modes propres instables. Une telle propriété permettra ensuite d'identifier les méthodes de reconstruction donnant des schémas stables.

L'étude théorique examine d'abord le schéma volumes finis le plus simple, c'est-à-dire le schéma avec reconstruction constante par cellule. Ensuite, il faut combiner ces résultats théoriques avec les connaissances empiriques obtenues par l'étude numérique sur les schémas d'ordre plus élevé. En particulier, il faut essayer d'expliquer par la théorie les corrélations observées entre propriétés locales de la reconstruction et les instabilités.

L'objectif final du travail est d'établir des conclusions pratiques pour la reconstruction dans des codes de volumes finis (comme CEDRE par exemple). Il faudra comprendre si certaines méthodes de reconstruction favorisent davantage la stabilité que d'autres. Il est également nécessaire de comprendre l'influence de la taille du voisinage de reconstruction sur la stabilité du schéma. Finalement, il est essentiel d'examiner la stabilité des différentes méthodes itératives de reconstruction proposées dans le chapitre 8.

La question de la stabilité des schémas Navier-Stokes sera discutée à l'exemple des calculs tridimensionnels dans les chapitres 13 et 14.

10.2. Construction des schémas semi-discrets

L'objectif de cette section est d'appliquer les méthodes de discrétisation spatiale développées dans les chapitres 6, 7, 8 et 9 à l'équation de convection linéaire avec vitesse constante $\boldsymbol{c} \in \mathbb{R}^d$

$$\partial_t u(\boldsymbol{x}, t) + \boldsymbol{c} \cdot \boldsymbol{\nabla} u(\boldsymbol{x}, t) = 0\,,\ (\boldsymbol{x}, t) \in \mathbb{R}^d \times \mathbb{R}_+\,. \tag{10.2.1}$$

Le flux conservatif associé à (10.2.1) est linéaire en u

$$\boldsymbol{f}(u) = \boldsymbol{c}u\,. \tag{10.2.2}$$

L'équation de bilan de la loi de conservation (10.2.1) est

$$\frac{d}{dt}\int_{\mathcal{T}} u\left(\boldsymbol{x},t\right)\,dx = -\int_{\partial\mathcal{T}} \left(\boldsymbol{\nu}\left(\boldsymbol{x}\right)\cdot\boldsymbol{c}\right)\,u\left(\boldsymbol{x},t\right)\,d\sigma \tag{10.2.3}$$

sur tout sous-volume suffisamment régulier $\mathcal{T}\subseteq\mathbb{R}^d$.

La discrétisation de (10.2.3) suit les lignes du chapitre 6. L'objectif est d'établir une équation semi-discrète du type (6.3.4)

$$\frac{du_\alpha\left(t\right)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \widetilde{f}_{\alpha\beta}\left(w_\alpha\left[\mathfrak{u}\left(t\right)\right]\left(\boldsymbol{x}\right), w_\beta\left[\mathfrak{u}\left(t\right)\right]\left(\boldsymbol{x}\right)\right)\,d\sigma\,,\ 1\leq\alpha\leq N$$

dont la solution $\mathfrak{u}\triangleq\left(u_1,\ldots,u_N\right)$ approche les moyennes $\left(\overline{u}_1,\ldots,\overline{u}_N\right)$ de la solution exacte u de (10.2.3). Les fonctions $w_\alpha\left[\mathfrak{u}\left(t\right)\right]$ sont les fonctions reconstruites à partir de $\mathfrak{u}\triangleq\left(u_1,\ldots,u_N\right)$ dans la cellule $\mathcal{T}_\alpha$. La fonction $\widetilde{f}_{\alpha\beta}$ est un flux numérique qui satisfait les critères de la définition 6.2.2.

Le choix du flux numérique $\widetilde{f}_{\alpha\beta}$ pour la discrétisation de l'équation (10.2.1) est le flux décentré classique [**79, 52, 73**]

$$\begin{aligned}\widetilde{f}^{(\mathrm{d})}_{\alpha\beta}\left(w_{\mathrm{int}},w_{\mathrm{ext}}\right) = \mathbf{c}\cdot\boldsymbol{\nu}_{\alpha\beta}\frac{w_{\mathrm{int}}+w_{\mathrm{ext}}}{2} + \left|\mathbf{c}\cdot\boldsymbol{\nu}_{\alpha\beta}\right|\frac{w_{\mathrm{int}}-w_{\mathrm{ext}}}{2} = \\ = \left(\mathbf{c}\cdot\boldsymbol{\nu}_{\alpha\beta}\right)_+ w_{\mathrm{int}} + \left(\mathbf{c}\cdot\boldsymbol{\nu}_{\alpha\beta}\right)_- w_{\mathrm{ext}}\,.\end{aligned} \tag{10.2.4}$$

Dans la formule (10.2.4), w_{int} désigne un état interpolé à l'intérieur de la cellule $\mathcal{T}_\alpha$ et w_{ext} un état interpolé à l'extérieur de la cellule $\mathcal{T}_\alpha$, c'est-à-dire dans la cellule $\mathcal{T}_\beta$. Les notations x_+ et x_- signifient respectivement la partie positive et la partie négative d'un nombre réel x.

Il est important de noter que le flux (10.2.4) est un *flux monotone* au sens de la définition 6.2.3. Un choix alternatif pour le flux numérique est le flux centré

$$\widetilde{f}^{(\mathrm{c})}_{\alpha\beta}\left(w_{\mathrm{int}},w_{\mathrm{ext}}\right) = \mathbf{c}\cdot\boldsymbol{\nu}_{\alpha\beta}\frac{w_{\mathrm{int}}+w_{\mathrm{ext}}}{2} \tag{10.2.5}$$

qui n'est pas monotone.

L'insertion de (10.2.4) dans le schéma (6.3.4) et l'utilisation de reconstructions constantes par cellule

$$w_\alpha\left[\mathfrak{u}\right]\left(\boldsymbol{x}\right) = u_\alpha\,,\ w_\beta\left[\mathfrak{u}\right]\left(\boldsymbol{x}\right) = u_\beta$$

donne le schéma semi-discret

$$|\mathcal{T}_\alpha|\,\frac{du_\alpha}{dt} = -\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \left[\left(\boldsymbol{c}\cdot\boldsymbol{\nu}\left(x\right)\right)_+ u_\alpha + \left(\boldsymbol{c}\cdot\boldsymbol{\nu}\left(x\right)\right)_- u_\beta\right]\,d\sigma\,,\ 1\leq\alpha\leq N. \tag{10.2.6}$$

Sous l'hypothèse que le signe de $\boldsymbol{c}\cdot\boldsymbol{\nu}\left(x\right)$ ne change pas le long de la face $\mathcal{A}_{\alpha\beta}$, il est possible de sortir le vecteur de vitesse $\boldsymbol{c}$ de l'intégrale en (10.2.6). L'équation différentielle (10.2.6) se simplifie alors grâce à la définition (5.3.3) du vecteur surface

$$\boldsymbol{a}_{\alpha\beta}\triangleq\int_{\mathcal{A}_{\alpha\beta}}\boldsymbol{\nu}\left(\boldsymbol{x}\right)\,d\sigma$$

pour donner

$$\frac{du_\alpha}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta\left[\left(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ u_\alpha + \left(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- u_\beta\right]\,,\ 1\leq\alpha\leq N. \tag{10.2.7}$$

Dans la suite, l'équation différentielle ordinaire (10.2.7) est appelée *schéma semi-discret d'ordre un*. La définition de l'opérateur $\widetilde{J}$

$$\widetilde{J}_{\alpha\beta}\triangleq -\frac{1}{|\mathcal{T}_\alpha|}\left(\sum_\gamma\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\gamma}\right)_+\right)\delta_{\alpha\beta} - \frac{1}{|\mathcal{T}_\alpha|}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- \tag{10.2.8}$$

permet d'écrire (10.2.7) sous la forme

$$\frac{du_\alpha(t)}{dt} = \sum_\beta \widetilde{J}_{\alpha\beta} u_\beta(t) \ , \ 1 \leq \alpha \leq N.$$

Considérons à présent la reconstruction linéaire, c'est-à-dire la reconstruction des polynômes de degré un. Dans ce cas, la forme générale des fonctions reconstruites (7.5.5) est

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \overline{u}_\alpha + \boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] \bullet (\boldsymbol{x} - \boldsymbol{x}_\alpha)$$

où

$$\boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(1)} \overline{u}_\beta$$

est un gradient consistant au sens de la définition 7.4.1. Dans la suite, le gradient $\boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}]$ est noté $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$ et les coefficients $\boldsymbol{w}_{\alpha\beta}^{(1)}$ s'appellent $\boldsymbol{\sigma}_{\alpha\beta} \triangleq \boldsymbol{w}_{\alpha\beta}^{(1)}$, ce qui permet d'écrire

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] \ = \ \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} u_\beta = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} (u_\beta - u_\alpha) \ . \tag{10.2.9}$$

Avec ces notations, les polynômes reconstruits dans les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ sont respectivement

$$\left.\begin{array}{l} w_\alpha[\mathfrak{u}](\boldsymbol{x}) = u_\alpha + \boldsymbol{\sigma}_\alpha[\mathfrak{u}] \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) \\ w_\beta[\mathfrak{u}](\boldsymbol{x}) = u_\beta + \boldsymbol{\sigma}_\beta[\mathfrak{u}] \cdot (\boldsymbol{x} - \boldsymbol{x}_\beta) \end{array}\right\} . \tag{10.2.10}$$

L'insertion des polynômes reconstruits (10.2.10) et du flux (10.2.4) dans (6.3.4) donne le schéma

$$\begin{aligned} |\mathcal{T}_\alpha| \frac{du_\alpha}{dt} = &- \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{c} \cdot \boldsymbol{\nu}(x))_+ \left\{ u_\alpha + (\boldsymbol{x} - \boldsymbol{x}_\alpha) \cdot \sum_\gamma \boldsymbol{\sigma}_{\alpha\gamma} (u_\gamma - u_\alpha) \right\} d\sigma - \\ &- \sum_\beta \int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{c} \cdot \boldsymbol{\nu}(x))_- \left\{ u_\beta + (\boldsymbol{x} - \boldsymbol{x}_\beta) \cdot \sum_\gamma \boldsymbol{\sigma}_{\beta\gamma} (u_\gamma - u_\beta) \right\} d\sigma \end{aligned} \tag{10.2.11}$$

où l'indice α parcourt toutes les cellules. Si la face $\mathcal{A}_{\alpha\beta}$ est plane, le vecteur normal unitaire $\boldsymbol{\nu}(x)$ est constant. Il est alors possible de sortir l'expression $\boldsymbol{c} \cdot \boldsymbol{\nu}(x)$ de l'intégrale dans (10.2.11) et d'appliquer la formule

$$\int_{\mathcal{A}_{\alpha\beta}} (\boldsymbol{c} \cdot \boldsymbol{\nu}(x))_\pm (\boldsymbol{x} - \boldsymbol{x}_\alpha) \, d\sigma = (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_\pm \boldsymbol{k}_{\alpha\beta} \tag{10.2.12}$$

à (10.2.11). Cela donne finalement le schéma semi-discret

$$\begin{aligned} \frac{du_\alpha}{dt} = &- \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_+ \left\{ u_\alpha + \boldsymbol{k}_{\alpha\beta} \cdot \sum_\gamma \boldsymbol{\sigma}_{\alpha\gamma} (u_\gamma - u_\alpha) \right\} - \\ &- \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \left\{ u_\beta + \boldsymbol{k}_{\beta\alpha} \cdot \sum_\gamma \boldsymbol{\sigma}_{\beta\gamma} (u_\gamma - u_\beta) \right\} , \ 1 \leq \alpha \leq N \, . \end{aligned} \tag{10.2.13}$$

Dans la suite, le schéma (10.2.13) est appelé *schéma semi-discret d'ordre deux* ou *schéma* MUSCL *semi-discret*. La définition

$$\boldsymbol{\sigma}_\alpha \triangleq \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \overline{u}_\beta$$

permet de définir la matrice de l'opérateur J, appelé *opérateur* MUSCL, par

$$J_{\alpha\beta} \triangleq -\frac{1}{|\mathcal{T}_\alpha|}\left\{\sum_\gamma (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\gamma})_+ \delta_{\alpha\beta} + \sum_\gamma (\boldsymbol{a}_{\alpha\gamma}\cdot\boldsymbol{c})_+ \boldsymbol{k}_{\alpha\gamma}\cdot\boldsymbol{\sigma}_{\alpha\beta} - \right.$$
$$- \sum_\gamma (\boldsymbol{a}_{\alpha\gamma}\cdot\boldsymbol{c})_+ \boldsymbol{k}_{\alpha\gamma}\cdot\boldsymbol{\sigma}_\alpha\,\delta_{\alpha\beta} + (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_- -$$
$$\left. - \sum_\gamma (\boldsymbol{a}_{\gamma\alpha}\cdot\boldsymbol{c})_+ \boldsymbol{k}_{\gamma\alpha}\cdot\boldsymbol{\sigma}_{\gamma\beta} + (\boldsymbol{a}_{\beta\alpha}\cdot\boldsymbol{c})_+ \boldsymbol{k}_{\beta\alpha}\cdot\boldsymbol{\sigma}_\beta \right\} \quad (10.2.14)$$

qui permet de formuler le schéma MUSCL semi-discret (10.2.13) sous la forme

$$\frac{du_\alpha(t)}{dt} = \sum_\beta J_{\alpha\beta} u_\beta(t)\ ,\ 1\le\alpha\le N.$$

Le procédé se généralise à la reconstruction des polynômes de degré $k=2$, c'est-à-dire à la reconstruction quadratique par cellule. Dans ce cas, la forme générale des fonctions reconstruites (7.5.5) est

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = u_\alpha + \boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}]\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha) + \frac{1}{2!}\boldsymbol{w}_\alpha^{(2)}[\mathfrak{u}]\bullet\left[(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 - \boldsymbol{x}_\alpha^{(2)}\right]$$

où

$$\boldsymbol{w}_\alpha^{(2)}[\mathfrak{u}] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(2)} u_\beta$$

est une dérivée seconde consistante au sens de la définition 7.4.1 et

$$\boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}] = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(1)} u_\beta$$

est un *gradient consistant de précision à l'ordre deux* au sens de la définition 7.4.1. Dans la suite, la dérivée seconde $\boldsymbol{w}_\alpha^{(2)}[\mathfrak{u}]$ est notée $\boldsymbol{\theta}_\alpha[\mathfrak{u}]$ et le gradient $\boldsymbol{w}_\alpha^{(1)}[\mathfrak{u}]$ est noté $\boldsymbol{\sigma}_\alpha[\mathfrak{u}]$. Les coefficients $\boldsymbol{w}_{\alpha\beta}^{(2)}$ s'appellent $\boldsymbol{\theta}_{\alpha\beta}\triangleq\boldsymbol{w}_{\alpha\beta}^{(2)}$ et les coefficients $\boldsymbol{w}_{\alpha\beta}^{(1)}$ sont notés $\boldsymbol{\sigma}_{\alpha\beta}\triangleq\boldsymbol{w}_{\alpha\beta}^{(1)}$, ce qui permet d'écrire

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} u_\beta = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\,(u_\beta - u_\alpha) \quad (10.2.15)$$

$$\boldsymbol{\theta}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\theta}_{\alpha\beta} u_\beta = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}\,(u_\beta - u_\alpha)\ . \quad (10.2.16)$$

Avec ces notations, les polynômes reconstruits dans les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$ sont respectivement

$$\left.\begin{array}{l} w_\alpha[\mathfrak{u}](\boldsymbol{x}) = u_\alpha + (\boldsymbol{x}-\boldsymbol{x}_\alpha)\cdot\boldsymbol{\sigma}_\alpha[\mathfrak{u}] + \dfrac{1}{2!}\left[(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 - \boldsymbol{x}_\alpha^{(2)}\right]\bullet\boldsymbol{\theta}_\alpha[\mathfrak{u}] \\ w_\beta[\mathfrak{u}](\boldsymbol{x}) = u_\beta + (\boldsymbol{x}-\boldsymbol{x}_\beta)\cdot\boldsymbol{\sigma}_\beta[\mathfrak{u}] + \dfrac{1}{2!}\left[(\boldsymbol{x}-\boldsymbol{x}_\beta)^2 - \boldsymbol{x}_\beta^{(2)}\right]\bullet\boldsymbol{\theta}_\beta[\mathfrak{u}] \end{array}\right\}. \quad (10.2.17)$$

Si toutes les faces sont planes, l'insertion de (10.2.17) et du flux (10.2.4) dans (6.3.4) donne le schéma semi-discret

$$|\mathcal{T}_\alpha|\frac{du_\alpha}{dt} = -\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+\left\{u_\alpha + \boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{\sigma}_\alpha[\mathfrak{u}] + \frac{1}{2!}\left[\boldsymbol{k}_{\alpha\beta}^{(2)} - \boldsymbol{x}_\alpha^{(2)}\right]\bullet\boldsymbol{\theta}_\alpha[\mathfrak{u}]\right\} -$$
$$- \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_-\left\{u_\beta + \boldsymbol{k}_{\beta\alpha}\cdot\boldsymbol{\sigma}_\beta[\mathfrak{u}] + \frac{1}{2!}\left[\boldsymbol{k}_{\beta\alpha}^{(2)} - \boldsymbol{x}_\beta^{(2)}\right]\bullet\boldsymbol{\theta}_\beta[\mathfrak{u}]\right\} \quad (10.2.18)$$

où l'indice α parcourt toutes les cellules du maillage.

Le cas de la reconstruction des polynômes de degré k est décrit par la

PROPOSITION 10.2.1 (Schéma semi-discret basé sur une reconstruction consistante de degré k). *On considère une reconstruction des polynômes de degré k basée sur des dérivées consistantes au sens de la définition 7.4.1*

$$\boldsymbol{w}_{\alpha}^{(l)}\left[\mathfrak{u}\right]=\sum_{\gamma}\boldsymbol{w}_{\alpha\gamma}^{(l)}\left(u_{\gamma}-u_{\alpha}\right)\ ,\ 1\leq l\leq k\,.$$

Si toutes les faces du maillage sont planes, la discrétisation spatiale de l'équation de convection linéaire à l'aide du flux décentré (10.2.4) donne l'équation différentielle

$$\begin{aligned}\left|\mathcal{T}_{\alpha}\right|\frac{du_{\alpha}}{dt}=&-\sum_{\beta}\left(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{+}\left\{u_{\alpha}+\sum_{l=1}^{k}\frac{1}{l!}\left[\boldsymbol{k}_{\alpha\beta}^{(l)}-\boldsymbol{x}_{\alpha}^{(l)}\right]\bullet\boldsymbol{w}_{\alpha}^{(l)}\left[\mathfrak{u}\right]\right\}-\\&-\sum_{\beta}\left(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{-}\left\{u_{\beta}+\sum_{l=1}^{k}\frac{1}{l!}\left[\boldsymbol{k}_{\beta\alpha}^{(l)}-\boldsymbol{x}_{\beta}^{(l)}\right]\bullet\boldsymbol{w}_{\beta}^{(l)}\left[\mathfrak{u}\right]\right\}\ ,\ 1\leq\alpha\leq N\,.\end{aligned}\tag{10.2.19}$$

DÉMONSTRATION. Dans le cas d'une reconstruction conservative des polynômes de degré k, les fonctions reconstruites sont données par (7.5.5). En particulier, les polynômes reconstruits dans les cellules $\mathcal{T}_{\alpha}$ et $\mathcal{T}_{\beta}$ sont respectivement

$$\left.\begin{aligned}w_{\alpha}\left[\mathfrak{u}\right]\left(\boldsymbol{x}\right)=&\ u_{\alpha}+\sum_{l=1}^{k}\frac{1}{l!}\left[\left(\boldsymbol{x}-\boldsymbol{x}_{\alpha}\right)^{l}-\boldsymbol{x}_{\alpha}^{(l)}\right]\bullet\boldsymbol{w}_{\alpha}^{(l)}\left[\mathfrak{u}\right]\\w_{\beta}\left[\mathfrak{u}\right]\left(\boldsymbol{x}\right)=&\ u_{\beta}+\sum_{l=1}^{k}\frac{1}{l!}\left[\left(\boldsymbol{x}-\boldsymbol{x}_{\beta}\right)^{l}-\boldsymbol{x}_{\beta}^{(l)}\right]\bullet\boldsymbol{w}_{\beta}^{(l)}\left[\mathfrak{u}\right]\end{aligned}\right\}\,.\tag{10.2.20}$$

Si toutes les faces sont planes, leurs normales unitaires sont constantes, ce qui donne les formules

$$\int_{\mathcal{A}_{\alpha\beta}}\left(\boldsymbol{c}\cdot\boldsymbol{\nu}\left(x\right)\right)_{\pm}\left(\boldsymbol{x}-\boldsymbol{x}_{\alpha}\right)^{l}\,d\sigma=\left(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{\pm}\boldsymbol{k}_{\alpha\beta}^{(l)}\tag{10.2.21}$$

car on peut alors sortir l'expression $\left(\boldsymbol{c}\cdot\boldsymbol{\nu}\left(x\right)\right)_{\pm}$ de l'intégrale de (10.2.21). L'insertion des fonctions (10.2.20), des formules (10.2.21) et du flux (10.2.4) dans le membre de gauche de (6.3.4)

$$\frac{du_{\alpha}\left(t\right)}{dt}=-\frac{1}{\left|\mathcal{T}_{\alpha}\right|}\sum_{\beta}\int_{\mathcal{A}_{\alpha\beta}}\widetilde{f}_{\alpha\beta}\left(w_{\alpha}\left[\mathfrak{u}\left(t\right)\right]\left(\boldsymbol{x}\right),w_{\beta}\left[\mathfrak{u}\left(t\right)\right]\left(\boldsymbol{x}\right)\right)\,d\sigma$$

donne le schéma semi-discret (10.2.19). □

La proposition 10.2.1 suggère d'introduire la

DÉFINITION 10.2.2 (Schéma et opérateur semi-discrets d'ordre $k+1$). Le système différentiel ordinaire (10.2.19) s'appelle *schéma semi-discret d'ordre* $k+1$ si l'équation a été établie à l'aide de la reconstruction consistante de degré k. Il s'agit d'un problème de Cauchy pour le vecteur des variables semi-discrètes $\mathfrak{u}=\left(u_{1},\ldots,u_{N}\right)$. Lorsqu'on écrit l'équation (10.2.19) sous la forme matricielle

$$\frac{d\mathfrak{u}\left(t\right)}{dt}=J\mathfrak{u}\left(t\right)\ ,\ \mathfrak{u}\left(t_{0}\right)=\mathfrak{u}_{0}\,,\tag{10.2.22}$$

la matrice J est appelée *opérateur semi-discret d'ordre* $k+1$. Pour $k+1=2$, l'équation est donnée par (10.2.13) et s'appelle *schéma* MUSCL *semi-discret.* L'opérateur (10.2.14) s'appelle dans ce cas *opérateur* MUSCL *semi-discret.* Pour $k=0$, l'équation est donnée par (10.2.7) et s'appelle *schéma semi-discret d'ordre un.* L'opérateur associé (10.2.8) est appelé *opérateur semi-discret d'ordre un.*

REMARQUE 10.2.3. Il faut souligner que la notion d'ordre dans la définition 10.2.2 est purement formelle. Cette définition suit la terminologie habituelle de la littérature mais ne dit rien sur le taux de convergence des schémas.

10.3. Notions de stabilité asymptotique

L'une des caractéristiques de l'équation de convection linéaire est de préserver la norme de toute condition initiale u_0. Considérons par exemple des conditions de périodicité dans $\Omega = [0,1]^d$. La forme de la solution $u(\boldsymbol{x},t) = u_0(\boldsymbol{x}-\boldsymbol{c}t)$ implique la conservation de toutes les normes L_p, $1 \leq p \leq \infty$,

$$\|u(.,t)\|_p = \|u_0\|_p \,. \tag{10.3.1}$$

Il faut examiner dans quelle mesure la propriété (10.3.1) peut être préservée par les schémas numériques du type (10.2.19).

On rappelle le résultat suivant pour un système linéaire homogène et autonome

$$\frac{d\mathfrak{u}(t)}{dt} = J\mathfrak{u}(t) \;,\; \mathfrak{u}(0) = \mathfrak{u}_0 \,,\; \mathfrak{u}(t) \in \mathbb{C}^N \,,\; J \in \mathbb{M}_N(\mathbb{C}) \tag{10.3.2}$$

dont la solution est

$$\mathfrak{u}(t) = \exp(tJ)\,\mathfrak{u}_0 \,,$$

voir [**47**, lm. 3.20 sur p.95 et th. 3.23 sur p. 97].

Proposition 10.3.1 (Stabilité asymptotique des systèmes linéaires en dimension finie). *Le système (10.3.2) est stable au sens que*

$$\mathrm{C} = \sup_{t\geq 0} \|\exp(tJ)\| < \infty \tag{10.3.3}$$

si et seulement si toutes les valeurs propres λ de J satisfont :

(i) $\Re(\lambda) \leq 0$ *où* $\Re(\lambda)$ *est la partie réelle de* λ.

(ii) *si* $\Re(\lambda) = 0$, *alors l'indice de Jordan* $\imath(\lambda) = 1$ *où* $\imath(\lambda)$ *est la dimension maximale du bloc de Jordan de* J *associé à* λ.

En particulier, la solution d'un système stable satisfait

$$\sup_{t\geq 0} \|\mathfrak{u}(t)\| \leq \mathrm{C}\,\|\mathfrak{u}_0\| \,. \tag{10.3.4}$$

La propriété (10.3.4) est l'équivalent de la propriété (10.3.1) dans le cas semi-discret. Par conséquent, (10.3.4) est un critère que tout opérateur semi-discret (10.2.22) doit satisfaire indépendamment de la vitesse de convection $\boldsymbol{c} \in \mathbb{R}^d$.

Définition 10.3.2 (Opérateur semi-discret stable). L'opérateur semi-discret J est dit *stable* si toutes ses valeur propres satisfont les propriétés (i) et (ii) de la proposition 10.3.1 pour *toutes* les vitesses de convection $\boldsymbol{c} \in \mathbb{R}^d$.

Le théorème suivant regroupe plusieurs résultats sur la localisation du spectre d'une matrice $A \in \mathbb{M}_N(\mathbb{C})$. Les démonstrations se trouvent dans [**110**, th. 1.1., p. 4, th. 3.7., p. 79], [**65**, th. 6.6.1, p. 344] et [**66**, propriété 1.2.6, p. 10].

Théorème 10.3.3 (Disques de Geršgorin, image numérique).

(i) *Le spectre* $\mathrm{sp}(A)$ *de* A *est contenu dans la réunion des disques de Geršgorin de* A

$$\mathrm{sp}(A) \subseteq \bigcup_{i=1}^{N} \Gamma_i(A)$$

où le i*-ème disque de Geršgorin de* A *est défini par*

$$\Gamma_i(A) = \left\{ z \in \mathbb{C} \,:\, |z - a_{ii}| \leq \sum_{j\neq i} |a_{ij}| \right\} .$$

(ii) *Le spectre de* A *est situé dans l'image numérique de* A

$$\mathrm{sp}(A) \subseteq \mathfrak{F}(A)$$

où $\mathfrak{F}(A)$ *est défini par*

$$\mathfrak{F}(A) = \left\{ (\boldsymbol{z}, A\boldsymbol{z}) \,:\, \boldsymbol{z} \in \mathbb{C}^N \,,\, (\boldsymbol{z},\boldsymbol{z}) = 1 \right\}$$

où $(.,.)$ *est le produit scalaire hermitien (5.2.2).*

■

Un théorème important est une variante du théorème classique de Lyapunov, voir [**66**, th. 2.2.1, p. 96].

THÉORÈME 10.3.4 (Théorème de Lyapunov étendu). *Soit* $J \in \mathbb{M}_N(\mathbb{C})$. *Les propriétés suivantes sont équivalentes :*

(i) *J satisfait les conditions de la proposition 10.3.1, c'est-à-dire toutes les valeurs propres λ de J satisfont $\Re(\lambda) \leq 0$ et si $\Re(\lambda) = 0$, alors l'indice de Jordan est $\imath(\lambda) = 1$.*

(ii) *Il existe une matrice définie positive G telle que la matrice $Q = GJ + J^*G$ est semi-définie négative.*

DÉMONSTRATION. On observe d'abord que les propriétés (i) et (ii) sont vraies pour une matrice J si et seulement si elles sont vraies pour toute matrice semblable à J. C'est évident pour la propriété (i) car le spectre est un invariant de similitude. Pour démontrer que la propriété (ii) est un invariant de similitude, on suppose que J a la propriété (ii) et que $\widehat{J} = S^{-1}JS$ est une matrice semblable à J. La multiplication de $GJ + J^*G = Q$ par S^* à gauche et par S à droite permet d'écrire

$$S^*GS\,S^{-1}JS + S^*J^*\,(S^*)^{-1}\;S^*GS = S^*QS\,.$$

La matrice $\widehat{Q} \triangleq S^*QS$ est semi-définie négative car elle est congruente à Q et la matrice $\widehat{G} \triangleq S^*GS$ est définie positive car elle est congruente à G. Puisque la matrice $\widehat{J}$ satisfait l'identité $\widehat{Q} = \widehat{G}\widehat{J} + \widehat{J}^*\widehat{G}$ où $\widehat{G}$ est définie positive et $\widehat{Q}$ est semi-définie négative, $\widehat{J}$ a également la propriété (ii).

(i) $\Rightarrow$(ii) : On suppose que J satisfait (i). Pour toute matrice J, il existe une matrice $\widehat{J}$ qui est semblable à J et qui se trouve sous la forme normale de Jordan, cf. [**65**, th. 3.1.5, p. 123]. D'après l'observation au début de la démonstration, il suffit de montrer que $\widehat{J}$ possède la propriété (ii). Pour tous les blocs de Jordan associés aux valeurs propres λ ayant $\Re(\lambda) < 0$, on peut d'ailleurs supposer que les éléments au-dessus de la diagonale sont tous égaux à un $\varepsilon > 0$ au lieu de 1, cf. [**65**, cor. 3.1.13, p. 128]. On peut supposer que ε satisfait

$$\varepsilon < \min\left\{|\Re(\lambda)|\ ,\ \lambda \in \mathrm{sp}(J)\ ,\ \Re(\lambda) < 0\right\}\ . \tag{10.3.5}$$

Tous les blocs de Jordan qui correspondent à des valeurs propres λ avec $\Re(\lambda) = 0$ sont diagonaux parce que $\imath(\lambda) = 1$ d'après l'hypothèse. La matrice $\widehat{Q} = \widehat{J} + \widehat{J}^*$ est hermitienne et diagonale par bloc. La relation (10.3.5) entraîne que les blocs diagonaux de $\widehat{Q}$ associés aux λ telles que $\Re(\lambda) < 0$ sont de dominance diagonale stricte. Les blocs de $\widehat{Q}$ associés aux λ telles que $\Re(\lambda) = 0$ sont nuls. Pour ces raisons, la matrice $\widehat{Q} = \widehat{J} + \widehat{J}^*$ est semi-définie négative, ce qui prouve que la forme normale de Jordan $\widehat{J}$ a la propriété (ii) si l'on choisit pour $\widehat{G}$ la matrice identité. Comme la matrice J est semblable à la matrice $\widehat{J}$, elle satisfait également la propriété (ii).

(ii) $\Rightarrow$(i) : Supposons que J satisfait (ii). La multiplication de la relation $GJ + J^*G = Q$ par la matrice $G^{-\frac{1}{2}}$ à droite et à gauche donne

$$G^{\frac{1}{2}}JG^{-\frac{1}{2}} + G^{-\frac{1}{2}}J^*G^{\frac{1}{2}} = G^{\frac{1}{2}}JG^{-\frac{1}{2}} + \left(G^{\frac{1}{2}}JG^{-\frac{1}{2}}\right)^* = G^{-\frac{1}{2}}QG^{-\frac{1}{2}}\,.$$

Cela signifie que la partie hermitienne de la matrice $G^{\frac{1}{2}}JG^{-\frac{1}{2}}$ est semi-définie négative. Pour cette raison, toutes les valeurs propres de la matrice $G^{\frac{1}{2}}JG^{-\frac{1}{2}}$ ont des parties réelles non positives. Comme la matrice J est semblable à la matrice $G^{\frac{1}{2}}JG^{-\frac{1}{2}}$, les valeurs propres de J ont également des parties réelles non positives. Il reste à démontrer que toute valeur propre purement imaginaire λ de J a un indice de Jordan $\imath(\lambda) = 1$. D'après l'observation au début de la preuve, on peut supposer que J est donnée sous la forme normale de Jordan. Supposons qu'il existe une valeur propre purement imaginaire $\lambda = i\mu$, $\mu \in \mathbb{R}$, associée au bloc de Jordan numéro l de de J, et appelons ce bloc $J^{(l)}$. Supposons par ailleurs que le bloc $J^{(l)}$ ne soit pas diagonalisable. En raison de la structure de la forme normale de Jordan de J, il est possible d'écrire la relation

$GJ + J^*G = Q$ bloc par bloc. Soit $G^{(l)}$ le bloc diagonal de G qui correspond à $J^{(l)}$. Le bloc $G^{(l)}$ est une matrice définie positive car $G^{(l)}$ est un bloc diagonal de G. On peut alors écrire la relation $G^{(l)}J^{(l)} + J^{(l)*}G^{(l)} = Q^{(l)}$ où $Q^{(l)}$ est le bloc diagonal de Q qui correspond à $J^{(l)}$. Le bloc $Q^{(l)}$ est une matrice semi-définie négative car $Q^{(l)}$ est un bloc diagonal de Q. La relation $G^{(l)}J^{(l)} + J^{(l)*}G^{(l)} = Q^{(l)}$ peut être formulée sous la forme explicite

$$\begin{pmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} i\mu & 1 & 0 & \cdots \\ 0 & i\mu & 1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} +$$

$$+ \begin{pmatrix} -i\mu & 0 & \cdots \\ 1 & -i\mu & \ddots \\ 0 & 1 & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} 0 & g_{11} & \cdots \\ g_{11} & g_{12}+g_{21} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} .$$

En particulier, il vient $Q^{(l)}_{11} = 0$ et $Q^{(l)}_{12} = g_{11}$. On a $g_{11} > 0$ car G est définie positive. Néanmoins, une matrice $Q^{(l)}$ ayant $Q^{(l)}_{11} = 0$ et $Q^{(l)}_{12} > 0$ ne peut pas être semi-définie négative. Cette contradiction démontre que le bloc de Jordan associé à λ doit être diagonal. □

La combinaison de la proposition 10.3.1 et du théorème 10.3.4 conduit au

COROLLAIRE 10.3.5. *Considérons le problème de Cauchy*

$$\frac{d\mathfrak{u}(t)}{dt} = J\mathfrak{u}(t)\ ,\ \mathfrak{u}(0) = \mathfrak{u}_0\ ,\ \mathfrak{u}(t) \in \mathbb{C}^N\ ,\ J \in \mathbb{M}_N(\mathbb{C})\ . \tag{10.3.6}$$

Il existe une constante C *telle que*

$$\|\mathfrak{u}(t)\| \leq \mathrm{C}\,\|\mathfrak{u}_0\| \ \textit{ pour tout } t \geq 0$$

si et seulement s'il existe une matrice positive définie G telle que la matrice $Q = GJ + J^*G$ est semi-définie négative.

10.4. Analyse de la stabilité du schéma d'ordre un

Cette section présente l'analyse de stabilité du schéma d'ordre un (10.2.7)

$$\frac{du_\alpha(t)}{dt} = \sum_\beta \widetilde{J}_{\alpha\beta} u_\beta(t)\ ,\ 1 \leq \alpha \leq N$$

défini par l'opérateur de discrétisation spatiale d'ordre un (10.2.8)

$$\widetilde{J}_{\alpha\beta} \quad = \quad -\frac{1}{|\mathcal{I}_\alpha|}\left(\sum_\gamma (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\gamma})_+\right)\delta_{\alpha\beta} - \frac{1}{|\mathcal{I}_\alpha|}(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_-\ .$$

Bien que la matrice de l'opérateur (10.2.8) ait des coefficients réels, ses valeurs propres sont en général complexes. Pour cette raison, il est préférable de travailler avec des vecteurs complexes $\mathfrak{u} \in \mathbb{C}^N$. Le conjugué complexe de $\mathfrak{u} \in \mathbb{C}^N$ est noté $\mathfrak{u}^*$.

Le théorème de Geršgorin conduit immédiatement au

THÉORÈME 10.4.1 (Localisation du spectre de (10.2.8) par les disques de Geršgorin). *Toutes les valeurs propres λ de $\widetilde{J}$ satisfont $\Re(\lambda) \leq 0$.*

DÉMONSTRATION. La démonstration consiste à montrer que tous les disques de Geršgorin de $\widetilde{J}$ satisfont

$$\Gamma_\alpha\left(\widetilde{J}\right) \subset \{\, z \in \mathbb{C} |\, \Re(z) \leq 0\}\ ,\ 1 \leq \alpha \leq N\,.$$

Pour $1 \leq \alpha \leq N$, $z_\alpha \in \mathbb{C}$ est le centre et $\rho_\alpha \in \mathbb{R}_+$ le rayon du disque de Geršgorin $\Gamma_\alpha\left(\widetilde{J}\right)$.

La relation (5.5.2) implique pour toutes les cellules $\mathcal{T}_\alpha$ l'existence d'une face $\mathcal{A}_{\alpha\beta}$ telle que $\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta} > 0$. Si ce n'était pas le cas, il viendrait $\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta} = 0$ pour toutes les faces $\mathcal{A}_{\alpha\beta}$ de la cellule $\mathcal{T}_\alpha$, ce qui est impossible. Les éléments sur la diagonale de $\widetilde{J}$ satisfont par conséquent

$$z_\alpha \triangleq \widetilde{J}_{\alpha\alpha} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\gamma \left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\gamma}\right)_+ < 0\,. \tag{10.4.1}$$

D'après (10.4.1), z_α est réel et strictement négatif. Les éléments hors de la diagonale de $\widetilde{J}$ sont non négatifs

$$\widetilde{J}_{\alpha\beta} = -\frac{1}{|\mathcal{T}_\alpha|}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- \geq 0\,,\ \text{si } \beta\neq\alpha\,. \tag{10.4.2}$$

Pour cette raison, le rayon ρ_α du disque de Geršgorin $\Gamma_\alpha\left(\widetilde{J}\right)$ est donné par

$$\rho_\alpha \triangleq \sum_{\beta\neq\alpha}\left|\widetilde{J}_{\alpha\beta}\right| = \sum_{\beta\neq\alpha}\widetilde{J}_{\alpha\beta} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_-\,.$$

La relation géométrique (5.5.2) démontre que $\rho_\alpha = -z_\alpha$. Cela prouve que tous les disques de Geršgorin $\Gamma_\alpha\left(\widetilde{J}\right)$ sont inclus dans le demi-plan gauche fermé des nombres complexes

$$\Gamma_\alpha\left(\widetilde{J}\right) = \{z\in\mathbb{C}\,:\,|z-z_\alpha|\leq\rho_\alpha\}\subset\{z\in\mathbb{C}\,:\,\Re(z)\leq 0\}\ \text{ pour } 1\leq\alpha\leq N\,.$$

Le point (i) du théorème 10.3.3 permet de conclure que cela est également vrai pour le spectre $\mathrm{sp}\left(\widetilde{J}\right)$ de $\widetilde{J}$. □

Le théorème 10.4.1 démontre que les solutions du système (10.2.7) ne peuvent pas croître de façon exponentielle. Le théorème suivant donne un résultat complet pour la stabilité de (10.2.8) qui exclut également des solutions de croissance polynomiale. Un résultat de stabilité équivalent pour le schéma volumes finis discret en espace et en temps se trouve dans [**46**].

THÉORÈME 10.4.2 (Stabilité du schéma d'ordre un). *Soit $\widetilde{J}$ l'opérateur de discrétisation spatiale associé au schéma d'ordre un (10.2.7). Toutes les solutions du système (10.2.7) satisfont*

$$\sup_{t\geq 0}\|\mathfrak{u}(t)\| \leq \mathrm{C}\,\|\mathfrak{u}_0\|$$

où C est une constante donnée par

$$\mathrm{C} = \sqrt{\frac{\max_\alpha|\mathcal{T}_\alpha|}{\min_\alpha|\mathcal{T}_\alpha|}}\,.$$

Ce résultat est valable sur des maillages arbitraires et pour toutes les vitesses de convection $\boldsymbol{c}\in\mathbb{R}^d$.

DÉMONSTRATION. Il suffit de démontrer que l'opérateur (10.2.8) a la propriété (ii) du théorème 10.3.4. Quelques identités simples rendent cela possible. La première identité est l'antisymétrie des vecteurs surface

$$\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- = \left(-\mathbf{c}\cdot\boldsymbol{a}_{\beta\alpha}\right)_- = -\left(\mathbf{c}\cdot\boldsymbol{a}_{\beta\alpha}\right)_+\,. \tag{10.4.3}$$

La deuxième identité est (5.5.2)

$$\sum_\beta\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ + \sum_\beta\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- = 0\,. \tag{10.4.4}$$

Les équations (10.4.3) et (10.4.4) donnent l'identité

$$\sum_\alpha|u_\alpha|^2\sum_\beta\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ = \sum_\alpha|u_\alpha|^2(-1)\sum_\beta\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- = \sum_\alpha|u_\alpha|^2\sum_\beta\left(\mathbf{c}\cdot\boldsymbol{a}_{\beta\alpha}\right)_+ = \sum_\beta|u_\beta|^2\sum_\alpha\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ \tag{10.4.5}$$

qui se trouve également dans [**46**]. Soit $\mathfrak{u}$ une solution de l'équation semi-discrète

$$\frac{d\mathfrak{u}(t)}{dt} = \widetilde{J}\mathfrak{u}(t)\ ,\ \mathfrak{u}(0) = \mathfrak{u}_0\,.$$

On définit la matrice diagonale définie positive $\widehat{G}$ dont les coefficients sont

$$\widehat{g}_{\alpha\beta} = |\mathcal{T}_\alpha|\,\delta_{\alpha\beta}\,. \tag{10.4.6}$$

Les identités (10.4.3) et (10.4.5) permettent d'écrire

$$\begin{aligned}
\frac{d}{dt}\left(\mathfrak{u},\widehat{G}\mathfrak{u}\right) &= \left(\mathfrak{u},\left[\widehat{G}\widetilde{J}+\widetilde{J}^*\widehat{G}\right]\mathfrak{u}\right) = 2\Re\left\{\left(\mathfrak{u},\widehat{G}\widetilde{J}\mathfrak{u}\right)\right\} = \\
&= -2\Re\left\{\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ u_\alpha^* u_\alpha + \sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- u_\alpha^* u_\beta\right\} = \\
&= -2\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ |u_\alpha|^2 - \sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_-\left(u_\alpha^* u_\beta + u_\alpha u_\beta^*\right) = \\
&= -\sum_{\alpha,\beta}\left\{\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+\left(|u|_\alpha^2+|u|_\beta^2\right) - \left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+\left(u_\alpha^* u_\beta + u_\alpha u_\beta^*\right)\right\} = \\
&= -\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ |u_\alpha - u_\beta|^2 = -\frac{1}{2}\sum_{\alpha,\beta}|\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}|\,|u_\alpha-u_\beta|^2 \le 0.
\end{aligned} \tag{10.4.7}$$

Par conséquent, $\left(\mathfrak{u}(t),\widehat{G}\mathfrak{u}(t)\right) \le \left(\mathfrak{u}(0),\widehat{G}\mathfrak{u}(0)\right)$ pour tous les $t \ge 0$. La définition de $\widehat{G}$ par (10.4.6) conduit aux bornes

$$\frac{\left(\mathfrak{u},\widehat{G}\mathfrak{u}\right)}{\max_\alpha|\mathcal{T}_\alpha|} \le (\mathfrak{u},\mathfrak{u}) \le \frac{\left(\mathfrak{u},\widehat{G}\mathfrak{u}\right)}{\min_\alpha|\mathcal{T}_\alpha|} \text{ pour tout } \mathfrak{u}\in\mathbb{C}^N.$$

Cela implique que

$$\sqrt{(\mathfrak{u}(t),\mathfrak{u}(t))} \le \sqrt{\frac{\max_\alpha|\mathcal{T}_\alpha|}{\min_\alpha|\mathcal{T}_\alpha|}}\sqrt{(\mathfrak{u}(0),\mathfrak{u}(0))} \text{ pour tout } t\ge 0.$$

□

10.5. Analyse de la stabilité du schéma d'ordre deux : le cadre général

Contrairement à l'opérateur d'ordre un (10.2.8), l'opérateur MUSCL (10.2.14) dépend de la reconstruction du gradient (10.2.9)

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\overline{u}_\beta$$

dans toutes les cellules $\mathcal{T}_\alpha$.

La question de la stabilité asymptotique de l'opérateur MUSCL (10.2.14) sur des maillages arbitraires peut se formuler de la façon suivante :

PROBLÈME 10.5.1. Pour un maillage non structuré donné, existe-t-il *une reconstruction consistante* du gradient au sens de la définition 7.4.1 telle que l'opérateur J soit stable dans le sens de la définition 10.3.2 ? Cette question s'exprime de façon explicite à l'aide du théorème 10.3.4 : existe-t-il des paramètres consistants de reconstruction $\boldsymbol{\sigma}_{\alpha\beta}$ tels que pour toute vitesse de convection $\boldsymbol{c}\in\mathbb{R}^d$ il existe une matrice définie positive G telle que la matrice $GJ+J^*G$ soit semi-définie négative ?

Dans la formulation du problème 10.5.1, on exige que la reconstruction garantisse la stabilité asymptotique indépendamment de la vitesse de convection. Cela est nécessaire pour les applications à la dynamique des gaz où la vitesse de convection n'est pas constante.

Il semble trop difficile de trouver la réponse au problème 10.5.1 en toute généralité. Une idée simple est de vérifier si la matrice diagonale $\widehat{G}$ définie par (10.4.6) donne un résultat général de stabilité dans le cas de l'opérateur MUSCL (10.2.14). Il s'avère que ce n'est pas le cas. Les expériences numériques montrent d'une part que $\widehat{G}J + J^*\widehat{G}$ n'est en général pas semi-définie négative sur des maillages non structurés, même si l'opérateur MUSCL (10.2.14) est stable.

D'autre part, il existe des maillages où certaines méthodes de reconstruction du gradient, par exemple la méthode des moindres carrés, rendent l'opérateur MUSCL (10.2.14) instable.

Malgré cela, il est intéressant et important de disposer d'une stratégie pour concevoir des schémas MUSCL semi-discrets avec les meilleures propriétés de stabilité possibles. Considérons la dérivée en temps de la forme quadratique $\left(\mathfrak{u}, \widehat{G}\mathfrak{u}\right)$ donnée par

$$\frac{d}{dt}\left(\mathfrak{u}, \widehat{G}\mathfrak{u}\right) = \left(\mathfrak{u}, \left[\widehat{G}J + J^*\widehat{G}\right]\mathfrak{u}\right) \tag{10.5.1}$$

où la matrice diagonale $\widehat{G}$ est définie par (10.4.6). L'objectif est de chercher des paramètres de reconstruction qui rendent le membre de droite de (10.5.1) aussi petit que possible. Cette approche ne donne pas de critère général pour la stabilité asymptotique de l'opérateur (10.2.14) mais elle permet d'identifier les méthodes de reconstruction qui minimisent la croissance de $\left(\mathfrak{u}, \widehat{G}\mathfrak{u}\right)$.

Avant de poursuivre dans cette voie, il faut réécrire l'expression (10.5.1). On a la

PROPOSITION 10.5.2 (Énergie du schéma MUSCL semi-discret). *Soit* $\mathfrak{u} : [0,T] \to \mathbb{C}^N$ *solution du schéma semi-discret (10.2.13). Dans ce cas, la fonction* $\mathfrak{u}$ *satisfait*

$$\frac{d}{dt}\left(\mathfrak{u}, \widehat{G}\mathfrak{u}\right) = \sum_{\alpha,\beta} \left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ \left(-\left|u_\beta - u_\alpha\right|^2 + 2\sum_\gamma \Re\left\{\left(u_\beta^* - u_\alpha^*\right)\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{\sigma}_{\alpha\gamma}\left(u_\gamma - u_\alpha\right)\right\}\right) \tag{10.5.2}$$

où $\widehat{G}$ *est la matrice diagonale définie par les éléments (10.4.6),* $g_{\alpha\beta} = |\mathcal{T}_\alpha|\,\delta_{\alpha\beta}$.

DÉMONSTRATION. Pour simplifier la forme de l'équation différentielle (10.2.13), on introduit les définitions

$$\left.\begin{aligned} u_{\alpha\beta} &\triangleq u_\alpha + \boldsymbol{k}_{\alpha\beta}\cdot\sum_\gamma \boldsymbol{\sigma}_{\alpha\gamma}\left(u_\gamma - u_\alpha\right) \\ u_{\beta\alpha} &\triangleq u_\beta + \boldsymbol{k}_{\beta\alpha}\cdot\sum_\gamma \boldsymbol{\sigma}_{\beta\gamma}\left(u_\gamma - u_\beta\right) \end{aligned}\right\} \tag{10.5.3}$$

pour les valeurs reconstruites au barycentre de la face $\mathcal{A}_{\alpha\beta}$ du côté de la cellule $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$. En général, on a $u_{\alpha\beta} \neq u_{\beta\alpha}$. L'équation différentielle (10.2.13) prend alors la forme simple

$$\frac{du_\alpha}{dt} = \sum_\beta J_{\alpha\beta}u_\beta = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \left\{\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ u_{\alpha\beta} + \left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- u_{\beta\alpha}\right\},\ 1 \le \alpha \le N\,. \tag{10.5.4}$$

La multiplication de (10.5.4) par $u_\alpha^*\,|\mathcal{T}_\alpha|$, où u_α^* est le conjugué complexe de u_α, et la sommation sur l'indice α permettent de démontrer l'identité

$$\begin{aligned} \frac{d}{dt}\left(\mathfrak{u}, \widehat{G}\mathfrak{u}\right) = \left(\mathfrak{u}, \left[\widehat{G}J + J^*\widehat{G}\right]\mathfrak{u}\right) &= 2\Re\left\{\left(\mathfrak{u}, \widehat{G}J\mathfrak{u}\right)\right\} = \\ &= -2\sum_{\alpha,\beta}\Re\left\{u_\alpha^*\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ u_{\alpha\beta} + u_\alpha^*\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- u_{\beta\alpha}\right\}. \end{aligned} \tag{10.5.5}$$

Dans le deuxième terme du membre de droite de (10.5.5), on remplace α par β et vice versa, et on utilise l'identité (10.4.3)

$$\left(\mathbf{c}\cdot\boldsymbol{a}_{\beta\alpha}\right)_- = \left(-\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_- = -\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+ .$$

Cela démontre que

$$\frac{d}{dt}\left(\mathfrak{u}, \widehat{G}\mathfrak{u}\right) = -2\sum_{\alpha,\beta}\Re\left\{\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_+\left(u_\alpha^* - u_\beta^*\right)u_{\alpha\beta}\right\}. \tag{10.5.6}$$

L'insertion des définitions (10.5.3) dans l'identité (10.5.6) donne

$$\frac{d}{dt}\left(\mathfrak{u},\widehat{G}\mathfrak{u}\right)=\left(\mathfrak{u},\left[\widehat{G}J+J^{*}\widehat{G}\right]\mathfrak{u}\right)= \\ =-2\sum_{\alpha,\beta}\Re\left\{\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{+}\left(u_{\alpha}^{*}-u_{\beta}^{*}\right)\left(u_{\alpha}+\boldsymbol{k}_{\alpha\beta}\cdot\sum_{\gamma}\boldsymbol{\sigma}_{\alpha\gamma}\left(u_{\gamma}-u_{\alpha}\right)\right)\right\}. \quad (10.5.7)$$

L'identité (10.4.5) permet de prouver

$$-2\sum_{\alpha,\beta}\Re\left\{\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{+}\left(u_{\alpha}^{*}-u_{\beta}^{*}\right)u_{\alpha}\right\}=-2\sum_{\alpha,\beta}\Re\left\{\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{+}\left(|u_{\alpha}|^{2}-u_{\beta}^{*}u_{\alpha}\right)\right\}= \\ =-\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{+}\left(|u_{\alpha}|^{2}+|u_{\beta}|^{2}-u_{\beta}^{*}u_{\alpha}-u_{\beta}u_{\alpha}^{*}\right)=-\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{+}|u_{\beta}-u_{\alpha}|^{2}. \quad (10.5.8)$$

Finalement, l'utilisation de (10.5.8) dans (10.5.7) donne le résultat (10.5.2)

$$\frac{d}{dt}\left(\mathfrak{u},\widehat{G}\mathfrak{u}\right)=\left(\mathfrak{u},\left[\widehat{G}J+J^{*}\widehat{G}\right]\mathfrak{u}\right)= \\ =\sum_{\alpha,\beta}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{+}\left[-|u_{\beta}-u_{\alpha}|^{2}+2\sum_{\gamma}\Re\left\{\left(u_{\beta}^{*}-u_{\alpha}^{*}\right)\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{\sigma}_{\alpha\gamma}\left(u_{\gamma}-u_{\alpha}\right)\right\}\right].$$

□

La forme de l'expression quadratique (10.5.2) justifie l'introduction de la

Définition 10.5.3 (Opérateur de reconstruction locale). Soit m_α le nombre de cellules dans le voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$ et soit l_α le nombre de premiers voisins $\mathcal{T}_\alpha$, c'est-à-dire le nombre de ses faces. *L'opérateur de reconstruction locale* de la cellule $\mathcal{T}_\alpha$ est la matrice R_α de l_α lignes et m_α colonnes dont les éléments sont donnés par

$$r_{\beta\gamma}^{(\alpha)}\triangleq\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{\sigma}_{\alpha\gamma}. \quad (10.5.9)$$

Remarque 10.5.4. L'opérateur de reconstruction locale détermine comment les fluctuations $u_\gamma-u_\alpha$ engendrent les fluctuations $u_{\alpha\beta}-u_\alpha$ où $u_{\alpha\beta}$ est la valeur reconstruite sur l'interface entre les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$, voir la figure 10.5.1. En notation matricielle, R_α peut être écrite sous la forme $R_\alpha=K_\alpha S_\alpha$ où K_α est la matrice de dimensions $l_\alpha\times d$ dont les colonnes sont les vecteurs $\boldsymbol{k}_{\alpha\beta}$

$$K_{\alpha}^{t}\quad\triangleq\quad\left[\boldsymbol{k}_{\alpha\beta_1},\boldsymbol{k}_{\alpha\beta_2},\ldots,\boldsymbol{k}_{\alpha\beta_{l_\alpha}}\right] \quad (10.5.10)$$

et $\mathbb{V}_\alpha\triangleq\{\beta_1,\ldots,\beta_{l_\alpha}\}$ est le premier voisinage de la cellule $\mathcal{T}_\alpha$. □

Remarque 10.5.5. Il est important de ne pas confondre l'opérateur de reconstruction locale $R_\alpha=K_\alpha S_\alpha$ avec la matrice de reconstruction du gradient S_α. La matrice R_α a l'importante propriété d'être une grandeur sans dimension et invariante par des transformations d'échelle du maillage. Elle décrit par conséquent la géométrie locale de la reconstruction dans la cellule $\mathcal{T}_\alpha$. □

Introduisons le vecteur $\delta\mathfrak{u}$ de composantes $\delta u_{\alpha\beta}=u_\beta-u_\alpha$. La définition 10.5.3 permet alors d'énoncer le

Corollaire 10.5.6 (Décomposition de l'énergie du schéma Muscl semi-discret). *L'expression (10.5.2) peut être écrite sous la forme d'une somme*

$$\frac{d}{dt}\left(\mathfrak{u},\widehat{G}\mathfrak{u}\right)=\sum_{\alpha=1}^{N}\left(\Theta_\alpha\left(\delta\mathfrak{u}\right)+\Phi_\alpha\left(\delta\mathfrak{u}\right)\right) \quad (10.5.11)$$

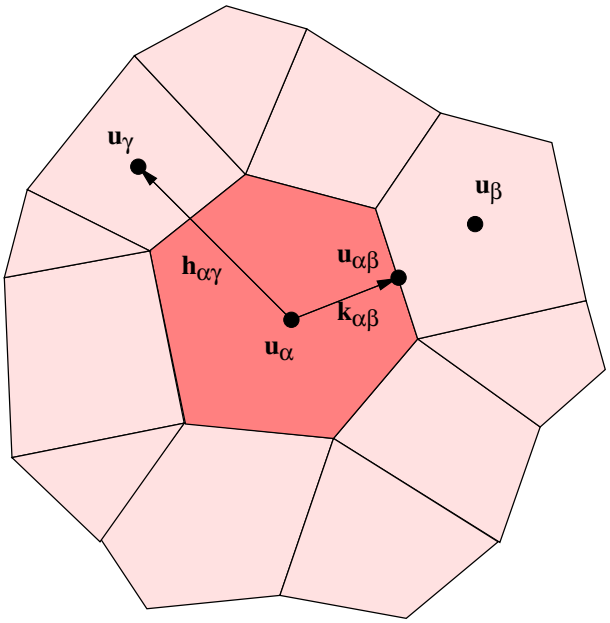


FIG. 10.5.1: Les fluctuations $u_{\alpha\beta} - u_\alpha$ dépendent linéairement des fluctuations $u_\gamma - u_\alpha$ via l'opérateur de reconstruction locale.

où les $\Theta_\alpha(\delta\mathfrak{u})$ *et* $\Phi_\alpha(\delta\mathfrak{u})$ *sont donnés par*

$$\Theta_\alpha(\delta\mathfrak{u}) \triangleq -\sum_\beta (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ |\delta u_{\alpha\beta}|^2 \tag{10.5.12}$$

$$\Phi_\alpha(\delta\mathfrak{u}) \triangleq 2\sum_\beta\sum_\gamma (\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \Re\left\{\delta u^*_{\alpha\beta} r^{(\alpha)}_{\beta\gamma}\delta u_{\alpha\gamma}\right\}. \tag{10.5.13}$$

Les termes dans la somme (10.5.11) satisfont les points suivants :

(i) *Les termes* $\Theta_\alpha(\delta\mathfrak{u})$ *sont indépendants des paramètres de reconstruction* $\boldsymbol{\sigma}_{\alpha\beta}$ *et satisfont* $\Theta_\alpha(\delta\mathfrak{u}) \le 0$ *pour tout* $\delta\mathfrak{u}$.

(ii) *Le terme* $\Phi_\alpha(\delta\mathfrak{u})$ *associé à la cellule* $\mathcal{T}_\alpha$ *dépend linéairement de la matrice de reconstruction locale* $R_\alpha = K_\alpha S_\alpha$ *dans la cellule* $\mathcal{T}_\alpha$.

(iii) *Les paramètres de reconstruction* $\boldsymbol{\sigma}_{\alpha\beta}$ *dans la cellule* $\mathcal{T}_\alpha$ *apparaissent exclusivement dans le terme* $\Phi_\alpha(\delta\mathfrak{u})$.

DÉMONSTRATION. Il suffit d'insérer la définition 10.5.3 dans le résultat de la proposition 10.5.2 et de conclure. □

L'équation (10.5.11) montre que la croissance de la norme discrète donnée par $\left(\mathfrak{u}, \widehat{G}\mathfrak{u}\right)$ dépend de façon affine des éléments $r^{(\alpha)}_{\beta\gamma}$ de la matrice de reconstruction locale R_α dans chaque cellule. Il est important de noter que la matrice de reconstruction R_α de la cellule $\mathcal{T}_\alpha$, et avec elle les paramètres de reconstruction $\boldsymbol{\sigma}_{\alpha\gamma}$ dans la cellule $\mathcal{T}_\alpha$, apparaissent uniquement dans le terme Φ_α dans (10.5.11). De cette façon, les paramètres de reconstruction qui minimisent (10.5.11) peuvent être choisis par une minimisation des Φ_α cellule par cellule.

10.6. Propriété minimisante de la méthode des moindres carrés

Le corollaire 10.5.6 permet de définir un critère pour déterminer les reconstructions consistantes S_α les plus susceptibles d'assurer la stabilité de l'opérateur MUSCL semi-discret (10.2.14). Ce critère est basé sur la décomposition (10.5.11) de l'énergie associée à l'opérateur (10.2.14).

D'après le corollaire 10.5.6, cette décomposition est donnée par

$$\frac{d}{dt}\left(\mathfrak{u},\widehat{G}\mathfrak{u}\right)=\sum_{\alpha=1}^{N}\left(\Theta_{\alpha}\left(\delta\mathfrak{u}\right)+\Phi_{\alpha}\left(\delta\mathfrak{u}\right)\right).$$

Le corollaire 10.5.6 montre que le terme $\Theta_{\alpha}\left(\delta\mathfrak{u}\right)$ est indépendant de la reconstruction et toujours non positif. Pour cette raison, il est possible de se concentrer uniquement sur le deuxième terme $\Phi_{\alpha}\left(\delta\mathfrak{u}\right)$ de (10.5.11) qui peut être positif et provoquer une croissance de l'énergie (10.5.11).

Le terme $\Phi_{\alpha}\left(\delta\mathfrak{u}\right)$ dépend linéairement de la matrice de reconstruction locale $R_{\alpha}=K_{\alpha}S_{\alpha}$. La matrice S_{α} satisfait la condition de consistance (7.6.3)

$$S_{\alpha}H_{\alpha}=\mathrm{I}_{d}\,.$$

D'après le théorème 7.8.1, la matrice R_{α} s'écrit dans la forme la plus générale

$$R_{\alpha}S_{\alpha}=K_{\alpha}\left(\widetilde{S}_{\alpha}+\Lambda_{\alpha}B_{\alpha}\right) \tag{10.6.1}$$

où $\widetilde{S}_{\alpha}$ est une solution particulière de (7.8.4), par exemple la reconstruction des moindres carrés, et Λ_{α} est une matrice arbitraire. La forme (10.6.1) de R_{α} montre que l'expression (10.5.13)

$$\Phi_{\alpha}\left(\delta\mathfrak{u}\right)=2\sum_{\beta}\sum_{\gamma}\left(\mathbf{c}\cdot\boldsymbol{a}_{\alpha\beta}\right)_{+}\Re\left\{\delta u_{\alpha\beta}^{*}r_{\beta\gamma}^{(\alpha)}\delta u_{\alpha\gamma}\right\}$$

dépend de façon affine de la matrice Λ_{α}.

L'objectif est de rendre la croissance de l'énergie (10.5.11) aussi petite que possible. Dans ce contexte, l'équation (10.6.1) suggère de choisir dans chaque cellule $\mathcal{T}_{\alpha}$ la matrice Λ_{α} de telle façon que le terme $\Phi_{\alpha}\left(\delta\mathfrak{u}\right)$ défini par (10.5.13) devienne aussi petit que possible.

Néanmoins, une minimisation directe de (10.5.11) pour des valeurs fixes de la solution $\mathfrak{u}$ et de la vitesse $\boldsymbol{c}\in\mathbb{R}^{d}$ rencontre deux problèmes.

(1) Premièrement, il n'est pas évident que l'infimum de $\Phi_{\alpha}\left(\delta\mathfrak{u}\right)$ par rapport à Λ_{α} existe car il est possible que la fonction $\Lambda_{\alpha}\mapsto\Phi_{\alpha}\left(\delta\mathfrak{u}\right)$ ne soit pas bornée inférieurement.

(2) Deuxièmement, si une matrice Λ_{α} existe qui minimise $\Phi_{\alpha}\left(\delta\mathfrak{u}\right)$ pour des valeurs fixes de la solution $\mathfrak{u}$ et de la vitesse $\boldsymbol{c}\in\mathbb{R}^{d}$, alors Λ_{α} dépend elle-même de $\mathfrak{u}$ et de $\boldsymbol{c}$. Ces dépendances ne sont pas souhaitables car le but final est l'application des résultats à la dynamique des gaz où la vitesse de convection n'est pas fixe.

Pour surmonter ces obstacles, il est nécessaire de proposer un critère alternatif qui mesure *de façon qualitative et approximative* l'influence de la reconstruction du gradient sur la croissance de l'énergie $\left(\mathfrak{u},\widehat{G}\mathfrak{u}\right)$. Ce critère s'appuie sur des normes matricielles de l'opérateur de reconstruction locale introduit par la définition 10.5.3. Dans la suite, une *norme matricielle* désigne toute norme vectorielle sur l'espace des matrices $\mathbb{M}_{n,k}\left(\mathbb{C}\right)$ vu comme un espace linéaire. Les normes matricielles utilisées ci-dessous ne satisfont donc pas nécessairement l'inégalité $\left\|AB\right\|\leq\left\|A\right\|\left\|B\right\|$.

Le critère est donné par la

Définition 10.6.1 (Reconstructions consistantes optimales). Soit $\left\|.\right\|_{M}$ une norme matricielle sur l'espace des matrices $\mathbb{M}_{n,k}\left(\mathbb{C}\right)$. Une reconstruction consistante du gradient donnée par une matrice $\breve{S}_{\alpha}$ s'appelle *optimale par rapport à la norme* $\left\|.\right\|_{M}$ si $\breve{S}_{\alpha}$ est solution du problème de minimisation

$$\left\|K_{\alpha}\breve{S}_{\alpha}\right\|_{M}=\min\left\{\left\|K_{\alpha}S_{\alpha}\right\|_{M}\;\middle|\;S_{\alpha}H_{\alpha}=\mathrm{I}_{d}\,,\;S_{\alpha}\in\mathbb{M}_{d,m}\left(\mathbb{C}\right)\right\}. \tag{10.6.2}$$

Pour toute norme matricielle $\left\|.\right\|_{M}$, le problème de minimisation (10.6.2) de la définition 10.6.1 est un problème d'optimisation convexe. En effet, la fonction

$$\Lambda_{\alpha}\longmapsto\left\|K_{\alpha}\left(\widetilde{S}_{\alpha}+\Lambda_{\alpha}B_{\alpha}\right)\right\|_{M}$$

est une fonction convexe bornée inférieurement. D'ailleurs, toute solution $\breve{S}_{\alpha}$ de (10.6.2) est indépendante de la solution $\mathfrak{u}$ et de la vitesse $\boldsymbol{c}\in\mathbb{R}^{d}$. Cette approche semble donc plus facile à généraliser aux applications de la dynamique des gaz où la vitesse n'est pas constante. L'importance de ce critère est soulignée par la

REMARQUE 10.6.2. Dans une cellule $\mathcal{T}_\alpha$, considérons le critère de la définition 10.6.1 pour la norme spectrale $\|.\|_2$. Notons $\mathbb{W}_\alpha$ l'ensemble des indices des cellules dans le voisinage de reconstruction de $\mathcal{T}_\alpha$. Soit $\mathbb{V}_\alpha$ l'ensemble des indices des premiers voisins de $\mathcal{T}_\alpha$ et $u_{\alpha\beta}$ la valeur reconstruite à la face $\mathcal{A}_{\alpha\beta}$ entre les cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$, cf. la figure 10.5.1. Si $\|K_\alpha S_\alpha\|_2 > 1$, il existe un $\mathfrak{u} = (u_1, \ldots, u_N)$ tel que

$$\sum_{\beta\in\mathbb{V}_\alpha} |u_{\alpha\beta} - u_\alpha|^2 > \sum_{\gamma\in\mathbb{W}_\alpha} |u_\gamma - u_\alpha|^2 .$$

Cela signifie que les fluctuations entre les valeurs aux faces et la valeur au barycentre de $\mathcal{T}_\alpha$ deviennent plus grandes en moyenne quadratique que celles entre les barycentres de $\mathcal{T}_\alpha$ et les cellules voisines $\mathcal{T}_\gamma$. Cette situation n'est pas souhaitable car elle risque de provoquer des fluctuations qui se renforcent elles-mêmes. Ce comportement a d'ailleurs été observé pour la reconstruction au premier voisinage sur des maillages de tétraèdres, cf. la section 10.9, et semble corrélé avec l'apparition des instabilités. Il faut donc tenter de rendre $\|K_\alpha S_\alpha\|_2$ aussi petit que possible. □

Le nouveau critère a toutefois le problème que toute solution Λ_α du problème de minimisation (10.6.2) dépend du choix de la norme $\|.\|_M$. Pour cette raison, le critère de la définition 10.6.1 semble uniquement pertinent s'il existe une solution commune du problème (10.6.2) pour au moins une famille de normes.

Une famille importante de normes est introduite par la

DÉFINITION 10.6.3 (Normes unitairement invariantes). Une norme matricielle $\|.\|_M : \mathbb{M}_{m,n}(\mathbb{C}) \to \mathbb{R}$ s'appelle *unitairement invariante* si $\|UAV\|_M = \|A\|_M$ pour toute matrice $A \in \mathbb{M}_{m,n}(\mathbb{C})$ et toutes les matrices unitaires $U \in \mathbb{M}_m(\mathbb{C})$ et $V \in \mathbb{M}_n(\mathbb{C})$.

La famille des normes unitairement invariantes inclut les normes suivantes :
– La norme de Frobenius, définie par (5.2.5)

$$\|A\|_F = \sqrt{\operatorname{trace}(A^*A)} .$$

– La norme spectrale, définie par (5.2.4)

$$\|A\|_2 = \sup_{\|\boldsymbol{x}\|_2} \|A\boldsymbol{x}\|_2 .$$

– La norme trace

$$\|A\|_{tr} = \operatorname{trace}\left(\sqrt{A^*A}\right) .$$

Un exemple d'une norme qui n'est pas unitairement invariante est donné par

$$\|A\|_{\mathcal{L}(2,\infty)} \triangleq \sup_{1\le\alpha\le l} \sup_{\|\mathfrak{z}\|=1} \left| \sum_{1\le\beta\le m} a_{\alpha\beta} z_\beta \right| = \sup_{1\le\alpha\le l} \sqrt{\sum_{1\le\beta\le m} |a_{\alpha\beta}|^2} . \tag{10.6.3}$$

Avant d'énoncer le résultat, il est utile de rappeler la décomposition en valeurs singulières, cf. [**66**, th. 3.1.1, p. 144].

THÉORÈME 10.6.4 (Décomposition en valeurs singulières). *Soit $A \in \mathbb{M}_{m,n}(\mathbb{C})$ et soit $q = \min(m,n)$. Il existe une matrice réelle $\Sigma \in \mathbb{M}_{m,n}(\mathbb{R})$ telle que $\Sigma_{ij} = 0$ si $i \neq j$ et $\Sigma_{11} \geq \Sigma_{22} \geq \cdots \geq \Sigma_{qq} \geq 0$ et il existe deux matrices unitaires V et W telles que $A = V\Sigma W^*$. Si $A \in \mathbb{M}_{m,n}(\mathbb{R})$, U et V peuvent être choisies réelles orthogonales. Les nombres $\varsigma_i \triangleq \Sigma_{ii}$, pour $1 \leq i \leq q$, s'appellent les valeurs singulières de A.*

REMARQUE 10.6.5. Les valeurs singulières sont habituellement notées σ_i. Dans ce chapitre, on utilise le symbole ς_i afin d'éviter la confusion avec les composantes $\sigma_{\alpha,i}$ du gradient de cellule.□

Les démonstrations de cette section nécessitent la

PROPOSITION 10.6.6. *Soit $A \in \mathbb{M}_{m,n}(\mathbb{C})$ et soit $q = \min(m,n)$. Notons $\lambda_1 \geq \cdots \geq \lambda_q \geq 0$ les q plus grandes valeurs propres de AA^* et A^*A. Alors la valeur singulière ς_i de A satisfait $\varsigma_i = \sqrt{\lambda_i}$.*

DÉMONSTRATION. Voir [**66**, p. 146]. □

Ces résultats permettent de prouver le

THÉORÈME 10.6.7 (Propriété minimisante de la méthode des moindres carrés). *Soit $R_\alpha = K_\alpha S_\alpha$ l'opérateur de la reconstruction locale dans la cellule $\mathcal{T}_\alpha$. Soit S_α une matrice de reconstruction du gradient qui satisfait la condition de consistance (7.8.4) donnée par $S_\alpha H_\alpha = \mathrm{I}_d$ et soit $\widetilde{S}_\alpha = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t$ la matrice de reconstruction des moindres carrés. Alors $\widetilde{S}_\alpha$ satisfait les points suivants :*

(i) *La matrice $\widetilde{S}_\alpha$ minimise chaque valeur singulière de $K_\alpha S_\alpha$:*

$$\varsigma_i\left(K_\alpha \widetilde{S}_\alpha\right) \leq \varsigma_i\left(K_\alpha S_\alpha\right), 1 \leq i \leq d\,.$$

(ii) *Si $\|.\|_M$ est une norme matricielle unitairement invariante, alors*

$$\left\|K_\alpha \widetilde{S}_\alpha\right\|_M \leq \left\|K_\alpha S_\alpha\right\|_M .$$

En particulier, $\widetilde{S}_\alpha$ minimise la norme spectrale, la norme de Frobenius et la norme trace de la matrice $K_\alpha S_\alpha$ parmi les matrices vérifiant $S_\alpha H_\alpha = \mathrm{I}_d$.

(iii) *Soit $\|.\|_M$ une norme qui s'exprime sous la forme $\|A\|_M = F\left(AA^*\right)$ où F est une fonction de matrices hermitiennes telle que $F(P) \leq F(P+Q)$ pour toute matrice hermitienne P et toute matrice semi-définie positive Q. Alors $\widetilde{S}_\alpha$ satisfait*

$$\left\|K_\alpha \widetilde{S}_\alpha\right\|_M \leq \left\|K_\alpha S_\alpha\right\|_M .$$

En particulier, $\widetilde{S}_\alpha$ minimise la norme (10.6.3) de $K_\alpha S_\alpha$.

DÉMONSTRATION. Le théorème 7.8.1 montre qu'une reconstruction consistante s'écrit comme $S_\alpha = \widetilde{S}_\alpha + \Lambda_\alpha B_\alpha$ où $\widetilde{S}_\alpha$ est la matrice de la reconstruction par la méthode des moindres carrés, la matrice B_α est une solution de rang maximal de $B_\alpha H_\alpha = 0$ et la matrice Λ_α représente les degrés de liberté de la reconstruction consistante. Les valeurs singulières de $K_\alpha S_\alpha$ sont les racines carrées des valeurs propres de la matrice

$$K_\alpha S_\alpha S_\alpha^t K_\alpha^t = \left(K_\alpha \widetilde{S}_\alpha + K_\alpha \Lambda_\alpha B_\alpha\right)\left(\widetilde{S}_\alpha^t K_\alpha^t + B_\alpha^t \Lambda_\alpha^t K_\alpha^t\right). \tag{10.6.4}$$

La matrice de la reconstruction par la méthode des moindres carrés est donnée par $\widetilde{S}_\alpha = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t$ et satisfait par conséquent $\widetilde{S}_\alpha B_\alpha^t = 0$. La matrice (10.6.4) est donc une somme de deux matrices semi-définies positives

$$K_\alpha S_\alpha S_\alpha^t K_\alpha^t = K_\alpha \widetilde{S}_\alpha \widetilde{S}_\alpha^t K_\alpha^t + K_\alpha \Lambda_\alpha B_\alpha B_\alpha^t \Lambda_\alpha^t K_\alpha^t . \tag{10.6.5}$$

La démonstration du point (i) fait usage de [**65**, cor. 4.4.3, p. 182] : Soient P et Q deux matrices hermitiennes et soit Q semi-définie positive. Si les valeurs propres de P et de $P+Q$ sont rangées par ordre croissant et si $\lambda_k(P)$ et $\lambda_k(P+Q)$ notent respectivement la k-ième valeur propre de P et de $P+Q$, ces valeurs propres satisfont $\lambda_k(P) \leq \lambda_k(P+Q)$. Cela prouve que les k-ièmes valeurs propres des matrices dans (10.6.5) satisfont

$$\lambda_k\left(K_\alpha \widetilde{S}_\alpha \widetilde{S}_\alpha^t K_\alpha^t\right) \leq \lambda_k\left(K_\alpha \widetilde{S}_\alpha \widetilde{S}_\alpha^t K_\alpha^t + K_\alpha \Lambda_\alpha B_\alpha B_\alpha^t \Lambda_\alpha^t K_\alpha^t\right). \tag{10.6.6}$$

Le vecteur des valeurs singulières de $K_\alpha S_\alpha$ est le vecteur des racines carrées des valeurs propres de $K_\alpha S_\alpha S_\alpha^t K_\alpha^t$ arrangées en ordre croissant, cf. la proposition 10.6.6. Par conséquent, (10.6.6) montre que la k-ième valeur singulière de $K_\alpha S_\alpha$ a un minimum en $\Lambda_\alpha = 0$, c'est-à-dire en la matrice $S_\alpha = \widetilde{S}_\alpha$. Cela démontre le premier point du théorème 10.6.7.

La démonstration du point (ii) utilise [**66**, déf. 3.5.17, p. 209, et th. 3.5.18, p. 210]. Pour toute norme matricielle qui est unitairement invariante, et en particulier pour les normes citées ci-dessus, il existe une fonction de jauge symétrique telle que la norme s'exprime comme la fonction de jauge du vecteur de valeurs singulières. Une fonction de jauge est également une norme vectorielle monotone. Comme la reconstruction des moindres carrés minimise chaque élément du vecteur des valeurs singulières de $K_\alpha S_\alpha$ parmi les reconstructions consistantes, elle minimise aussi les fonctions de jauge des valeurs singulières et par conséquent la norme matricielle.

La démonstration de (iii) procède de la même manière. La norme (10.6.3) s'écrit sous la forme

$$\|K_\alpha S_\alpha\|_{\mathcal{L}(2,\infty)} = \sup_{\beta\in\mathbb{V}_\alpha} \sup_{\|\mathfrak{z}\|_2=1} \left|\sum_\gamma \boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{\sigma}_{\alpha\gamma} z_\gamma\right| =$$

$$= \sup_{\beta\in\mathbb{V}_\alpha} \sqrt{\sum_{1\le\gamma\le m_\alpha} \left(K_\alpha\widetilde{S}_\alpha + K_\alpha\Lambda_\alpha B_\alpha\right)_{\beta\gamma} \left(\widetilde{S}_\alpha^t K_\alpha^t + B_\alpha^t\Lambda_\alpha^t K_\alpha^t\right)_{\gamma\beta}}. \quad (10.6.7)$$

Comme démontré dans le point (i), l'expression (10.6.7) prend la forme

$$\|K_\alpha S_\alpha\|_{\mathcal{L}(2,\infty)} = \sup_{\beta\in\mathbb{V}_\alpha} \sqrt{\left(K_\alpha\widetilde{S}_\alpha\widetilde{S}_\alpha^t K_\alpha^t\right)_{\beta\beta} + \left(K_\alpha\Lambda_\alpha B_\alpha B_\alpha^t\Lambda_\alpha^t K_\alpha^t\right)_{\beta\beta}}$$

qui a un minimum en $\Lambda_\alpha = 0$, c'est-à-dire en la reconstruction des moindres carrés. Cela prouve la propriété minimisante pour la norme (10.6.3). Considérons une norme matricielle de la forme $\|A\|_M = F(AA^*)$ où F est une fonction ayant la propriété énoncée dans le point (iii). L'argument utilisé ci-dessus montre que

$$\left\|K_\alpha\widetilde{S}_\alpha\right\|_M = F\left(K_\alpha\widetilde{S}_\alpha\widetilde{S}_\alpha^t K_\alpha^t\right) \le F\left(K_\alpha\widetilde{S}_\alpha\widetilde{S}_\alpha^t K_\alpha^t + K_\alpha\Lambda_\alpha B_\alpha B_\alpha^t\Lambda_\alpha^t K_\alpha^t\right)$$

ce qui prouve que la norme matricielle en question admet un minimum à $S_\alpha = \widetilde{S}_\alpha$. □

Le théorème 10.6.7 montre que la reconstruction des moindres carrés minimise la norme de l'opérateur de reconstruction locale simultanément pour toute une famille de normes matricielles. Pour ces normes, la reconstruction des moindres carrés s'avère optimale dans le sens de la définition 10.6.1.

Remarque 10.6.8. Le théorème 7.7.2 de la section 7.7 devient simplement un corollaire du théorème 10.6.7. Pour le prouver, il suffit de remplacer la matrice K_α par la matrice identité et d'appliquer le théorème 10.6.7 au cas particulier de la norme de Frobenius.□

L'étape suivante consiste à analyser l'influence de la taille du voisinage de reconstruction sur l'opérateur de reconstruction locale et sur la stabilité de l'opérateur Muscl (10.2.14). Pour la reconstruction des moindres carrés, le théorème suivant montre l'influence d'un élargissement du voisinage sur les normes de l'opérateur de reconstruction locale.

Théorème 10.6.9 (Influence de la taille du voisinage sur la reconstruction des moindres carrés). *Considérons un voisinage de reconstruction fixe $\mathbb{W}_\alpha$ dans la cellule $\mathcal{T}_\alpha$. Pour ce $\mathbb{W}_\alpha$, soit H_α la matrice géométrique introduite par la définition 7.6.3 et soit $\widetilde{S}_\alpha = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t$ la matrice de la méthode des moindres carrés. On élargit le voisinage de reconstruction $\mathbb{W}_\alpha$ de la cellule $\mathcal{T}_\alpha$ en y ajoutant un nombre de $l \ge 1$ cellules avec les indices $\overline{\mathbb{W}}_\alpha = \{\beta_1,\dots,\beta_l\}$. Soit $\overline{H}_\alpha \in \mathbb{M}_{l,d}(\mathbb{R})$ la matrice dont les lignes sont les l nouveau vecteurs $\{\boldsymbol{h}_{\alpha\beta_1},\dots,\boldsymbol{h}_{\alpha\beta_l}\}$ définis par $\boldsymbol{h}_{\alpha\beta} = \boldsymbol{x}_\beta - \boldsymbol{x}_\alpha$. Pour le voisinage élargi $\mathbb{W}_\alpha \cup \overline{\mathbb{W}}_\alpha$, la matrice géométrique est donnée par*

$$\widehat{H}_\alpha^t = \left[\, H_\alpha^t \middle|\, \overline{H}_\alpha^t \,\right].$$

Notons $\widehat{S}_\alpha = \left(\widehat{H}_\alpha^t\widehat{H}_\alpha\right)^{-1}\widehat{H}_\alpha^t$ la matrice de la méthode des moindres carrés sur le voisinage élargi $\mathbb{W}_\alpha \cup \overline{\mathbb{W}}_\alpha$. Alors $\widetilde{S}_\alpha$ et $\widehat{S}_\alpha$ satisfont les points suivants :

(i) *Les valeurs singulières de $K_\alpha\widehat{S}_\alpha$ et de $K_\alpha\widetilde{S}_\alpha$ satisfont les estimations*

$$\varsigma_j\left(K_\alpha\widehat{S}_\alpha\right) \le \varsigma_j\left(K_\alpha\widetilde{S}_\alpha\right), 1 \le j \le d.$$

Soit $\|.\|_M$ une norme matricielle unitairement invariante, la norme (10.6.3) ou une norme qui s'écrit sous la forme $\|A\|_M = F(AA^*)$ où F est une fonction de matrices hermitiennes telle que $F(P) \le F(P+Q)$ pour toute matrice hermitienne P et toute matrice semi-définie positive Q. Alors $K_\alpha\widehat{S}_\alpha$ et $K_\alpha\widetilde{S}_\alpha$ satisfont l'estimation

$$\left\|K_\alpha\widehat{S}_\alpha\right\|_M \le \left\|K_\alpha\widetilde{S}_\alpha\right\|_M.$$

(ii) *Supposons que la matrice* $\overline{H}_\alpha$ *soit de rang d. Si* $\varsigma_j\left(K_\alpha\widetilde{S}_\alpha\right) > 0$ *pour un* $j \in \{1,\dots,d\}$, *alors* $K_\alpha\widehat{S}_\alpha$ *et* $K_\alpha\widetilde{S}_\alpha$ *satisfont l'estimation stricte*

$$\varsigma_j\left(K_\alpha\widehat{S}_\alpha\right) < \varsigma_j\left(K_\alpha\widetilde{S}_\alpha\right).$$

Soit $\|.\|_M$ *l'une des normes citées dans le point (i). Pour les normes qui peuvent être écrites sous la forme* $\|A\|_M = F(AA^*)$, *on suppose de plus que* $F(P) < F(P+Q')$ *pour toute matrice hermitienne* P *et toute matrice définie positive* Q'. *Alors* $K_\alpha\widehat{S}_\alpha$ *et* $K_\alpha\widetilde{S}_\alpha$ *satisfont l'estimation stricte*

$$\left\|K_\alpha\widehat{S}_\alpha\right\|_M < \left\|K_\alpha\widetilde{S}_\alpha\right\|_M.$$

Démonstration. La démonstration est basée sur l'identité matricielle de Sherman-Morrison-Woodbury, cf. [**54**, p. 3] ou [**67**, p. 124]. Soient A, U, C et V des matrices complexes de dimensions respectives $n\times n$, $n\times k$, $k\times k$ et $k\times n$. Alors

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}$$

pourvu que les matrices inverses existent.

D'abord, on démontre la propriété (i) pour les valeurs singulières. Dans le cas de la méthode des moindres carrés, les valeurs singulières de la matrice $K_\alpha\widetilde{S}_\alpha$ sont données par les racines carrées des valeurs propres de la matrice

$$K_\alpha\widetilde{S}_\alpha\widetilde{S}_\alpha^t K_\alpha^t = K_\alpha\left(H_\alpha^t H_\alpha\right)^{-1}K_\alpha^t. \tag{10.6.8}$$

Il est par conséquent suffisant de démontrer chaque inégalité pour les valeurs propres de la matrice réelle symétrique (10.6.8). Si $\overline{H}_\alpha$ est la matrice dont les lignes sont les nouveaux vecteurs $\{\boldsymbol{h}_{\alpha\beta_1},\dots,\boldsymbol{h}_{\alpha\beta_l}\}$, alors la matrice $\left(H_\alpha^t H_\alpha\right)^{-1}$ dans (10.6.8) est remplacée par

$$\left(\widehat{H}_\alpha^t\widehat{H}_\alpha\right)^{-1} = \left(H_\alpha^t H_\alpha + \overline{H}_\alpha^t\overline{H}_\alpha\right)^{-1}. \tag{10.6.9}$$

La matrice K_α reste inchangée car le nombre de faces de $\mathcal{T}_\alpha$ ne change pas. L'application de l'identité matricielle de Woodbury à (10.6.9) ainsi que la multiplication par K_α à gauche et par K_α^t à droite donne la relation

$$\begin{aligned} K_\alpha\left(H_\alpha^t H_\alpha\right)^{-1}K_\alpha^t - K_\alpha\left(H_\alpha^t H_\alpha + \overline{H}_\alpha^t\overline{H}_\alpha\right)^{-1}K_\alpha^t = \\ = K_\alpha\left(H_\alpha^t H_\alpha\right)^{-1}\overline{H}_\alpha^t\left(\mathrm{I}_l + \overline{H}_\alpha\left(H_\alpha^t H_\alpha\right)^{-1}\overline{H}_\alpha^t\right)^{-1}\overline{H}_\alpha\left(H_\alpha^t H_\alpha\right)^{-1}K_\alpha^t \end{aligned} \tag{10.6.10}$$

où $\mathrm{I}_l \in \mathbb{M}_l(\mathbb{R})$ est la matrice identité.

La démonstration procède maintenant de la même manière que la preuve du théorème 10.6.7. L'application du corollaire [**65**, cor. 4.4.3, p. 182] à l'équation (10.6.10) montre que les k-ièmes valeurs propres des matrices dans (10.6.10) satisfont

$$\lambda_k\left(K_\alpha\left(H_\alpha^t H_\alpha + \overline{H}_\alpha^t\overline{H}_\alpha\right)^{-1}K_\alpha^t\right) \le \lambda_k\left(K_\alpha\left(H_\alpha^t H_\alpha\right)^{-1}K_\alpha^t\right). \tag{10.6.11}$$

D'après la proposition 10.6.6, les valeurs singulières de $K_\alpha\widetilde{S}_\alpha$ et $K_\alpha\widehat{S}_\alpha$ s'écrivent sous la forme

$$\left.\begin{aligned} \varsigma_k\left(K_\alpha\widetilde{S}_\alpha\right) &= \sqrt{\lambda_k\left(K_\alpha\left(H_\alpha^t H_\alpha\right)^{-1}K_\alpha^t\right)} \\ \varsigma_k\left(K_\alpha\widehat{S}_\alpha\right) &= \sqrt{\lambda_k\left(K_\alpha\left(H_\alpha^t H_\alpha + \overline{H}_\alpha^t\overline{H}_\alpha\right)^{-1}K_\alpha^t\right)} \end{aligned}\right\} 1\le k\le d.$$

L'estimation (10.6.11) permet alors de démontrer le point (i) pour les valeurs singulières.

Toutes les normes unitairement invariantes peuvent être représentées comme des normes vectorielles monotones du vecteur des valeurs singulières, cf. l'argument dans la démonstration du théorème 10.6.7. Pour cette raison, l'estimation (10.6.11) permet de prouver le point (i) pour les normes unitairement invariantes.

Supposons maintenant que $\overline{H}_\alpha$ soit de rang d. Dans ce cas, la matrice

$$\left(H_\alpha^t H_\alpha\right)^{-1} \overline{H}_\alpha^t \left(\mathrm{I}_l + \overline{H}_\alpha \left(H_\alpha^t H_\alpha\right) \overline{H}_\alpha^t\right)^{-1} \overline{H}_\alpha \left(H_\alpha^t H_\alpha\right)^{-1}$$

est définie positive sur $\mathbb{R}^d$. Pour cette raison, la matrice dans le membre de droite de (10.6.10) est définie positive sur le complément orthogonal de $\mathrm{Ker}\left(K_\alpha^t\right)$. Les trois matrices dans l'identité (10.6.10) ont le même noyau qui est égal à $\mathrm{Ker}\left(K_\alpha^t\right)$. Il est maintenant suffisant d'appliquer le théorème de Weyl, cf. [**65**, th. 4.3.1, p. 181], à la restriction des matrices hermitiennes dans (10.6.10) au complément orthogonal de $\mathrm{Ker}\left(K_\alpha^t\right)$. Cela prouve l'estimation stricte

$$\lambda_k\left(K_\alpha \left(H_\alpha^t H_\alpha + \overline{H}_\alpha^t \overline{H}_\alpha\right)^{-1} K_\alpha^t\right) < \lambda_k\left(K_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} K_\alpha^t\right) \tag{10.6.12}$$

pour toutes les valeurs propres de $K_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} K_\alpha^t$ qui sont strictement positives. Cet argument démontre l'estimation stricte du point (ii) pour les valeurs singulières.

La démonstration pour les valeurs singulières entraîne immédiatement la propriété (ii) pour toutes les normes matricielles unitairement invariantes, cf. le théorème 10.6.7.

Soit maintenant $\|.\|_M$ une norme qui s'exprime sous la forme $\|A\| = F\left(AA^*\right)$ où F est une fonction de matrices hermitiennes telle que $F(P) \leq F(P+Q)$ pour P hermitienne et Q semi-définie positive et $F(P) < F(P+Q')$ pour P hermitienne et Q' définie positive. Les propriétés (i) et (ii) sont alors des conséquences directes des inégalités (10.6.10) et (10.6.12).

Finalement, il reste à prouver les points (i) et (ii) pour la norme (10.6.3). Pour cela, on rappelle la définition des vecteurs (5.3.7) comme $\boldsymbol{k}_{\alpha\beta} \triangleq \boldsymbol{x}_{\alpha\beta} - \boldsymbol{x}_\alpha$. Soit $\widetilde{S}_\alpha$ la matrice de la méthode des moindres carrés. La norme (10.6.3) de $K_\alpha \widetilde{S}_\alpha$ est alors donnée par la formule

$$\left\|K_\alpha \widetilde{S}_\alpha\right\|_{\mathcal{L}(\infty,2)} = \sup_{\beta \in \mathbb{V}_\alpha} \sqrt{\boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta}}\,. \tag{10.6.13}$$

L'identité de Woodbury entraîne l'identité

$$\begin{aligned}\left(H_\alpha^t H_\alpha\right)^{-1} - \left(H_\alpha^t H_\alpha + \overline{H}_\alpha^t \overline{H}_\alpha\right)^{-1} = \\ = \left(H_\alpha^t H_\alpha\right)^{-1} \overline{H}_\alpha^t \left(\mathrm{I}_l + \overline{H}_\alpha \left(H_\alpha^t H_\alpha\right)^{-1} \overline{H}_\alpha^t\right)^{-1} \overline{H}_\alpha \left(H_\alpha^t H_\alpha\right)^{-1}\,.\end{aligned} \tag{10.6.14}$$

Cela démontre le point (i) pour la norme (10.6.3) car

$$\boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha + \overline{H}_\alpha^t \overline{H}_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta} \leq \boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta} \tag{10.6.15}$$

dans (10.6.13). Si $\overline{H}_\alpha$ est de rang égal à d, alors la matrice dans le membre de droite de (10.6.14) devient définie positive. Cela implique

$$\boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha + \overline{H}_\alpha^t \overline{H}_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta} < \boldsymbol{k}_{\alpha\beta}^t \left(H_\alpha^t H_\alpha\right)^{-1} \boldsymbol{k}_{\alpha\beta} \tag{10.6.16}$$

pour toutes les vecteurs $\boldsymbol{k}_{\alpha\beta}$, ce qui prouve le point (ii) pour la norme (10.6.3). □

10.7. Conclusions pratiques de l'étude théorique pour le schéma d'ordre deux

La définition 10.6.1 introduit un nouveau critère pour évaluer l'impact de la reconstruction du gradient sur la stabilité du schéma MUSCL. Ce critère définit une *mesure qualitative et approchée* pour identifier les méthodes de reconstruction qui augmentent la robustesse du schéma. Un tel critère est indispensable parce que la relation exacte entre les valeurs propres de l'opérateur MUSCL et la reconstruction du gradient est trop difficile à analyser en maillage non structuré. Le critère repose sur une propriété locale dans chaque cellule, l'opérateur de reconstruction locale introduit par la définition 10.5.3. Cet opérateur est une matrice adimensionnelle et invariante par des changements d'échelle du maillage.

Les théorèmes 10.6.7 et 10.6.9 permettent de tirer deux conclusions importantes pour le choix et la conception des méthodes de reconstruction :

(1) Le théorème 10.6.7 montre que la reconstruction des moindres carrés minimise le critère de la définition 10.6.1 pour une large famille de normes. Ce résultat suggère que si la méthode des moindres carrés donne un schéma instable, alors toute autre méthode de reconstruction consistante est également susceptible de produire un schéma instable. Ce résultat permet de voir la méthode des moindres carrés comme une méthode de stabilité optimale mais cette interprétation n'est évidemment pas totalement rigoureuse. La remarque 10.6.8 montre par ailleurs que le théorème 7.7.2 de la section 7.7 devient simplement un corollaire du théorème 10.6.7.

(2) Le résultat du théorème 10.6.9 suggère que des voisinages de reconstruction plus grands conduisent à des schémas plus robustes, au moins dans le cas de la reconstruction des moindres carrés.

Ces deux points ont été examinés en détail dans les tests numériques décrits dans la section 10.9.

10.8. Généralisation de l'étude aux schémas d'ordre trois et quatre

La définition de l'opérateur de reconstruction locale se généralise aux reconstructions d'ordre trois et quatre. Pour cela, on rappelle le schéma général (10.2.19)

$$|\mathcal{T}_\alpha| \frac{du_\alpha}{dt} = -\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \left\{ u_\alpha + \sum_{l=1}^{k} \frac{1}{l!} \left[\boldsymbol{k}_{\alpha\beta}^{(l)} - \boldsymbol{x}_\alpha^{(l)}\right] \bullet \boldsymbol{w}_\alpha^{(l)}\,[\mathfrak{u}] \right\} - \\ - \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{ u_\beta + \sum_{l=1}^{k} \frac{1}{l!} \left[\boldsymbol{k}_{\beta\alpha}^{(l)} - \boldsymbol{x}_\beta^{(l)}\right] \bullet \boldsymbol{w}_\beta^{(l)}\,[\mathfrak{u}] \right\},\ 1 \le \alpha \le N$$

où les coefficients

$$\boldsymbol{w}_\alpha^{(l)}\,[\mathfrak{u}] = \sum_\gamma \boldsymbol{w}_{\alpha\gamma}^{(l)}\,(u_\gamma - u_\alpha)\ ,\ 1 \le l \le k$$

sont des dérivées consistantes au sens de la définition 7.4.1. La forme du schéma (10.2.19) montre immédiatement qu'un opérateur de reconstruction locale, semblable à celui de la définition 10.5.3, peut être défini par

$$r_{\beta\gamma}^{(\alpha)} \triangleq \sum_{l=1}^{k} \frac{1}{l!} \left[\boldsymbol{k}_{\alpha\beta}^{(l)} - \boldsymbol{x}_\alpha^{(l)}\right] \bullet \boldsymbol{w}_{\alpha\gamma}^{(l)}\,. \tag{10.8.1}$$

Avec la définition (10.8.1), le schéma (10.2.19) s'écrit sous la forme

$$|\mathcal{T}_\alpha| \frac{du_\alpha}{dt} = -\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \left\{ u_\alpha + \sum_\gamma r_{\beta\gamma}^{(\alpha)}\,(u_\gamma - u_\alpha) \right\} - \\ - \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{ u_\beta + \sum_\gamma r_{\alpha\gamma}^{(\beta)}\,(u_\gamma - u_\beta) \right\},\ 1 \le \alpha \le N\,,$$

ce qui montre que la proposition 10.5.2 et le corollaire 10.5.6 restent valables pour l'opérateur de reconstruction locale défini par (10.8.1).

On s'attend donc à ce que les conclusions pratiques de la section 10.7 restent valables pour la reconstruction des polynômes de degré deux et trois. On devrait notamment observer que l'élargissement des voisinages de reconstruction augmente la stabilité du schéma et que la reconstruction des moindres carrés, introduite par la définition 7.7.1, constitue une méthode de stabilité optimale dans le sens approximatif de la section 10.7. Des expériences numériques sont indispensables pour vérifier ces hypothèses, ce qui fait l'objet des sections 10.9.3 et 10.9.4.

10.9. Étude numérique

L'objectif de cette section est de compléter et confirmer les conclusions de la section 10.7 par le calcul numérique de spectres sur un échantillon de maillages structurés et non structurés. Un objectif majeur est de mettre en évidence une relation entre l'opérateur de reconstruction locale R_α de la définition 10.5.3 et la stabilité asymptotique du schéma semi-discret (10.2.19).

(1) La section 10.9.2 présente les résultats numériques pour les méthodes de reconstruction linéaire par morceaux, c'est-à-dire les méthodes de reconstruction des polynômes de degré un. Ces méthodes donnent le schéma (10.2.13) qui est formellement d'ordre deux.

(2) La section 10.9.3 présente les résultats numériques pour les méthodes de reconstruction de degré deux, c'est-à-dire les schémas qui sont formellement d'ordre trois.

(3) La section 10.9.4 présente les résultats numériques pour les méthodes de reconstruction de degré trois, c'est-à-dire les schémas qui sont formellement d'ordre quatre.

10.9.1. Description des cas tests. Afin d'isoler l'influence du type de maillage, de la méthode de reconstruction et du voisinage de reconstruction sur la stabilité des schémas du type (10.2.19), il est utile de considérer l'équation de convection linéaire dans le contexte le plus simple, c'est-à-dire sur un carré ou un cube avec des conditions de périodicité au bord. Pour chaque cas test, un programme construit la matrice de l'opérateur J du schéma (10.2.19) et calcule son spectre par des algorithmes numériques. De plus, le programme détermine pour chaque cellule $\mathcal{T}_\alpha$ la valeur de $\|R_\alpha\|_{\mathcal{L}(2,\infty)} = \|K_\alpha S_\alpha\|_{\mathcal{L}(2,\infty)}$ et calcule la moyenne, le médian, le minimum et le maximum ainsi que le $90^{\text{ème}}$, $95^{\text{ème}}$ et $99^{\text{ème}}$ centile de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ pour chaque maillage.

Le critère numérique pour identifier des discrétisations instables est l'abscisse spectrale de J, définie par

$$\omega_J = \max\left\{\Re(\lambda) \middle| \lambda \in \operatorname{sp}(J)\right\}.$$

D'après la proposition 10.3.1, il est nécessaire pour la stabilité que $\omega_J \leq 0$. Si $\omega_J > 0$, l'opérateur est instable dans le sens de la définition 10.3.2.

Les cas tests comprennent des maillages de triangles et de tétraèdres, des maillages hybrides, des maillages cartésiens ainsi que des maillages cartésiens déformés. Tous les cas tests ont les vitesses de convection fixes $\boldsymbol{c} = \frac{1}{\sqrt{8}}\left(-\sqrt{5}, \sqrt{3}\right)$ en dimension deux et $\boldsymbol{c} = \frac{1}{\sqrt{14}}(1, -3, 2)$ en dimension trois.

10.9.2. Résultats numériques pour le schéma d'ordre deux. Les tests du schéma d'ordre deux (10.2.13) couvrent plusieurs méthodes en dimension deux et trois :

(1) La reconstruction des moindres carrés, introduite par la définition 7.7.1 et décrite dans la section 7.9.1 pour le cas $k = 1$, sur le premier et deuxième voisinage.

(2) La reconstruction de Green, décrite dans la section 7.9.2, sur le premier voisinage.

(3) La reconstruction d'ordre deux sur le deuxième voisinage. Cette méthode consiste à reconstruire une dérivée seconde (10.2.16) et un gradient de précision à l'ordre deux (10.2.15), mais de n'utiliser que le gradient pour une reconstruction linéaire par cellule.

(4) La reconstruction d'un gradient moyenné sur un voisinage élargi, par l'algorithme 8.4.1 de la section 8.4. Cette méthode reconstruit un gradient sur le deuxième voisinage.

Le tableau 10.9.1 donne un résumé des résultats. Dans la suite, il s'agit d'examiner point par point les observations et de les mettre en relation avec les résultats théoriques de la section 10.6. L'étude tente également de mettre en évidence une corrélation entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)} = \|K_\alpha S_\alpha\|_{\mathcal{L}(2,\infty)}$ et la stabilité du schéma d'ordre deux (10.2.13). Rappelons que la matrice $K_\alpha S_\alpha$ est invariante par rapport aux changements d'échelle du maillage. Cela justifie la comparaison de valeurs de $K_\alpha S_\alpha$ entre différents maillages.

(1) *Reconstruction par la méthode des moindres carrés sur le premier voisinage en dimension deux* : les tests numériques ne révèlent aucune instabilité pour cette méthode en dimension deux. Le tableau 10.9.2 affiche les abscisses spectrales ω_J et les statistiques

TAB. 10.9.1: Résumé des résultats pour les schémas d'ordre deux (MUSCL)

Dimension	Méthode de reconstruction	Voisinage	Instabilité
2	Moindres carrés	Premier voisinage	Non
2	Green	Premier voisinage	Oui, mais faible
2	Moindres carrés	Deuxième voisinage	Non
2	Ordre deux	Deuxième voisinage	Non
3	Moindres carrés	Premier voisinage	Oui
3	Green	Premier voisinage	Oui
3	Moindres carrés	Deuxième voisinage	Non
3	Ordre deux	Deuxième voisinage	Oui
3	Algorithme 8.4.1	Deuxième voisinage	Non

TAB. 10.9.2: Reconstruction de degré un par la méthode des moindres carrés sur le premier voisinage en dimension deux : abscisse spectrale ω_J et statistiques de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$

maillage	abscisse spectrale	moyenne	maximum	$90^{\text{ème}}$ centile
triangles 1	-0.39404e-9	0.43621	0.54114	0.49491
triangles 2	0.23357e-9	0.42221	0.55833	0.45913
triangles 3	0.18482e-9	0.42372	0.59746	0.46769
triangles 4	0.39867e-10	0.41897	0.55139	0.44342
hybride 1	0.23081e-9	0.42646	0.63910	0.52123
hybride 2	0.63704e-10	0.41479	0.62273	0.49499
hybride 3	-0.16952e-9	0.41004	0.61284	0.49313
hybride 4	-0.32351e-10	0.40816	0.58758	0.47695
cartésien déformé 1	0.15821e-9	0.42557	0.65135	0.51010
cartésien déformé 2	0.52366e-10	0.43035	0.63335	0.51623
cartésien déformé 3	-0.21041e-9	0.43152	0.63201	0.51990
cartésien déformé 4	-0.32943e-9	0.43145	0.65605	0.51863

les plus significatives de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ pour cette méthode de reconstruction. On constate que les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ restent nettement inférieures à 1.

(2) *Reconstruction par la méthode de Green sur le premier voisinage en dimension deux* : la méthode de Green de la section 7.9.2 produit des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ qui sont plus grandes que celles de la méthode des moindres carrés. Dans certaines cellules du premier maillage hybride, les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ approchent même la valeur $1,5$. Les interfaces entre ces cellules montrent une forte courbure, c'est-à-dire le centre de la face se situe loin de l'axe qui relie les deux barycentres des cellules voisines. Cela pose un problème pour la méthode de Green. L'opérateur correspondant est stable pour la vitesse $\boldsymbol{c} = \frac{1}{\sqrt{8}}\left(-\sqrt{5}, \sqrt{3}\right)$ mais devient instable pour la vitesse $\boldsymbol{c} = (1, 0)$ avec une abscisse spectrale légèrement positive de $\omega_J \approx 0.0002$. Dans ce cas particulier, la reconstruction des moindres carrés produit un schéma stable avec des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ qui sont nettement plus petites que celles de la méthode de Green. Cette observation soutient la conclusion de l'étude théorique qui prévoit que la méthode des moindres carrés favorise la stabilité par rapport à d'autres méthodes.

(3) *Reconstruction par la méthode des moindres carrés sur le deuxième voisinage en dimension deux* : les opérateurs MUSCL sont stables sur tous les maillages testés. Sur le deuxième voisinage, la méthode des moindres carrés produit des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$

TAB. 10.9.3: Reconstruction de degré un par la méthode des moindres carrés sur le premier voisinage en dimension trois : abscisse spectrale ω_J et statistiques de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$

maillage	abscisse spectrale	moyenne	maximum	$90^{\text{ème}}$ centile
tétraèdres 1	1.6539	0.57376	0.99051	0.67154
tétraèdres 2	-0.46968e-10	0.57143	1.0878	0.67217
tétraèdres 3	5.7716	0.56804	1.0533	0.65979
tétraèdres 4	7.5288	0.57435	1.0888	0.67144
hybride 1	2.1612	0.54796	1.0820	0.65732
hybride 2	5.5859	0.55320	1.0702	0.68159
hybride 3	6.5645	0.53307	1.0962	0.66178
hybride 4	7.2591	0.52921	1.1547	0.64271
cartésien déformé 1	-0.17017e-9	0.40784	0.54825	0.45403
cartésien déformé 2	0.42669e-10	0.41018	0.54821	0.45641
cartésien déformé 3	-0.63580e-10	0.41188	0.58191	0.46029
cartésien déformé 4	-0.55940e-10	0.41334	0.56309	0.46198

plus petites que sur le premier voisinage, ce qui correspond au résultat du théorème 10.6.9.

(4) *Reconstruction par la méthode d'ordre deux sur le deuxième voisinage en dimension deux* : Cette méthode produit des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ qui sont plus grandes que celles de la méthode des moindres carrés, ce qui correspond au résultat du théorème 10.6.7. L'opérateur MUSCL est néanmoins stable pour cette méthode en dimension deux.

(5) *Reconstruction par la méthode des moindres carrés sur le premier voisinage en dimension trois* : en maillage de tétraèdres, cette méthode de reconstruction génère des modes propres instables. En même temps, le médian et la moyenne de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ sont supérieurs à $0,5$ et le maximum de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ est supérieur à 1. Cette observation suggère qu'il existe un lien entre l'apparition des instabilités et les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$. Le tableau 10.9.3 montre les abscisses spectrales et les statistiques les plus importantes de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ pour soutenir cette hypothèse. La figure 10.9.1 montre le spectre pour le maillage de tétraèdres numéro 3 : deux modes propres instables sont visibles à droite de l'axe imaginaire. Ce résultat indique que le premier voisinage est trop petit pour la reconstruction de degré un en maillage de tétraèdres. En maillage cartésien et maillage cartésien déformé, cette méthode ne rencontre pas de problèmes d'instabilité. Toutes ces observations correspondent exactement aux expériences faites avec CEDRE : si les limiteurs sont désactivés, la reconstruction de degré un sur le premier voisinage est instable en maillage de tétraèdres, alors qu'elle est stable en maillage de hexaèdres.

(6) *Reconstruction par la méthode des moindres carrés sur le deuxième voisinage en dimension trois :* sur le deuxième voisinage, la reconstruction des moindres carrés conduit à des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ nettement plus petites que sur le premier voisinage, ce qui correspond au résultat du théorème 10.6.9. En même temps, les modes propres instables disparaissent. Cette observation met à nouveau en évidence une relation entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ et l'apparition de modes instables. La figure 10.9.2 montre le spectre stable pour le maillage de tétraèdres numéro 3. Il s'agit du même maillage que dans la figure 10.9.1.

(7) *Reconstruction par la méthode du gradient d'ordre deux sur le deuxième voisinage en dimension trois :* cette méthode de reconstruction génère un opérateur instable pour le maillage de tétraèdres numéro 2 où elle produit dans certaines cellules des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ supérieures à 1. Cette observation suggère à nouveau l'existence d'un lien entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ et l'apparition de modes propres instables. La comparaison de cette méthode avec la reconstruction des moindres carrés montre que

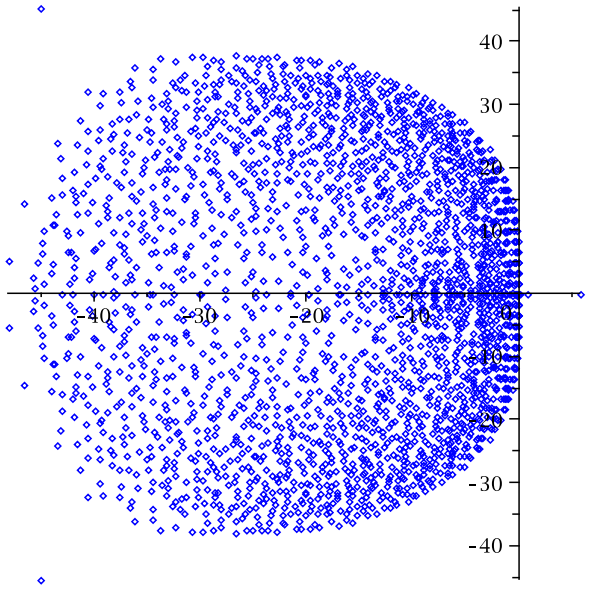


FIG. 10.9.1: Spectre instable pour la reconstruction de degré un sur le premier voisinage (maillage de tétraèdres numéro 3) : deux valeurs propres instables sont visibles à droite de l'axe imaginaire.

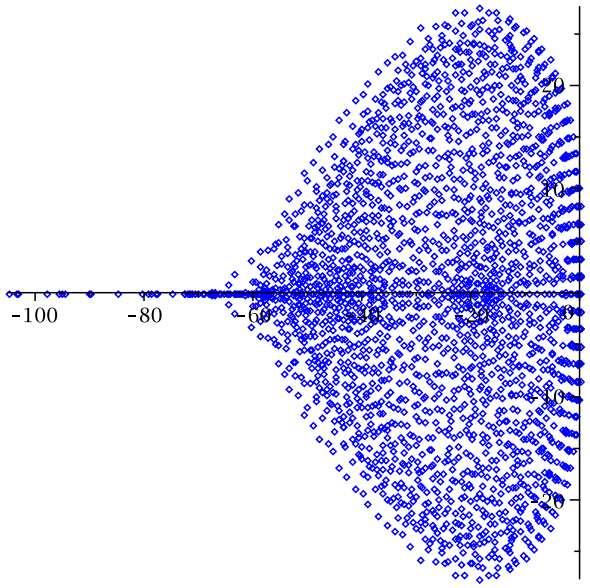


FIG. 10.9.2: Spectre stable pour la reconstruction de degré un sur le deuxième voisinage (maillage de tétraèdres numéro 3) : les modes instables de la figure 10.9.1 ont disparu.

la méthode des moindres carrés fournit des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ nettement inférieures et donne en même temps une discrétisation stable. Cette observation suggère à nouveau un lien entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ et l'apparition de modes instables.

(8) *Reconstruction du gradient sur un voisinage élargi* : la reconstruction du gradient par l'algorithme 8.4.1 fournit un schéma stable dans tous les cas testés. Les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ sont nettement inférieures à celles produites par la reconstruction des

moindres carrés sur le premier voisinage, ce qui peut éventuellement expliquer la disparition des instabilités.

(9) *Maillages cartésiens* : les opérateurs MUSCL sont stables sur les maillages cartésiens. Sur ces maillages, on a $\|R_\alpha\|_{\mathcal{L}(2,\infty)} = 0.35355$ pour la reconstruction en dimension deux et trois. Les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ sont donc plus petites en maillage cartésien qu'en maillage de tétraèdres, ce qui peut s'expliquer par le théorème 10.6.9 et le fait que le premier voisinage d'une cellule est plus grand en maillage cartésien qu'en maillage de tétraèdres. Cela pourrait expliquer l'absence de modes instables sur de tels maillages.

(10) *Maillages cartésiens déformés* : sur ces maillages, les tests n'ont révélé aucune instabilité. Les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ sont supérieures à celles des maillages cartésiens mais nettement inférieures à celles des maillages de tétraèdres. Cela pourrait s'expliquer par le fait que le premier voisinage de chaque cellule est plus grand qu'en maillage de tétraèdres.

(11) *Influence du type de maillage sur la stabilité* : Les résultats des tests laissent supposer que les instabilités émergent seulement sur des maillages de tétraèdres dans le cas de la reconstruction sur le premier voisinage. Les tétraèdres et les prismes sont les cellules pour lesquelles la taille du premier voisinage est le plus petit et les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ sont les plus grandes.

(12) *Nombre de modes instables.* Le nombre de valeurs propres avec une partie réelle positive semble toujours assez petit, souvent moins d'un pour cent du nombre total des valeurs propres. Cela ne remédie pas au problème car des erreurs d'arrondi introduisent toujours ces modes dans la solution numérique.

(13) *Relation entre la matrice de reconstruction locale et la stabilité asymptotique.* Les expériences numériques montrent une forte corrélation entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ et l'existence de valeurs propres instables. Ces valeurs propres apparaissent uniquement sur des maillages où les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ sont proches de ou supérieures à 1 dans certaines cellules. Il manque cependant une preuve théorique générale pour cette relation.

10.9.3. Résultats numériques pour le schéma d'ordre trois. Cette section présente les résultats numériques pour les méthodes de reconstruction des polynômes de degré deux. Les tests couvrent les cinq méthodes suivantes :

(1) La reconstruction des polynômes de degré deux par la méthode des moindres carrés, ou méthode de la pseudo-inverse, introduite par la définition 7.7.1.

(2) La reconstruction des polynômes de degré deux par la méthode des moindres carrés couplés par itérations, ou méthode MCCI, définie par l'algorithme 8.2.2. Le nombre d'itérations est fixé à trois.

(3) La reconstruction des polynômes de degré deux par la méthode des moindres carrés couplés par itérations sur un voisinage élargi, ou méthode MCCIE, définie par l'algorithme 8.4.3. Le nombre d'itérations est fixé à trois.

(4) La reconstruction des polynômes de degré deux par la méthode des corrections successives, ou méthode CS, définie par l'algorithme 8.3.8.

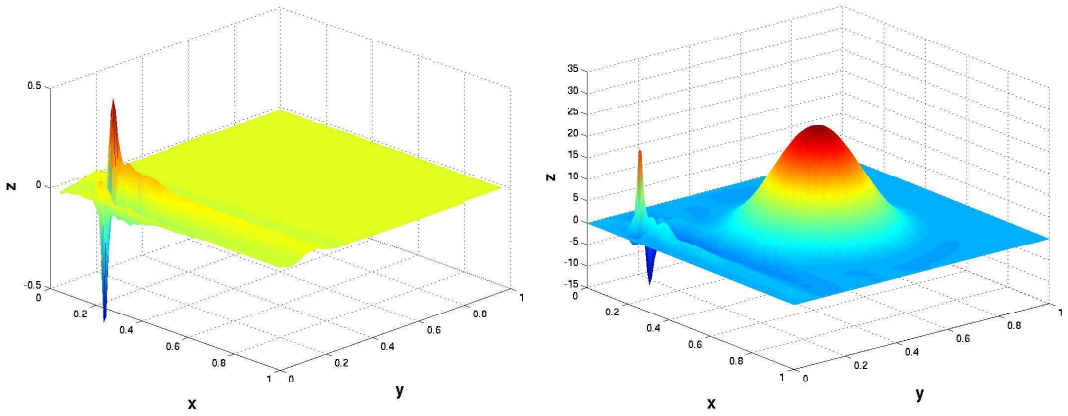
(5) La reconstruction des polynômes de degré deux par la méthode des corrections successives sur un voisinage élargi, ou méthode CSE, définie par l'algorithme 8.4.4.

Le tableau 10.9.4 montre un résumé des résultats.

(1) *Reconstruction des polynômes de degré deux par la méthode des moindres carrés sur le deuxième voisinage en dimension deux* : les tests numériques ne révèlent aucune instabilité pour la reconstruction des moindres carrés sur les maillages de triangles, les maillages hybrides et les maillages cartésiens. Les valeurs observées pour la norme $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ sont inférieures à $0,3$ sur les maillages de triangles.

TAB. 10.9.4: Résumé des résultats pour la reconstruction de degré deux.

Dimension	Méthode de reconstruction	Voisinage	Instabilité
2	Moindres carrés (pseudo-inverse)	2	Non
2	Moindres carrés couplés MCCI	4	Oui
2	Moindres carrés couplés sur voisinage élargi MCCIE	5	Non
2	Corrections successives CS	2	Oui
2	Corrections successives sur voisinage élargi CSE	4	Non
3	Moindres carrés (pseudo-inverse)	2	Oui
3	Moindres carrés (pseudo-inverse)	3	Non
3	Moindres carrés couplés MCCI	4	Oui
3	Moindres carrés couplés sur voisinage élargi MCCIE	5	Non
3	Corrections successives CS	2	Oui
3	Corrections successives sur voisinage élargi CSE	4	Non



(a) Mode propre réel

(b) Apparition du mode instable lors de la convection d'une gaussienne

FIG. 10.9.3: Exemple d'un mode instable pour la reconstruction de degré deux par la méthode CS sur le maillage de triangles numéro trois. La valeur propre de ce mode est réelle positive : $\lambda = 0,94518$.

(2) *Reconstruction des polynômes de degré deux par les méthodes MCCI et CS en dimension deux* : ces méthodes s'avèrent instables sur certains maillages de triangles où elles donnent des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ nettement plus grandes que la reconstruction des moindres carrés. Cette observation, qui correspond au résultat du théorème 10.6.7, suggère un lien entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ et l'apparition d'instabilités. La figure 10.9.3 montre le résultat de la convection d'une fonction gaussienne par la méthode CS sur le maillage de triangles numéro trois avec une vitesse de convection $\boldsymbol{c} = (1, 0)$: le mode instable apparaît au bout d'un certain temps car il est toujours présent dans la solution numérique. Sa croissance exponentielle conduit à l'arrêt du calcul. Sur des maillages hybrides, sur des maillages cartésiens et sur des maillages cartésiens déformés, ces méthodes sont stables.

(3) *Reconstruction des polynômes de degré deux par les méthodes MCCIE et CSE sur des voisinages élargis en dimension deux* : ces méthodes de reconstruction sont stables dans tous les cas testés et donnent des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ nettement inférieures à celles

TAB. 10.9.5: Reconstruction de degré deux par la méthode des moindres carrés sur le deuxième voisinage en dimension trois : abscisse spectrale ω_J et statistiques de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$

maillage	abscisse spectrale	moyenne	maximum	90ème centile
tétraèdres 1	0.10622e-9	0.30011	0.91670	0.39174
tétraèdres 2	12.708	0.30903	1.5275	0.41208
tétraèdres 3	0.29960e-9	0.30021	2.0787	0.38466
tétraèdres 4	-0.26388e-10	0.30059	3.2774	0.38622
hybride 1	0.17447e-9	0.26283	0.90139	0.35064
hybride 2	0.11756e-9	0.28307	1.1724	0.39445
hybride 3	-0.46136e-9	0.25690	1.0797	0.35191
hybride 4	-0.23562e-9	0.26962	1.0959	0.38779
cartésien déformé 1	0.25041e-10	0.15156	0.20832	0.17452
cartésien déformé 2	-0.39756e-10	0.15247	0.21253	0.17467
cartésien déformé 3	-0.71528e-10	0.15301	0.21993	0.17564
cartésien déformé 4	-0.40058e-10	0.15390	0.22172	0.17678

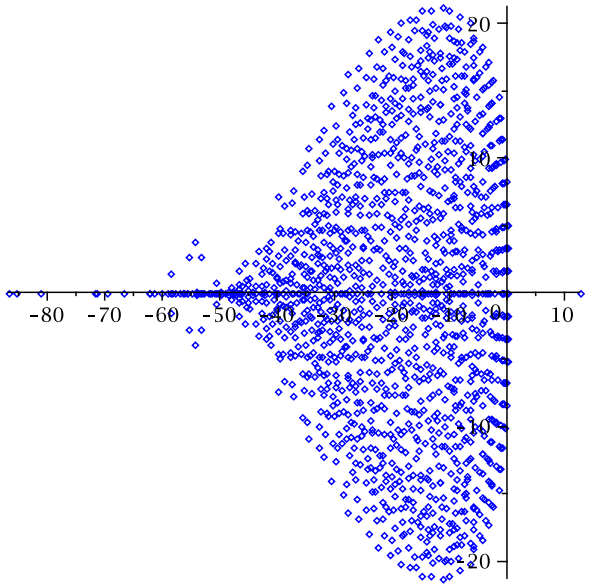


FIG. 10.9.4: Spectre instable pour la reconstruction des polynômes de degré deux par la méthode des moindres carrés sur le deuxième voisinage (maillage de tétraèdres numéro deux). Une valeur propre réelle positive est visible à droite de l'axe imaginaire.

des méthodes MCCI et CS. Ce résultat démontre la pertinence de l'élargissement des voisinages de reconstruction et souligne à nouveau le lien entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ et la stabilité.

(4) *Reconstruction des polynômes de degré deux par la méthode des moindres carrés sur le deuxième voisinage en dimension trois.* Sur le deuxième voisinage, ce type d'opérateur peut être instable sur des maillages de tétraèdres. Dans ce cas particulier, le maximum de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ peut devenir supérieur à 1. Le tableau 10.9.5 montre les abscisses spectrales et les statistiques les plus importantes de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ pour cette méthode de reconstruction. La figure 10.9.4 montre le spectre instable pour le maillage de tétraèdres

TAB. 10.9.6: Reconstruction de degré deux par la méthode des moindres carrés sur le troisième voisinage en dimension trois : abscisse spectrale ω_J et statistiques de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$

maillage	abscisse spectrale	moyenne	maximum	$90^{\text{ème}}$ centile
tétraèdres 1	0.26108e-9	0.10093	0.16102	0.12224
tétraèdres 2	-0.90014e-10	0.10345	0.24880	0.12631
tétraèdres 3	0.26784e-9	0.10026	0.21860	0.11935
tétraèdres 4	-0.18319e-9	0.10210	0.20339	0.12175
hybride 1	0.22848e-10	0.97704e-1	0.16284	0.11634
hybride 2	-0.18282e-9	0.98169e-1	0.21361	0.12321
hybride 3	-0.23332e-9	0.97287e-1	0.16045	0.12040
hybride 4	-0.21137e-10	0.97685e-1	0.20481	0.12385
cartésien déformé 1	0.58025e-10	0.68114e-1	0.94082e-1	0.77703e-1
cartésien déformé 2	-0.72853e-10	0.68504e-1	0.98941e-1	0.78130e-1
cartésien déformé 3	-0.11688e-9	0.68763e-1	0.97924e-1	0.78538e-1
cartésien déformé 4	-0.13138e-10	0.69083e-1	0.10249	0.79103e-1

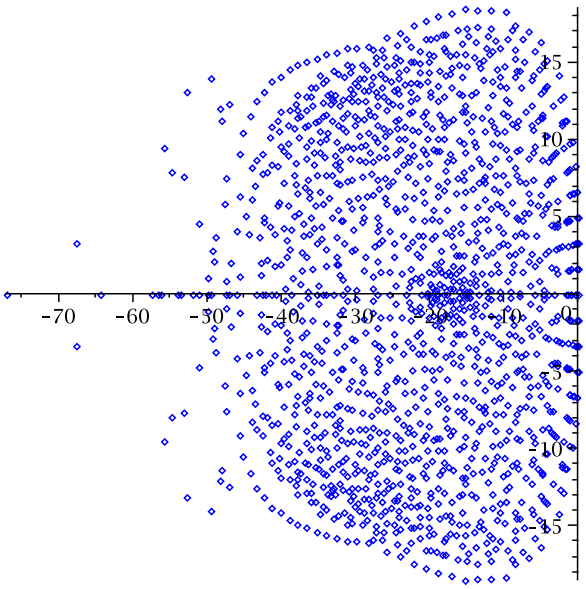


FIG. 10.9.5: Spectre stable pour la reconstruction des polynômes de degré deux par la méthode des moindres carrés sur le troisième voisinage (maillage de tétraèdres numéro deux). La valeur propre instable de la figure 10.9.4 a disparu.

numéro 2. Un mode propre instable est visible à droite de l'axe imaginaire. Ce résultat numérique prouve que le deuxième voisinage est trop petit pour la reconstruction quadratique sur des tétraèdres.

(5) *Reconstruction des polynômes de degré deux par la méthode des moindres carrés sur le troisième voisinage en dimension trois.* La reconstruction des moindres carrés sur le troisième voisinage produit des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ inférieures à celles du deuxième voisinage, ce qui correspond au résultat du théorème 10.6.9. En même temps, les tests ne révèlent aucune instabilité. Le tableau 10.9.6 montre les abscisses spectrales et les statistiques les plus importantes de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ pour cette méthode de reconstruction.La figure 10.9.5 montre le spectre stable pour le maillage de tétraèdres numéro 2. Ce

résultat confirme l'idée que l'élargissement des voisinages de reconstruction est la bonne approche pour supprimer les instabilités.

(6) *Reconstruction des polynômes de degré deux par les méthodes MCCI et CS en dimension* trois. Sur des maillages de tétraèdres, les reconstructions par les méthodes MCCI et CS donnent systématiquement des schémas instables. Cela montre que ces méthodes doivent être remplacées par les méthodes MCCIE et CSE. L'instabilité de la méthode CS sur le deuxième voisinage en maillage de tétraèdres s'explique de manière qualitative par le théorème 10.6.7 : la méthode des moindres carrés est instable sur ce voisinage et d'après l'interprétation du résultat du théorème, si cette méthode est instable, alors les autres méthodes consistantes sont également instables. La situation est plus compliquée pour la méthode MCCI, car elle utilise de plus grands voisinages. La reconstruction par ces méthodes conduit à des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ qui sont sensiblement plus grandes que celles engendrées par la méthode des moindres carrés, ce qui correspond au résultat du théorème 10.6.9.

(7) *Reconstruction des polynômes de degré deux par les méthodes MCCIE et CSE sur un voisinage élargi en dimension* trois. Ces méthodes fournissent des valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ qui sont nettement inférieures à celles des méthodes MCCI et CS. En même temps, ces méthodes donnent des schémas stables dans tous les cas testés. Cette observation confirme l'hypothèse de l'existence d'une corrélation entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ et l'apparition d'instabilités.

(8) *Maillages cartésiens* et *maillages cartésiens déformés*. Toutes les méthodes sont stables sur ces types de maillages.

(9) *Influence du type de maillage sur la stabilité*. Les résultats des tests laissent supposer que les instabilités émergent sur les maillages de triangles et de tétraèdres, dans le cas des méthodes MCCI et CS. La méthode des moindres carrés peut devenir instable pour la reconstruction sur le deuxième voisinage en maillage de tétraèdres, ce qui montre que ce voisinage est trop petit pour une reconstruction de degré deux en dimension trois. Les méthodes MCCIE et CSE sur des voisinages élargis s'avèrent toujours stables. Cela peut s'expliquer par le fait qu'elles utilisent de grands voisinages de reconstruction.

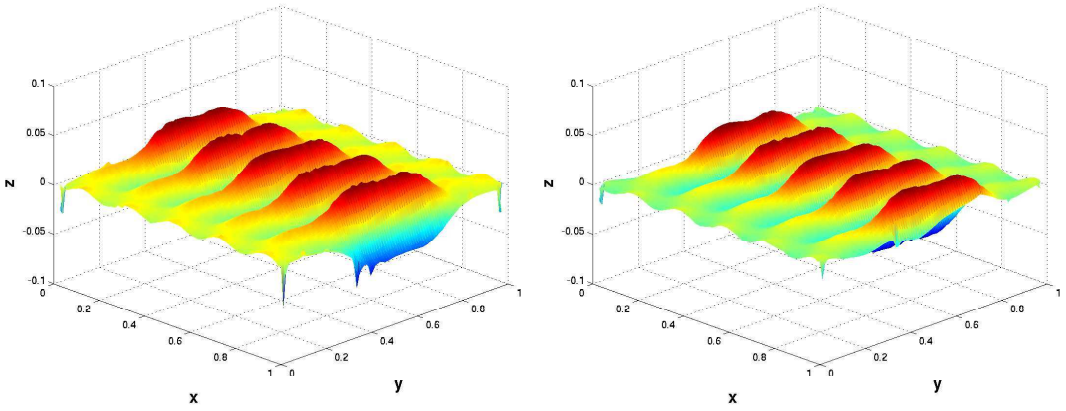
(10) *Nombre de modes instables*. Le nombre de valeurs propres avec une partie réelle positive est toujours assez petit, souvent moins d'un pourcent du nombre total des valeurs propres.

(11) *Relation entre la matrice de reconstruction locale et la stabilité asymptotique*. L'évidence numérique montre une forte corrélation entre les valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ et l'existence de valeurs propres instables. Ces valeurs propres apparaissent uniquement sur des maillages où certaines valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$ sont proches de ou supérieures à 1.

Les conclusions de l'étude numérique sont les suivantes :

- pour les schémas basés sur une reconstruction de degré deux, qui sont donc formellement d'ordre trois, les modes propres instables ressemblent beaucoup à ceux des schémas d'ordre deux, car ces modes apparaissent uniquement sur des maillages de tétraèdres et de triangles et leur nombre est très petit ;
- on constate que la reconstruction des moindres carrés est stable, sauf en maillage de tétraèdres sur le deuxième voisinage. Il est donc nécessaire d'utiliser cette méthode sur le troisième voisinage en maillages de tétraèdres ;
- les méthodes MCCI et CS ne peuvent pas être utilisées telles quelles sur les maillages de triangles et de tétraèdres ; elles doivent être remplacées par les méthodes MCCIE et CSE qui s'avèrent stables partout.

10.9.4. Résultats numériques pour le schéma d'ordre quatre. Cette section présente les tests numériques pour la reconstruction des polynômes de degré trois, c'est-à-dire pour les schémas qui sont formellement d'ordre quatre. L'étude numérique est restreinte à la méthode des moindres carrés (pseudo-inverse), introduite par la définition 7.7.1.



(a) Partie réelle (b) Partie imaginaire

FIG. 10.9.6: Partie réelle et imaginaire d'un mode propre instable « lent » du schéma semi-discret (10.2.19) basé sur une reconstruction de degré $k = 3$ sur le maillage de triangles numéro cinq. Le mode est associé à la valeur propre $\lambda = 0,07949 - 31,49454\,i$ et la vitesse de convection est $\boldsymbol{c} = (1, 0)$.

Dans certains aspects, les résultats ressemblent à ceux des schémas d'ordre deux et trois. On constate surtout que ce type de reconstruction doit utiliser le quatrième voisinage en maillage de tétraèdres, car la reconstruction cubique sur le troisième voisinage produit des modes instables et en même temps de très grandes valeurs de $\|R_\alpha\|_{\mathcal{L}(2,\infty)}$.

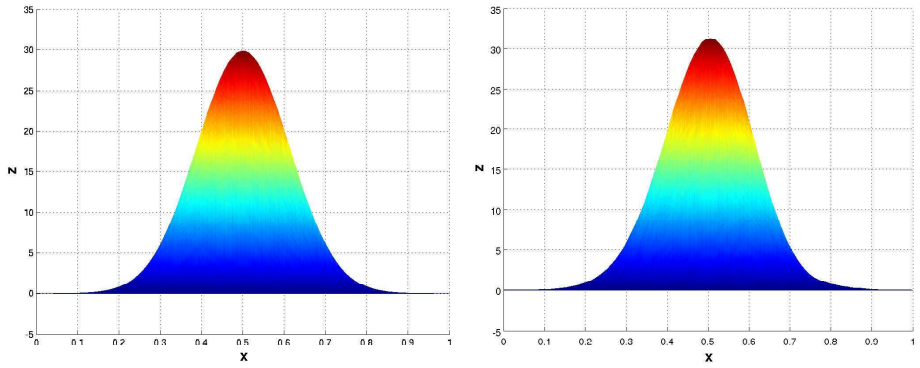
Par d'autres aspects, les résultats sont différents de ceux des schémas d'ordre deux et trois. Dans le cas des schémas d'ordre quatre, les valeurs propres de l'opérateur J se rapprochent de plus en plus de l'axe imaginaire et une partie des valeurs propres se situe directement sur cet axe. Pour cette raison, ces modes propres ne subissent pratiquement pas de dissipation. Cependant, un certain nombre de ces modes ont une partie réelle légèrement positive, ce qui amène une croissance exponentielle de ces modes. Ces modes instables, typiques des schémas d'ordre quatre, sont d'une croissance lente et à grande longueur d'onde, ce qui les différencie des modes instables observés sur les schémas d'ordre deux et trois. La figure 10.9.6 montre l'exemple d'un tel mode sur un maillage de triangles. L'élargissement du voisinage de reconstruction diminue la partie réelle positive de ces modes instables mais ne les fait pas complètement disparaître. La figure 10.9.7 montre la croissance d'une fonction gaussienne lorsqu'elle est transportée sur un maillage triangulaire. Cette croissance est induite par une instabilité de type « lent ». Finalement, la figure 10.9.8 montre le spectre de l'opérateur de discrétisation spatiale associé au schéma semi-discret (10.2.19) reposant sur une reconstruction de degré $k = 3$ sur le maillage de triangles numéro cinq. La figure montre qu'un certain nombre de valeurs propres sont très proches de l'axe imaginaire. Ce phénomène explique l'absence de dissipation constatée dans les expériences numériques de convection présentées dans le chapitre 11, voir les figures 11.3.3 et 11.3.4. Certaines de ces valeurs propres ont toutefois une partie réelle légèrement positive, comme le mode présenté dans la figure 10.9.6, ce qui provoque la croissance de la fonction gaussienne montrée dans la figure 10.9.7.

10.9.5. Conclusions de l'étude numérique. Les résultats des expériences numériques soutiennent les conclusions pratiques de l'étude théorique, présentées dans la section 10.7 :

(1) La méthode des moindres carrés semble être plus stable que les autres méthodes testées.

(2) L'élargissement des voisinages de reconstruction paraît être un moyen efficace pour supprimer les instabilités.

Par ailleurs, le schéma d'ordre quatre présente un type de mode instable qui est différent des instabilités des schémas d'ordre deux et trois. Il s'agit de modes instables à croissance plus lente qui montrent une variation spatiale lisse, ce qui laisse penser que les limiteurs agissent moins



(a) Condition initiale (b) Résultat de la convection

FIG. 10.9.7: Croissance d'une fonction gaussienne (11.3.2) sur le maillage de triangles numéro cinq avec la reconstruction de degré 3 sur le troisième voisinage : vue du plan $y = 0$.

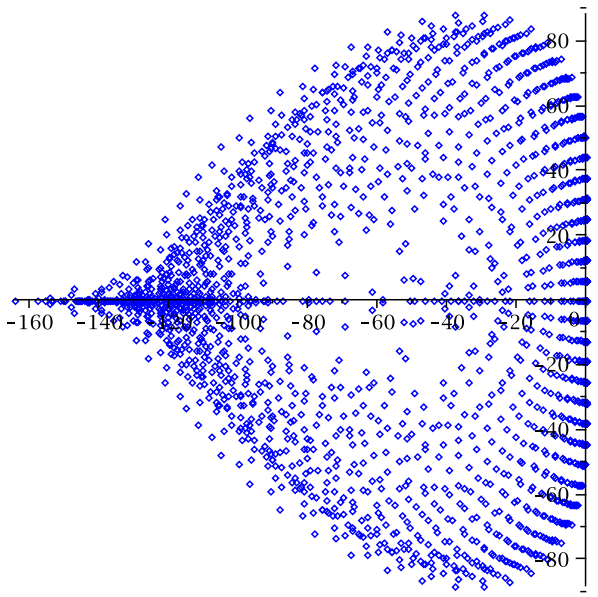


FIG. 10.9.8: Spectre du schéma semi-discret (10.2.19) basé sur une reconstruction de degré $k = 3$ sur le maillage de triangles numéro cinq.

efficacement sur ces modes. L'étude de ce phénomène mérite une étude plus approfondie, mais il semble à première vue que la suppression de ces modes nécessite une dissipation numérique artificielle ou des méthodes d'intégration en temps adaptées.

10.10. Bilan du chapitre

10.10.1. Méthodologie de l'étude. Le sujet de ce chapitre est l'influence de la reconstruction sur la robustesse et la stabilité des schémas volumes finis en maillage non structuré. Pour isoler cette influence, l'étude s'est déroulée dans le contexte suivant :

(1) Pour éliminer les effets de la discrétisation en temps et mettre en évidence ceux de la discrétisation spatiale, l'étude se concentre sur le schéma semi-discret, c'est-à-dire sur le schéma discret en espace et continu en temps.

(2) Les limiteurs modifient la reconstruction aux interfaces entre les cellules. Afin de mieux mettre en évidence l'influence de la reconstruction sur la stabilité, le schéma a été étudié sans limiteurs.

(3) La question de la stabilité a été analysée sur l'exemple de l'équation de convection linéaire à vitesse constante qui est l'exemple le plus simple d'une loi de conservation. Pour cette raison, une méthode numérique dédiée aux lois de conservation doit fournir une discrétisation stable de l'équation de convection linéaire. La linéarisation des équations d'Euler fournit par ailleurs un système d'équations de convection linéaires, ce qui signifie que l'étude de l'équation de convection linéaire est importante pour l'étude des équations d'Euler.

Dans ces conditions, le schéma volumes finis transforme l'équation de la convection linéaire à vitesse constante

$$\frac{\partial u\left(\boldsymbol{x},t\right)}{\partial t}=-\boldsymbol{c}\cdot\boldsymbol{\nabla}u\left(\boldsymbol{x},t\right)$$

en un système dynamique linéaire

$$\frac{d\mathfrak{u}\left(t\right)}{dt}=J\mathfrak{u}\left(t\right)\,,\,\mathfrak{u}\left(0\right)=\mathfrak{u}_{0}\,,\,\mathfrak{u}\left(t\right)\in\mathbb{C}^{N} \tag{10.10.1}$$

où $\mathfrak{u}=(\overline{u}_1,\ldots,\overline{u}_N)$ est le vecteur des moyennes de cellule et J est la matrice qui correspond à la discrétisation spatiale de l'opérateur $-\boldsymbol{c}\cdot\boldsymbol{\nabla}$. La solution générale de (10.10.1) est

$$\mathfrak{u}\left(t\right)=\exp\left(tJ\right)\mathfrak{u}_{0}\,.$$

La notion de stabilité considérée ici est la stabilité asymptotique du système linéaire (10.10.1). Ce système est appelé *stable* s'il existe une constante $\mathrm{C}>0$ telle que

$$\left\|\exp\left(tJ\right)\right\|\leq\mathrm{C}<+\infty\text{ pour tout }t>0\,.$$

La proposition 10.3.1 démontre que cette condition est satisfaite si et seulement si toutes les valeurs propres de J ont une partie réelle non positive et les valeurs propres sur l'axe imaginaire ont un indice de Jordan égal à un. Il s'agit donc d'une condition pour le spectre $\mathrm{sp}\left(J\right)$ de J. Cette condition est formalisée par la définition 10.3.2.

10.10.2. Étude de la stabilité des schémas d'ordre un et deux. L'étude théorique commence par une analyse du schéma d'ordre un dans le cas de la convection linéaire à vitesse constante. Dans ce contexte, le théorème 10.4.2 permet de prouver la stabilité asymptotique du schéma volumes finis semi-discret d'ordre un *sur tous les maillages non structurés*. Ce résultat de stabilité très général paraît étroitement lié au fait que ce schéma engendre une très forte dissipation numérique, ce que l'étude du chapitre 11 a bien mis en évidence.

L'analyse pour le schéma MUSCL, c'est-à-dire le schéma d'ordre deux, procède en deux étapes :

(1) Le chapitre introduit d'abord un nouveau critère pour évaluer l'impact de la reconstruction du gradient sur la stabilité du schéma MUSCL. Ce critère est une *mesure approchée et qualitative* pour identifier les méthodes de reconstruction qui favorisent la robustesse et la stabilité du schéma MUSCL. Un tel critère est indispensable parce que la relation exacte entre les valeurs propres de l'opérateur MUSCL et la reconstruction du gradient est trop difficile à analyser en maillage non structuré. Le critère se base sur une propriété locale dans chaque cellule, l'opérateur de reconstruction locale introduit par la définition 10.5.3. Cet opérateur est une matrice adimensionnelle et invariante par des changements d'échelle du maillage. La valeur du critère est donnée par certaines normes matricielles de l'opérateur de reconstruction locale. On peut alors dire de manière qualitative que les méthodes qui minimisent ces normes sont celles qui favorisent la stabilité.

(2) La deuxième étape a consisté à chercher des méthodes qui minimisent le critère local de stabilité. Pour une importante famille de normes, il est possible d'identifier la méthode des moindres carrés comme un minimum, cf. le théorème 10.6.7. De plus, il a été possible de montrer qu'un élargissement du voisinage de reconstruction ne peut pas augmenter

la valeur de ces normes et conduit même à une diminution stricte de leur valeur, si une simple condition géométrique est satisfaite.

L'analyse et les théorèmes de la section 10.6 offrent deux conclusions pratiques pour le choix et la conception de méthodes de reconstruction.

(1) Le théorème 10.6.7 démontre que la méthode des moindres carrés minimise le nouveau critère de la définition 10.6.1 pour une importante famille de normes. Le résultat suggère que si la méthode des moindres carrés donne un schéma instable, alors toute autre méthode consistante de reconstruction est également susceptible de produire un schéma instable. Ce résultat peut être interprété de manière qualitative comme un résultat de stabilité optimale de la méthode des moindres carrés mais cette interprétation n'est évidemment pas complètement rigoureuse. Cette conclusion est confirmé par le fait que la méthode de reconstruction de Green de la section 7.9.2 a du être modifiée dans Cedre parce qu'elle produisait des gradients très grands sur certains maillages non structurés, contrairement à la méthode des moindres carrés qui ne présentait pas ce problème.

(2) Le résultat du théorème 10.6.9 suggère que des voisinages de reconstruction plus grands conduisent à des schémas plus robustes, au moins dans le cas de la reconstruction des moindres carrés. Dans la section 10.9.1, cette hypothèse a été testée numériquement et elle s'avère particulièrement pertinente pour des maillages en dimension trois. Cela pourrait expliquer pourquoi les maillages d'hexaèdres présentent de meilleures propriétés de stabilité que des maillages de tétraèdres.

Des calculs numériques de spectres d'opérateurs Muscl sur des échantillons variés de maillages ont complété et confirmé les conclusions de l'étude théorique. Ils montrent une forte corrélation entre certaines normes de l'opérateur de reconstruction locale et l'apparition de modes propres instables.

La discrétisation semble stable en dimension deux, quel que soit le type de maillage. C'est en dimension trois que des instabilités peuvent apparaître sur des maillages de tétraèdres si le voisinage de reconstruction est le premier voisinage. Il s'avère dans les tests que la méthode des moindres carrés paraît effectivement optimale parmi les méthodes de reconstruction consistantes dans le sens suivant :

(1) Si une autre méthode de reconstruction est stable, la méthode des moindres carrés est également stable.

(2) Si la méthode des moindres carrés est instable, les autres méthodes testées s'avèrent également instables.

Ces deux points appuient les conclusions du théorème 10.6.7.

Une observation importante de cette étude est que l'élargissement des voisinages de reconstruction fait en général disparaître les instabilités. Ce résultat est en accord avec les renseignements qualitatifs qu'on peut tirer du théorème 10.6.9. En particulier, les algorithmes 8.4.1 et 8.4.2 de la section 8.4 permettent de supprimer les instabilités dans le cas de la convection linéaire. L'algorithme 8.4.1 a été implémenté dans Cedre et testé avec succès dans les calculs décrits dans les chapitres 13 et 14.

10.10.3. Étude de la stabilité des schémas d'ordre trois et quatre. La section 10.8 montre que la notion d'opérateur de reconstruction locale, et avec lui le critère de stabilité de la définition 10.6.1, se généralise aux reconstructions des polynômes de degré deux et trois. Cela laisse supposer que les conclusions de l'étude théorique, présentées dans la section 10.7, restent valables pour les schémas d'ordre trois et quatre. Pour vérifier cette hypothèse, il a été nécessaire de mener une série d'expériences numériques afin de valider les deux points suivants :

(1) La méthode des moindres carrés constitue grossièrement une approche optimale pour la robustesse et la stabilité.

(2) L'élargissement des voisinages de reconstruction augmente la stabilité des schémas volumes finis.

Pour la reconstruction des polynômes de degré deux, les tests numériques couvrent cinq méthodes différentes :

(1) La reconstruction des polynômes de degré deux par la méthode des moindres carrés, introduite par la définition 7.7.1.

(2) La reconstruction des polynômes de degré deux par la méthode des moindres carrés couplés par itération, ou méthode MCCI. Cette méthode est définie par l'algorithme 8.2.2.

(3) La reconstruction des polynômes de degré deux par la méthode des moindres carrés couplés par itération sur un voisinage élargi, ou méthode MCCIE. Cette méthode est définie par l'algorithme 8.4.3.

(4) La reconstruction des polynômes de degré deux par la méthode des corrections successives, ou méthode CS. Cette méthode est définie par l'algorithme 8.3.8.

(5) La reconstruction des polynômes de degré deux par la méthode des corrections successives sur un voisinage élargi, ou méthode CSE. Cette méthode est définie par l'algorithme 8.4.4.

En dimension deux, la reconstruction de degré deux par la méthode des moindres carrés est stable sur le deuxième voisinage. En dimension trois, sur des maillages de tétraèdres, cette méthode nécessite le troisième voisinage pour éviter l'apparition d'instabilités.

Les reconstructions itératives présentées dans le chapitre 8, appelées les méthodes MCCI et CS, se sont avérées instables sur des maillages de triangles et tétraèdres. Ce phénomène était prévisible en dimension trois, car, sur des maillages de tétraèdres, les deux méthodes sont basées sur une reconstruction du gradient sur le premier voisinage qui est déjà instable selon l'étude numérique de la section 10.9.2. En contraste, les méthodes MCCIE et CSE, définies par les algorithmes 8.4.3 et 8.4.4, s'avèrent stable sur tous les types de maillages en dimension deux et trois. Cela signifie qu'en maillage non structuré, il faut remplacer les méthodes MCCI et CS par les méthodes MCCIE et CSE, définies par les algorithmes 8.4.3 et 8.4.4.

Finalement, la reconstruction des polynômes de degré trois par la méthode des moindres carrés fait apparaître un nouveau type d'instabilité. Ces modes instables croissent lentement et leur forme spatiale est lisse. Cela les différencie des modes instables observés pour les schémas d'ordre deux et trois, dont la forme semble accidentée et concentrée en certains endroits et dont la croissance est plus rapide.

La forme des modes instables « rapides » indique que les limiteurs peuvent supprimer ces modes. Cette idée est confirmée par le fait que la reconstruction par la méthode des moindres carrés, qui est instable sur le premier voisinage en maillage de tétraèdres, est stabilisée par l'utilisation des limiteurs dans Cedre. La forme des modes instables « lents » est lisse, ce qui laisse penser que les limiteurs agiront sur ces modes d'une manière moins efficace. Seule une implémentation de la reconstruction des polynômes de degré trois dans un code de volumes finis, comme par exemple Cedre, pourra fournir une réponse définitive à cette question.

CHAPITRE 11

Caractérisation et évaluation des erreurs numériques en maillage non structuré général

11.1. Objectif du chapitre

Les méthodes consistantes de reconstruction développées dans le chapitre 7 permettent la définition de schémas volumes finis avec des erreurs de reconstruction d'ordre deux, trois et quatre. L'objectif du présent chapitre est d'évaluer et caractériser les erreurs numériques de ces schémas, c'est-à-dire l'écart entre la solution exacte et la solution approchée.

Il faut d'abord fixer l'équation modèle pour laquelle on évalue la précision des schémas. L'équation de convection linéaire à vitesse constante se révèle en pratique le modèle le plus simple qui permet de tester les propriétés fondamentales des schémas pour la dynamique des gaz. Les solutions exactes de cette équation sont d'ailleurs connues, ce qui permet de quantifier l'erreur et facilite aussi l'évaluation qualitative des solutions.

Une fois l'équation modèle fixée, il est nécessaire de choisir un critère de précision. On peut par exemple considérer l'ordre de troncature expliqué dans le chapitre 6. Ce critère caractérise le comportement de l'erreur numérique lorsque le diamètre des cellules tend vers zéro. Cette caractérisation est insuffisante car elle ne donne pas assez de renseignements sur des solutions obtenues en pratique.

Une meilleure évaluation peut se faire selon deux axes, l'un théorique, l'autre numérique.

(1) La théorie des différences finies en maillage cartésien utilise la technique de l'équation modifiée [**119, 79**]. Elle prédit une classification des erreurs numériques en erreurs sur l'amplitude, appelées *erreurs dissipatives*, et en erreurs sur la vitesse, appelées *erreurs dispersives*. Le but de l'étude théorique est de retrouver cette classification à l'aide de la méthode de l'équation modifiée. À la connaissance de l'auteur, cette méthode n'a pas encore été appliquée ni même définie dans le cas de maillages non structurés de type général. L'idée est de supposer l'existence d'une fonction régulière qui résout exactement l'équation approchée obtenue par une méthode de discrétisation spatiale. Un développement de Taylor permet d'établir pour cette fonction une équation d'évolution qui est la somme de l'équation initiale exacte et d'une série infinie de dérivées partielles spatiales. Une dérivée d'ordre pair correspond à une erreur dissipative alors qu'une dérivée d'ordre impair correspond à une erreur dispersive. Notons que la question difficile de la convergence de la série ne fait pas l'objet de cette étude. L'étude théorique est présentée dans la section 11.2.

(2) La connaissance des solutions exactes de l'équation de convection linéaire permet de calculer l'erreur numérique du schéma pour des conditions initiales arbitraires. Il est en particulier possible d'étudier l'évolution de l'erreur en fonction du diamètre du maillage. Il faut cependant ajouter que les tests numériques nécessitent une méthode d'intégration en temps qui doit être assez précise afin de ne pas ajouter trop d'erreur de discrétisation en temps. L'étude numérique fait l'objet de la section 11.3.

L'étude permettra également de comparer les erreurs en maillage uniforme cartésien et les erreurs en maillage non structuré pour une même méthode de discrétisation.

11.2. Définition de l'équation modifiée d'un schéma MUSCL semi-discret

L'approche de l'équation modifiée [**119, 79**] est une méthode qui fournit des renseignements qualitatifs sur les solutions d'un schéma numérique. Le principe est de supposer l'existence d'une

solution suffisamment régulière des équations du schéma et d'établir, par un développement asymptotique en h, une équation aux dérivées partielles pour cette solution. L'équation ainsi construite s'appelle *l'équation modifiée du schéma numérique* et sa forme permet de tirer des conclusions sur le comportement des solutions du schéma.

11.2.1. L'équation modifiée du schéma d'ordre un. Pour établir l'équation modifiée, on suppose l'existence d'une fonction $v : \Omega \times [0,T] \to \mathbb{R}$ suffisamment régulière dont les moyennes satisfont le schéma (10.2.7)

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \left\{ (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \overline{v}_\alpha(t) + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \overline{v}_\beta(t) \right\} . \tag{11.2.1}$$

On définit

$$\psi_\alpha^{(0)}(v) \triangleq \overline{v}_\alpha(t) - v(\boldsymbol{x}_\alpha, t) \tag{11.2.2}$$

$$\psi_\beta^{(0)}(v) \triangleq \overline{v}_\beta(t) - v(\boldsymbol{x}_\beta, t) . \tag{11.2.3}$$

La formule (5.8.2) permet de conclure que

$$\begin{aligned} \psi_\alpha^{(0)}(v) &= \frac{1}{2!} D^{(2)} v\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{x}_\alpha^{(2)} + \mathrm{O}\left(h^3\right) = \mathrm{O}\left(h^2\right) \\ \psi_\beta^{(0)}(v) &= \frac{1}{2!} D^{(2)} v\Big|_{\boldsymbol{x}_\beta} \bullet \boldsymbol{x}_\beta^{(2)} + \mathrm{O}\left(h^3\right) = \mathrm{O}\left(h^2\right) . \end{aligned}$$

La définition

$$\mathfrak{B}_\alpha^{(0)}(v) \triangleq \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \left\{ (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \psi_\alpha^{(0)}(v) + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \psi_\beta^{(0)}(v) \right\} \tag{11.2.4}$$

permet d'écrire (11.2.1) comme

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \left\{ (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ v(\boldsymbol{x}_\alpha, t) + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- v(\boldsymbol{x}_\beta, t) \right\} - \mathfrak{B}_\alpha^{(0)}(v) . \tag{11.2.5}$$

La définition

$$\mathfrak{H}_\alpha^{(0)}(v) \triangleq \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{ v(\boldsymbol{x}_\beta, t) - v(\boldsymbol{x}_\alpha, t) \right\} - \boldsymbol{c}\cdot D^{(1)} v\Big|_{\boldsymbol{x}_\alpha} \tag{11.2.6}$$

permet d'écrire (11.2.5) comme

$$\begin{aligned} \frac{d\overline{v}_\alpha(t)}{dt} = -\boldsymbol{c}\cdot D^{(1)} v\Big|_{\boldsymbol{x}_\alpha} - \mathfrak{B}_\alpha^{(0)}(v) - \mathfrak{H}_\alpha^{(0)}(v) - \\ - \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \left\{ (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ v(\boldsymbol{x}_\alpha, t) + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- v(\boldsymbol{x}_\alpha, t) \right\} . \end{aligned} \tag{11.2.7}$$

L'identité géométrique (5.5.2)

$$\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ + \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- = 0$$

permet de simplifier (11.2.7) pour donner

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\boldsymbol{c}\cdot D^{(1)} v\Big|_{\boldsymbol{x}_\alpha} - \mathfrak{H}_\alpha^{(0)}(v) - \mathfrak{B}_\alpha^{(0)}(v) . \tag{11.2.8}$$

La définition

$$\mathfrak{F}_\alpha^{(0)}(v) \triangleq \frac{d\overline{v}_\alpha(t)}{dt} - \frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} \tag{11.2.9}$$

permet d'énoncer la

PROPOSITION 11.2.1 (Équation modifiée du schéma d'ordre un). *L'équation modifiée associée au schéma (10.2.7) en maillage non structuré général est de la forme*

$$\left.\frac{\partial v}{\partial t}\right|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot \left.D^{(1)}v\right|_{\boldsymbol{x}_\alpha} = -\mathfrak{F}_\alpha^{(0)}(v) - \mathfrak{B}_\alpha^{(0)}(v) - \mathfrak{H}_\alpha^{(0)}(v) = \mathrm{O}(1)\ . \tag{11.2.10}$$

(i) *Le terme $\mathfrak{H}_\alpha^{(0)}(v)$, défini par (11.2.6), admet le développement asymptotique*

$$\mathfrak{H}_\alpha^{(0)}(v) = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \Big((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta} - (\boldsymbol{c}\cdot\boldsymbol{a}_{\beta\alpha})_+ \boldsymbol{k}_{\beta\alpha}\Big)\cdot \left.D^{(1)}v\right|_{\boldsymbol{x}_\alpha} + \\ + \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \frac{1}{2!}\boldsymbol{h}_{\alpha\beta}^2 \bullet \left.D^{(2)}v\right|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^2\right) = \mathrm{O}(1)\ . \tag{11.2.11}$$

(ii) *Le terme $\mathfrak{B}_\alpha^{(0)}(v)$, défini par (11.2.4), satisfait le développement asymptotique*

$$\mathfrak{B}_\alpha^{(0)}(v) \triangleq \frac{1}{2!\,|\mathcal{T}_\alpha|}\sum_\beta \left.D^{(2)}v\right|_{\boldsymbol{x}_\alpha} \bullet \Big((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{x}_\alpha^{(2)} + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{x}_\beta^{(2)}\Big) + \mathrm{O}\left(h^2\right) = \mathrm{O}(h)\ . \tag{11.2.12}$$

(iii) *Le terme $\mathfrak{F}_\alpha^{(0)}(v)$, défini par (11.2.9), admet le développement asymptotique*

$$\mathfrak{F}_\alpha^{(0)}(v) = \frac{1}{2!}\, D^{(2)}\left.\frac{\partial v}{\partial t}\right|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{x}_\alpha^{(2)} + \mathrm{O}\left(h^3\right) = \mathrm{O}\left(h^2\right)\ . \tag{11.2.13}$$

DÉMONSTRATION. Notons d'abord que
- $\boldsymbol{a}_{\alpha\beta} = \mathrm{O}\left(h^{d-1}\right)$
- $\boldsymbol{k}_{\alpha\beta} = \mathrm{O}(h)$
- $\boldsymbol{h}_{\alpha\beta} = \mathrm{O}(h)$
- $|\mathcal{T}_\alpha| = \mathrm{O}\left(h^d\right)$.

(i) Considérons le terme $\mathfrak{H}_\alpha^{(0)}(v)$. Le développement de Taylor de v au point $\boldsymbol{x}_\alpha$ donne

$$v(\boldsymbol{x}_\beta, t) - v(\boldsymbol{x}_\alpha, t) = \left.D^{(1)}v\right|_{\boldsymbol{x}_\alpha}\cdot \boldsymbol{h}_{\alpha\beta} + \frac{1}{2!}\left.D^{(2)}v\right|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{h}_{\alpha\beta}^2 + \mathrm{O}\left(h^3\right)\ .$$

L'identité (5.3.9) $\boldsymbol{h}_{\alpha\beta} = \boldsymbol{k}_{\alpha\beta} - \boldsymbol{k}_{\beta\alpha}$, l'antisymétrie (5.3.4) des vecteurs surface

$$(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- = (-\boldsymbol{c}\cdot\boldsymbol{a}_{\beta\alpha})_- = -(\boldsymbol{c}\cdot\boldsymbol{a}_{\beta\alpha})_+$$

et l'identité géométrique (5.9.8) permettent d'écrire

$$\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{h}_{\alpha\beta} = \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- (\boldsymbol{k}_{\alpha\beta} - \boldsymbol{k}_{\beta\alpha}) = \\ = \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- (\boldsymbol{k}_{\alpha\beta} - \boldsymbol{k}_{\beta\alpha}) + \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ (\boldsymbol{k}_{\alpha\beta} - \boldsymbol{k}_{\alpha\beta}) = \\ = |\mathcal{T}_\alpha|\,\boldsymbol{c} - \sum_\beta \Big\{(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta} - (\boldsymbol{c}\cdot\boldsymbol{a}_{\beta\alpha})_+ \boldsymbol{k}_{\beta\alpha}\Big\}\ .$$

Cette identité permet de développer $\mathfrak{H}_\alpha^{(0)}(v)$ comme

$$\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \Big\{\left.D^{(1)}v\right|_{\boldsymbol{x}_\alpha}\cdot\boldsymbol{h}_{\alpha\beta} + \frac{1}{2!}\left.D^{(2)}v\right|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{h}_{\alpha\beta}^2 + \mathrm{O}\left(h^3\right)\Big\} - \boldsymbol{c}\cdot\left.D^{(1)}v\right|_{\boldsymbol{x}_\alpha} = \\ = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \Big((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta} - (\boldsymbol{c}\cdot\boldsymbol{a}_{\beta\alpha})_+ \boldsymbol{k}_{\beta\alpha}\Big)\cdot\left.D^{(1)}v\right|_{\boldsymbol{x}_\alpha} + \\ + \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \frac{1}{2!}\boldsymbol{h}_{\alpha\beta}^2 \bullet \left.D^{(2)}v\right|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^2\right)$$

ce qui prouve (11.2.11). L'application des comparaisons asymptotiques de $\boldsymbol{k}_{\alpha\beta}$, $\boldsymbol{a}_{\alpha\beta}$ et $|\mathcal{T}_\alpha|$ au premier terme de (11.2.11) permet de conclure que

$$\mathfrak{H}_\alpha^{(0)}(v) = \mathrm{O}\left(\frac{h^{d-1}}{h^d}h\right) = \mathrm{O}(1)\,.$$

(ii) Le développement asymptotique de $\mathfrak{B}_\alpha^{(0)}(v)$ s'obtient directement par l'insertion des définitions (11.2.2) de $\psi_\alpha^{(0)}(v)$ et (11.2.3) de $\psi_\beta^{(0)}(v)$ dans (11.2.4). Il suffit de remplacer les dérivées de v en $\boldsymbol{x}_\beta$ à l'aide du développement

$$D^{(l)}v\Big|_{\boldsymbol{x}_\beta} = D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} + D^{(l+1)}v\Big|_{\boldsymbol{x}_\alpha}\cdot\boldsymbol{h}_{\alpha\beta} + \mathrm{O}\left(h^2\right)\,.$$

(iii) Le développement de $\mathfrak{F}_\alpha^{(0)}(v)$ vient de l'application du développement de Taylor (5.8.2) à la dérivée partielle en temps de v et de l'identité $\boldsymbol{x}_\alpha^{(1)} = 0$

$$\frac{d\overline{v}_\alpha(t)}{dt} = \frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha}\frac{\partial v(\boldsymbol{x},t)}{\partial t}\,dx = \frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \frac{1}{2!}\boldsymbol{x}_\alpha^{(2)}\bullet D^{(2)}\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^3\right) = \frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^2\right)\,.$$

□

La proposition 11.2.1 montre que la solution v de l'équation modifiée (11.2.10) satisfait une équation de convection-diffusion si les termes d'ordre $\mathrm{O}\left(h^2\right)$ sont négligés. Les termes responsables de la diffusion sont les termes de $\mathfrak{H}_\alpha^{(0)}(v)$ et de $\mathfrak{B}_\alpha^{(0)}(v)$ qui contiennent la dérivée seconde de v. Le premier terme de $\mathfrak{B}_\alpha^{(0)}(v)$

$$\frac{1}{2!\,|\mathcal{T}_\alpha|}\sum_\beta D^{(2)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+\,\boldsymbol{x}_\alpha^{(2)} + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_-\,\boldsymbol{x}_\beta^{(2)}\right) = \mathrm{O}(h)$$

est nul si les moments d'ordre deux des cellules sont égaux.

Il faut porter une attention particulière au deuxième terme de $\mathfrak{H}_\alpha^{(0)}(v)$ dans (11.2.11). Avec la définition du tenseur

$$\boldsymbol{\vartheta}_\alpha^{(2)} \triangleq \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_-\,\frac{1}{2!}\boldsymbol{h}_{\alpha\beta}^2\,, \tag{11.2.14}$$

le deuxième terme de $\mathfrak{H}_\alpha^{(0)}(v)$ s'écrit sous la forme

$$\boldsymbol{\vartheta}_\alpha^{(2)}\bullet D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} = \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_-\,\frac{1}{2!}\boldsymbol{h}_{\alpha\beta}^2\bullet D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} = \mathrm{O}(h)\,. \tag{11.2.15}$$

Il est important de noter que le tenseur (11.2.14) est semi-défini négatif et que l'expression (11.2.15) est en général non nulle. En maillage cartésien, on voit facilement que cela provoque un effet dissipatif dans l'équation modifiée (11.2.10). On s'attend à ce que le même effet se manifeste en maillage non structuré.

Une conclusion importante de la proposition 11.2.1 est que l'équation modifiée (11.2.10) présente une erreur $\mathrm{O}(1)$ qui provient du terme $\mathfrak{H}_\alpha^{(0)}(v)$. Cela montre que, *même si les erreurs d'ordre* $\mathrm{O}(h)$ *sont négligées*, la solution v du schéma (11.2.1) ne satisfait pas l'équation de convection linéaire au point $\boldsymbol{x}_\alpha$

$$\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = 0\,.$$

Par conséquent, les solutions numériques du schéma (10.2.7) présentent une erreur de vitesse qui est indépendante du diamètre h du maillage et qui ne diminue pas lorsque $h\to 0$.

Il est intéressant de déterminer les maillages sur lesquels cette erreur disparaît. L'erreur d'ordre $\mathrm{O}(1)$ dans (11.2.1) est nulle indépendamment de $\boldsymbol{c}$ et de v si et seulement si dans la cellule $\mathcal{T}_\alpha$

$$\sum_\beta(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+\,\boldsymbol{k}_{\alpha\beta} = \sum_\gamma(\boldsymbol{c}\cdot\boldsymbol{a}_{\gamma\alpha})_+\,\boldsymbol{k}_{\gamma\alpha}\ \text{pour tout}\ \boldsymbol{c}\in\mathbb{R}^d. \tag{11.2.16}$$

Une condition suffisante pour (11.2.16) est donnée par la

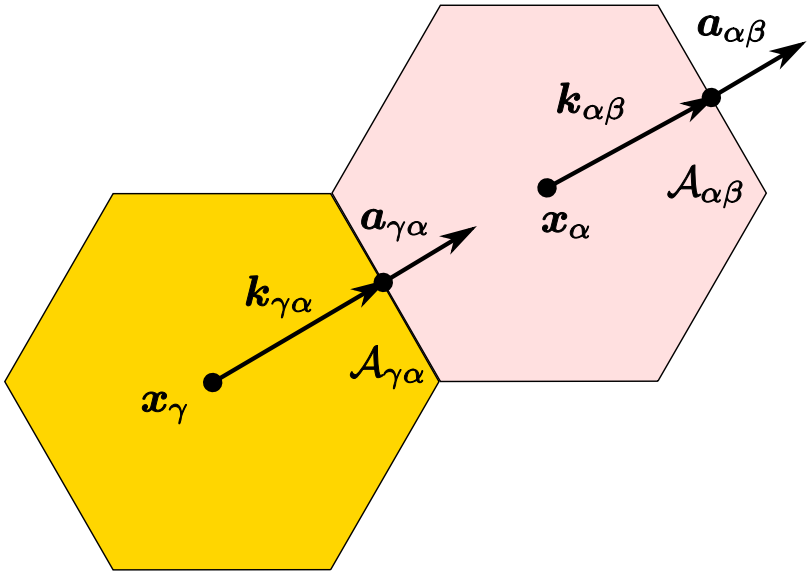


FIG. 11.2.1: Compensation géométrique des erreurs par symétrie entre les faces $\mathcal{A}_{\alpha\beta}$ et $\mathcal{A}_{\gamma\alpha}$.

PROPOSITION 11.2.2 (Compensation géométrique de l'erreur d'ordre O (1)). *La condition (11.2.16) est satisfaite dans la cellule $\mathcal{T}_\alpha$ si pour chaque face $\mathcal{A}_{\alpha\beta}$ de $\mathcal{T}_\alpha$ il existe une et une seule face $\mathcal{A}_{\alpha\gamma}$ telle que $\boldsymbol{k}_{\alpha\beta} = \boldsymbol{k}_{\gamma\alpha}$ et $\boldsymbol{a}_{\alpha\beta} = \boldsymbol{a}_{\gamma\alpha}$.*

DÉMONSTRATION. Si la condition de la proposition 11.2.2 est satisfaite, les sommes dans (11.2.16) sont égales. □

La proposition 11.2.2 implique le

COROLLAIRE 11.2.3. *La condition (11.2.16) est par exemple satisfaite dans les cas suivants :*

(i) *Les maillages cartésiens.*

(ii) *Les pavages du plan par des hexagones réguliers.*

(iii) *Les pavages de l'espace par des dodécaèdres réguliers.*

La figure 11.2.1 montre un exemple de cellules qui satisfont la proposition 11.2.2.

11.2.2. L'équation modifiée du schéma d'ordre deux. Cette section présente l'équation modifiée pour le schéma semi-discret (10.2.13) d'ordre deux. On suppose une fonction $v : \Omega \times [0,T] \to \mathbb{R}$ suffisamment régulière dont les moyennes $\mathfrak{v} = (\overline{v}_1, \ldots, \overline{v}_N)$ satisfont le schéma (10.2.13)

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \left\{\overline{v}_\alpha(t) + \boldsymbol{k}_{\alpha\beta}\cdot\sum_\gamma \boldsymbol{\sigma}_{\alpha\gamma}\overline{v}_\gamma(t)\right\} - \\ - \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{\overline{v}_\beta(t) + \boldsymbol{k}_{\beta\alpha}\cdot\sum_\gamma \boldsymbol{\sigma}_{\beta\gamma}\overline{v}_\gamma(t)\right\} \quad (11.2.17)$$

où

$$\boldsymbol{\sigma}_\alpha[\mathfrak{v}] = \sum_\gamma \boldsymbol{\sigma}_{\alpha\gamma}\overline{v}_\gamma = \sum_\gamma \boldsymbol{\sigma}_{\alpha\gamma}(\overline{v}_\gamma - \overline{v}_\alpha)$$

est une reconstruction consistante du gradient au sens de la définition 7.4.1. On définit l'erreur sur la dérivée de v par

$$\boldsymbol{\varepsilon}_{\alpha}^{(1)}(v) \triangleq \boldsymbol{\sigma}_{\alpha}[\mathfrak{v}] - D^{(1)}v\Big|_{\boldsymbol{x}_{\alpha}} \tag{11.2.18}$$

$$\boldsymbol{\varepsilon}_{\beta}^{(1)}(v) \triangleq \boldsymbol{\sigma}_{\beta}[\mathfrak{v}] - D^{(1)}v\Big|_{\boldsymbol{x}_{\beta}} . \tag{11.2.19}$$

D'après la définition 7.4.1, les erreurs d'approximation admettent le développement

$$\boldsymbol{\varepsilon}_{\alpha}^{(1)}(v) = \sum_{\gamma} \boldsymbol{\sigma}_{\alpha\gamma}\left[\frac{1}{2!}\, D^{(2)}v\Big|_{\boldsymbol{x}_{\alpha}} \bullet \boldsymbol{z}_{\alpha\gamma}^{(2)} + \mathrm{O}\left(h^3\right)\right] = \mathrm{O}(h)$$

$$\boldsymbol{\varepsilon}_{\beta}^{(1)}(v) = \sum_{\gamma} \boldsymbol{\sigma}_{\beta\gamma}\left[\frac{1}{2!}\, D^{(2)}v\Big|_{\boldsymbol{x}_{\beta}} \bullet \boldsymbol{z}_{\beta\gamma}^{(2)} + \mathrm{O}\left(h^3\right)\right] = \mathrm{O}(h)$$

où les tenseurs $\boldsymbol{z}_{\alpha\beta}^{(2)}$ sont définis par (5.7.10), $\boldsymbol{z}_{\alpha\beta}^{(2)} = \boldsymbol{h}_{\alpha\beta}^{2} + \boldsymbol{x}_{\beta}^{(2)}$.

L'insertion de (11.2.18) et (11.2.19) dans (11.2.17) donne

$$\frac{d\overline{v}_{\alpha}(t)}{dt} = -\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{+}\left\{\overline{v}_{\alpha}(t) + \boldsymbol{k}_{\alpha\beta}\cdot\left(D^{(1)}v\Big|_{\boldsymbol{x}_{\alpha}} + \boldsymbol{\varepsilon}_{\alpha}^{(1)}(v)\right)\right\} -$$
$$-\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{-}\left\{\overline{v}_{\beta}(t) + \boldsymbol{k}_{\beta\alpha}\cdot\left(D^{(1)}v\Big|_{\boldsymbol{x}_{\beta}} + \boldsymbol{\varepsilon}_{\beta}^{(1)}(v)\right)\right\} . \tag{11.2.20}$$

Pour simplifier l'écriture, on définit

$$\psi_{\alpha}^{(1)}(v) \triangleq \overline{v}_{\alpha}(t) - v(\boldsymbol{x}_{\alpha},t) - \boldsymbol{x}_{\alpha}^{(1)} \bullet D^{(1)}v\Big|_{\boldsymbol{x}_{\alpha}} \tag{11.2.21}$$

$$\psi_{\beta}^{(1)}(v) \triangleq \overline{v}_{\beta}(t) - v(\boldsymbol{x}_{\beta},t) - \boldsymbol{x}_{\beta}^{(1)} \bullet D^{(1)}v\Big|_{\boldsymbol{x}_{\beta}} . \tag{11.2.22}$$

L'identité (5.7.4), $\boldsymbol{x}_{\beta}^{(1)} = \boldsymbol{x}_{\alpha}^{(1)} = 0$, permet de simplifier (11.2.21) et (11.2.22)

$$\psi_{\alpha}^{(1)}(v) = \overline{v}_{\alpha}(t) - v(\boldsymbol{x}_{\alpha},t)$$
$$\psi_{\beta}^{(1)}(v) = \overline{v}_{\beta}(t) - v(\boldsymbol{x}_{\beta},t) .$$

La formule (5.8.2) permet de développer $\psi_{\alpha}^{(1)}(v)$ et $\psi_{\beta}^{(1)}(v)$ sous la forme

$$\psi_{\alpha}^{(1)}(v) = \frac{1}{2!}\, D^{(2)}v\Big|_{\boldsymbol{x}_{\alpha}} \bullet \boldsymbol{x}_{\alpha}^{(2)} + \frac{1}{3!}\, D^{(3)}v\Big|_{\boldsymbol{x}_{\alpha}} \bullet \boldsymbol{x}_{\alpha}^{(3)} + \mathrm{O}\left(h^4\right) = \mathrm{O}\left(h^2\right)$$

$$\psi_{\beta}^{(1)}(v) = \frac{1}{2!}\, D^{(2)}v\Big|_{\boldsymbol{x}_{\beta}} \bullet \boldsymbol{x}_{\beta}^{(2)} + \frac{1}{3!}\, D^{(3)}v\Big|_{\boldsymbol{x}_{\beta}} \bullet \boldsymbol{x}_{\beta}^{(3)} + \mathrm{O}\left(h^4\right) = \mathrm{O}\left(h^2\right) .$$

L'insertion de (11.2.21) et de (11.2.22) dans (11.2.20) donne

$$\frac{d\overline{v}_{\alpha}(t)}{dt} = -\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{+}\left\{v(\boldsymbol{x}_{\alpha},t) + \psi_{\alpha}^{(1)}(v) + \boldsymbol{k}_{\alpha\beta}\cdot\left(D^{(1)}v\Big|_{\boldsymbol{x}_{\alpha}} + \boldsymbol{\varepsilon}_{\alpha}^{(1)}(v)\right)\right\} -$$
$$-\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{-}\left\{v(\boldsymbol{x}_{\beta},t) + \psi_{\beta}^{(1)}(v) + \boldsymbol{k}_{\beta\alpha}\cdot\left(D^{(1)}v\Big|_{\boldsymbol{x}_{\beta}} + \boldsymbol{\varepsilon}_{\beta}^{(1)}(v)\right)\right\} . \tag{11.2.23}$$

Les définitions

$$\mathfrak{E}_{\alpha}^{(1)}(v) \triangleq \frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}\left\{(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{+}\,\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{\varepsilon}_{\alpha}^{(1)}(v) + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{-}\,\boldsymbol{k}_{\beta\alpha}\cdot\boldsymbol{\varepsilon}_{\beta}^{(1)}(v)\right\} \tag{11.2.24}$$

et

$$\mathfrak{B}_{\alpha}^{(1)}(v) \triangleq \frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}\left\{(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{+}\,\psi_{\alpha}^{(1)}(v) + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{-}\,\psi_{\beta}^{(1)}(v)\right\} \tag{11.2.25}$$

permettent d'écrire (11.2.23) sous la forme

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \left\{ v(\boldsymbol{x}_\alpha,t) + \boldsymbol{k}_{\alpha\beta}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha}\right\} - \\ - \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{ v(\boldsymbol{x}_\beta,t) + \boldsymbol{k}_{\beta\alpha}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\beta}\right\} - \mathfrak{E}_\alpha^{(1)}(v) - \mathfrak{B}_\alpha^{(1)}(v)\ . \quad (11.2.26)$$

La définition

$$\mathfrak{H}_\alpha^{(1)}(v) \triangleq \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{ v(\boldsymbol{x}_\beta,t) + \boldsymbol{k}_{\beta\alpha}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\beta} - v(\boldsymbol{x}_\alpha,t) - D^{(1)}v\Big|_{\boldsymbol{x}_\alpha}\cdot\boldsymbol{k}_{\alpha\beta}\right\} \quad (11.2.27)$$

permet d'écrire (11.2.26) sous la forme

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \left\{ v(\boldsymbol{x}_\alpha,t) + \boldsymbol{k}_{\alpha\beta}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha}\right\} - \\ - \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{ v(\boldsymbol{x}_\alpha,t) + \boldsymbol{k}_{\alpha\beta}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha}\right\} - \mathfrak{E}_\alpha^{(1)}(v) - \mathfrak{B}_\alpha^{(1)}(v) - \mathfrak{H}_\alpha^{(1)}(v)\ . \quad (11.2.28)$$

L'identité (5.5.2) entraîne

$$-\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ v(\boldsymbol{x}_\alpha,t) - \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- v(\boldsymbol{x}_\alpha,t) = 0 \quad (11.2.29)$$

et l'identité (5.9.8) permet la simplification

$$-\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ D^{(1)}v\Big|_{\boldsymbol{x}_\alpha}\cdot\boldsymbol{k}_{\alpha\beta} - \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- D^{(1)}v\Big|_{\boldsymbol{x}_\alpha}\cdot\boldsymbol{k}_{\alpha\beta} = -\boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha}\ . \quad (11.2.30)$$

Les identités (11.2.29), (11.2.30) et la définition

$$\mathfrak{F}_\alpha^{(1)}(v) \triangleq \frac{d\overline{v}_\alpha(t)}{dt} - \frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} \quad (11.2.31)$$

permettent de formuler la

PROPOSITION 11.2.4 (Équation modifiée pour le schéma d'ordre deux). *L'équation modifiée associée au schéma (10.2.13) en maillage non structuré général est*

$$\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = -\mathfrak{E}_\alpha^{(1)}(v) - \mathfrak{F}_\alpha^{(1)}(v) - \mathfrak{B}_\alpha^{(1)}(v) - \mathfrak{H}_\alpha^{(1)}(v) = \mathrm{O}(h)\ . \quad (11.2.32)$$

(i) *Le terme $\mathfrak{H}_\alpha^{(1)}(v)$, défini par (11.2.27), admet le développement*

$$\mathfrak{H}_\alpha^{(1)}(v) = \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{ \frac{1}{2!}D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} \bullet \left(\boldsymbol{h}_{\alpha\beta}^2 + 2\boldsymbol{k}_{\beta\alpha}\odot\boldsymbol{h}_{\alpha\beta}\right) + \right. \\ \left. + \frac{1}{3!}D^{(3)}v\Big|_{\boldsymbol{x}_\alpha} \bullet \left(\boldsymbol{h}_{\alpha\beta}^3 + 3\boldsymbol{k}_{\beta\alpha}\odot\boldsymbol{h}_{\alpha\beta}^2\right)\right\} + \mathrm{O}\left(h^3\right) = \mathrm{O}(h)\ . \quad (11.2.33)$$

(ii) *Le terme $\mathfrak{B}_\alpha^{(1)}(v)$, défini par (11.2.25), admet le développement*

$$\mathfrak{B}_\alpha^{(1)}(v) = \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \left\{ \frac{1}{2}\left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{x}_\alpha^{(2)} + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{x}_\beta^{(2)}\right) \bullet D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} + \right. \\ \left. + \frac{1}{3}\left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{x}_\alpha^{(3)} + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left(\boldsymbol{x}_\beta^{(3)} + \frac{3}{2}\boldsymbol{h}_{\alpha\beta}\odot\boldsymbol{x}_\beta^{(2)}\right)\right) \bullet D^{(3)}v\Big|_{\boldsymbol{x}_\alpha}\right\} + \mathrm{O}\left(h^3\right) = \mathrm{O}(h)\ . \quad (11.2.34)$$

(iii) *Le terme* $\mathfrak{F}_\alpha^{(1)}(v)$*, défini par (11.2.31), admet le développement asymptotique*

$$\mathfrak{F}_\alpha^{(1)}(v) = \frac{1}{2!}\, D^{(3)}v\Big|_{\boldsymbol{x}_\alpha} \bullet \left(\boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(2)}\right) + \mathrm{O}\left(h^3\right) = \mathrm{O}\left(h^2\right) . \tag{11.2.35}$$

(iv) *Le terme* $\mathfrak{E}_\alpha^{(1)}(v)$*, défini par (11.2.24), regroupe les erreurs engendrées par la reconstruction polynomiale et admet un développement*

$$\mathfrak{E}_\alpha^{(1)}(v) = \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta \sum_\gamma \Big\{ (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ (\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{\sigma}_{\alpha\gamma})\, \boldsymbol{z}_{\alpha\gamma}^{(2)} + \\ + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- (\boldsymbol{k}_{\beta\alpha}\cdot\boldsymbol{\sigma}_{\beta\gamma})\, \boldsymbol{z}_{\beta\gamma}^{(2)} \Big\} \bullet D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^2\right) = \mathrm{O}(h) . \tag{11.2.36}$$

Démonstration. La dérivée de v en $\boldsymbol{x}_\beta$ peut être développée autour de $\boldsymbol{x}_\alpha$, ce qui donne, avec le produit tensoriel symétrique (5.2.7), les développements asymptotiques

$$v(\boldsymbol{x}_\beta, t) = v(\boldsymbol{x}_\alpha, t) + \boldsymbol{h}_{\alpha\beta}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} + \frac{1}{2!}\boldsymbol{h}_{\alpha\beta}^2 \bullet D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} + \frac{1}{3!}\boldsymbol{h}_{\alpha\beta}^3 \bullet D^{(3)}v\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^3\right) \tag{11.2.37}$$

et

$$\boldsymbol{k}_{\beta\alpha}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\beta} = \boldsymbol{k}_{\beta\alpha}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} + (\boldsymbol{k}_{\beta\alpha}\odot\boldsymbol{h}_{\alpha\beta}) \bullet D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} + \\ + \frac{1}{2!}\left(\boldsymbol{k}_{\beta\alpha}\odot\boldsymbol{h}_{\alpha\beta}^2\right) \bullet D^{(3)}v\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^3\right) . \tag{11.2.38}$$

L'identité (5.3.9), $\boldsymbol{k}_{\alpha\beta} = \boldsymbol{h}_{\alpha\beta} + \boldsymbol{k}_{\beta\alpha}$, simplifie la somme de (11.2.37) et de (11.2.38). Cela permet de développer l'expression dans la définition (11.2.27) de $\mathfrak{H}_\alpha^{(1)}(v)$ comme

$$v(\boldsymbol{x}_\beta, t) + \boldsymbol{k}_{\beta\alpha}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\beta} - v(\boldsymbol{x}_\alpha, t) - \boldsymbol{k}_{\alpha\beta}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = \\ = \frac{1}{2}\, D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} \bullet \left(\boldsymbol{h}_{\alpha\beta}^2 + 2\boldsymbol{k}_{\beta\alpha}\odot\boldsymbol{h}_{\alpha\beta}\right) + \frac{1}{3!}\left(\boldsymbol{h}_{\alpha\beta}^3 + 3\boldsymbol{k}_{\beta\alpha}\odot\boldsymbol{h}_{\alpha\beta}^2\right) \bullet D^{(3)}v\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^3\right) . \tag{11.2.39}$$

L'insertion de ce développement dans la définition (11.2.27) de $\mathfrak{H}_\alpha^{(1)}(v)$ prouve (11.2.33). Les termes entre les accolades de (11.2.33) sont $\mathrm{O}\left(h^2\right)$. La multiplication par $\boldsymbol{a}_{\alpha\beta} = \mathrm{O}\left(h^{d-1}\right)$ et la division par $|\mathcal{T}_\alpha| = \mathrm{O}\left(h^d\right)$ entraîne

$$\mathfrak{H}_\alpha^{(1)}(v) = \mathrm{O}(h)$$

ce qui démontre le point (i) de la proposition 11.2.4.

Le développement asymptotique de $\mathfrak{B}_\alpha^{(1)}(v)$ s'obtient directement par l'insertion des définitions (11.2.21) de $\psi_\alpha^{(1)}(v)$ et (11.2.22) de $\psi_\beta^{(1)}(v)$ dans (11.2.25). Il suffit de remplacer les dérivées de v en $\boldsymbol{x}_\beta$ à l'aide du développement

$$D^{(l)}v\Big|_{\boldsymbol{x}_\beta} = D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} + D^{(l+1)}v\Big|_{\boldsymbol{x}_\alpha} \cdot \boldsymbol{h}_{\alpha\beta} + \mathrm{O}\left(h^2\right)$$

ce qui prouve le point (ii) de la proposition 11.2.4.

De façon analogue, le développement asymptotique de $\mathfrak{E}_\alpha^{(1)}(v)$ s'obtient directement par l'insertion des définitions (11.2.18) de $\varepsilon_\alpha^{(1)}(v)$ et (11.2.19) de $\varepsilon_\beta^{(1)}(v)$ dans (11.2.24). Cela démontre le point (iv) de la proposition 11.2.4.

Le développement (11.2.35) de $\mathfrak{F}_\alpha^{(1)}(v)$ vient de l'application du développement de Taylor (5.8.2) à la dérivée partielle en temps de v, ce qui donne avec l'identité $\boldsymbol{x}_\alpha^{(1)} = 0$

$$\frac{d\overline{v}_\alpha(t)}{dt} = \frac{1}{|\mathcal{T}_\alpha|}\int_{\mathcal{T}_\alpha} \frac{\partial v(\boldsymbol{x}, t)}{\partial t}\, dx = \frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \frac{1}{2!}\boldsymbol{x}_\alpha^{(2)} \bullet D^{(2)}\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^3\right) . \tag{11.2.40}$$

La combinaison des développements (11.2.33), (11.2.34), (11.2.36) et (11.2.40) montre que

$$\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = \mathrm{O}(h) . \tag{11.2.41}$$

La comparaison asymptotique (11.2.41) permet de remplacer la dérivée partielle en temps de v dans (11.2.40) par la dérivée spatiale. Cela démontre le point (iii) de la proposition 11.2.4. □

L'identité (11.2.32) de la proposition 11.2.4 montre que les solutions v du schéma (10.2.13) satisfont une équation de convection-diffusion si les termes contenant des dérivées d'ordre trois et supérieur sont négligés. Contrairement au schéma d'ordre un, cf. la proposition 11.2.1, le schéma d'ordre deux ne présente pas d'erreur $\mathrm{O}(1)$ sur la vitesse de convection. Les premiers termes d'erreurs sont $\mathrm{O}(h)$ et dépendent linéairement de la dérivée seconde de v.

Il est intéressant d'examiner les conditions dans lesquelles ces termes $\mathrm{O}(h)$ sont nuls. Le terme d'erreur $\mathfrak{E}_\alpha^{(1)}(v)$ provient de la reconstruction du gradient. Si la reconstruction du gradient est de précision à l'ordre deux au sens de la définition 7.4.1, la reconstruction du gradient satisfait la condition de précision à l'ordre deux (8.3.8)

$$\sum_\gamma \boldsymbol{\sigma}_{\alpha\gamma} \otimes \boldsymbol{z}_{\alpha\gamma}^{(2)} = 0$$

et le premier terme dans le développement (11.2.36) de $\mathfrak{E}_\alpha^{(1)}(v)$ devient nul. Dans ce cas, le développement (11.2.36) de $\mathfrak{E}_\alpha^{(1)}(v)$ ne contient que les dérivées d'ordre au moins trois de v.

Les termes d'erreur $\mathfrak{B}_\alpha^{(1)}(v)$ et $\mathfrak{H}_\alpha^{(1)}(v)$ sont d'origine géométrique et indépendants de la reconstruction du gradient. L'identité géométrique (5.5.2) montre que le premier terme de

$$\mathfrak{B}_\alpha^{(1)}(v) = \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta \left\{ \frac{1}{2} \left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{x}_\alpha^{(2)} + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{x}_\beta^{(2)} \right) \bullet D^{(2)} v \Big|_{\boldsymbol{x}_\alpha} \right\} + \mathrm{O}\left(h^2\right)$$

est nul si les moments d'ordre deux des cellules sont égaux $\boldsymbol{x}_\alpha^{(2)} = \boldsymbol{x}_\beta^{(2)}$. Examinons le premier terme du développement

$$\mathfrak{H}_\alpha^{(1)}(v) = \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \frac{1}{2} \left(\boldsymbol{h}_{\alpha\beta}^2 + 2\boldsymbol{k}_{\beta\alpha} \odot \boldsymbol{h}_{\alpha\beta} \right) \bullet D^{(2)} v \Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^2\right). \tag{11.2.42}$$

La formule (5.7.8) permet de démontrer l'identité

$$\boldsymbol{h}_{\alpha\beta}^2 + 2\boldsymbol{k}_{\beta\alpha} \odot \boldsymbol{h}_{\alpha\beta} = \boldsymbol{h}_{\alpha\beta}^2 + 2\boldsymbol{k}_{\beta\alpha} \odot \boldsymbol{h}_{\alpha\beta} + \boldsymbol{k}_{\beta\alpha}^{(2)} - \boldsymbol{k}_{\beta\alpha}^{(2)} = \boldsymbol{k}_{\alpha\beta}^{(2)} - \boldsymbol{k}_{\beta\alpha}^{(2)}. \tag{11.2.43}$$

L'identité (5.9.11) donne la relation

$$\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(2)} + \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{k}_{\alpha\beta}^{(2)} = 0$$

ce qui entraîne

$$\begin{aligned} \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left(\boldsymbol{h}_{\alpha\beta}^2 + 2\boldsymbol{k}_{\beta\alpha} \odot \boldsymbol{h}_{\alpha\beta} \right) &= \sum_\beta \left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{k}_{\alpha\beta}^{(2)} - (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{k}_{\beta\alpha}^{(2)} \right) = \\ &= -\sum_\beta \left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(2)} + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{k}_{\beta\alpha}^{(2)} \right) = -\sum_\beta \left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(2)} - (\boldsymbol{c}\cdot\boldsymbol{a}_{\beta\alpha})_+ \boldsymbol{k}_{\beta\alpha}^{(2)} \right). \end{aligned} \tag{11.2.44}$$

La distributivité du produit $\odot$ permet d'écrire

$$\boldsymbol{h}_{\alpha\beta}^2 + 2\boldsymbol{k}_{\beta\alpha} \odot \boldsymbol{h}_{\alpha\beta} = \boldsymbol{h}_{\alpha\beta} \odot \boldsymbol{h}_{\alpha\beta} + 2\boldsymbol{k}_{\beta\alpha} \odot \boldsymbol{h}_{\alpha\beta} = (\boldsymbol{h}_{\alpha\beta} + 2\boldsymbol{k}_{\beta\alpha}) \odot \boldsymbol{h}_{\alpha\beta}. \tag{11.2.45}$$

La combinaison des identités (11.2.44) et (11.2.45) permet de démontrer l'identité

$$\begin{aligned} \mathfrak{H}_\alpha^{(1)} &= \frac{1}{2\,|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left((\boldsymbol{h}_{\alpha\beta} + 2\boldsymbol{k}_{\beta\alpha}) \odot \boldsymbol{h}_{\alpha\beta} \right) \bullet D^{(2)} v \Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^2\right) = \\ &= -\frac{1}{2\,|\mathcal{T}_\alpha|} \sum_\beta \left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(2)} - (\boldsymbol{c}\cdot\boldsymbol{a}_{\beta\alpha})_+ \boldsymbol{k}_{\beta\alpha}^{(2)} \right) \bullet D^{(2)} v \Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^2\right). \end{aligned} \tag{11.2.46}$$

Il est intéressant de déterminer les maillages sur lesquels le premier terme d'erreur de (11.2.46) disparaît. L'erreur d'ordre $\mathrm{O}(h)$ dans (11.2.1) est nulle indépendamment de $\boldsymbol{c}$ et de v si et seulement si dans la cellule $\mathcal{T}_\alpha$

$$\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}^{(2)}_{\alpha\beta} = \sum_\gamma (\boldsymbol{c}\cdot\boldsymbol{a}_{\gamma\alpha})_+ \boldsymbol{k}^{(2)}_{\gamma\alpha} \text{ pour tout } \boldsymbol{c}\in\mathbb{R}^d. \tag{11.2.47}$$

Deux conditions suffisantes pour (11.2.47) sont fournies par la

PROPOSITION 11.2.5 (Compensation géométrique de l'erreur d'ordre $\mathrm{O}(h)$). *La condition (11.2.47) est satisfaite dans la cellule $\mathcal{T}_\alpha$ si l'une des deux conditions est vraie :*

(i) *Pour chaque face $\mathcal{A}_{\alpha\beta}$ de $\mathcal{T}_\alpha$ il existe une et une seule face $\mathcal{A}_{\alpha\gamma}$ telle que $\boldsymbol{k}^{(2)}_{\alpha\beta} = \boldsymbol{k}^{(2)}_{\gamma\alpha}$ et $\boldsymbol{a}_{\alpha\beta} = \boldsymbol{a}_{\gamma\alpha}$.*

(ii) *Le barycentre $\boldsymbol{x}_{\alpha\beta}$ de chaque face $\mathcal{A}_{\alpha\beta}$ de $\mathcal{T}_\alpha$ se situe exactement au milieu du segment joignant les barycentres des cellules $\mathcal{T}_\alpha$ et $\mathcal{T}_\beta$. Cela s'exprime par la condition géométrique*

$$\boldsymbol{h}_{\alpha\beta} = \frac{1}{2}\boldsymbol{k}_{\alpha\beta} = -\frac{1}{2}\boldsymbol{k}_{\beta\alpha}\,. \tag{11.2.48}$$

DÉMONSTRATION. Si la première condition est satisfaite, les sommes dans (11.2.47) sont égales. Si la deuxième condition est satisfaite, l'identité (11.2.44) montre que la condition (11.2.47) est vérifiée. □

La discussion ci-dessus démontre la

PROPOSITION 11.2.6 (Compensation des erreurs d'ordre $\mathrm{O}(h)$). *Les termes d'erreur d'ordre $\mathrm{O}(h)$ dans le membre de droite de (11.2.32) sont nuls si les conditions suivantes sont satisfaites :*

(i) *La reconstruction du gradient est de précision à l'ordre deux au sens de la définition 7.4.1.*

(ii) *Les cellules ont les mêmes moments d'ordre deux $\boldsymbol{x}^{(2)}_\alpha$.*

(iii) *L'une des conditions géométriques de la proposition 11.2.5 est satisfaite.*

11.2.3. L'équation modifiée pour le schéma d'ordre élevé. Dans le cas d'une reconstruction par des polynômes de degré k, le schéma semi-discret pour la convection linéaire est donné par (10.2.19). On suppose ici l'existence d'une fonction v dont les moyennes $\mathfrak{v}(t) = (\overline{v}_1(t),\dots,\overline{v}_N(t))$ satisfont le système dynamique

$$\begin{aligned}\frac{d\overline{v}_\alpha(t)}{dt} = &-\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \left\{\overline{v}_\alpha(t) + \sum_{l=1}^{k}\frac{1}{l!}\left(\boldsymbol{k}^{(l)}_{\alpha\beta} - \boldsymbol{x}^{(l)}_\alpha\right)\bullet \boldsymbol{w}^{(l)}_\alpha[\mathfrak{v}(t)]\right\} - \\ &-\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \left\{\overline{v}_\beta(t) + \sum_{l=1}^{k}\frac{1}{l!}\left(\boldsymbol{k}^{(l)}_{\beta\alpha} - \boldsymbol{x}^{(l)}_\beta\right)\bullet \boldsymbol{w}^{(l)}_\beta[\mathfrak{v}(t)]\right\}.\end{aligned} \tag{11.2.49}$$

Dans la suite, on suppose la fonction v suffisamment régulière pour tous les développements asymptotiques nécessaires. Les tenseurs $\boldsymbol{w}^{(l)}_\beta[\mathfrak{v}]$ sont des dérivées consistantes d'ordre l et de précision à l'ordre $k+1-l$ au sens de la définition 7.4.1. Les erreurs de reconstruction de $\boldsymbol{w}^{(l)}_\alpha[\mathfrak{v}]$ et de $\boldsymbol{w}^{(l)}_\beta[\mathfrak{v}]$ sont définies par

$$\boldsymbol{\varepsilon}^{(l)}_\alpha(v) \triangleq \boldsymbol{w}^{(l)}_\alpha[\mathfrak{v}] - D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} \tag{11.2.50}$$

$$\boldsymbol{\varepsilon}^{(l)}_\beta(v) \triangleq \boldsymbol{w}^{(l)}_\beta[\mathfrak{v}] - D^{(l)}v\Big|_{\boldsymbol{x}_\beta}\,. \tag{11.2.51}$$

Sous l'hypothèse de la régularité, elles satisfont le développement asymptotique

$$\begin{aligned}\boldsymbol{\varepsilon}_{\alpha}^{(l)}(v) &= \sum_{\gamma}\frac{\boldsymbol{w}_{\alpha\gamma}^{(l)}}{(k+1)!}\,D^{(k+1)}v\Big|_{\boldsymbol{x}_{\alpha}}\bullet\boldsymbol{z}_{\alpha\gamma}^{(k+1)}+\mathrm{O}\left(h^{k-l+2}\right)=\mathrm{O}\left(h^{k-l+1}\right)\\ \boldsymbol{\varepsilon}_{\beta}^{(l)}(v) &= \sum_{\gamma}\frac{\boldsymbol{w}_{\beta\gamma}^{(l)}}{(k+1)!}\,D^{(k+1)}v\Big|_{\boldsymbol{x}_{\beta}}\bullet\boldsymbol{z}_{\beta\gamma}^{(k+1)}+\mathrm{O}\left(h^{k-l+2}\right)=\mathrm{O}\left(h^{k-l+1}\right)\end{aligned}$$

où les tenseurs $\boldsymbol{z}_{\alpha\beta}^{(k)}$ sont définis par (5.7.10)

$$\boldsymbol{z}_{\alpha\beta}^{(k)}\triangleq\sum_{l=0}^{k}\binom{k}{l}\boldsymbol{y}_{\alpha\beta}^{(k-l|l)}=\sum_{l=0}^{k}\binom{k}{l}\boldsymbol{h}_{\alpha\beta}^{k-l}\odot\boldsymbol{x}_{\beta}^{(l)}\,.$$

L'insertion de (11.2.50) et de (11.2.51) dans (11.2.49) donne

$$\begin{aligned}\frac{d\overline{v}_{\alpha}(t)}{dt}=&-\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{+}\left\{\overline{v}_{\alpha}(t)+\sum_{l=1}^{k}\frac{1}{l!}\left(\boldsymbol{k}_{\alpha\beta}^{(l)}-\boldsymbol{x}_{\alpha}^{(l)}\right)\bullet\left(D^{(l)}v\Big|_{\boldsymbol{x}_{\alpha}}+\boldsymbol{\varepsilon}_{\alpha}^{(l)}(v)\right)\right\}-\\&-\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{-}\left\{\overline{v}_{\beta}(t)+\sum_{l=1}^{k}\frac{1}{l!}\left(\boldsymbol{k}_{\beta\alpha}^{(l)}-\boldsymbol{x}_{\beta}^{(l)}\right)\bullet\left(D^{(l)}v\Big|_{\boldsymbol{x}_{\beta}}+\boldsymbol{\varepsilon}_{\beta}^{(l)}(v)\right)\right\}.\end{aligned}\tag{11.2.52}$$

Pour simplifier l'écriture, on définit

$$\psi_{\alpha}^{(k)}(v)\triangleq\overline{v}_{\alpha}(t)-\sum_{i=0}^{k}\frac{1}{i!}\,D^{(i)}v\Big|_{\boldsymbol{x}_{\alpha}}\bullet\boldsymbol{x}_{\alpha}^{(i)}\tag{11.2.53}$$

$$\psi_{\beta}^{(k)}(v)\triangleq\overline{v}_{\beta}(t)-\sum_{i=0}^{k}\frac{1}{i!}\,D^{(i)}v\Big|_{\boldsymbol{x}_{\beta}}\bullet\boldsymbol{x}_{\beta}^{(i)}\,.\tag{11.2.54}$$

La formule (5.8.2) permet de développer $\psi_{\alpha}^{(k)}(v)$ et $\psi_{\beta}^{(k)}(v)$ sous la forme

$$\begin{aligned}\psi_{\alpha}^{(k)}(v) &= \frac{1}{(k+1)!}\,D^{(k+1)}v\Big|_{\boldsymbol{x}_{\alpha}}\bullet\boldsymbol{x}_{\alpha}^{(k+1)}+\mathrm{O}\left(h^{k+2}\right)=\mathrm{O}\left(h^{k+1}\right)\\ \psi_{\beta}^{(k)}(v) &= \frac{1}{(k+1)!}\,D^{(k+1)}v\Big|_{\boldsymbol{x}_{\beta}}\bullet\boldsymbol{x}_{\beta}^{(k+1)}+\mathrm{O}\left(h^{k+2}\right)=\mathrm{O}\left(h^{k+1}\right).\end{aligned}$$

L'insertion de (11.2.53) et de (11.2.54) dans (11.2.52) donne

$$\begin{aligned}\frac{d\overline{v}_{\alpha}(t)}{dt}=&-\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{+}\left\{\sum_{l=0}^{k}\frac{1}{l!}\left[\boldsymbol{k}_{\alpha\beta}^{(l)}\bullet D^{(l)}v\Big|_{\boldsymbol{x}_{\alpha}}+\left(\boldsymbol{k}_{\alpha\beta}^{(l)}-\boldsymbol{x}_{\alpha}^{(l)}\right)\bullet\boldsymbol{\varepsilon}_{\alpha}^{(l)}(v)\right]+\psi_{\alpha}^{(k)}(v)\right\}-\\&-\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{-}\left\{\sum_{l=0}^{k}\frac{1}{l!}\left[\boldsymbol{k}_{\beta\alpha}^{(l)}\bullet D^{(l)}v\Big|_{\boldsymbol{x}_{\beta}}+\left(\boldsymbol{k}_{\beta\alpha}^{(l)}-\boldsymbol{x}_{\beta}^{(l)}\right)\bullet\boldsymbol{\varepsilon}_{\beta}^{(l)}(v)\right]+\psi_{\beta}^{(k)}(v)\right\}.\end{aligned}\tag{11.2.55}$$

Pour simplifier (11.2.55), on introduit les définitions

$$\begin{aligned}\mathfrak{E}_{\alpha}^{(k)}(v)\triangleq\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}\sum_{l=0}^{k}\frac{1}{l!}\Big\{&(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{+}\left(\boldsymbol{k}_{\alpha\beta}^{(l)}-\boldsymbol{x}_{\alpha}^{(l)}\right)\bullet\boldsymbol{\varepsilon}_{\alpha}^{(l)}(v)+\\&+(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{-}\left(\boldsymbol{k}_{\beta\alpha}^{(l)}-\boldsymbol{x}_{\beta}^{(l)}\right)\bullet\boldsymbol{\varepsilon}_{\beta}^{(l)}(v)\Big\}\end{aligned}\tag{11.2.56}$$

et

$$\mathfrak{B}_{\alpha}^{(k)}(v)\triangleq\frac{1}{|\mathcal{T}_{\alpha}|}\sum_{\beta}\left\{(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{+}\,\psi_{\alpha}^{(k)}(v)+(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_{-}\,\psi_{\beta}^{(k)}(v)\right\}.\tag{11.2.57}$$

Les termes entre les accolades de (11.2.56) et (11.2.57) sont d'ordre $O\left(h^{k+1}\right)$. Comme ces termes sont multipliés par $\boldsymbol{a}_{\alpha\beta} = \mathrm{O}\left(h^{d-1}\right)$ et divisés par $|\mathcal{T}_\alpha| = \mathrm{O}\left(h^d\right)$, leur contribution au développement asymptotique est a priori

$$\mathfrak{E}_\alpha^{(k)}(v) = O\left(h^k\right) \text{ et } \mathfrak{B}_\alpha^{(k)}(v) = O\left(h^k\right). \tag{11.2.58}$$

Les définitions (11.2.56) et (11.2.57) permettent d'écrire (11.2.55) sous la forme

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\mathfrak{E}_\alpha^{(k)}(v) - \mathfrak{B}_\alpha^{(k)}(v) - \\ - \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \sum_{l=0}^{k} \frac{1}{l!}\left\{(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{k}_{\beta\alpha}^{(l)} \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\beta}\right\}. \tag{11.2.59}$$

La définition

$$\mathfrak{H}_\alpha^{(k)}(v) \triangleq \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \sum_{l=0}^{k}\frac{1}{l!}\left\{\boldsymbol{k}_{\beta\alpha}^{(l)} \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\beta} - \boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha}\right\} - \\ - \frac{1}{k!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(k)}\right) \bullet D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha} \tag{11.2.60}$$

permet de remplacer dans l'équation (11.2.55) les dérivées de v en $\boldsymbol{x}_\beta$ par celles en $\boldsymbol{x}_\alpha$

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\mathfrak{E}_\alpha^{(k)}(v) - \mathfrak{B}_\alpha^{(k)}(v) - \mathfrak{H}_\alpha^{(k)}(v) - \frac{1}{k!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(k)}\right) \bullet D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha} - \\ - \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta\sum_{l=0}^{k}\frac{1}{l!}\left((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(l)} + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- \boldsymbol{k}_{\alpha\beta}^{(l)}\right) \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha}. \tag{11.2.61}$$

L'identité $(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ + (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_- = \boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}$ permet d'écrire (11.2.61) sous la forme

$$\frac{d\overline{v}_\alpha(t)}{dt} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}\sum_{l=0}^{k}\frac{1}{l!}\boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} - \frac{1}{k!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(k)}\right) \bullet D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha} - \\ - \mathfrak{E}_\alpha^{(k)}(v) - \mathfrak{B}_\alpha^{(k)}(v) - \mathfrak{H}_\alpha^{(k)}(v). \tag{11.2.62}$$

La somme dans le membre de droite de (11.2.62) s'écrit

$$\sum_\beta \boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}\sum_{l=0}^{k}\frac{1}{l!}\boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} = \\ = \sum_\beta \boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta}\left\{v(\boldsymbol{x}_\alpha) + \boldsymbol{k}_{\alpha\beta}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} + \sum_{l=2}^{k}\frac{1}{l!}\boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha}\right\}. \tag{11.2.63}$$

Il est possible de simplifier les termes entre les accolades dans la deuxième ligne de (11.2.63). Le premier terme est nul en raison de l'identité (5.5.2). Le deuxième terme devient, grâce à l'identité (5.9.7),

$$\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})\,\boldsymbol{k}_{\alpha\beta}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha}. \tag{11.2.64}$$

Le troisième terme est une somme dont les termes peuvent être reformulés grâce à la proposition 5.9.2

$$\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})\,\frac{1}{l!}\boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} = \frac{1}{(l-1)!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(l-1)}\right) \bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha}. \tag{11.2.65}$$

L'insertion de (11.2.64) et de (11.2.65) dans (11.2.62) conduit à l'identité

$$\frac{d\overline{v}_\alpha(t)}{dt} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = -\sum_{l=2}^{k}\frac{1}{(l-1)!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(l-1)}\right)\bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} - \\ -\frac{1}{k!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(k)}\right)\bullet D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha} - \mathfrak{E}_\alpha^{(k)}(v) - \mathfrak{B}_\alpha^{(k)}(v) - \mathfrak{H}_\alpha^{(k)}(v)\,. \quad (11.2.66)$$

La somme dans le membre de droite de (11.2.66) devient, après un changement d'indices et avec l'utilisation de $\boldsymbol{x}_\alpha^{(1)} = 0$

$$\sum_{l=2}^{k}\frac{1}{(l-1)!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(l-1)}\right)\bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} + \frac{1}{k!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(k)}\right)\bullet D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha} = \\ = \sum_{l=2}^{k+1}\frac{1}{(l-1)!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(l-1)}\right)\bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha} = \sum_{l=2}^{k}\frac{1}{l!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(l)}\right)\bullet D^{(l+1)}v\Big|_{\boldsymbol{x}_\alpha}\,. \quad (11.2.67)$$

La définition

$$\mathfrak{F}_\alpha^{(k)}(v) \triangleq \frac{d\overline{v}_\alpha(t)}{dt} - \frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} - \sum_{l=2}^{k}\frac{1}{l!}\boldsymbol{x}_\alpha^{(l)}\bullet D^{(l)}\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} \quad (11.2.68)$$

et l'identité (11.2.67) permettent d'écrire le développement (11.2.66) comme

$$\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = -\mathfrak{E}_\alpha^{(k)}(v) - \mathfrak{F}_\alpha^{(k)}(v) - \mathfrak{B}_\alpha^{(k)}(v) - \mathfrak{H}_\alpha^{(k)}(v) - \\ -\sum_{l=2}^{k}\frac{1}{l!}\boldsymbol{x}_\alpha^{(l)}\bullet D^{(l)}\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} - \sum_{l=2}^{k}\frac{1}{l!}\left(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(l)}\right)\bullet D^{(l+1)}v\Big|_{\boldsymbol{x}_\alpha}\,. \quad (11.2.69)$$

L'identité (11.2.69) permet d'établir la

PROPOSITION 11.2.7 (Équation modifiée du schéma d'ordre k). *Une solution régulière v du schéma (10.2.19) satisfait l'identité*

$$\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = -\sum_{l=2}^{k}\frac{1}{l!}\boldsymbol{x}_\alpha^{(l)}\bullet D^{(l)}\left(\frac{\partial v}{\partial t} + \boldsymbol{c}\cdot D^{(1)}v\right)\Big|_{\boldsymbol{x}_\alpha} - \\ - \mathfrak{E}_\alpha^{(k)}(v) - \mathfrak{F}_\alpha^{(k)}(v) - \mathfrak{B}_\alpha^{(k)}(v) - \mathfrak{H}_\alpha^{(k)}(v) \quad (11.2.70)$$

où les quatre termes $\mathfrak{E}_\alpha^{(k)}(v)$, $\mathfrak{F}_\alpha^{(k)}(v)$, $\mathfrak{B}_\alpha^{(k)}(v)$ et $\mathfrak{H}_\alpha^{(k)}(v)$ sont donnés par les définitions respectives (11.2.56), (11.2.68), (11.2.57) et (11.2.60). Le développement asymptotique de v en $\boldsymbol{x}_\alpha$ donne l'équation modifiée associée au schéma (10.2.19) en maillage non structuré général

$$\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = \\ = -\mathfrak{E}_\alpha^{(k)}(v) - \mathfrak{F}_\alpha^{(k)}(v) - \mathfrak{B}_\alpha^{(k)}(v) - \mathfrak{H}_\alpha^{(k)}(v) + O\left(h^{k+2}\right) = O\left(h^k\right) \quad (11.2.71)$$

où le membre de droite ne contient que les dérivées d'ordre $k+1$ ou supérieur de v en $\boldsymbol{x}_\alpha$.

(i) *Le terme $\mathfrak{H}_\alpha^{(k)}(v)$ admet le développement*

$$\mathfrak{H}_\alpha^{(k)}(v) = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta\frac{1}{(k+1)!}\Big((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+\boldsymbol{k}_{\alpha\beta}^{(k+1)} - \\ -(\boldsymbol{c}\cdot\boldsymbol{a}_{\beta\alpha})_+\boldsymbol{k}_{\beta\alpha}^{(k+1)}\Big)\bullet D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k+1}\right) = O\left(h^k\right)\,. \quad (11.2.72)$$

(ii) *Le terme* $\mathfrak{E}_\alpha^{(k)}(v)$ *regroupe les erreurs engendrées par la reconstruction polynomiale et admet le développement*

$$\mathfrak{E}_\alpha^{(k)}(v) \triangleq \frac{1}{|\mathcal{T}_\alpha|}\sum_\beta\sum_\gamma\sum_{l=0}^{k}\frac{1}{l!\,(k+1)!}\Big((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+\,\boldsymbol{w}_{\alpha\gamma}^{(l)}\bullet\Big(\boldsymbol{k}_{\alpha\beta}^{(l)}-\boldsymbol{x}_\alpha^{(l)}\Big)\,\boldsymbol{z}_{\alpha\gamma}^{(k+1)}+$$
$$+(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_-\,\boldsymbol{w}_{\beta\gamma}^{(l)}\bullet\Big(\boldsymbol{k}_{\beta\alpha}^{(l)}-\boldsymbol{x}_\beta^{(l)}\Big)\,\boldsymbol{z}_{\beta\gamma}^{(k+1)}\Big)\bullet D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha}+\mathrm{O}\left(h^{k+1}\right)=O\left(h^k\right). \quad (11.2.73)$$

(iii) *Le terme* $\mathfrak{B}_\alpha^{(k)}(v)$ *admet le développement*

$$\mathfrak{B}_\alpha^{(k)}(v)=\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta\frac{1}{(k+1)!}\Big((\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+\,\boldsymbol{x}_\alpha^{(k+1)}+$$
$$+(\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_-\,\boldsymbol{x}_\beta^{(k+1)}\Big)\bullet D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha}+\mathrm{O}\left(h^{k+1}\right)=O\left(h^k\right). \quad (11.2.74)$$

(iv) *Le terme* $\mathfrak{F}_\alpha^{(k)}(v)$ *admet le développement*

$$\mathfrak{F}_\alpha^{(k)}(v)=\frac{1}{(k+1)!}\,D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\Big(\boldsymbol{c}\odot\boldsymbol{x}_\alpha^{(k+1)}\Big)+\mathrm{O}\left(h^{k+2}\right)=\mathrm{O}\left(h^{k+1}\right). \quad (11.2.75)$$

Démonstration. L'équation (11.2.70) résulte du regroupement des termes de (11.2.69).

(i) Pour établir le développement asymptotique de $\mathfrak{H}_\alpha^{(k)}(v)$, il est nécessaire de développer les dérivées de v en $\boldsymbol{x}_\beta$ dans (11.2.60) autour de $\boldsymbol{x}_\alpha$. Soit k' un entier plus grand que k. Pour v suffisamment régulière, il vient

$$D^{(l)}v\Big|_{\boldsymbol{x}_\beta}\bullet\boldsymbol{k}_{\beta\alpha}^{(l)}=\sum_{j=0}^{k'-l}\frac{1}{j!}\,D^{(j+l)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\left[\boldsymbol{h}_{\alpha\beta}^{j}\odot\boldsymbol{k}_{\beta\alpha}^{(l)}\right]+\mathrm{O}\left(h^{k'+1}\right). \quad (11.2.76)$$

Le développement (11.2.76) permet de développer la somme

$$\sum_{l=0}^{k'}\frac{1}{l!}\boldsymbol{k}_{\beta\alpha}^{(l)}\bullet D^{(l)}v\Big|_{\boldsymbol{x}_\beta}=\sum_{l=0}^{k'}\sum_{j=0}^{k'-l}\frac{1}{l!}\frac{1}{j!}\,D^{(l+j)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\left[\boldsymbol{h}_{\alpha\beta}^{j}\odot\boldsymbol{k}_{\beta\alpha}^{(l)}\right]+\mathrm{O}\left(h^{k'+1}\right). \quad (11.2.77)$$

Le domaine de la double somme dans le membre de droite de (11.2.77) peut être exprimé par

$$\big\{(l,j)\in\mathbb{Z}\times\mathbb{Z}|\,0\le l\,,\,l\le k'\,,\,0\le j\,,\,j\le k'-l\big\}=$$
$$=\big\{(l,j)\in\mathbb{Z}\times\mathbb{Z}|\,0\le l\,,\,0\le l+j\,,\,l\le l+j\,,\,j+l\le k'\big\}=$$
$$=\big\{(i-j,j)\in\mathbb{Z}\times\mathbb{Z}|\,j\le i\,,\,0\le i\,,\,0\le j\,,\,i\le k'\big\}. \quad (11.2.78)$$

La nouvelle variable $i\triangleq l+j$, introduite dans (11.2.78), et l'identité (5.7.8)

$$\boldsymbol{k}_{\alpha\beta}^{(i)}=\sum_{j=0}^{i}\binom{i}{j}\boldsymbol{k}_{\beta\alpha}^{(j)}\odot\boldsymbol{h}_{\alpha\beta}^{i-j}=\sum_{j=0}^{i}\binom{i}{j}\boldsymbol{k}_{\beta\alpha}^{(i-j)}\odot\boldsymbol{h}_{\alpha\beta}^{j}$$

permettent d'exprimer la double somme dans (11.2.77) comme

$$\sum_{l=0}^{k'}\sum_{j=0}^{k'-l}\frac{1}{l!}\frac{1}{j!}\,D^{(l+j)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\left[\boldsymbol{h}_{\alpha\beta}^{j}\odot\boldsymbol{k}_{\beta\alpha}^{(l)}\right]=$$
$$=\sum_{i=0}^{k'}\frac{1}{i!}\sum_{j=0}^{i}\frac{i!}{(i-j)!j!}\,D^{(i)}v\Big|_{\boldsymbol{x}_\alpha}\bullet\left[\boldsymbol{h}_{\alpha\beta}^{j}\odot\boldsymbol{k}_{\beta\alpha}^{(i-j)}\right]=\sum_{i=0}^{k'}\frac{1}{i!}\boldsymbol{k}_{\alpha\beta}^{(i)}\bullet D^{(i)}v\Big|_{\boldsymbol{x}_\alpha}. \quad (11.2.79)$$

L'équation (11.2.79) simplifie (11.2.77) et donne

$$\sum_{l=0}^{k'}\frac{1}{l!}\boldsymbol{k}_{\beta\alpha}^{(l)}\bullet D^{(l)}v\Big|_{\boldsymbol{x}_\beta}=\sum_{l=0}^{k'}\frac{1}{l!}\boldsymbol{k}_{\alpha\beta}^{(l)}\bullet D^{(l)}v\Big|_{\boldsymbol{x}_\alpha}+\mathrm{O}\left(h^{k'+1}\right). \quad (11.2.80)$$

Pour $k' = k+1$, la comparaison asymptotique (11.2.80) peut s'exprimer comme

$$\sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{k}_{\beta\alpha}^{(l)} \bullet D^{(l)} v \Big|_{\boldsymbol{x}_\beta} = \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)} v \Big|_{\boldsymbol{x}_\alpha} + \\ + \frac{1}{(k+1)!} \left(\boldsymbol{k}_{\alpha\beta}^{(k+1)} - \boldsymbol{k}_{\beta\alpha}^{(k+1)} \right) \bullet D^{(k+1)} v \Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k+2} \right) \quad (11.2.81)$$

car

$$\boldsymbol{k}_{\beta\alpha}^{(k+1)} \bullet D^{(k+1)} v \Big|_{\boldsymbol{x}_\beta} = \boldsymbol{k}_{\beta\alpha}^{(k+1)} \bullet D^{(k+1)} v \Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k+2} \right) .$$

La formule (11.2.81) permet de développer la somme

$$\frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{k}_{\beta\alpha}^{(l)} \bullet D^{(l)} v \Big|_{\boldsymbol{x}_\beta} = \\ = \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)} v \Big|_{\boldsymbol{x}_\alpha} + \\ + \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \left\{ \frac{1}{(k+1)!} \left(\boldsymbol{k}_{\alpha\beta}^{(k+1)} - \boldsymbol{k}_{\beta\alpha}^{(k+1)} \right) \bullet D^{(k+1)} v \Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k+2} \right) \right\} . \quad (11.2.82)$$

Le terme dans la troisième ligne de (11.2.82) peut être réécrit. L'antisymétrie (5.3.4) des vecteurs surface

$$(\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- = (-\boldsymbol{c} \cdot \boldsymbol{a}_{\beta\alpha})_- = -(\boldsymbol{c} \cdot \boldsymbol{a}_{\beta\alpha})_+$$

et l'identité (5.9.21) de la proposition 5.9.2 entraînent

$$\frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \left(\boldsymbol{k}_{\alpha\beta}^{(k+1)} - \boldsymbol{k}_{\beta\alpha}^{(k+1)} \right) = \\ = \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \left(\boldsymbol{k}_{\alpha\beta}^{(k+1)} - \boldsymbol{k}_{\beta\alpha}^{(k+1)} \right) + \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_+ \left(\boldsymbol{k}_{\alpha\beta}^{(k+1)} - \boldsymbol{k}_{\alpha\beta}^{(k+1)} \right) = \\ = (k+1)\, \boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(k)} - \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta \left[(\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(k+1)} - (\boldsymbol{c} \cdot \boldsymbol{a}_{\beta\alpha})_+ \boldsymbol{k}_{\beta\alpha}^{(k+1)} \right] . \quad (11.2.83)$$

L'identité (11.2.83) permet d'écrire (11.2.82) sous la forme

$$\frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{k}_{\beta\alpha}^{(l)} \bullet D^{(l)} v \Big|_{\boldsymbol{x}_\beta} = \\ = \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta (\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_- \sum_{l=0}^{k} \frac{1}{l!} \boldsymbol{k}_{\alpha\beta}^{(l)} \bullet D^{(l)} v \Big|_{\boldsymbol{x}_\alpha} + \frac{1}{k!} \left(\boldsymbol{c} \odot \boldsymbol{x}_\alpha^{(k)} \right) \bullet D^{(k+1)} v \Big|_{\boldsymbol{x}_\alpha} - \\ - \frac{1}{|\mathcal{T}_\alpha|} \sum_\beta \frac{1}{(k+1)!} \left((\boldsymbol{c} \cdot \boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(k+1)} - (\boldsymbol{c} \cdot \boldsymbol{a}_{\beta\alpha})_+ \boldsymbol{k}_{\beta\alpha}^{(k+1)} \right) \bullet D^{(k+1)} v \Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k+1} \right) . \quad (11.2.84)$$

L'insertion de (11.2.84) dans la définition (11.2.60) de $\mathfrak{H}_\alpha^{(k)}(v)$ prouve le point (i) de la proposition 11.2.7.

(ii) De façon analogue, le développement asymptotique de $\mathfrak{E}_\alpha^{(k)}(v)$ s'obtient directement par l'insertion des définitions (11.2.50) de $\varepsilon_\alpha^{(k)}(v)$ et (11.2.51) de $\varepsilon_\beta^{(k)}(v)$ dans (11.2.56). Cela prouve le point (ii) de la proposition 11.2.7.

(iii) Le développement asymptotique de $\mathfrak{B}_\alpha^{(k)}(v)$ s'obtient par l'insertion des définitions (11.2.53) de $\psi_\alpha^{(k)}(v)$ et (11.2.54) de $\psi_\beta^{(k)}(v)$ dans (11.2.57). Il suffit ensuite de remplacer les dérivées de v en $\boldsymbol{x}_\beta$ à l'aide de

$$D^{(k+1)}v\Big|_{\boldsymbol{x}_\beta} = D^{(k+1)}v\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}(h)$$

ce qui démontre le point (iii) de la proposition 11.2.7.

(iv) D'après la formule (5.8.2), la dérivée en temps de la moyenne $\overline{v}_\alpha(t)$ dans le membre de gauche de (11.2.66) admet un développement asymptotique

$$\frac{d\overline{v}_\alpha(t)}{dt} = \frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \sum_{l=2}^{k+1} \frac{1}{l!}\boldsymbol{x}_\alpha^{(l)} \bullet D^{(l)}\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k+2}\right). \tag{11.2.85}$$

Le développement (11.2.85) montre que le terme $\mathfrak{F}_\alpha^{(k)}(v)$ satisfait

$$\mathfrak{F}_\alpha^{(k)}(v) = \frac{1}{(k+1)!}\boldsymbol{x}_\alpha^{(k+1)} \bullet D^{(k+1)}\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^{k+2}\right) = \mathrm{O}\left(h^{k+1}\right). \tag{11.2.86}$$

La comparaison asymptotique

$$\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} = -\boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} + \mathrm{O}\left(h^k\right)$$

permet de remplacer la dérivée partielle en temps dans (11.2.86) par la dérivée spatiale, ce qui prouve le point (iv) de la proposition 11.2.7.

La forme de l'équation (11.2.70) suggère de l'insérer de façon récursive dans elle-même, ce qui donne, pour la première itération

$$\begin{aligned}\frac{\partial v}{\partial t}\Big|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = &-\mathfrak{E}_\alpha^{(k)}(v) - \mathfrak{F}_\alpha^{(k)}(v) - \mathfrak{B}_\alpha^{(k)}(v) - \mathfrak{H}_\alpha^{(k)}(v) - \\ &- \sum_{l=2}^{k}\frac{1}{l!}\sum_{l=2}^{k}\frac{1}{j!}\,\boldsymbol{x}_\alpha^{(l)} \bullet D^{(l)}\left(\boldsymbol{x}_\alpha^{(j)} \bullet D^{(j)}\left(\frac{\partial v}{\partial t} + \boldsymbol{c}\cdot D^{(1)}v\right)\right)\Big|_{\boldsymbol{x}_\alpha} - \\ &- \sum_{l=2}^{k}\frac{1}{l!}\boldsymbol{x}_\alpha^{(l)} \bullet D^{(l)}\left(\mathfrak{E}_\alpha^{(k)}(v) + \mathfrak{F}_\alpha^{(k)}(v) + \mathfrak{B}_\alpha^{(k)}(v) + \mathfrak{H}_\alpha^{(k)}(v)\right)\Big|_{\boldsymbol{x}_\alpha}. \end{aligned}\tag{11.2.87}$$

D'après l'estimation (11.2.58), $\mathfrak{E}_\alpha^{(k)}(v) = \mathrm{O}\left(h^k\right)$ et $\mathfrak{B}_\alpha^{(k)}(v) = \mathrm{O}\left(h^k\right)$. Par ailleurs, la démonstration des points (i) et (iv) de la proposition 11.2.7 montre que $\mathfrak{H}_\alpha^{(k)}(v)$ et $\mathfrak{F}_\alpha^{(k)}(v)$ sont d'ordre au moins $\mathrm{O}\left(h^k\right)$. Le terme dans la troisième ligne de (11.2.87) est d'ordre au moins $O\left(h^{k+2}\right)$ car $\mathfrak{E}_\alpha^{(k)}(v)$, $\mathfrak{F}_\alpha^{(k)}(v)$, $\mathfrak{B}_\alpha^{(k)}(v)$ et $\mathfrak{H}_\alpha^{(k)}(v)$ sont multipliés par $\boldsymbol{x}_\alpha^{(l)} = \mathrm{O}\left(h^l\right)$ pour $l \geq 2$. Ce terme peut donc être négligé pour un développement à l'ordre $\mathrm{O}\left(h^k\right)$. Le terme dans la deuxième ligne de (11.2.87) est au moins d'ordre $\mathrm{O}\left(h^4\right)$ car il contient le produit de $\boldsymbol{x}_\alpha^{(l)}$ et $\boldsymbol{x}_\alpha^{(j)}$ pour $j, l \geq 2$. Une nouvelle insertion de (11.2.70) dans le terme de la deuxième ligne de (11.2.87) donne une équation qui a la même forme que (11.2.87) mais où la somme dans la deuxième ligne est d'ordre $\mathrm{O}\left(h^6\right)$. Un nombre d'insertions récursives supérieur ou égal à $\frac{k+2}{2}$ permet par conséquent d'établir le développement asymptotique (11.2.71). □

Il est intéressant de déterminer les maillages sur lesquels le premier terme d'erreur de $\mathfrak{H}_\alpha^{(k)}(v)$ disparaît. Cette erreur, qui est d'ordre $\mathrm{O}\left(h^k\right)$ dans (11.2.72), est nulle indépendamment de $\boldsymbol{c}$ et de v si et seulement si dans la cellule $\mathcal{T}_\alpha$

$$\sum_\beta (\boldsymbol{c}\cdot\boldsymbol{a}_{\alpha\beta})_+ \boldsymbol{k}_{\alpha\beta}^{(k+1)} = \sum_\gamma (\boldsymbol{c}\cdot\boldsymbol{a}_{\gamma\alpha})_+ \boldsymbol{k}_{\gamma\alpha}^{(k+1)} \text{ pour tout } \boldsymbol{c}\in\mathbb{R}^d. \tag{11.2.88}$$

Une condition suffisante pour (11.2.88) est donnée par la

PROPOSITION 11.2.8 (Compensation géométrique de l'erreur d'ordre $\mathrm{O}\left(h^k\right)$). *La condition (11.2.88) est satisfaite dans la cellule $\mathcal{T}_\alpha$ si pour chaque face $\mathcal{A}_{\alpha\beta}$ de $\mathcal{T}_\alpha$ il existe une et une seule face $\mathcal{A}_{\alpha\gamma}$ telle que $\boldsymbol{k}_{\alpha\beta}^{(k+1)} = \boldsymbol{k}_{\gamma\alpha}^{(k+1)}$ et $\boldsymbol{a}_{\alpha\beta} = \boldsymbol{a}_{\gamma\alpha}$.*

DÉMONSTRATION. Si la condition de la proposition 11.2.8 est satisfaite, les sommes dans (11.2.88) sont égales. □

La proposition 11.2.8 implique le

COROLLAIRE 11.2.9. *La condition (11.2.88) est satisfaite dans les cas suivants :*

(i) *Les maillages cartésiens.*

(ii) *Les pavages du plan par des hexagones réguliers.*

(iii) *Les pavages de l'espace par des dodécaèdres réguliers.*

La discussion ci-dessus démontre la

PROPOSITION 11.2.10 (Compensation des erreurs d'ordre $\mathrm{O}\left(h^k\right)$). *Les termes d'erreur d'ordre $\mathrm{O}\left(h^k\right)$ dans le membre de droite de (11.2.71) sont nuls si les conditions suivantes sont satisfaites :*

(i) *La reconstruction des dérivées consistantes $\boldsymbol{w}_{\alpha\gamma}^{(l)}$, pour $1 \leq l \leq k$, est de précision à l'ordre $k+1$ au sens de la définition 7.4.1.*

(ii) *Les cellules ont les mêmes moments $\boldsymbol{x}_\alpha^{(k+1)}$.*

(iii) *La condition géométrique de la proposition 11.2.8 est satisfaite.*

11.2.4. Conclusion de l'étude de l'équation modifiée. L'étude théorique des sections 11.2.1, 11.2.2 et 11.2.3 a permis d'établir l'équation modifiée des schémas d'ordre $k \geq 1$ pour l'équation de convection linéaire en maillage non structuré général. Pour la reconstruction des polynômes de degré k, les différents termes d'erreur dans l'équation modifiée sont donnés par :

(1) Le terme $\mathfrak{E}_\alpha^{(k)}(v) = \mathrm{O}\left(h^k\right)$, défini par (11.2.56), représente l'erreur de la reconstruction des polynômes de degré k. Le premier terme d'erreur dans $\mathfrak{E}_\alpha^{(k)}(v)$ disparaît si la reconstruction est de précision à l'ordre $k+1$ au sens de la définition 7.4.1.

(2) Le terme $\mathfrak{B}_\alpha^{(k)}(v) = \mathrm{O}\left(h^k\right)$, défini par (11.2.57), provient du fait que les variables fondamentales du schéma sont les moyennes de cellule. Il disparaît si les moments d'ordre k des cellules sont égaux.

(3) Le terme $\mathfrak{H}_\alpha^{(k)}(v) = \mathrm{O}\left(h^k\right)$, défini par (11.2.60), peut s'interpréter comme une contribution du décentrement des flux.

(4) Le terme $\mathfrak{F}_\alpha^{(k)}(v) = \mathrm{O}\left(h^{k+1}\right)$, défini par (11.2.68), vient du développement de Taylor de la moyenne de cellule.

Les développements asymptotiques ont permis d'expliciter les premiers termes d'erreur dans $\mathfrak{B}_\alpha^{(k)}(v)$, $\mathfrak{E}_\alpha^{(k)}(v)$, $\mathfrak{H}_\alpha^{(k)}(v)$, et $\mathfrak{F}_\alpha^{(k)}(v)$. Il s'avère que les premiers termes d'erreurs dans $\mathfrak{B}_\alpha^{(k)}(v)$, $\mathfrak{E}_\alpha^{(k)}(v)$ et $\mathfrak{H}_\alpha^{(k)}(v)$ disparaissent dans certaines conditions géométriques. Ces conditions géométriques sont des conditions de symétrie entre les cellules et les faces, qui sont satisfaites, par exemple, en maillage cartésien. Dans ce cas, les erreurs $\mathfrak{B}_\alpha^{(k)}(v)$, $\mathfrak{E}_\alpha^{(k)}(v)$ et $\mathfrak{H}_\alpha^{(k)}(v)$ sont d'ordre $\mathrm{O}\left(h^{k+1}\right)$, comme $\mathfrak{F}_\alpha^{(k)}(v)$.

Le fait que certaines conditions géométriques suppriment les termes d'erreur d'ordre $\mathrm{O}\left(h^k\right)$ dans $\mathfrak{B}_\alpha^{(k)}(v)$, $\mathfrak{E}_\alpha^{(k)}(v)$ et $\mathfrak{H}_\alpha^{(k)}(v)$ laisse supposer que les termes prépondérants sont les deuxièmes termes d'erreur, qui sont d'ordre $\mathrm{O}\left(h^{k+1}\right)$. Cette hypothèse est soutenue par les résultats numériques. Les deuxièmes termes d'erreur n'ont pas encore été explicités ici mais la présente étude permettra de le faire sans trop d'effort.

11.3. Étude numérique en dimension deux

11.3.1. Description des tests. Des tests numériques sont indispensables pour évaluer la précision des schémas volumes finis en maillage non structuré. L'équation modèle pour les tests est l'équation de convection linéaire à vitesse constante qui est l'exemple le plus simple d'une loi de conservation. Cette équation permet d'évaluer beaucoup d'aspects de schémas numériques qui sont importants pour les équations de Navier-Stokes.

En dimension trois, les maillages de tétraèdres nécessitent un grand nombre de cellules pour fournir une résolution suffisante de la condition initiale. Puisqu'il n'a pas été possible de développer des outils de test suffisamment performants pour la dimension trois, l'étude numérique se restreint à la dimension deux, où des maillages de quelques centaines ou milliers de triangles peuvent donner une bonne résolution.

Les tests consistent à transporter une condition initiale sur le carré

$$\Omega = \left\{ (x,y) \in \mathbb{R}^2 \middle| \, 0 \leq x \leq 1 \, , \, 0 \leq y \leq 1 \right\}$$

avec des conditions limites de périodicité, ce qui permet d'éliminer l'influence des conditions limites sur la précision du schéma. Les schémas volumes finis transforment l'équation de la convection linéaire à vitesse constante

$$\frac{\partial u\left(\boldsymbol{x},t\right)}{\partial t} = -\boldsymbol{c} \cdot \boldsymbol{\nabla} u\left(\boldsymbol{x},t\right)$$

en un système dynamique linéaire

$$\frac{d\mathfrak{u}\left(t\right)}{dt} = J\mathfrak{u}\left(t\right) \, , \, \mathfrak{u}\left(0\right) = \mathfrak{u}_0 \, , \, \mathfrak{u}\left(t\right) \in \mathbb{C}^N \tag{11.3.1}$$

où $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$ est le vecteur des moyennes de cellule et J est la matrice qui correspond à la discrétisation spatiale de l'opérateur $-\boldsymbol{c} \cdot \boldsymbol{\nabla}$. Un schéma de Runge-Kutta explicite en temps d'ordre quatre permet d'intégrer l'équation (11.3.1) avec une précision temporelle suffisante pour faire ressortir l'erreur de discrétisation spatiale, voir par exemple [**69**].

Les maillages de test sont neuf maillages de triangles et neuf maillages hybrides, composés de triangles et quadrangles irréguliers. Ces maillages ont été générés par le logiciel Gmsh. Quatre maillages cartésiens permettent la comparaison entre calculs sur maillage régulier et irrégulier.

La figure 11.3.1 montre les deux conditions initiales testées. La première condition initiale est une fonction gaussienne

$$u_0(x,y) = 30 \exp\left(-40 \left(x - \frac{1}{2} \right)^2 - 40 \left(y - \frac{1}{2} \right)^2 \right) \tag{11.3.2}$$

qui s'annule approximativement aux bords du carré. La deuxième condition initiale est la fonction périodique

$$u_0(x,y) = \sin\left(2\pi x\right) \sin\left(2\pi y\right) \, . \tag{11.3.3}$$

La vitesse de convection est $\boldsymbol{c} = (1,0)$, c'est-à-dire parallèle à l'axe x, ce qui ne restreint pas la généralité des résultats car les maillages testés sont non structurés. La solution est transportée pendant un temps $t_{\text{fin}} = 10$, ce qui permet à la condition initiale de traverser le carré dix fois. Les variables fondamentales sont les moyennes de cellule. Il faut donc d'abord calculer les moyennes de cellule des conditions initiales (11.3.2) et (11.3.3)

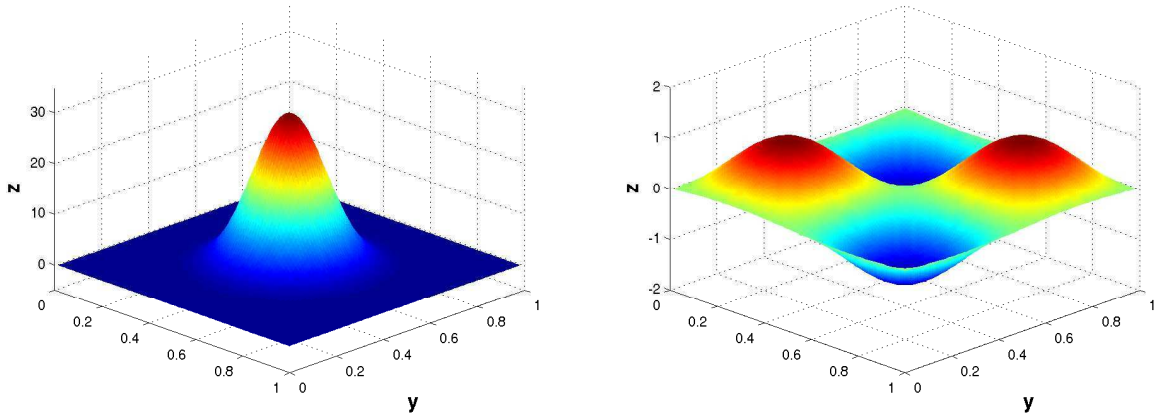
$$\mathfrak{u}_0 = (\overline{u}_1, \ldots, \overline{u}_N) \tag{11.3.4}$$

avec une précision qui est du même ordre que le schéma utilisé. Dans le cas présent, il faut des formules capables d'intégrer exactement les polynômes de degré trois car les schémas testés reposent sur la reconstruction des polynômes de degré inférieur ou égal à trois.

À $t_{\text{fin}} = 10$, la solution exacte de l'équation de convection avec la vitesse $\boldsymbol{c} = (1,0)$ est égale à la condition initiale. Notons

$$\mathfrak{v} = (\overline{v}_1, \ldots, \overline{v}_N) \tag{11.3.5}$$

le vecteur des moyennes de cellule de la solution numérique à $t_{\text{fin}} = 10$.



(a) Condition initiale (11.3.2). (b) Condition initiale (11.3.3).

FIG. 11.3.1: Conditions initiales pour la convection.

Une méthode pour évaluer la précision d'un schéma volumes finis consiste à calculer l'écart entre la solution exacte et la solution numérique dans une norme. Il faut choisir une norme en tenant compte du fait que seules les moyennes de cellule (11.3.5) de la solution numérique sont connues. Une norme simple est la norme discrète

$$\|\mathfrak{v}\|_{\ell_2(\Omega)} = \sqrt{\sum_{\alpha=1}^{N} |\mathcal{T}_\alpha| \, |\overline{v}_\alpha|^2} \, . \tag{11.3.6}$$

Dans la norme (11.3.6), l'erreur relative de la solution numérique par rapport à la solution exacte est

$$\varepsilon = \frac{\|\mathfrak{v} - \mathfrak{u}\|_{\ell_2(\Omega)}}{\|\mathfrak{u}\|_{\ell_2(\Omega)}} \, . \tag{11.3.7}$$

Une autre possibilité pour calculer l'écart entre la solution exacte et la solution numérique consiste à approcher la solution numérique dans chaque cellule par une reconstruction consistante de degré trois : dans la cellule $\mathcal{T}_\alpha$, on reconstruit le polynôme de degré trois

$$w_\alpha \left[\mathfrak{v}\right](\boldsymbol{x}) = \overline{v}_\alpha + \sum_\beta \sum_{l=1}^{3} \frac{1}{l!} \boldsymbol{w}_{\alpha\beta}^{(l)} \bullet \left[(\boldsymbol{x} - \boldsymbol{x}_\alpha)^l - \boldsymbol{x}_\alpha^{(l)} \right] \overline{v}_\beta$$

à partir des moyennes de cellule (11.3.5) par la méthode des moindres carrés, cf. la définition 7.7.1. Cela permet de calculer l'écart relatif entre la solution numérique et la solution exacte par la formule

$$\varepsilon = \frac{\sqrt{\sum_{\alpha=1}^{N} \int_{\mathcal{T}_\alpha} |w_\alpha \left[\mathfrak{v}\right](\boldsymbol{x}) - u_0(\boldsymbol{x})|^2 \, dx}}{\sqrt{\sum_{\alpha=1}^{N} \int_{\mathcal{T}_\alpha} |u_0(\boldsymbol{x})|^2 \, dx}} \, . \tag{11.3.8}$$

Il s'avère que l'utilisation de la formule (11.3.8) n'a pas d'effet significatif sur l'erreur mesurée par rapport à la formule simple (11.3.7). Cela signifie que l'erreur introduite par l'utilisation de la formule (11.3.7) est négligeable par rapport à l'erreur numérique engendré par le schéma volumes finis.

Les tests couvrent les schémas suivants :

(1) La reconstruction constante par cellule, ce qui donne le schéma (10.2.7) qui est formellement d'ordre un. Cette reconstruction est une reconstruction des polynômes de degré $k = 0$, la méthode associée est appelée *degré 0* dans la légende.

(2) La reconstruction des polynômes de degré un par la méthode des moindres carrés, ou méthode de la pseudo-inverse, cf. la définition 7.7.1, ce qui donne le schéma (10.2.13) qui est *formellement* d'ordre deux. Ce schéma est appelé *degré 1* dans la légende.

(3) La reconstruction des polynômes de degré deux par la méthode des moindres carrés, introduite par la définition 7.7.1. Cela donne un schéma qui est *formellement* d'ordre trois (10.2.18). Ce schéma est appelé *degré 2* dans la légende.

(4) La reconstruction des polynômes de degré deux par la méthode des moindres carrés couplés par itérations, ou méthode MCCI. Cette méthode est définie par l'algorithme 8.2.2 et s'appelle *degré 2 MCCI* dans la légende.

(5) La reconstruction des polynômes de degré deux par la méthode des moindres carrés couplés par itérations sur un voisinage élargi, ou méthode MCCIE. Cette méthode, appelée *degré 2 MCCIE* dans la légende, est définie par l'algorithme 8.4.3.

(6) La reconstruction des polynômes de degré deux par la méthode des corrections successives, ou méthode CS. Cette méthode est définie par l'algorithme 8.3.8 et s'appelle *degré 2 CS* dans la légende.

(7) La reconstruction des polynômes de degré deux par la méthode des corrections successives sur un voisinage élargi, ou méthode CSE. Cette méthode est définie par l'algorithme 8.4.4 et s'appelle *degré 2 CSE* dans la légende.

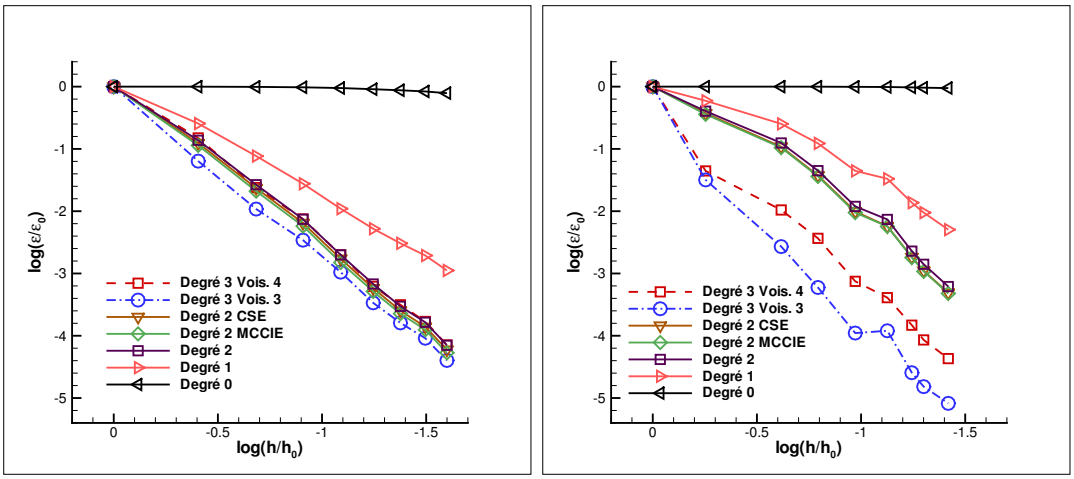
(8) La reconstruction des polynômes de degré trois sur le troisième et quatrième voisinage par la méthode des moindres carrés, introduite par la définition 7.7.1. Cette méthode s'appelle *degré 3* dans la légende.

11.3.2. Discussion des résultats. Une première évaluation des schémas consiste à analyser l'évolution de l'écart (11.3.7) en fonction du diamètre h des maillages. Pour le maillage le plus grossier, notons l'erreur (11.3.7) ε_0 et le diamètre du maillage h_0. Lorsqu'on dessine $\log(\varepsilon/\varepsilon_0)$ en fonction de $\log(h/h_0)$ pour les différents maillages testés, on obtient une courbe dont la pente est une mesure pour le taux de convergence de l'erreur numérique en fonction du diamètre des mailles. L'étude de cette courbe se restreint à la condition initiale sinusoïde car elle est plus lisse et permet donc de mieux mettre en évidence le comportement des erreurs numériques. Notons que les schémas *degré 2 MCCI* et *degré 2 CS* montrent des instabilités en maillage de triangles et ne sont donc pas présentés dans les résultats. La figure 11.3.2 montre les courbes pour les maillages de triangles, les maillages hybrides et les maillages cartésiens. Ces diagrammes permettent de tirer plusieurs conclusions.

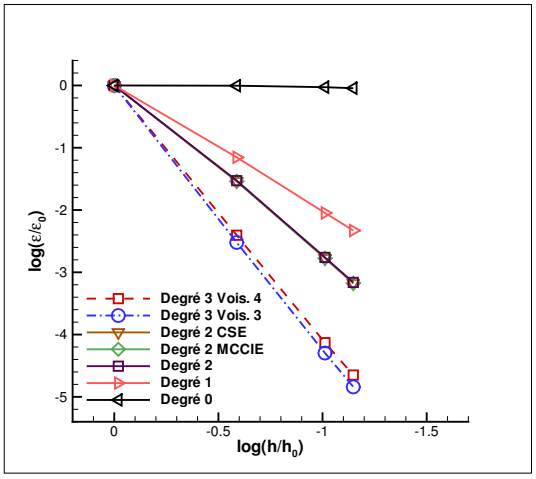
(1) On constate d'abord que le taux de convergence en maillage cartésien est de $k+1$ pour une reconstruction de degré k, sauf dans le cas de la reconstruction constante par morceaux ($k=0$). Les courbes sont par ailleurs presque des droites, ce qui indique que le taux de convergence est indépendant de h. Tout cela signifie qu'en maillage structuré, le taux de convergence correspond presque parfaitement à l'ordre de la reconstruction. L'étude théorique de la section 11.2 a effectivement montré que le premier terme d'erreur dans l'équation modifié est $\mathrm{O}\left(h^{k+1}\right)$ pour ce type de maillage, ce qui pourrait expliquer les résultats observés.

(2) Pour la reconstruction des polynômes de degré un et deux, le taux de convergence se dégrade en maillage non structuré. La reconstruction de degré un donne un schéma dont le taux de convergence se situe entre un et deux. Le taux de convergence des méthodes reposant sur une reconstruction de degré deux se situe entre deux et trois. On constate néanmoins que la différence entre les reconstructions de degré un et deux est clairement visible.

(3) Pour ce test particulier, la reconstruction des polynômes de degré trois constitue un cas à part. Dans le cas des maillages cartésiens, le taux de convergence est de quatre et la différence entre les reconstructions de degré trois et deux est clairement visible. Dans le cas des maillages hybrides, le taux de convergence pour cette reconstruction est plus faible que quatre, mais il est plus grand que celui pour la reconstruction de degré deux.



(a) Maillage de triangles (b) Maillage hybride



(c) Maillage cartésien

FIG. 11.3.2: Taux de convergence de l'erreur numérique (11.3.7) dans le cas de la condition initiale sinusoïde (11.3.3).

En maillage de triangles, par contre, l'erreur numérique pour la reconstruction de degré trois ne semble pas converger plus vite que celui pour la reconstruction quadratique. Ce phénomène nécessite une étude beaucoup plus approfondie qui n'a pas pu être menée dans le cadre de cette thèse.

(4) L'erreur du schéma d'ordre un, appelé *degré 0,* est pratiquement insensible à la diminution du diamètre h du maillage. Pour les maillages non structurés, cette observation correspond au résultat de la proposition 11.2.1 qui suggère que l'erreur numérique peut être $\mathrm{O}(1)$ par rapport à h. Ce comportement s'observe par contre également en maillage cartésien. Cela peut s'expliquer par le fait que la solution numérique de ce schéma satisfait grossièrement une équation de convection-diffusion de la forme

$$\left.\frac{\partial v}{\partial t}\right|_{\boldsymbol{x}_\alpha} + \boldsymbol{c}\cdot D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} = -\boldsymbol{\vartheta}_\alpha^{(2)} \bullet D^{(2)}v\Big|_{\boldsymbol{x}_\alpha}$$

où le tenseur d'ordre deux $\boldsymbol{\vartheta}_\alpha^{(2)}$, défini par (11.2.14), est semi-défini négatif, ce qui conduit à une diminution rapide de l'amplitude de la solution.

Il est également instructif de regarder les valeurs absolues de l'erreur numérique. Les tableaux suivants montrent les erreurs numériques pour les neuf schémas testés :

(1) Les tableaux 11.3.1 et 11.3.2 montrent l'erreur (11.3.7) pour les neuf maillages triangulaires et les neuf maillages hybrides, dans le cas de la condition initiale (11.3.2).
(2) Les tableaux 11.3.3 et 11.3.4 montrent l'erreur (11.3.7) pour les neuf maillages triangulaires et les neuf maillages hybrides, dans le cas de la condition initiale (11.3.3).

On constate les points suivants :

(1) Même si les reconstructions de degré deux et trois ont à peu près le même taux de convergence en maillage de triangles, on constate néanmoins que la valeur absolue de l'erreur (11.3.7) est plus petite pour les reconstructions de degré trois. Les reconstructions de degré trois apportent donc une amélioration par rapport à celles de degré deux.
(2) Les erreurs sont plus grandes pour la condition initiale (11.3.2). Cela peut éventuellement s'expliquer par le fait que la fonction gaussienne a des dérivées plus grandes que la sinusoïde (11.3.3) qui est plus lisse.
(3) Comme le diamètre h des maillages hybrides est plus grand que celui des maillages triangulaires, les erreurs numériques sont un peu plus grandes pour les premiers.
(4) Les schémas *degré 2 MCCI* et *degré 2 CS* sont instables sur certains maillages de triangles, ce qui explique les erreurs excessivement grandes qu'on constate parfois pour ces schémas.

La discussion ci-dessus est centrée sur des critères purement quantitatifs. Pour les besoins de la simulation numérique avec des codes comme Cedre, il est souvent aussi important de regarder des aspects qualitatifs de la solution numérique. Pour cela, il est instructif d'examiner visuellement les déformations subies par la condition initiale car cela permet de caractériser les différents types d'erreurs. On analyse ici le cas de la condition initiale gaussienne (11.3.2) car les effets sont plus marqués pour ce type de fonction.

(1) La figure 11.3.3 montre le résultat de la convection de la condition initiale gaussienne (11.3.2) sur le maillage hybride numéro cinq avec $h \approx 0.06583$ et $N = 1250$ cellules.
(2) La figure 11.3.4 montre le résultat de la convection de la condition initiale gaussienne (11.3.2) sur le maillage triangulaire numéro neuf avec $h \approx 0.02799$ et $N = 5742$ cellules.

Les figures 11.3.3 et 11.3.4 permettent de comparer les résultats entre les différents schémas et entre les deux maillages. Cela permet de tirer plusieurs conclusions :

(1) Sur le maillage plus grossier, cf. la figure 11.3.3, les effets de la reconstruction se montrent très clairement. Sur le maillage plus fin, cf. la figure 11.3.4, les schémas numériques donnent des résultats qui se ressemblent beaucoup plus, à l'exception de la reconstruction constante par cellule. Sur ce maillage, même le schéma reposant sur une reconstruction de degré un donne une solution satisfaisante, ce qui peut expliquer le succès des schémas d'ordre deux dans des codes comme Cedre.
(2) Sur le maillage plus grossier, on distingue aisément le phénomène suivant : les solutions pour les reconstructions d'ordre pair, degré 2 et degré 0, ont un aspect symétrique par rapport à l'axe $x = 0,5$, alors que les solutions pour les reconstructions de degré impair, degré 3 et degré 1, montrent un aspect asymétrique par rapport à cet axe. Cette observation peut s'interpréter de manière qualitative à l'aide de la théorie de l'équation modifiée. Considérons l'équation modifiée pour l'équation de convection linéaire en dimension un. Si le terme d'erreur dominant contient une dérivée impaire, la solution numérique montre surtout une erreur dispersive, ce qui signifie que différentes composantes de la solution sont transportées à des vitesses différentes. Ce phénomène pourrait expliquer le caractère asymétrique de certaines déformations. Si le terme d'erreur dominant contient une dérivée paire, la solution montre d'abord une erreur dissipative, c'est-à-dire une erreur sur l'amplitude de la solution qui affecte toutes les composantes de la solution d'une manière similaire. Cela pourrait expliquer le caractère symétrique de certaines déformations. La proposition 11.2.7 montre qu'une reconstruction de degré k conduit à un premier terme d'erreur d'ordre $\mathrm{O}\left(h^k\right)$ dans l'équation modifiée. Ce

		Méthode	Degré 3	Degré 3	Degré 2 CS	Degré 2 MCCI	Degré 2 CSE	Degré 2 MCCIE	Degré 2	Degré 1	Degré 0
		Voisinage	3	4	2	2	4	4	2	2	0
Maillage	h	N									
1	0.13858	242	2.6212e-01	3.9549e-01	4.3024e-01	3.4740e-01	5.5395e-01	5.2764e-01	5.7402e-01	6.3571e-01	9.0842e-01
2	0.09239	548	1.9036e-01	1.8662e-01	2.6162e-01	1.8219e-01	3.9596e-01	3.7073e-01	4.1758e-01	4.8016e-01	8.9986e-01
3	0.06995	940	1.0309e-01	9.0118e-02	9.2993e-01	1.0123e-01	2.6897e-01	2.4583e-01	2.9051e-01	3.6798e-01	8.8132e-01
4	0.05586	1428	7.1098e-02	5.8794e-02	1.1854e-01	7.2793e-02	2.0772e-01	1.8892e-01	2.2666e-01	3.0396e-01	8.6918e-01
5	0.04662	2128	3.8045e-02	2.9374e-02	6.3721e-02	3.4830e-02	1.2602e-01	1.1145e-01	1.4017e-01	2.1436e-01	8.5001e-01
6	0.03988	2896	2.5554e-02	2.0597e-02	2.3339e+12	5.3141e+02	9.1247e-02	7.9192e-02	1.0234e-01	1.7168e-01	8.3472e-01
7	0.03500	3698	1.8021e-02	1.5084e-02	4.1667e+35	9.8890e+05	6.9123e-02	6.0637e-02	7.8085e-02	1.4232e-01	8.1119e-01
8	0.03105	4500	1.3582e-02	1.1576e-02	2.7134e-02	1.4339e-02	5.8118e-02	5.0894e-02	6.6299e-02	1.2254e-01	8.1090e-01
9	0.02799	5742	8.8938e-03	7.8273e-03	1.3528e+25	9.2991e-03	3.9379e-02	3.3872e-02	4.5054e-02	9.3349e-02	7.8282e-01

TAB. 11.3.1: Erreurs numériques sur les maillages triangulaires pour la condition initiale gaussienne (11.3.2).

		Méthode	Degré 3	Degré 3	Degré 2 CS	Degré 2 MCCI	Degré 2 CSE	Degré 2 MCCIE	Degré 2	Degré 1	Degré 0
		Voisinage	3	4	2	2	4	4	2	2	0
Maillage	h	N									
1	0.17401	139	5.2711e-01	6.2240e-01	6.0855e-01	6.1062e-01	7.5686e-01	7.5364e-01	7.7072e-01	8.0539e-01	9.1351e-01
2	0.13501	308	3.1293e-01	4.3571e-01	4.2879e-01	4.2677e-01	5.7891e-01	5.7290e-01	5.9237e-01	6.5852e-01	9.1327e-01
3	0.09394	543	1.8779e-01	2.9232e-01	3.2242e-01	3.2421e-01	4.7106e-01	4.6760e-01	4.8513e-01	5.4217e-01	9.0946e-01
4	0.07874	830	1.1397e-01	1.8778e-01	2.4675e-01	2.4849e-01	3.9627e-01	3.9352e-01	4.1188e-01	4.7900e-01	9.0452e-01
5	0.06583	1250	7.4674e-02	1.1288e-01	1.6252e-01	1.6330e-01	2.9297e-01	2.8971e-01	3.0854e-01	3.9317e-01	8.9759e-01
6	0.05644	1655	5.3982e-02	8.4733e-02	1.4589e-01	1.4602e-01	2.6375e-01	2.6269e-01	2.8039e-01	3.6025e-01	8.9619e-01
7	0.05017	2133	3.8862e-02	4.9916e-02	1.0287e-01	1.0318e-01	2.0007e-01	1.9884e-01	2.1456e-01	2.9082e-01	8.7603e-01
8	0.04740	2618	3.2281e-02	3.5542e-02	8.0063e-02	8.0145e-02	1.6731e-01	1.6551e-01	1.7959e-01	2.6075e-01	8.7539e-01
9	0.04212	3296	2.1713e-02	2.1898e-02	5.6989e-02	5.7273e-02	1.2516e-01	1.2393e-01	1.3580e-01	2.1274e-01	8.5918e-01

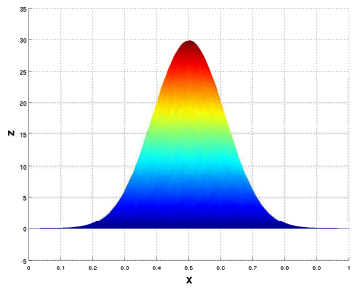
TAB. 11.3.2: Erreurs numériques sur les maillages hybrides pour la condition initiale gaussienne (11.3.2).

Méthode de reconstruction			Degré 3	Degré 3	Degré 2 CS	Degré 2 MCCI	Degré 2 CSE	Degré 2 MCCIE	Degré 2	Degré 1	Degré 0
		Voisinage	3	4	2	2	4	4	2	2	0
Maillage	h	N									
1	0.13858	242	1.3177e-01	1.5739e-01	3.1342e-01	1.8349e-01	5.8712e-01	5.2216e-01	6.3602e-01	8.0149e-01	1.0000e-00
2	0.09239	548	5.7844e-02	4.7601e-02	1.0744e-01	5.7711e-02	2.3707e-01	2.0307e-01	2.6879e-01	4.4331e-01	9.9927e-01
3	0.06995	940	2.6046e-02	2.2048e-02	3.8688e+00	2.7248e-02	1.1510e-01	9.7486e-02	1.3175e-01	2.6293e-01	9.9592e-01
4	0.05586	1428	1.5740e-02	1.3375e-02	2.8416e-02	1.5710e-02	6.5668e-02	5.5618e-02	7.5674e-02	1.6857e-01	9.8912e-01
5	0.04662	2128	8.8366e-03	8.0063e-03	1.6010e-02	8.2097e-03	3.7034e-02	3.1007e-02	4.2709e-02	1.1290e-01	9.7750e-01
6	0.03988	2896	5.3328e-03	4.8893e-03	4.5402e+11	2.2826e+01	2.3048e-02	1.9486e-02	2.6750e-02	8.1864e-02	9.5946e-01
7	0.03500	3698	3.9693e-03	3.5432e-03	4.4735e+35	2.3718e+06	1.6094e-02	1.3641e-02	1.8680e-02	6.4764e-02	9.4309e-01
8	0.03105	4500	3.0361e-03	2.7701e-03	5.3516e-03	2.9608e-03	1.2406e-02	1.0521e-02	1.4457e-02	5.3242e-02	9.2551e-01
9	0.02799	5742	2.0816e-03	1.9393e-03	2.2555e+24	2.3903e-03	8.5949e-03	7.2689e-03	9.9978e-03	4.1878e-02	9.0055e-01

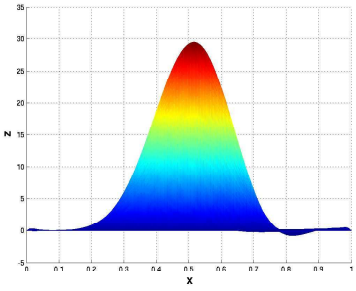
TAB. 11.3.3: Erreurs numériques sur les maillages triangulaires pour la condition initiale sinusoïde (11.3.3).

Méthode de reconstruction			Degré 3	Degré 3	Degré 2 CS	Degré 2 MCCI	Degré 2 CSE	Degré 2 MCCIE	Degré 2	Degré 1	Degré 0
		Voisinage	3	4	2	2	4	4	2	2	0
Maillage	h	N									
1	0.17401	139	4.4118e-01	7.9444e-01	6.9503e-01	6.9739e-01	9.6326e-01	9.6104e-01	9.7852e-01	9.8645e-01	1.0000e-00
2	0.13501	308	1.1393e-01	1.7770e-01	3.0361e-01	3.0173e-01	6.2759e-01	6.1585e-01	6.5990e-01	7.8625e-01	1.0000e-00
3	0.09394	543	6.0884e-02	6.0904e-02	1.5692e-01	1.5866e-01	3.6695e-01	3.6067e-01	3.9610e-01	5.4266e-01	9.9987e-01
4	0.07874	830	3.8625e-02	3.1594e-02	9.4054e-02	9.5060e-02	2.3104e-01	2.2737e-01	2.5399e-01	3.9670e-01	9.9936e-01
5	0.06583	1250	1.9366e-02	1.5193e-02	5.2430e-02	5.2705e-02	1.2975e-01	1.2712e-01	1.4302e-01	2.5517e-01	9.9669e-01
6	0.05644	1655	1.4897e-02	1.5762e-02	4.3036e-02	4.3329e-02	1.0240e-01	1.0173e-01	1.1599e-01	2.2466e-01	9.9416e-01
7	0.05017	2133	9.5754e-03	8.0361e-03	2.4822e-02	2.4868e-02	6.2335e-02	6.1754e-02	6.9744e-02	1.5304e-01	9.8799e-01
8	0.04740	2618	7.5556e-03	6.4270e-03	1.9953e-02	2.0158e-02	5.0083e-02	4.9503e-02	5.6470e-02	1.3053e-01	9.8342e-01
9	0.04212	3296	5.5976e-03	4.9136e-03	1.3776e-02	1.3823e-02	3.5173e-02	3.4714e-02	3.9492e-02	9.9287e-02	9.7700e-01

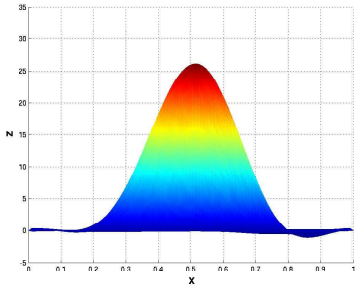
TAB. 11.3.4: Erreurs numériques sur les maillages hybrides pour la condition initiale sinusoïde (11.3.3).



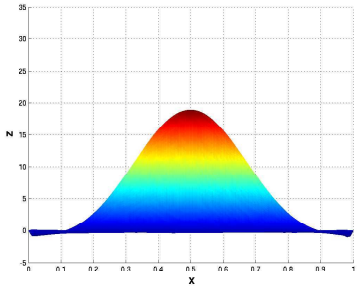
(a) Condition initiale



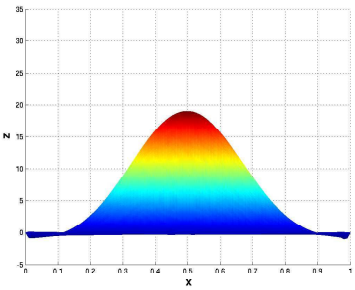
(b) Degré 3 sur 3ème vois.



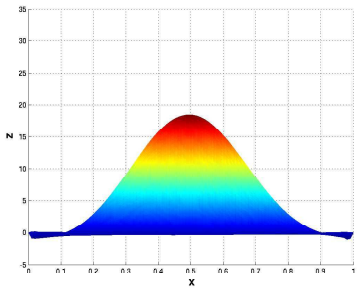
(c) Degré 3 sur 4ème vois.



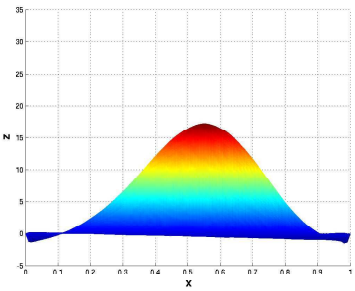
(d) Degré 2 CSE



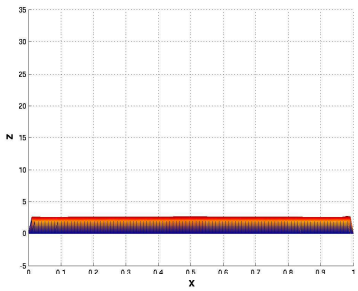
(e) Degré 2 MCCIE



(f) Degré 2 sur 2ème vois.

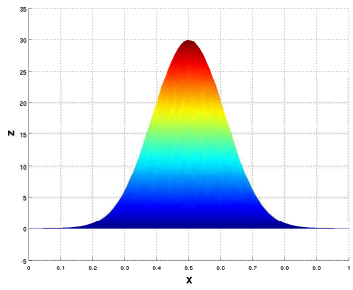


(g) Degré 1 sur 2ème vois.

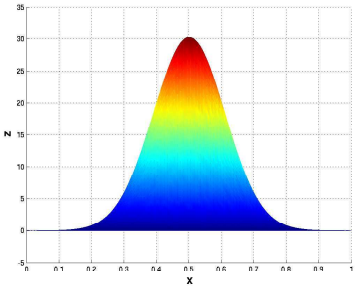


(h) Degré 0

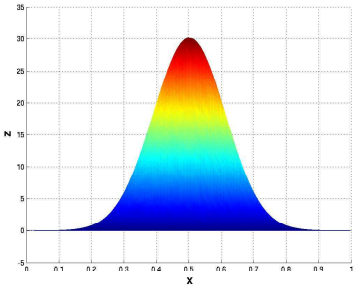
FIG. 11.3.3: Résultats pour la convection de la fonction gaussienne (11.3.2) sur le maillage hybride numéro cinq avec $h \approx 0.06583$ et $N = 1250$: vue du plan $y = 0$.



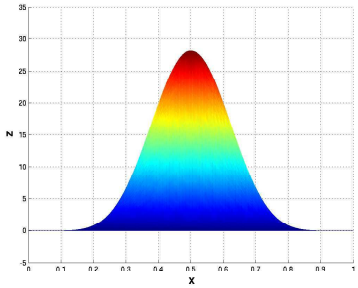
(a) Condition initiale



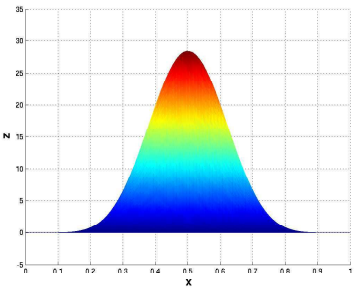
(b) Degré 3 sur 3ème vois.



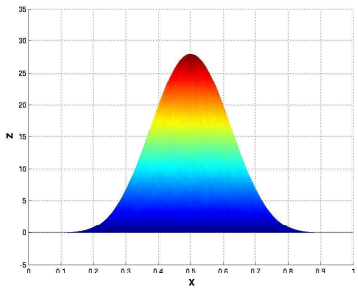
(c) Degré 3 sur 4ème vois.



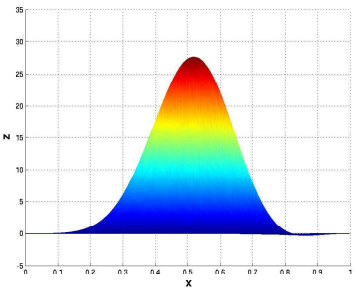
(d) Degré 2 CSE



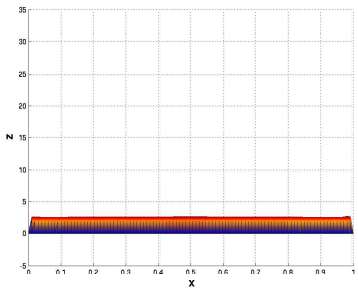
(e) Degré 2 MCCIE



(f) Degré 2 sur 2ème vois.



(g) Degré 1 sur 2ème vois.



(h) Degré 0

FIG. 11.3.4: Résultats pour la convection de la fonction gaussienne (11.3.2) sur le maillage de triangles numéro neuf avec $h \approx 0.02799$ et $N = 5742$: vue du plan $y = 0$.

terme, qui contient la dérivée d'ordre $k+1$ de la solution numérique, disparaît dans certaines conditions géométriques et en particulier en maillage cartésien. Le deuxième terme d'erreur, qui est d'ordre $\mathrm{O}\left(h^{k+1}\right)$, existe même en maillage cartésien et contient la dérivée d'ordre $k+2$ de la solution numérique. Pour une reconstruction de degré un, ce deuxième terme d'erreur contient donc la dérivée troisième. Sous cet angle, on peut interpréter l'erreur pour la reconstruction de degré un comme l'action d'un terme qui contient la dérivée troisième et provoque la déformation asymétrique. Pour les reconstructions de degré deux, la solution devient symétrique autour de l'axe $x = 0,5$. On peut supposer que le terme dominant dans l'équation modifiée pour le degré deux contient la dérivée quatrième, ce qui provoque une erreur dissipative qui est également présente pour la reconstruction de degré un. Pour la reconstruction de degré trois, cette erreur dissipative disparaît visiblement car la gaussienne conserve presque sa hauteur initiale. On aperçoit de nouveau une erreur asymétrique qui peut cette fois-ci être reliée à un terme d'erreur dominant qui contient la dérivée cinquième.

(3) Dans le cas de la reconstruction de degré trois, la dissipation est plus grande pour la reconstruction sur le quatrième voisinage que pour celle sur le troisième voisinage. Cela peut s'expliquer par le théorème 10.6.9 de la section 10.6 qui suggère que la dissipation numérique augmente lorsqu'on élargit le voisinage de reconstruction.

11.4. Bilan du chapitre

L'étude théorique de la section 11.2 a permis d'établir l'équation modifiée des schémas d'ordre $k \geq 1$ pour l'équation de convection linéaire en maillage non structuré général. La proposition 11.2.7 met en évidence quatre termes d'erreur différents dont les trois premiers sont d'ordre $\mathrm{O}\left(h^k\right)$ en maillage non structuré général :

(1) Le terme $\mathfrak{E}_\alpha^{(k)}(v)$ provient de la reconstruction des dérivées.

(2) Le terme $\mathfrak{B}_\alpha^{(k)}(v)$ provient de l'utilisation des moyennes de cellule comme variables fondamentales.

(3) Le terme $\mathfrak{H}_\alpha^{(k)}(v)$ vient du décentrement des flux.

(4) Le terme $\mathfrak{F}_\alpha^{(k)}(v)$ est d'ordre $\mathrm{O}\left(h^{k+1}\right)$ et provient de la dérivée en temps de la moyenne de cellule dans le membre de gauche de l'équation.

Un développement asymptotique montre que le premier terme d'erreur de $\mathfrak{B}_\alpha^{(k)}(v)$, $\mathfrak{E}_\alpha^{(k)}(v)$ et $\mathfrak{H}_\alpha^{(k)}(v)$ disparaît si le maillage satisfait certaines conditions de symétrie. Dans ce cas particulier, $\mathfrak{B}_\alpha^{(k)}(v)$, $\mathfrak{E}_\alpha^{(k)}(v)$ et $\mathfrak{H}_\alpha^{(k)}(v)$ sont d'ordre $\mathrm{O}\left(h^{k+1}\right)$. Le fait que les termes d'erreur d'ordre $\mathrm{O}\left(h^k\right)$ s'annulent sur certains maillages laisse supposer que les termes d'erreur d'ordre $\mathrm{O}\left(h^{k+1}\right)$ sont prépondérants. Pour le démontrer de façon définitive, il faudrait encore expliciter ces termes d'ordre $\mathrm{O}\left(h^{k+1}\right)$, ce qui n'a pas été fait dans le cadre de cette thèse.

L'étude numérique a permis d'évaluer les erreurs pour les reconstructions de degré $k = 0, 1, 2, 3$ sur l'équation de convection linéaire à vitesse constante en dimension deux sur des maillages suivants :

- des maillages de triangles ;
- des maillages hybrides composés de triangles et de quadrangles ;
- des maillages cartésiens.

Une analyse quantitative a consisté à évaluer le taux de convergence de l'erreur numérique en fonction du diamètre des mailles :

(1) En maillage cartésien, le taux de convergence pour les reconstructions de degré $k = 1, 2, 3$ est presque exactement $k+1$. Cela correspond au résultat de la proposition 11.2.10 qui prévoit que le premier terme d'erreur dans l'équation modifiée est d'ordre $\mathrm{O}\left(h^{k+1}\right)$ pour les maillages cartésiens.

(2) En maillage non structuré, le taux de convergence se dégrade par rapport au maillage cartésien et se situe entre k et $k+1$ pour les reconstructions de degré un et deux. Cela

pourrait s'expliquer par les termes d'erreur d'ordre $\mathrm{O}\left(h^k\right)$ mis en évidence par l'étude théorique en maillage non structuré.

(3) En maillage non structuré, la reconstruction de degré trois constitue un cas à part. En maillage de triangles, le taux de convergence est pratiquement égal à celui de la reconstruction de degré deux, alors qu'il est supérieur en maillage cartésien et hybride. Ce phénomène particulier demande une étude approfondie qui n'a pas pu être menée dans le cadre de cette thèse.

(4) L'erreur pour la reconstruction de degré $k = 0$ est presque insensible au diamètre des mailles, au moins pour les maillages testés. Ce phénomène nécessite également une étude plus détaillée.

L'étude numérique a également mis en évidence les différents types de déformation que la condition initiale gaussienne (11.3.2) subit lorsqu'elle est transportée par le schéma numérique. Les types de déformation peuvent être rapprochés avec les termes d'erreur dans l'équation modifiée. La présence d'un terme dominant contenant une dérivée d'ordre impair est reliée à une erreur dispersive sur la vitesse de propagation, ce qui provoque une déformation asymétrique de la solution. La présence d'un terme dominant d'ordre pair est reliée à une erreur dissipative sur l'amplitude de la solution, ce qui engendre une déformation symétrique.

Cette interprétation qualitative des résultats suggère que le passage de la reconstruction de degré un à la reconstruction de degré deux permet surtout de supprimer la dispersion, alors que le passage à la reconstruction de degré trois élimine l'erreur dissipative.

Un résultat numérique très important est le fait que les méthodes de reconstruction MCCIE et CSE fournissent un résultat de la même qualité que la reconstruction de degré deux par la méthode des moindres carrés standard.

CHAPITRE 12

Reconstruction monotone en maillage non structuré

12.1. Objectif du chapitre

Les études de précision et de stabilité des chapitres 10 et 11 reposent sur une analyse de l'équation de convection linéaire et sur l'hypothèse de solutions suffisamment régulières. Lorsque les solutions présentent des discontinuités et des chocs avec des gradients importants, ces approches ne sont plus suffisantes. En général, les schémas d'ordre élevé engendrent dans ces cas des oscillations qui sont des artefacts numériques sans connexion avec la physique de l'écoulement, cf. [**62, 63, 79**]. Un schéma précis doit donc supprimer ces oscillations. De plus, dans le cas des équations de Navier-Stokes, il est possible que ces oscillations, lorsqu'elles sont suffisamment fortes, entraînent des états non physiques, par exemple une température négative dans certaines cellules. Ces oscillations affectent donc également la robustesse et la stabilité des schémas.

Une stratégie très répandue pour supprimer ces oscillations consiste à utiliser des limiteurs de reconstruction et en particulier des limiteurs de gradient dans le contexte du schéma MUSCL. De tels mécanismes étaient déjà présents dans les travaux de Van Leer [**108, 109**]. Les limiteurs de reconstruction sont basés sur l'observation que les oscillations sont en général provoquées par des valeurs reconstruites qui dépassent les minima et maxima au voisinage de la cellule. La méthode consiste à restreindre les valeurs reconstruites aux interfaces à des bornes admissibles, choisies de façon à ce que les oscillations artificielles soient supprimées. Un algorithme de limitation comporte de ce fait deux aspects distincts.

(1) Il faut imposer un *critère de monotonie* aux valeurs reconstruites aux faces. Ce critère de monotonie spécifie sur chaque face les valeurs minimales et maximales admissibles.

(2) Il faut ensuite modifier la fonction reconstruite dans la cellule afin que les valeurs reconstruites aux interfaces respectent les bornes du critère de monotonie.

On constate qu'il existe relativement peu de critères de monotonie qui sont mathématiquement rigoureux et applicables aux maillages non structurés généraux. Un exemple important est le principe du maximum [**8, 10**] qui spécifie des bornes de reconstruction qui assurent que la solution reste confinée entre des minima et maxima locaux au pas de temps suivant. Ce principe est intéressant parce qu'il s'applique de façon générique aux maillages non structurés généraux et découle d'une analyse du schéma semi-discret en espace.

L'étude de ce chapitre se concentre sur le deuxième aspect de l'algorithme de la limitation. L'objectif est de chercher des méthodes efficaces pour modifier les fonctions reconstruites afin qu'elles respectent le critère du principe du maximum. L'étude se restreint au cas de la reconstruction des polynômes de degré un, c'est-à-dire à la reconstruction linéaire par morceaux du schéma MUSCL.

Il faut noter que dans la pratique des simulations numériques, un algorithme de limitation pour la reconstruction de degré un doit respecter un critère supplémentaire : si la solution est une fonction linéaire dans le voisinage de la cellule, le gradient reconstruit ne doit pas être limité. Ce critère concerne essentiellement la géométrie de la reconstruction autour de la cellule. Par manque de temps, une étude approfondie de cette question n'a pas été possible dans cette étude.

Des expériences numériques sont nécessaires pour évaluer ces méthodes de limitation. Un test très simple consiste à isoler une cellule et ses cellules voisines et à générer des valeurs aléatoires de la solution sur ce voisinage. Ces valeurs permettent de reconstruire un gradient dans la cellule auquel on applique l'algorithme de limitation. Il est alors possible de mesurer la différence entre le gradient reconstruit et le gradient limité. La moyenne de ces différences sur un grand nombre

de réalisations permet d'évaluer de façon qualitative et approchée la précision de l'algorithme de limitation.

Pour les équations de Navier-Stokes, l'évaluation des limiteurs se fera sur l'exemple des calculs tridimensionnels décrits dans les chapitres 13 et 14.

12.2. Présentation d'un critère de monotonie : le principe du maximum

Cette section donne une explication rapide du principe du maximum présenté dans [**8, 10**]. Ce résultat spécifie un critère de monotonie qui est adapté de façon générique aux maillages non structurés.

Dans [**8, 10**], le principe du maximum est formulé pour le schéma discret en temps et en espace, bien que sa démonstration repose sur le schéma semi-discret (6.3.7). Il est donc nécessaire d'introduire des variables discrètes en temps et en espace. La solution au temps $t = t_n$ est notée $\mathfrak{u}^n = (u_1^n, \ldots, u_N^n)$.

Pour obtenir un schéma discret en espace et en temps, on applique le schéma d'Euler explicite en temps au schéma semi-discret (6.3.7), ce qui donne le schéma discret

$$u_\alpha^{n+1} = u_\alpha^n - \frac{\Delta t}{|\mathcal{T}_\alpha|} \sum_\beta \sum_q \omega_q \, \widetilde{f}_{\alpha\beta;q} \left(w_\alpha \left[\mathfrak{u}^n \right] (\boldsymbol{x}_{\alpha\beta;q}) , w_\beta \left[\mathfrak{u}^n \right] (\boldsymbol{x}_{\alpha\beta;q}) \right) . \tag{12.2.1}$$

Dans la formule (12.2.1), les $\boldsymbol{x}_{\alpha\beta;q}$ sont les points de Gauss et les ω_q sont les poids de la quadrature.

Le principe du maximum est valable pour des reconstructions linéaires par morceaux. Pour cette raison, l'étude se restreint aux fonctions reconstruites $w_\alpha \left[\mathfrak{u}\,(t) \right] (\boldsymbol{x})$ et $w_\beta \left[\mathfrak{u}\,(t) \right] (\boldsymbol{x})$ qui sont linéaires en $\boldsymbol{x}$ et s'écrivent par conséquent sous la forme (10.2.10)

$$\begin{aligned} w_\alpha \left[\mathfrak{u} \right] (\boldsymbol{x}) = &\quad u_\alpha + \boldsymbol{\sigma}_\alpha \left[\mathfrak{u} \right] \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) \\ w_\beta \left[\mathfrak{u} \right] (\boldsymbol{x}) = &\quad u_\beta + \boldsymbol{\sigma}_\beta \left[\mathfrak{u} \right] \cdot (\boldsymbol{x} - \boldsymbol{x}_\beta) \end{aligned} .$$

Les gradients de cellule

$$\begin{aligned} \boldsymbol{\sigma}_\alpha \left[\mathfrak{u} \right] = &\quad \sum_\gamma \boldsymbol{\sigma}_{\alpha\gamma} \left(u_\gamma - u_\alpha \right) \\ \boldsymbol{\sigma}_\beta \left[\mathfrak{u} \right] = &\quad \sum_\gamma \boldsymbol{\sigma}_{\beta\gamma} \left(u_\gamma - u_\beta \right) \end{aligned}$$

sont reconstruits à partir des moyennes de cellule $\mathfrak{u} = (u_1, \ldots, u_N)$. En chaque point de Gauss, on note la valeur reconstruite

$$u_{\alpha\beta;q} \triangleq u_\alpha + \boldsymbol{\sigma}_\alpha \left[\mathfrak{u} \right] \cdot (\boldsymbol{x}_{\alpha\beta;q} - \boldsymbol{x}_\alpha) . \tag{12.2.2}$$

La notion des voisinages augmentés $\widehat{\mathbb{V}}_\alpha$, définis par (5.4.2), permet d'écrire les minima et maxima locaux pour chaque cellule $\mathcal{T}_\alpha$ comme

$$\left. \begin{aligned} u_\alpha^{\max} \triangleq &\quad \max_{\beta \in \widehat{\mathbb{V}}_\alpha} \{u_\beta\} \\ u_\alpha^{\min} \triangleq &\quad \min_{\beta \in \widehat{\mathbb{V}}_\alpha} \{u_\beta\} \end{aligned} \right\} . \tag{12.2.3}$$

Le principe du maximum nécessite la définition d'un coefficient géométrique $\Gamma_\alpha^{\text{geom}}$ dans chaque cellule. Pour la définition exacte de $\Gamma_\alpha^{\text{geom}}$, on se réfère à [**8**, p. 32] ou à [**10**, p. 21]. Il suffit de préciser que $\Gamma_\alpha^{\text{geom}}$ dépend uniquement de la géométrie de la cellule $\mathcal{T}_\alpha$ et que sa valeur est $d+1$ pour le simplexe en $\mathbb{R}^d$ et 2 pour le parallélogramme et le parallélépipède.

Avec ces définitions, il est possible d'énoncer le résultat suivant.

Théorème 12.2.1 (Principe du maximum pour le schéma Muscl). *On suppose que le flux numérique $\widetilde{f}_{\alpha\beta;q}$ dans (12.2.1) satisfait les critères des définitions 6.2.2 et 6.2.3. De plus, on suppose que les coefficients de quadrature ω_q dans (12.2.1) sont tous positifs et que le pas de temps Δt satisfait la condition CFL*

$$\Gamma_\alpha^{\text{geom}} \frac{\Delta t}{|\mathcal{T}_\alpha|} \sum_\beta \sum_q \omega_q \sup \left\{ \left| \frac{\partial}{\partial w_{\text{int}}} \left[\widetilde{f}_{\alpha\beta;q} \left(\widetilde{w}_{\text{int}}, \widetilde{w}_{\text{ext}} \right) \right] \right| \; ; \; \widetilde{w}_{\text{int}}, \widetilde{w}_{\text{ext}} \in \left[u_\alpha^{\min}, u_\alpha^{\max} \right] \right\} \leq 1 .$$

Si toutes les valeurs $u^n_{\alpha\beta;q}$ satisfont le critère de monotonie

$$\max\left\{u_\alpha^{\min,n}, u_\beta^{\min,n}\right\} \leq u^n_{\alpha\beta;q} \leq \min\left\{u_\alpha^{\max,n}, u_\beta^{\max,n}\right\}\ ,\ \beta \in \mathbb{V}_\alpha\,, \tag{12.2.4}$$

alors la solution u_α^{n+1} à $t_{n+1} = t_n + \Delta t$ appartient à l'intervalle

$$u_\alpha^{\min,n} \leq u_\alpha^{n+1} \leq u_\alpha^{\max,n}\,. \tag{12.2.5}$$

□

La démonstration du théorème 12.2.1 se trouve dans [**10**, p. 21]. Ce théorème nécessite quelques remarques importantes :

(1) La monotonie du flux numérique, énoncée dans la définition 6.2.3, joue un rôle essentiel dans la preuve du principe du maximum.

(2) Dans un sens rigoureux, le résultat n'est démontré que dans le cas d'une loi de conservation scalaire. Il est cependant intéressant de l'appliquer aux systèmes de lois de conservation.

(3) Le théorème 12.2.1 est valable pour le schéma d'Euler explicite en temps. Cette méthode d'intégration en temps n'est pas adaptée à la discrétisation spatiale d'ordre élevé. Ce problème a motivé un effort de recherche très important pour trouver des méthodes d'intégration en temps qui sont d'ordre élevé mais qui restent compatibles avec le principe du maximum, cf. [**95, 56**]. Ces méthodes s'appellent *strong stability preserving methods* ou SSP.

(4) Une reconstruction constante par morceaux respecte les conditions (12.2.4) en raison de la définition (12.2.3).

Finalement, le critère (12.2.4) peut être remplacé par le critère

$$\min\left\{u^n_\alpha, u^n_\beta\right\} \leq u^n_{\alpha\beta;q} \leq \max\left\{u^n_\alpha, u^n_\beta\right\}\ ,\ \beta \in \mathbb{V}_\alpha\,. \tag{12.2.6}$$

Le critère de monotonie (12.2.6) est plus restrictif que (12.2.4) car

$$\left.\begin{aligned}\max\left\{u_\alpha^{\min,n}, u_\beta^{\min,n}\right\} &\leq \min\left\{u^n_\alpha, u^n_\beta\right\}\\ \min\left\{u_\alpha^{\max,n}, u_\beta^{\max,n}\right\} &\geq \max\left\{u^n_\alpha, u^n_\beta\right\}\end{aligned}\right\}\ ,\ \beta \in \mathbb{V}_\alpha\,,$$

mais il est plus simple à implémenter.

Dans le cadre de cette thèse, les deux critères (12.2.4) et (12.2.6) ont été implémentés et testés dans Cedre.

12.3. Interprétation géométrique du critère de monotonie

Le théorème 12.2.1 permet d'établir une stratégie simple pour supprimer des oscillations artificielles dans le voisinage de chocs et de discontinuités. Il suffit que les valeurs reconstruites $u^n_{\alpha\beta}$ à $t = t_n$ respectent le critère de monotonie (12.2.4) pour que la solution u_α^{n+1} à $t = t_n + \Delta t$ reste confinée entre les minima et maxima locaux au temps $t = t_n$, cf. l'équation (12.2.5). Ce mécanisme empêche en particulier l'apparition de nouveaux extrema.

Dans le contexte de la reconstruction linéaire par morceaux, les valeurs reconstruites $u^n_{\alpha\beta;q}$ sont déterminées par le gradient de cellule $\boldsymbol{\sigma}_\alpha[\mathfrak{u}^n]$, d'après la formule (12.2.2). Le principe de tout algorithme de limitation basé sur le principe du maximum est donc de remplacer le gradient $\boldsymbol{\sigma}_\alpha[\mathfrak{u}^n]$, reconstruit à partir des moyennes de cellule $\mathfrak{u}^n = (u_1^n, \ldots, u_N^n)$, par un gradient $\widetilde{\boldsymbol{\sigma}}_\alpha$ qui respecte les conditions (12.2.4). Dans la suite, le gradient initialement reconstruit $\boldsymbol{\sigma}_\alpha[\mathfrak{u}^n]$ est appelé *gradient brut* et tout gradient $\widetilde{\boldsymbol{\sigma}}_\alpha$ qui respecte les inégalités (12.2.4) est appelé *gradient monotone*.

Il est toujours possible de trouver un gradient monotone car le gradient nul $\widetilde{\boldsymbol{\sigma}}_\alpha = 0$ satisfait le critère de monotonie (12.2.4), cf. la remarque à la fin de la section 12.2. Il est cependant évident que le choix du gradient nul est mauvais car il dégrade la précision du schéma numérique de façon excessive. L'objectif d'un algorithme de limitation en maillage non structuré consiste donc à déterminer un gradient monotone qui est le plus proche possible du gradient brut.

Pour atteindre cet objectif, il est utile d'interpréter la contrainte (12.2.4) comme une condition géométrique. Pour simplifier l'analyse, on se restreint au cas où chaque face $\mathcal{A}_{\alpha\beta}$ n'a qu'un seul point de Gauss, ce qui est en général suffisant dans le cas d'une reconstruction linéaire par morceaux. On peut supposer que ce point coïncide avec le barycentre $\boldsymbol{x}_{\alpha\beta}$ de la face et on note la valeur reconstruite en $\boldsymbol{x}_{\alpha\beta}$

$$u^n_{\alpha\beta} \triangleq u^n_\alpha + \boldsymbol{\sigma}_\alpha [\mathfrak{u}^n] \cdot \boldsymbol{k}_{\alpha\beta} .$$

Dans la suite, on omet l'indexation temporelle t_n et on écrit $u_{\alpha\beta}$ au lieu de $u^n_{\alpha\beta}$. On omet également la dépendance de $\boldsymbol{\sigma}_\alpha [\mathfrak{u}]$ de $\mathfrak{u} = (u_1, \ldots, u_N)$ et on écrit $\boldsymbol{\sigma}_\alpha$ au lieu de $\boldsymbol{\sigma}_\alpha [\mathfrak{u}]$.

Dans le contexte de la reconstruction linéaire par morceaux, les conditions (12.2.4) s'écrivent sous la forme de conditions imposées au gradient de cellule $\boldsymbol{\sigma}_\alpha$, données par

$$\max \left\{ u^{\min}_\alpha, u^{\min}_\beta \right\} \leq u_\alpha + \boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} \leq \min \left\{ u^{\max}_\alpha, u^{\max}_\beta \right\} , \ \beta \in \mathbb{V}_\alpha . \tag{12.3.1}$$

On introduit les définitions

$$\left.\begin{array}{rl} w^{\max}_{\alpha\beta} \triangleq & \min \left\{ u^{\max}_\alpha, u^{\max}_\beta \right\} - u_\alpha \\ w^{\min}_{\alpha\beta} \triangleq & \max \left\{ u^{\min}_\alpha, u^{\min}_\beta \right\} - u_\alpha \end{array}\right\} \tag{12.3.2}$$

qui permettent d'écrire les conditions du critère de monotonie (12.3.1) dans la cellule $\mathcal{T}_\alpha$ comme

$$w^{\min}_{\alpha\beta} \leq \boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} \leq w^{\max}_{\alpha\beta} , \ \beta \in \mathbb{V}_\alpha . \tag{12.3.3}$$

Les définitions (12.2.3) et (12.3.2) impliquent

$$w^{\min}_{\alpha\beta} \leq 0 \leq w^{\max}_{\alpha\beta} , \ \beta \in \mathbb{V}_\alpha \tag{12.3.4}$$

ce qui montre que le vecteur nul $\boldsymbol{\sigma}_\alpha = 0$ satisfait toujours les conditions (12.3.3).

Pour expliciter la géométrie des contraintes (12.3.3), on introduit les sous-ensembles de $\mathbb{R}^d$ définis par

$$\mathcal{S}_{\alpha\beta} \triangleq \left\{ \boldsymbol{\sigma} \in \mathbb{R}^d \,\middle|\, w^{\min}_{\alpha\beta} \leq \boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta} \leq w^{\max}_{\alpha\beta} \right\} , \ \beta \in \mathbb{V}_\alpha . \tag{12.3.5}$$

Pour un indice $\beta \in \mathbb{V}_\alpha$ donné, la région $\mathcal{S}_{\alpha\beta}$ est un sous-ensemble fermé et convexe de $\mathbb{R}^d$ qui est délimité par les hyperplans affines $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta} = w^{\max}_\alpha$ et $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta} = w^{\min}_\alpha$. Les inégalités (12.3.4) montrent que chaque $\mathcal{S}_{\alpha\beta}$ est non vide car il contient au moins le gradient nul $\boldsymbol{\sigma}_\alpha = 0$.

La définition (12.3.5) montre qu'un gradient de cellule $\boldsymbol{\sigma}_\alpha$ satisfait les conditions (12.3.3) si et seulement si $\boldsymbol{\sigma}_\alpha$ est dans tous les $\mathcal{S}_{\alpha\beta}$ pour $\beta \in \mathbb{V}_\alpha$. L'intersection des $\mathcal{S}_{\alpha\beta}$ pour $\beta \in \mathbb{V}_\alpha$ joue donc un rôle important pour la limitation des gradients. Cela justifie l'introduction de la notion suivante qui est également employée dans [**68, 86, 22**].

Définition 12.3.1 (Région du principe du maximum). On appelle *région du principe du maximum* l'intersection des ensembles $\mathcal{S}_{\alpha\beta}$

$$\mathcal{S}_\alpha \triangleq \bigcap_{\beta \in \mathbb{V}_\alpha} \mathcal{S}_{\alpha\beta} .$$

□

Pour la suite, il est utile de caractériser l'ensemble $\mathcal{S}_\alpha$ par la

Proposition 12.3.2 (Propriétés de la région du principe du maximum). *Sous l'hypothèse que la famille de vecteurs $\boldsymbol{k}_{\alpha\beta}$, $\beta \in \mathbb{V}_\alpha$, contient d vecteurs libres, l'ensemble $\mathcal{S}_\alpha$ est un sous-ensemble convexe et compact de $\mathbb{R}^d$ qui contient l'origine.*

Démonstration. Les inégalités (12.3.4) entraînent que chaque $\mathcal{S}_{\alpha\beta}$ contient l'origine $\boldsymbol{\sigma} = 0$, ce qui prouve que $\mathcal{S}_\alpha$ contient l'origine. Chaque ensemble $\mathcal{S}_{\alpha\beta}$ est fermé et convexe parce que chaque $\mathcal{S}_{\alpha\beta}$ est défini par deux inégalités linéaires dans (12.3.3). Puisque $\mathcal{S}_\alpha$ est l'intersection des $\mathcal{S}_{\alpha\beta}$, $\mathcal{S}_\alpha$ est lui-même fermé et convexe. Pour montrer que $\mathcal{S}_\alpha$ est compact, il suffit donc de montrer que $\mathcal{S}_\alpha$ est borné.

Supposons le contraire, c'est-à-dire que $\mathcal{S}_\alpha$ ne soit pas borné. Dans ce cas, il existe une suite de vecteurs $\left\{\widehat{\boldsymbol{\sigma}}^{(n)} \in \mathbb{R}^d\right\}_{n\in\mathbb{N}}$ telle que $n \leq \left\|\widehat{\boldsymbol{\sigma}}^{(n)}\right\|_2$ et $\widehat{\boldsymbol{\sigma}}^{(n)} \in \mathcal{S}_\alpha$. Comme $\mathcal{S}_\alpha$ est convexe et contient l'origine, la suite de vecteurs

$$\boldsymbol{\sigma}^{(n)} = \frac{\widehat{\boldsymbol{\sigma}}^{(n)}}{\left\|\widehat{\boldsymbol{\sigma}}^{(n)}\right\|_2}$$

est également contenue dans $\mathcal{S}_\alpha$. Elle satisfait par construction

$$\left\|\boldsymbol{\sigma}^{(n)}\right\|_2 = 1$$

et

$$\frac{w_{\alpha\beta}^{\min}}{n} \leq \frac{w_{\alpha\beta}^{\min}}{\left\|\widehat{\boldsymbol{\sigma}}^{(n)}\right\|_2} \leq \boldsymbol{\sigma}^{(n)} \cdot \boldsymbol{k}_{\alpha\beta} \leq \frac{w_{\alpha\beta}^{\max}}{\left\|\widehat{\boldsymbol{\sigma}}^{(n)}\right\|_2} \leq \frac{w_{\alpha\beta}^{\max}}{n}\,,\ \beta \in \mathbb{V}_\alpha \tag{12.3.6}$$

car $w_{\alpha\beta}^{\min} \leq 0$ et $0 \leq w_{\alpha\beta}^{\max}$. Comme la sphère unité est compacte, il existe une sous-suite de $\left\{\boldsymbol{\sigma}^{(n)}\right\}_{n\in\mathbb{N}}$ qui converge vers un $\boldsymbol{\sigma}_0 \in \mathbb{R}^d$ tel que $\|\boldsymbol{\sigma}_0\|_2 = 1$. Par passage à la limite dans (12.3.6), il vient

$$\boldsymbol{\sigma}_0 \cdot \boldsymbol{k}_{\alpha\beta} = 0 \text{ pour tout } \beta \in \mathbb{V}_\alpha\,.$$

D'après l'hypothèse, la famille des $\boldsymbol{k}_{\alpha\beta}$ contient d vecteurs libres. Il s'ensuit donc que $\boldsymbol{\sigma}_0 = 0$, ce qui est en contradiction avec $\|\boldsymbol{\sigma}_0\|_2 = 1$. L'ensemble $\mathcal{S}_\alpha$ est par conséquent borné. Comme il est également fermé, il est compact. □

La définition 12.3.1 et la proposition 12.3.2 permettent de voir le problème de la limitation du gradient comme un problème géométrique. Ce problème consiste à trouver le gradient monotone $\widetilde{\boldsymbol{\sigma}}_\alpha \in \mathcal{S}_\alpha$ qui approche au mieux le gradient brut $\boldsymbol{\sigma}_\alpha \notin \mathcal{S}_\alpha$. Dans la suite, on appelle les algorithmes de limitation qui se basent sur cette approche géométrique des méthodes de *limitation directionnelle* car ils prennent en compte la géométrie multidimensionnelle des conditions de monotonie (12.3.3).

Une idée naturelle est de choisir une norme vectorielle $\|.\|_V$ dans $\mathbb{R}^d$ et de chercher un $\widetilde{\boldsymbol{\sigma}}_\alpha \in \mathcal{S}_\alpha$ qui est une solution du problème de minimisation

$$\|\widetilde{\boldsymbol{\sigma}}_\alpha - \boldsymbol{\sigma}_\alpha\|_V = \min\left\{\|\widehat{\boldsymbol{\sigma}}_\alpha - \boldsymbol{\sigma}_\alpha\|_V\ ;\ \widehat{\boldsymbol{\sigma}}_\alpha \in \mathcal{S}_\alpha\right\}\,. \tag{12.3.7}$$

Cette approche est proposée dans [**13**] pour la norme

$$\|\boldsymbol{\sigma}\|_1 \triangleq \sum_{i=1}^{d} |\sigma_i|$$

et dans [**86**] pour la norme $\|.\|_2$.

L'avantage de cette approche est qu'elle est bien posée sur le plan mathématique. Toute norme vectorielle $\|.\|_V$ dans $\mathbb{R}^d$ est une fonction convexe et continue qui est bornée inférieurement. Comme $\mathcal{S}_\alpha$ est par ailleurs un ensemble convexe et compact, le problème (12.3.7) admet des solutions dans $\mathcal{S}_\alpha$. Si $\|.\|_V$ est la norme $\|.\|_2$ induite par le produit scalaire canonique dans $\mathbb{R}^d$, la solution de (12.3.7) est unique.

L'approche de la minimisation (12.3.7) présente, par contre, deux désavantages :

(1) Premièrement, un algorithme de résolution pour (12.3.7) peut être coûteux en temps calcul car il faut déterminer la solution dans chaque cellule ou le gradient brut n'est pas monotone.

(2) Deuxièmement, de tels algorithmes peuvent être difficiles à implémenter sur des calculateurs vectoriels. En particulier, le nombre de faces et, avec lui, le nombre de contraintes peut varier d'une cellule à l'autre, ce qui peut empêcher la vectorisation de certaines boucles.

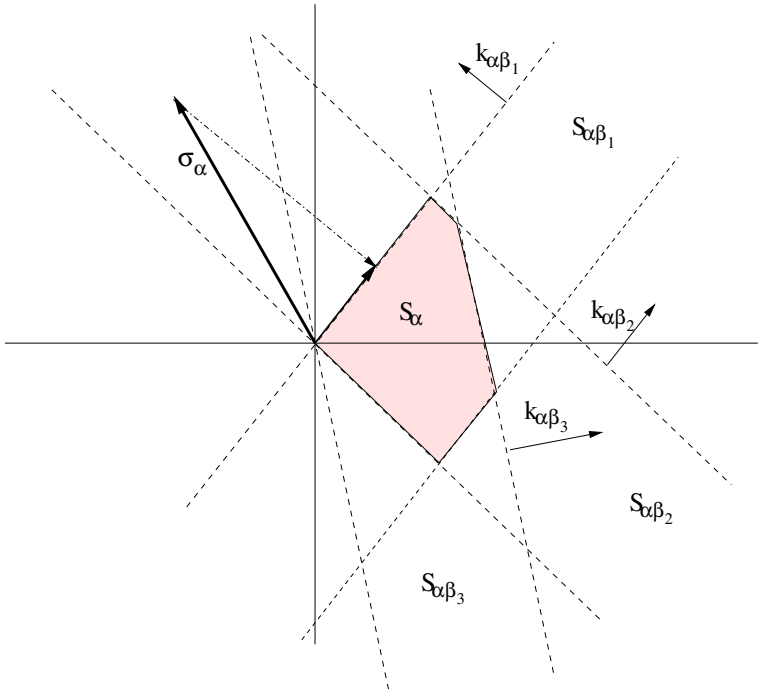


FIG. 12.3.1: Exemple d'une région du principe du maximum

Pour ces raisons, on propose souvent dans la littérature des stratégies plus simples pour obtenir un gradient monotone.

La plus simple consiste à multiplier le gradient brut $\boldsymbol{\sigma}_\alpha$ par un facteur $\xi_\alpha \in [0,1]$ de façon à ce que $\xi_\alpha \boldsymbol{\sigma}_\alpha \in \mathcal{S}_\alpha$. Dans la suite, on appelle ce type de méthode *limitation non directionnelle*. Les méthodes non directionnelles se distinguent entre elles uniquement par la façon de calculer ξ_α, voir [**8, 10**] pour des exemples en maillage non structuré. Elles permettent toujours de trouver au moins un gradient monotone car la valeur $\xi_\alpha = 0$ donne le gradient nul qui est monotone. Le grand désavantage de cette famille de méthodes réside dans le fait que le gradient peut être réduit de façon inappropriée. La figure 12.3.1 montre l'exemple d'une région du principe du maximum pour une cellule triangulaire dans $\mathbb{R}^2$. Les lignes pointillées délimitent les trois ensembles $\mathcal{S}_{\alpha\beta_1}$, $\mathcal{S}_{\alpha\beta_2}$ et $\mathcal{S}_{\alpha\beta_3}$ qui correspondent aux trois faces du triangle. La figure montre un gradient brut $\boldsymbol{\sigma}_\alpha$ à l'extérieur de la région du principe du maximum. Dans le cas spécifique montré ici, toute méthode de limitation non directionnelle donne un gradient nul. La figure révèle cependant qu'il existe un gradient non nul qui est plus proche de $\boldsymbol{\sigma}_\alpha$ dans la norme $\|.\|_2$ que le gradient nul. Dans ce cas particulier, il s'agit de la projection orthogonale de $\boldsymbol{\sigma}_\alpha$ sur la face la plus proche de la région du principe du maximum.

Cet exemple spécifique montre qu'il est nécessaire de réfléchir à des algorithmes approchés qui représentent un bon compromis entre la limitation non directionnelle et la solution exacte du problème (12.3.7).

12.4. Algorithmes approchés pour la limitation directionnelle

L'objectif de cette section est d'explorer des algorithmes pour résoudre le problème de minimisation (12.3.7) de façon approchée. Un tel algorithme doit respecter deux critères.

(1) Premièrement, il ne devrait pas consommer plus de deux ou trois fois le temps d'une limitation non directionnelle.

(2) Deuxièmement, toutes les boucles de l'algorithme doivent être vectorisables.

Les deux exigences ci-dessus restreignent fortement le choix des méthodes. Il faut tenir compte du fait que même une limitation non directionnelle du gradient consomme une part relativement importante en temps calcul, en particulier avec un schéma explicite en temps. Cela vient du fait

que l'algorithme de limitation doit calculer des minima et maxima locaux, comparer les valeurs aux faces avec les bornes admissibles et ajuster ces valeurs si nécessaire. Cela prend en général plus de temps que l'étape de reconstruction du gradient qui implique seulement des sommes sur des vecteurs et leur multiplication par des matrices calculées une fois pour toutes en début de calcul.

Un premier exemple d'un algorithme approché s'inspire de la discussion dans [**68**]. Il s'agit de résoudre de façon très approchée le problème de minimisation (12.3.7) pour la norme euclidienne $\|.\|_2$ dans $\mathbb{R}^d$. Puisque $\mathcal{S}_\alpha$ est l'intersection des ensembles $\mathcal{S}_{\alpha\beta}$ pour $\beta \in \mathbb{V}_\alpha$, l'idée est de chercher successivement pour chaque $\mathcal{S}_{\alpha\beta}$ le gradient $\widetilde{\boldsymbol{\sigma}}_\alpha \in \mathcal{S}_{\alpha\beta}$ qui est une solution du problème de minimisation

$$\|\widetilde{\boldsymbol{\sigma}}_\alpha - \boldsymbol{\sigma}_\alpha\|_2 = \min\left\{\|\widehat{\boldsymbol{\sigma}}_\alpha - \boldsymbol{\sigma}_\alpha\|_2 \ ; \ \widehat{\boldsymbol{\sigma}}_\alpha \in \mathcal{S}_{\alpha\beta}\right\} . \tag{12.4.1}$$

Puisque $\mathcal{S}_{\alpha\beta}$ est un ensemble fermé et convexe, le problème (12.4.1) possède une solution unique pour tout $\boldsymbol{\sigma}_\alpha \in \mathbb{R}^d$.

Cela permet d'introduire la

DÉFINITION 12.4.1 (Projection sur $\mathcal{S}_{\alpha\beta}$). Le problème (12.4.1) définit une application $\mathfrak{P}_{\alpha\beta} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ appelée *projection* sur $\mathcal{S}_{\alpha\beta}$. Pour tout $\boldsymbol{\sigma} \in \mathbb{R}^d$, $\mathfrak{P}_{\alpha\beta}(\boldsymbol{\sigma})$ est défini comme la solution du problème

$$\|\mathfrak{P}_{\alpha\beta}(\boldsymbol{\sigma}) - \boldsymbol{\sigma}\|_2 = \min\left\{\|\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2 \ ; \ \widehat{\boldsymbol{\sigma}} \in \mathcal{S}_{\alpha\beta}\right\} .$$

□

La définition des ensembles $\mathcal{S}_{\alpha\beta}$ permet d'expliciter la solution du problème (12.4.1). Si $\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} > w^{\max}_{\alpha\beta}$, le point $\mathfrak{P}_{\alpha\beta}(\boldsymbol{\sigma}_\alpha)$ est la projection orthogonale de $\boldsymbol{\sigma}_\alpha$ sur l'hyperplan $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta} = w^{\max}_{\alpha\beta}$ donnée par

$$\mathfrak{P}_{\alpha\beta}(\boldsymbol{\sigma}_\alpha) = \boldsymbol{\sigma}_\alpha - \boldsymbol{k}_{\alpha\beta}\frac{\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} - w^{\max}_{\alpha\beta}}{\|\boldsymbol{k}_{\alpha\beta}\|_2^2} . \tag{12.4.2}$$

Si $\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} < w^{\min}_{\alpha\beta}$, le point $\mathfrak{P}_{\alpha\beta}(\boldsymbol{\sigma}_\alpha)$ est la projection orthogonale de $\boldsymbol{\sigma}_\alpha$ sur l'hyperplan $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta} = w^{\min}_{\alpha\beta}$ donnée par

$$\mathfrak{P}_{\alpha\beta}(\boldsymbol{\sigma}_\alpha) = \boldsymbol{\sigma}_\alpha - \boldsymbol{k}_{\alpha\beta}\frac{\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} - w^{\min}_{\alpha\beta}}{\|\boldsymbol{k}_{\alpha\beta}\|_2^2} . \tag{12.4.3}$$

Ces observations permettent d'implémenter un algorithme peu coûteux en temps calcul qui est défini de la façon suivante.

ALGORITHME 12.4.2 (Limitation directionnelle approchée I). *Pour chaque cellule $\mathcal{T}_\alpha$, l'algorithme consiste à parcourir les conditions de monotonie (12.3.3)*

$$w^{\min}_{\alpha\beta} \leq \boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} \leq w^{\max}_{\alpha\beta} \ , \ \beta \in \mathbb{V}_\alpha$$

dans un ordre arbitraire. Soit $\beta_1 \in \mathbb{V}_\alpha$ le premier indice tel que $\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta_1} < w^{\min}_{\alpha\beta_1}$ ou $\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta_1} > w^{\max}_{\alpha\beta_1}$. Dans ce cas, l'algorithme remplace le gradient brut $\boldsymbol{\sigma}_\alpha$ par sa projection $\boldsymbol{\sigma}^{(\beta_1)}_\alpha = \mathfrak{P}_{\alpha\beta_1}(\boldsymbol{\sigma}_\alpha)$ sur la région $\mathcal{S}_{\alpha\beta_1}$, donnée selon le cas par (12.4.2) ou (12.4.3).

L'algorithme parcourt ensuite les contraintes restantes pour vérifier s'il existe un indice $\beta_2 \in \mathbb{V}_\alpha$ tel que $\boldsymbol{\sigma}^{(\beta_1)}_\alpha \cdot \boldsymbol{k}_{\alpha\beta_2} < w^{\min}_{\alpha\beta_2}$ ou $\boldsymbol{\sigma}^{(\beta_1)}_\alpha \cdot \boldsymbol{k}_{\alpha\beta_2} > w^{\max}_{\alpha\beta_2}$. Si l'algorithme rencontre un tel indice, il remplace $\boldsymbol{\sigma}^{(\beta_1)}_\alpha$ par sa projection $\boldsymbol{\sigma}^{(\beta_2)}_\alpha = \mathfrak{P}_{\alpha\beta_2}\left(\boldsymbol{\sigma}^{(\beta_1)}_\alpha\right)$ sur $\mathcal{S}_{\alpha\beta_2}$, donnée selon le cas par (12.4.2) ou (12.4.3). L'opération est répétée jusqu'à ce que toutes les faces aient été parcourues.

Puisque cette méthode ne garantit pas que le gradient résultant est monotone, il faut limiter le gradient obtenu par une limitation non directionnelle qui peut être réalisée par une seule boucle sur les faces. L'algorithme nécessite donc deux boucles sur les faces du maillage, l'une pour l'étape directionnelle, l'autre pour l'étape non directionnelle.□

L'avantage de cet algorithme réside dans le fait qu'il parcourt les faces dans un ordre arbitraire et que chacune des étapes concerne uniquement une face à la fois. Cet algorithme est

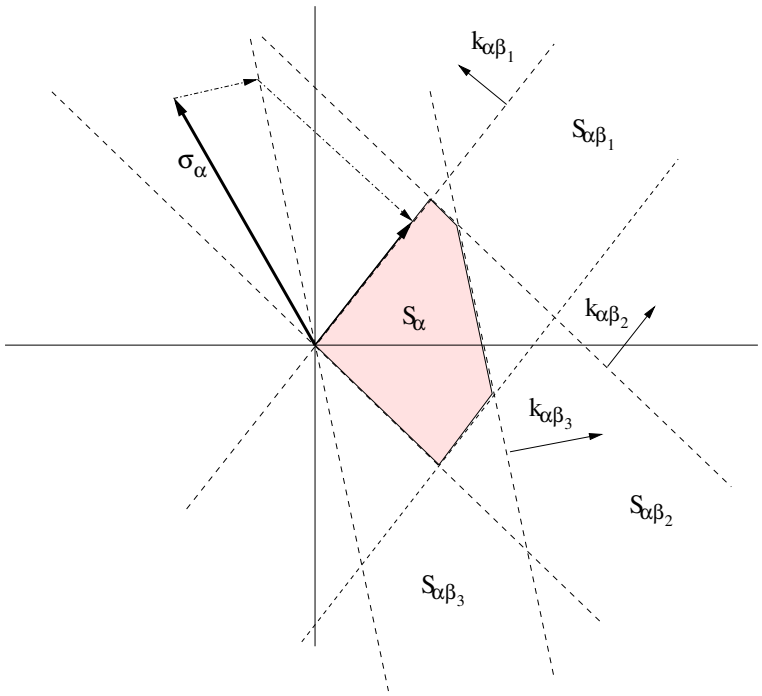


FIG. 12.4.1: Exemple d'une région du principe du maximum

donc facilement vectorisable en maillage non structuré général où **tout algorithme doit être réalisé comme une boucle sur les faces** du maillage.

Le résultat de cet algorithme dépend par contre de l'ordre dans lequel il parcourt les faces. Les figures 12.3.1 et 12.4.1 illustrent cette situation à l'aide d'un exemple.

La figure 12.3.1 montre un premier cas où le gradient brut $\boldsymbol{\sigma}_\alpha$ est d'abord projeté sur l'hyperplan $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta_1} = w^{\min}_{\alpha\beta_1}$. Dans cet exemple, le gradient obtenu est monotone et l'algorithme se termine. La figure 12.4.1 montre un deuxième cas où le gradient brut $\boldsymbol{\sigma}_\alpha$ est d'abord projeté sur l'hyperplan affine $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta_3} = w^{\min}_{\alpha\beta_3}$. Le gradient ainsi obtenu n'est pas encore monotone mais il est ensuite projeté sur $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta_1} = w^{\min}_{\alpha\beta_1}$, ce qui donne un gradient monotone situé sur le bord de la région du principe du maximum. On constate que cette méthode ne donne pas le même résultat suivant l'ordre des projections. Elle fournit cependant dans chaque cas un gradient monotone non nul alors que toute limitation non directionnelle donnerait un gradient limité nul.

Une variante de cet algorithme vient de l'idée de modifier la projection des gradients sur les hyperplans $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta} = w^{\min}_{\alpha\beta}$ ou $\boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta} = w^{\max}_{\alpha\beta}$. On observe que la projection du gradient brut $\boldsymbol{\sigma}_\alpha$ sur l'un des $\mathcal{S}_{\alpha\beta}$ peut augmenter la valeur absolue de certaines composantes de $\boldsymbol{\sigma}_\alpha$. En d'autres termes, il peut exister un $i \in \{1, \ldots, d\}$ tel que

$$\left|(\boldsymbol{\sigma}_\alpha)_i\right| < \left|\left(\widetilde{\boldsymbol{\sigma}}^{(\beta)}_\alpha\right)_i\right| .$$

Pour éviter ce phénomène, il faut déterminer le $\breve{\boldsymbol{\sigma}}^{(\beta)}_\alpha \in \mathcal{S}_{\alpha\beta}$ qui minimise la distance à $\boldsymbol{\sigma}_\alpha$ sous les contraintes

$$\min\{\sigma_{\alpha,i}, 0\} \leq \left(\breve{\boldsymbol{\sigma}}^{(\beta)}_\alpha\right)_i \leq \max\{\sigma_{\alpha,i}, 0\} \ , \ 1 \leq i \leq d \, .$$

Le gradient $\breve{\boldsymbol{\sigma}}^{(\beta)}_\alpha \in \mathcal{S}_{\alpha\beta}$ appartient donc à la région

$$\mathcal{C}_\alpha \triangleq \left\{ \boldsymbol{s} \in \mathbb{R}^d \middle| \min\{\sigma_{\alpha,i}, 0\} \leq s_i \leq \max\{\sigma_{\alpha,i}, 0\} \ , \ 1 \leq i \leq d \right\}$$

qui est un sous-ensemble convexe et compact de $\mathbb{R}^d$ et contient l'origine.

Le gradient cherché $\breve{\boldsymbol{\sigma}}^{(\beta)}_\alpha \in \mathcal{S}_{\alpha\beta} \cap \mathcal{C}_\alpha$ est alors solution du problème de minimisation

$$\left\| \breve{\boldsymbol{\sigma}}^{(\beta)}_\alpha - \boldsymbol{\sigma}_\alpha \right\|_2 = \min\left\{ \|\widehat{\boldsymbol{\sigma}}_\alpha - \boldsymbol{\sigma}_\alpha\|_2 \ ; \ \widehat{\boldsymbol{\sigma}}_\alpha \in \mathcal{S}_{\alpha\beta} \cap \mathcal{C}_\alpha \right\} . \tag{12.4.4}$$

Le problème (12.4.4) admet une unique solution car $\mathcal{S}_{\alpha\beta} \cap \mathcal{C}_\alpha$ est un sous-ensemble convexe et compact de $\mathbb{R}^d$.

Cela permet d'introduire la

DÉFINITION 12.4.3 (Projection sur $\mathcal{S}_{\alpha\beta} \cap \mathcal{C}_\alpha$). Le problème (12.4.4) définit une application $\breve{\mathfrak{P}}_{\alpha\beta} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, appelée *projection* sur $\mathcal{S}_{\alpha\beta} \cap \mathcal{C}_\alpha$. Pour tout $\boldsymbol{\sigma} \in \mathbb{R}^d$, $\breve{\mathfrak{P}}_{\alpha\beta}(\boldsymbol{\sigma})$ est défini comme la solution du problème

$$\left\| \breve{\mathfrak{P}}_{\alpha\beta}(\boldsymbol{\sigma}) - \boldsymbol{\sigma} \right\|_2 = \min \left\{ \|\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2 \, ; \, \widehat{\boldsymbol{\sigma}} \in \mathcal{S}_{\alpha\beta} \cap \mathcal{C}_\alpha \right\} .$$

□

La forme du problème de minimisation (12.4.4) garantit que les valeurs absolues des composantes de $\breve{\mathfrak{P}}_{\alpha\beta}(\boldsymbol{\sigma})$ ne dépassent pas celles de $\boldsymbol{\sigma}$.

Ces notions permettent de formuler une variante de l'algorithme 12.4.2.

ALGORITHME 12.4.4 (Limitation directionnelle approchée II). *L'algorithme consiste à parcourir les conditions de monotonie (12.3.3)*

$$w_{\alpha\beta}^{\min} \leq \boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} \leq w_{\alpha\beta}^{\max} \, , \, \beta \in \mathbb{V}_\alpha$$

de la même façon que l'algorithme 12.4.2, mais en remplaçant les projections $\mathfrak{P}_{\alpha\beta}$ de la définition 12.4.1 par les projections $\breve{\mathfrak{P}}_{\alpha\beta}$ de la définition 12.4.3.

Puisque cette méthode ne garantit pas que le gradient résultant soit monotone, il faut limiter le gradient obtenu par une limitation non directionnelle qui peut être réalisée par une seule boucle sur les faces. L'algorithme nécessite donc deux boucles sur les faces du maillage, l'une pour l'étape directionnelle, l'autre pour l'étape non directionnelle.□

Il est nécessaire de calculer les projections $\breve{\mathfrak{P}}_{\alpha\beta}$ pour implémenter l'algorithme 12.4.4. Une formule approchée pour $\breve{\mathfrak{P}}_{\alpha\beta}$ a été implémentée et testée dans CEDRE, cf. la section 12.5.

Une autre idée s'inspire de l'algorithme de limitation donné dans [**86**]. Un problème de l'algorithme 12.4.2 vient du fait que les projections successives s'effectuent de manière indépendante. Supposons que le gradient brut $\boldsymbol{\sigma}_\alpha$ n'appartienne pas à la région $\mathcal{S}_{\alpha\beta_1}$ pour un β_1 donné. Dans ce cas, l'algorithme remplace $\boldsymbol{\sigma}_\alpha$ par sa projection sur $\mathcal{S}_{\alpha\beta_1}$ qui est donnée par $\boldsymbol{\sigma}_\alpha^{(\beta_1)} = \mathfrak{P}_{\alpha\beta}(\boldsymbol{\sigma}_\alpha)$, d'après la définition 12.4.1. Supposons alors qu'il existe un autre indice β_2 tel que $\boldsymbol{\sigma}_\alpha^{(\beta_1)} \notin \mathcal{S}_{\alpha\beta_2}$. Dans ce cas, on a $w_{\alpha\beta_2}^{\min} > \boldsymbol{\sigma}_\alpha^{(\beta_1)} \cdot \boldsymbol{k}_{\alpha\beta_2}$ ou $\boldsymbol{\sigma}_\alpha^{(\beta_1)} \cdot \boldsymbol{k}_{\alpha\beta_2} > w_{\alpha\beta_2}^{\max}$. L'algorithme remplace dans ce cas $\boldsymbol{\sigma}_\alpha^{(\beta_1)}$ par sa projection sur $\mathcal{S}_{\alpha\beta_2}$, notée $\boldsymbol{\sigma}_\alpha^{(\beta_2)} = \mathfrak{P}_{\alpha\beta}\left(\boldsymbol{\sigma}_\alpha^{(\beta_1)}\right)$, mais il ne peut pas garantir que $\boldsymbol{\sigma}_\alpha^{(\beta_2)}$ appartienne à $\mathcal{S}_{\alpha\beta_1}$. On constate que la projection du gradient sur une région $\mathcal{S}_{\alpha\beta_2}$ peut le sortir d'une région $\mathcal{S}_{\alpha\beta_1}$ à laquelle il appartenait précédemment.

Un remède consiste à remplacer les projections $\mathfrak{P}_{\alpha\beta}$ de la définition 12.4.1 par des projections qui s'effectuent dans une direction perpendiculaire à un vecteur $\boldsymbol{k}_{\alpha\gamma}$ donné.

DÉFINITION 12.4.5 (Projection sur $\mathcal{S}_{\alpha\beta}$ perpendiculaire à $\boldsymbol{k}_{\alpha\gamma}$). Soient $\mathcal{S}_{\alpha\beta}$ et $\mathcal{S}_{\alpha\gamma}$ deux régions données par les inégalités respectives

$$\left.\begin{array}{l} w_{\alpha\beta}^{\min} \leq \boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\beta} \leq w_{\alpha\beta}^{\max} \\ w_{\alpha\gamma}^{\min} \leq \boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\gamma} \leq w_{\alpha\gamma}^{\max} \end{array}\right\} .$$

On définit une application $\mathfrak{P}_{\alpha\beta}[\boldsymbol{k}_{\alpha\gamma}] : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, appelée *projection sur $\mathcal{S}_{\alpha\beta}$ perpendiculaire à $\boldsymbol{k}_{\alpha\gamma}$*, de la manière suivante : pour tout $\boldsymbol{\sigma} \in \mathbb{R}^d$, $\mathfrak{P}_{\alpha\beta}[\boldsymbol{k}_{\alpha\gamma}](\boldsymbol{\sigma})$ est la solution du problème

$$\|\mathfrak{P}_{\alpha\beta}[\boldsymbol{k}_{\alpha\gamma}](\boldsymbol{\sigma}) - \boldsymbol{\sigma}\|_2 = \min \left\{ \|\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2 \, ; \, \widehat{\boldsymbol{\sigma}} \in \mathcal{S}_{\alpha\beta} \, , \, \widehat{\boldsymbol{\sigma}} \cdot \boldsymbol{k}_{\alpha\gamma} = \boldsymbol{\sigma} \cdot \boldsymbol{k}_{\alpha\gamma} \right\} . \tag{12.4.5}$$

□

La définition 12.4.5 assure que

$$\boldsymbol{\sigma} \in \mathcal{S}_{\alpha\gamma} \text{ entraîne } \mathfrak{P}_{\alpha\beta}[\boldsymbol{k}_{\alpha\gamma}](\boldsymbol{\sigma}) \in \mathcal{S}_{\alpha\gamma} \, .$$

La projection de $\boldsymbol{\sigma}$ sur $\mathcal{S}_{\alpha\beta}$ *perpendiculaire à* $\boldsymbol{k}_{\alpha\gamma}$ conserve donc l'appartenance de $\boldsymbol{\sigma}$ à $\mathcal{S}_{\alpha\gamma}$.

Si $\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} > w_{\alpha\beta}^{\max}$, le point $\mathfrak{P}_{\alpha\beta}\left[\boldsymbol{k}_{\alpha\gamma}\right](\boldsymbol{\sigma}_\alpha)$ est donné par

$$\mathfrak{P}_{\alpha\beta}\left[\boldsymbol{k}_{\alpha\gamma}\right](\boldsymbol{\sigma}_\alpha) = \boldsymbol{\sigma}_\alpha - \frac{\left(\boldsymbol{k}_{\alpha\beta}\left\|\boldsymbol{k}_{\alpha\gamma}\right\|_2^2 - \boldsymbol{k}_{\alpha\gamma}\left(\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{k}_{\alpha\gamma}\right)\right)\left(\boldsymbol{\sigma}_\alpha\cdot\boldsymbol{k}_{\alpha\beta} - w_{\alpha\beta}^{\max}\right)}{\left\|\boldsymbol{k}_{\alpha\beta}\right\|_2^2\left\|\boldsymbol{k}_{\alpha\gamma}\right\|_2^2 - \left(\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{k}_{\alpha\gamma}\right)^2}\,. \tag{12.4.6}$$

Si $\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} < w_{\alpha\beta}^{\min}$, le point $\mathfrak{P}_{\alpha\beta}\left[\boldsymbol{k}_{\alpha\gamma}\right](\boldsymbol{\sigma}_\alpha)$ est donné par

$$\mathfrak{P}_{\alpha\beta}\left[\boldsymbol{k}_{\alpha\gamma}\right](\boldsymbol{\sigma}_\alpha) = \boldsymbol{\sigma}_\alpha - \frac{\left(\boldsymbol{k}_{\alpha\beta}\left\|\boldsymbol{k}_{\alpha\gamma}\right\|_2^2 - \boldsymbol{k}_{\alpha\gamma}\left(\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{k}_{\alpha\gamma}\right)\right)\left(\boldsymbol{\sigma}_\alpha\cdot\boldsymbol{k}_{\alpha\beta} - w_{\alpha\beta}^{\min}\right)}{\left\|\boldsymbol{k}_{\alpha\beta}\right\|_2^2\left\|\boldsymbol{k}_{\alpha\gamma}\right\|_2^2 - \left(\boldsymbol{k}_{\alpha\beta}\cdot\boldsymbol{k}_{\alpha\gamma}\right)^2}\,. \tag{12.4.7}$$

Ces formules permettent d'implémenter une variante de l'algorithme 12.4.2.

ALGORITHME 12.4.6 (Limitation directionnelle approchée III). *Pour chaque cellule $\mathcal{T}_\alpha$, l'algorithme consiste à parcourir les conditions de monotonie (12.3.3)*

$$w_{\alpha\beta}^{\min} \leq \boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta} \leq w_{\alpha\beta}^{\max}\,,\ \beta \in \mathbb{V}_\alpha$$

dans un ordre arbitraire. Soit $\beta_1 \in \mathbb{V}_\alpha$ le premier indice tel que $\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta_1} < w_{\alpha\beta_1}^{\min}$ ou $\boldsymbol{\sigma}_\alpha \cdot \boldsymbol{k}_{\alpha\beta_1} > w_{\alpha\beta_1}^{\max}$. Dans ce cas, l'algorithme remplace le gradient brut $\boldsymbol{\sigma}_\alpha$ par sa projection $\boldsymbol{\sigma}_\alpha^{(\beta_1)} = \mathfrak{P}_{\alpha\beta_1}(\boldsymbol{\sigma}_\alpha)$ sur la région $\mathcal{S}_{\alpha\beta_1}$, donnée selon le cas par (12.4.2) ou (12.4.3). L'algorithme stocke le vecteur $\boldsymbol{k}_{\alpha\beta_1}$ dans un tableau de travail pour mémoriser qu'il ne faut plus modifier la composante de $\boldsymbol{\sigma}_\alpha^{(\beta_1)}$ qui est parallèle à $\boldsymbol{k}_{\alpha\beta_1}$.

Ensuite, l'algorithme continue de parcourir les contraintes restantes pour vérifier s'il existe un indice $\beta_2 \in \mathbb{V}_\alpha$ tel que $\boldsymbol{\sigma}_\alpha^{(\beta_1)} \cdot \boldsymbol{k}_{\alpha\beta_2} < w_{\alpha\beta_2}^{\min}$ ou $\boldsymbol{\sigma}_\alpha^{(\beta_1)} \cdot \boldsymbol{k}_{\alpha\beta_2} > w_{\alpha\beta_2}^{\max}$. Si l'algorithme rencontre un tel indice, il regarde d'abord dans le tableau de travail s'il contient un vecteur qui indique une direction $\boldsymbol{k}_{\alpha\beta_1}$ dans laquelle il ne faut plus modifier le gradient.

Il trouve le vecteur $\boldsymbol{k}_{\alpha\beta_1}$ et remplace donc $\boldsymbol{\sigma}_\alpha^{(\beta_1)}$ par sa projection $\boldsymbol{\sigma}_\alpha^{(\beta_2)} = \mathfrak{P}_{\alpha\beta_2}\left[\boldsymbol{k}_{\alpha\beta_1}\right]\left(\boldsymbol{\sigma}_\alpha^{(\beta_1)}\right)$ sur $\mathcal{S}_{\alpha\beta_2}$ perpendiculaire à $\boldsymbol{k}_{\alpha\beta_1}$, donnée selon le cas par (12.4.6) ou (12.4.7). Cela garantit que $\boldsymbol{\sigma}_\alpha^{(\beta_2)} \in \mathcal{S}_{\alpha\beta_1} \cap \mathcal{S}_{\alpha\beta_2}$. Il stocke ensuite le vecteur $\boldsymbol{k}_{\alpha\beta_2}$ à la place du vecteur $\boldsymbol{k}_{\alpha\beta_1}$ dans le tableau de travail. L'algorithme oublie de ce fait qu'il faut préserver la direction $\boldsymbol{k}_{\alpha\beta_1}$ et mémorise qu'il faut préserver la direction $\boldsymbol{k}_{\alpha\beta_2}$. L'opération est répétée jusqu'à ce que toutes les faces aient été parcourues.

On peut envisager une variante de cet algorithme qui stocke jusqu'à deux directions à préserver simultanément. Pour des raisons de simplicité, cette variante n'a pas été explicitée.

Puisque cette méthode ne garantit pas que le gradient résultant soit monotone, il faut limiter le gradient obtenu par une limitation non directionnelle qui peut être réalisée par une seule boucle sur les faces. L'algorithme nécessite donc deux boucles sur les faces du maillage, l'une pour l'étape directionnelle, l'autre pour l'étape non directionnelle.□

L'algorithme 12.4.6 nécessite plus de mémoire que les algorithmes 12.4.2 et 12.4.4 car il faut allouer un tableau de travail.

12.5. Étude numérique

Comme expliqué dans la section 12.1, un test très simple consiste à isoler une cellule $\mathcal{T}_\alpha$ et ses cellules voisines $\mathcal{T}_\beta$, $\beta \in \mathbb{V}_\alpha$, et de générer sur ce voisinage des valeurs aléatoires u_β pour $\beta \in \widehat{\mathbb{V}}_\alpha$. Ces valeurs permettent de reconstruire un gradient auquel on applique ensuite l'algorithme de limitation. La différence entre le gradient reconstruit et le gradient limité donne des renseignements qualitatifs sur les algorithmes de limitation.

Le test a été effectué pour les algorithmes 12.4.2 et 12.4.4 à l'aide du logiciel MAPLE. Ce logiciel dispose de fonctions pour déterminer la solution exacte du problème de minimisation (12.3.7) pour la norme $\|.\|_2$. Celle-ci peut donc être comparée avec les résultats des algorithmes approchés 12.4.2 et 12.4.4 afin d'évaluer l'efficacité de ces algorithmes.

	Critère (12.2.4)	Critère (12.2.6)
Solution exacte du problème (12.3.7)	0.21	0.28
Algorithme 12.4.2	0.23	0.39
Algorithme 12.4.4	0.22	0.37
Limitation non directionelle	0.27	0.57

TAB. 12.5.1: Moyenne de l'erreur relative (12.5.1) pour 10000 réalisations.

Notons $\boldsymbol{\sigma}_\alpha$ le gradient « brut », c'est-à-dire le gradient non limité, et $\widetilde{\boldsymbol{\sigma}}_\alpha$ le gradient monotone, c'est-à-dire le gradient limité. On s'intéresse à l'erreur relative

$$\varepsilon \triangleq \frac{\|\widetilde{\boldsymbol{\sigma}}_\alpha - \boldsymbol{\sigma}_\alpha\|_2}{\|\boldsymbol{\sigma}_\alpha\|_2}\,. \tag{12.5.1}$$

Pour le test, on isole une cellule tétraédrique et ses cellules voisines. Un générateur de nombres aléatoires produit 10000 réalisations pour le vecteur $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$ des moyennes de cellule. Ces moyennes de cellule permettent de reconstruire le gradient $\boldsymbol{\sigma}_\alpha$ qui est ensuite limité pour donner $\widetilde{\boldsymbol{\sigma}}_\alpha$. On calcule ensuite la moyenne de l'erreur relative (12.5.1) sur les 10000 réalisations, ce qui donne une première évaluation qualitative de la précision des algorithmes.

Le tableau 12.5.1 montre les résultats pour le critère de monotonie (12.2.4) et le critère de monotonie simplifié (12.2.6). On constate que les valeurs moyennes des erreurs relatives (12.5.1) sont très proches pour les algorithmes 12.4.2 et 12.4.4. Puisque l'algorithme 12.4.2 est plus simple et moins coûteux que l'algorithme 12.4.4, le résultat du test suggère d'implémenter l'algorithme 12.4.2. On remarque également que le choix du critère de monotonie a une grande influence sur l'erreur. Le critère de monotonie simplifié (12.2.6) engendre une erreur nettement plus élevée que le critère (12.2.4). Le résultat du test indique clairement que le critère (12.2.4) est préférable au critère (12.2.6). L'une des conclusions du test est donc que le critère de monotonie peut avoir une grande influence sur la précision de l'algorithme de limitation, ce qui suggère de poursuivre l'étude par la recherche de critères de monotonie moins restrictifs.

12.6. Résumé des méthodes de limitation existantes dans CEDRE

L'objectif de cette section est de présenter rapidement les méthodes de limitation du gradient qui étaient disponibles dans le code CEDRE avant le présent travail. Les simulations instationnaires décrites aux chapitres 13 et 14 permettent de comparer ces méthodes, qui sont plus simples, avec les nouveaux algorithmes développés dans la section 12.4. On obtient ainsi une idée du gain apporté par les nouvelles méthodes directionnelles.

La discrétisation spatiale implémentée dans CEDRE inclut un algorithme de limitation qui repose sur un critère de monotonie et une limitation non directionnelle. Le critère de monotonie est le suivant : dans une cellule $\mathcal{T}_\alpha$, on calcule les valeurs

$$\left.\begin{aligned} v_\alpha^{\max} &\triangleq \max_{\beta\in\mathbb{V}_\alpha}\{u_\beta\} \\ v_\alpha^{\min} &\triangleq \min_{\beta\in\mathbb{V}_\alpha}\{u_\beta\} \end{aligned}\right\} \tag{12.6.1}$$

et

$$\left.\begin{aligned} w_\alpha^{\max} &\triangleq \max_{\beta\in\mathbb{V}_\alpha}\{u_{\alpha\beta}\} \\ w_\alpha^{\min} &\triangleq \min_{\beta\in\mathbb{V}_\alpha}\{u_{\alpha\beta}\} \end{aligned}\right\} \tag{12.6.2}$$

où $u_{\alpha\beta}$ est la valeur reconstruite sur la face $\mathcal{A}_{\alpha\beta}$. Dans (12.6.1) et (12.6.2), $\mathbb{V}_\alpha$ désigne le premier voisinage de la cellule $\mathcal{T}_\alpha$ défini par (5.4.1), c'est-à-dire $\alpha \notin \mathbb{V}_\alpha$.

Le gradient monotone $\widetilde{\boldsymbol{\sigma}}_\alpha$ est alors défini par

$$\widetilde{\boldsymbol{\sigma}}_\alpha = \xi_\alpha \boldsymbol{\sigma}_\alpha$$

où $\boldsymbol{\sigma}_\alpha$ est le gradient brut et $\xi_\alpha \in [0,1]$ est défini par

$$\xi_\alpha = \min\left\{\max\left\{\frac{v_\alpha^{\max}-u_\alpha}{w_\alpha^{\max}-u_\alpha},0\right\},\max\left\{\frac{v_\alpha^{\min}-u_\alpha}{w_\alpha^{\min}-u_\alpha},0\right\},1\right\}.$$

À part ce limiteur de gradient, CEDRE offre un mécanisme de limitation appelé *limiteur de faces*. Cet algorithme de limitation agit uniquement sur les valeurs reconstruites aux faces et non pas sur le gradient lui-même. Pour cette raison, il ne permet pas d'établir de bornes du type (12.2.5) pour la solution. Ici, on explicite les formules uniquement pour la variante des limiteurs de face effectivement utilisée dans le cadre de cette thèse.

Le limiteur de face repose sur la définition d'une moyenne barycentrique sur chaque face $\mathcal{A}_{\alpha\beta}$

$$u_{\alpha\beta}^{(\mathrm{b})} \triangleq \frac{\|j_{\alpha\beta}\|}{\|h_{\alpha\beta}\|}u_\alpha + \left(1-\frac{\|j_{\alpha\beta}\|}{\|h_{\alpha\beta}\|}\right)u_\beta\,.$$

Le principe de limitation consiste alors à remplacer les valeurs reconstruites $u_{\alpha\beta}$ et $u_{\beta\alpha}$ par les valeurs ajustées respectives

$$\left.\begin{aligned}\widetilde{u}_{\alpha\beta} &= u_{\alpha\beta}-\phi\left(u_{\alpha\beta}^{(\mathrm{b})}-u_{\alpha\beta},u_{\alpha\beta}-u_\alpha\right)\left(u_{\alpha\beta}^{(\mathrm{b})}-u_{\alpha\beta}\right)\\ \widetilde{u}_{\beta\alpha} &= u_{\beta\alpha}-\phi\left(u_{\alpha\beta}^{(\mathrm{b})}-u_{\beta\alpha},u_{\beta\alpha}-u_\beta\right)\left(u_{\alpha\beta}^{(\mathrm{b})}-u_{\beta\alpha}\right)\end{aligned}\right\} \tag{12.6.3}$$

où la fonction ϕ est définie par la formule

$$\phi(v,w) = \frac{|u+v|-|u-v|}{|u+v|+|u-v|}\,.$$

12.7. Bilan du chapitre

Puisque les schémas d'ordre élevé génèrent des oscillations artificielles, il faut implémenter des algorithmes de monotonie pour supprimer ces oscillations parasites. Un moyen d'y parvenir est la limitation de la reconstruction qui consiste à contrôler les valeurs reconstruites aux interfaces et à les ajuster si nécessaire. Dans le cadre de cette thèse, on se restreint au cas de la reconstruction linéaire par cellule, c'est-à-dire à la reconstruction des polynômes de degré un. La reconstruction des valeurs aux faces est alors déterminée par la reconstruction d'un gradient dans chaque cellule. Dans ce cas particulier, il faut donc limiter le gradient reconstruit.

Un algorithme de limitation comporte de ce fait deux aspects distincts.

(1) Le critère de monotonie spécifie sur chaque face les valeurs minimales et maximales admissibles.

(2) Il faut ensuite modifier la fonction reconstruite dans la cellule afin que les valeurs reconstruites aux interfaces respectent les bornes du critère de monotonie.

L'étude commence par la présentation d'un critère de monotonie, le principe du maximum, exprimé par le théorème 12.2.1, qui garantit que la solution reste confinée entre les minima et maxima locaux au voisinage de la cellule. Ce principe, qui a été repris de la littérature et dont la preuve détaillée se trouve dans [**10**], empêche la formation de nouveaux extrema locaux.

L'étude théorique se concentre alors sur la façon dont il faut modifier la reconstruction aux faces des cellules afin que celle-ci satisfasse le principe du maximum. Les méthodes les plus répandues sont les méthodes de limitation non directionnelle, qui réduisent le gradient reconstruit sans changer sa direction. Ces méthodes ont l'avantage d'être très rapides mais peuvent dégrader la précision du schéma de façon inappropriée.

C'est pourquoi cette étude cherche des algorithmes de limitation directionnelle, qui modifient également la direction du gradient. Pour cela, le principe du maximum a été interprété comme une contrainte géométrique : le gradient $\tilde{\boldsymbol{\sigma}} \in \mathbb{R}^d$ qui satisfait le principe du maximum est inclus dans un ensemble convexe et compact du $\mathbb{R}^d$. Dans la norme $\|.\|_2$, le problème de trouver un gradient monotone est alors un problème de minimisation convexe.

La résolution exacte de ce problème de minimisation est certes possible, mais cette approche augmenterait le temps de calcul de façon excessive car il faut résoudre le problème dans chaque

cellule à chaque pas de temps. C'est pourquoi l'étude se concentre sur la recherche d'algorithmes simplifiés qui sont rapides mais fournissent néanmoins un meilleur résultat que des méthodes non directionnelles. La section 12.4 présente les algorithmes 12.4.2, 12.4.4 et 12.4.6 dont seuls les deux premiers ont été implémentés dans Cedre.

Une étude numérique simplifiée, présentée dans la section 12.5, a permis une première évaluation des algorithmes 12.4.2 et 12.4.4. L'algorithme 12.4.2, plus simple que l'algorithme 12.4.4, fournit des résultats presque aussi bons. Pour cette raison, l'algorithme 12.4.2 a été utilisé pour les calculs présentés dans les chapitres 13 et 14.

L'étude numérique montre également que le critère de monotonie peut avoir une grande influence sur la précision de l'algorithme de limitation. Il faudrait donc compléter la présente étude par une recherche de critères de monotonie moins restrictifs.

CHAPITRE 13

Simulation des grandes échelles d'un écoulement subsonique au-dessus d'une cavité profonde

13.1. Objectif du chapitre

L'objectif est d'appliquer certaines évolutions de la discrétisation spatiale, développées dans ce document, à des simulations d'écoulements. L'étude de ce chapitre est dédiée au cas subsonique alors que le chapitre 14 est consacré au cas supersonique. Ces simulations servent à analyser les différents aspects de la discrétisation spatiale sur des maillages non structurés en dimension trois. Comme les reconstructions d'ordre élevé décrites au chapitre 7 ne sont pas encore disponibles dans Cedre, l'étude se restreint aux schémas qui sont formellement d'ordre deux. Ces schémas reposent sur une reconstruction linéaire par cellule et les options testées sont donc les options de reconstruction et de limitation du gradient.

Un cas test intéressant est l'écoulement subsonique instationnaire au-dessus d'une cavité profonde. Ce type de configuration a fait l'objet d'une étude expérimentale menée à l'Onera [**50**] qui a fourni une base de données expérimentale permettant en particulier de valider des méthodes numériques instationnaires. L'écoulement présente des caractéristiques physiques intéressantes qu'on souhaite mieux comprendre par des simulations numériques. Il s'agit notamment de l'interaction entre la couche de mélange en amont et au-dessus de la cavité et des ondes acoustiques générées par l'impact de la couche de mélange sur le coin aval de la cavité. Ces mécanismes engendrent des oscillations de pression auto-entretenues qui sont typiques pour les cavités profondes. L'intérêt pour ce type de configuration vient en partie du fait que ces oscillations sont à l'origine d'importantes nuisances dans plusieurs domaines d'application.

Cette configuration expérimentale a fait l'objet de deux études numériques instationnaires s'appuyant sur la simulation des grandes échelles.

- L. Larchevêque et al. [**76, 75**] ont effectué plusieurs simulations des grandes échelles sur différents maillages structurés. Ces études ont non seulement permis de valider l'approche de la simulation des grandes échelles pour ce type de configuration, mais également de mieux analyser les phénomènes physiques de l'écoulement à l'aide des résultats numériques.
- N. Bertier [**17**] a effectué des simulations des grandes échelles avec le logiciel Cedre sur des maillages structurés et non structurés. L'objectif était d'évaluer la pertinence de l'approche de la simulation des grandes échelles en maillage non structuré et de comparer ces résultats avec ceux obtenus sur des maillages structurés.

Les calculs réalisés par N. Bertier *sur des maillages de tétraèdres* ont mis en évidence un certain nombre de problèmes [**17**, p. 144] :

(1) Le schéma volumes finis utilisé est instable sans limiteur de gradient. L'instabilité se manifeste en général par une température négative dans une ou plusieurs cellules, ce qui provoque l'arrêt du calcul.

(2) Si le limiteur de gradient est activé, le schéma devient stable mais une dissipation numérique excessive dégrade la précision par rapport aux résultats en maillage structuré.

La discrétisation spatiale utilisée par N. Bertier repose sur une méthode des volumes finis avec une reconstruction du gradient sur le premier voisinage de la cellule. L'étude de stabilité du chapitre 10 a démontré que ce schéma fournit une discrétisation instable de la convection linéaire en maillage de tétraèdres. Le limiteur de gradient arrive à contenir cette instabilité, mais au prix d'une précision détériorée. Ce calcul permet donc de vérifier si une reconstruction stable du gradient, et en particulier la méthode définie par l'algorithme 8.4.1, permet de réaliser la

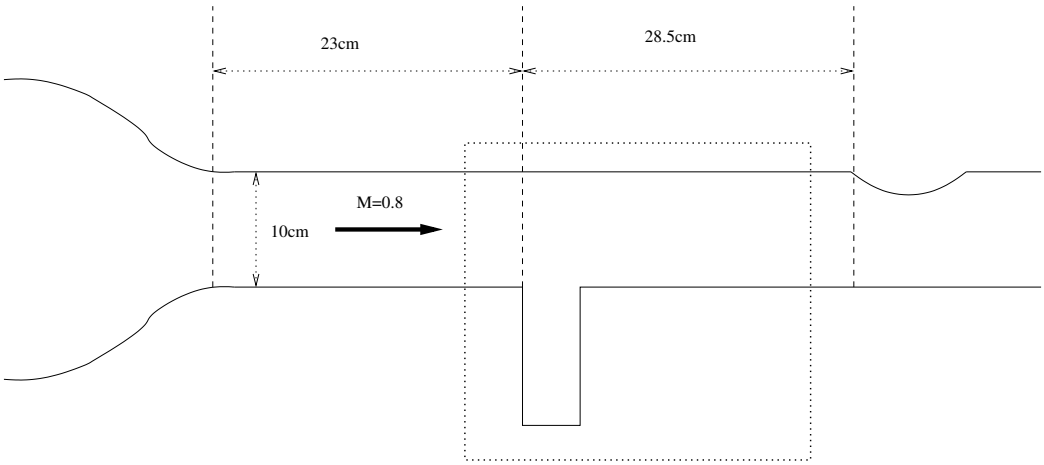


FIG. 13.2.1: Schéma du dispositif expérimental. La partie du canal qui constitue le domaine du calcul est indiquée par un rectangle pointillé.

simulation de la cavité sans limiteurs. Cela constituerait la preuve que les problèmes de robustesse constatés par N. Bertier proviennent d'une instabilité linéaire.

13.2. Résumé de l'étude expérimentale

Cette section fournit une présentation succincte de l'étude expérimentale menée par Forestier et al. [**50**]. Le résumé donné ci-dessous reprend en grande partie la description dans [**75**] et [**17**].

13.2.1. Description du dispositif expérimental. Le dispositif expérimental est constitué d'une maquette en forme de canal dont le plancher contient la cavité. Ce canal est placé dans une veine d'essai comme montré dans la figure 13.2.1. La maquette du canal qui constitue le domaine de calcul pour la simulation numérique est indiquée par un rectangle pointillé. La soufflerie injecte de l'air desséché à une température d'arrêt de $T_i \approx 293\,\mathrm{K}$ et une pression d'arrêt de $P_i \approx 10^5\,\mathrm{Pa}$. Le nombre de Mach est fixé à $M = 0,8$, ce qui correspond, dans ces conditions, à une vitesse de l'écoulement de $U_\infty \approx 260\,\frac{\mathrm{m}}{\mathrm{s}}$ au centre du canal. Le nombre de Reynolds basé sur la longueur de la cavité $L = 5\,\mathrm{cm}$ est dans ce cas $\mathrm{Re}_L \approx 8,6\,10^5$.

Une bande rugueuse placée à 20 cm en amont de la cavité sert à rendre la couche limite incidente suffisamment turbulente.

La figure 13.2.2 présente la géométrie du canal et de la cavité. Toutes les longueurs sont spécifiées en multiples de la longueur caractéristique du calcul $L = 5\,\mathrm{cm}$. La cavité est rectangulaire de longueur L dans la direction de l'écoulement ainsi que de profondeur $D = 2,4L$. Le canal est également rectangulaire de longueur totale $6L$ et de hauteur $2L$. La maquette du canal ne possède pas de parois latérales et la largeur du canal et de la cavité est égale à celle de la veine d'essai, c'est-à-dire $D = 2,4L$.

Le choix des axes est le suivant : la direction de l'écoulement est parallèle à l'axe x, l'axe y est situé perpendiculairement à l'écoulement dans le plancher du canal de telle façon que les plans $y = \pm 1,2L$ coïncident avec les parois latérales du canal et de la cavité. Le plan $z = 0$ coïncide avec le plancher du canal, le plan $z = 2L$ avec le plafond du canal et le plan $z = -2,4L$ avec le plancher de la cavité. La paroi en amont de la cavité est placée dans le plan $x = 0$ et la paroi en aval dans le plan $x = L$.

Cinq capteurs de pression se trouvent sur la paroi en amont de la cavité à la hauteur $z = -0,7L$ et aux positions latérales $y = 0$, $y = \pm 0,2L$ et $y = \pm 0,4L$.

13.2.2. Résumé des résultats expérimentaux. L'importance de cette étude expérimentale réside dans le fait que les données fournies permettent d'évaluer des méthodes numériques instationnaires comme la simulation des grandes échelles. Les données expérimentales collectées

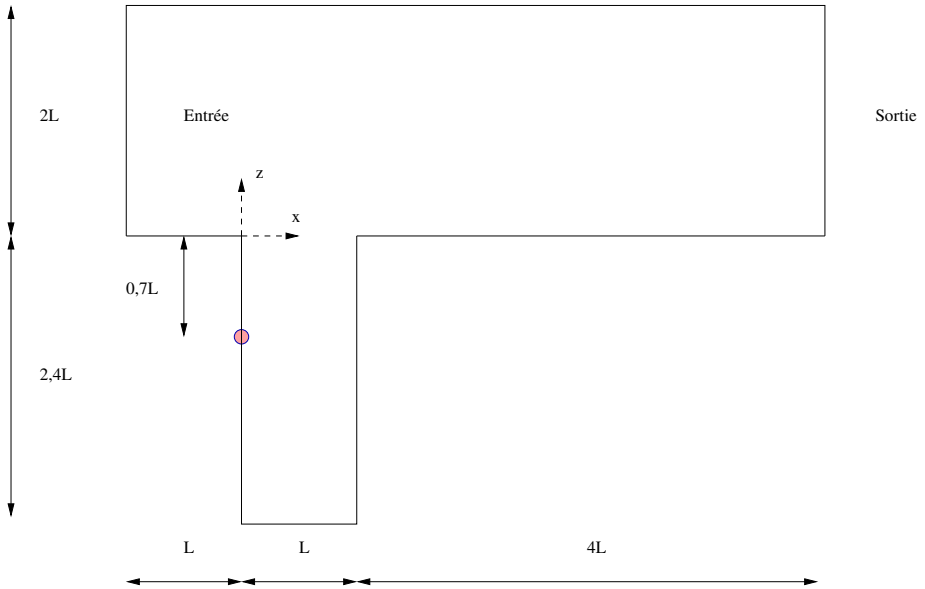


FIG. 13.2.2: Géométrie du canal. Vue du plan $y = 0$. La position du capteur est indiquée par un cercle.

incluent des mesures acoustiques, effectuées à l'aide de capteurs dans la cavité, ainsi que des mesures de vitesse fournies par un dispositif de vélocimétrie.

L'analyse spectrale des signaux de pression dans la cavité met en évidence le caractère auto-oscillant de l'écoulement. La densité spectrale de puissance est dominée par des harmoniques d'une fréquence fondamentale $f_1 \approx 1975$ Hz qui a une amplitude de 155 dB.

La strioscopie de la couche de mélange révèle que celle-ci s'enroule en trois structures cohérentes tourbillonaires bien identifiées, appelées S_1, S_2 et S_3. Ces structures se détachent de façon répétitive du bord aval de la cavité.

- La première, S_1, rentre dans la cavité lorsqu'elle s'approche du bord aval de celle-ci.
- La deuxième, S_2, impacte le bord aval de la cavité en déclenchant une onde de pression qui remonte vers l'aval.
- La troisième, S_3, passe au-dessus de la cavité.

La cinématique de ces structures révèle le caractère bidimensionnel à grande échelle de l'écoulement. La strioscopie permet également de détecter la propagation d'un système d'ondes qui se réfléchissent sur les parois du canal et forment des motifs en forme de losange.

Le dispositif de vélocimétrie fournit les profils de la vitesse et des tensions de Reynolds dans la couche de mélange au-dessus de la cavité. L'étude permet non seulement d'obtenir des moyennes de Reynolds mais aussi des moyennes de phase, conditionnées par le signal acoustique d'un capteur dans la cavité qui sert de signal de référence. Cela permet de séparer les composantes périodique et aléatoire du champ de vitesse.

L'analyse de l'épaisseur de quantité de mouvement incompressible de la couche de mélange met en évidence trois zones. Dans la zone où les structures se forment et se détachent, l'épaisseur de quantité de mouvement croît plus rapidement que dans les couches de mélange libres turbulentes. Plus en aval, la croissance revient vers des valeurs typiques. Dans la zone proche du bord aval de la cavité, l'épaisseur de quantité de mouvement décroît.

13.3. Description des calculs

13.3.1. Présentation de la géométrie et du maillage. La géométrie du domaine de calcul correspond à celle de la figure 13.2.2. La seule exception est la largeur du domaine de calcul qui est restreinte à $L = 5\,\text{cm}$. L'étude de Larchevêque et al. [**76, 75**] a montré que cette simplification est justifiée et pertinente en raison du caractère bidimensionnel à grande échelle

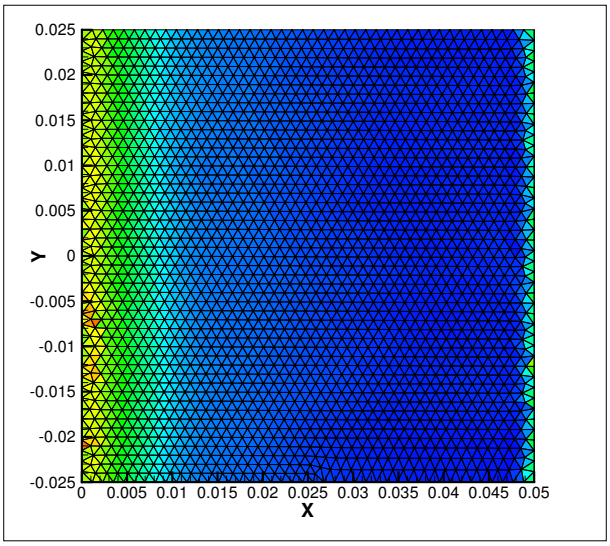


FIG. 13.3.1: Aperçu d'une coupe du maillage au-dessus de la cavité, dans le plan $z = 0$. Les longueurs sont spécifiées en mètres.

de l'écoulement. Le domaine de calcul est donc délimité par deux plans latéraux données par $y = \pm 0,5\, L$.

Les calculs ont été réalisés sur un maillage non structuré composé de tétraèdres dans le canal et dans la cavité. Le maillage comprend à peu près 1,5 millions de cellules. Le plafond et le plancher du canal sont recouverts d'une couche de prismes avec des bases triangulaires afin de mieux capter la couche limite.

La couche de prismes sur le plancher du canal a les caractéristiques suivantes : elle est constituée de 25 sous-couches de prismes, d'une épaisseur initiale de $0,008\, L$ avec un facteur d'étirement de $1,05$. La couche est donc d'une épaisseur totale de $0,356\, L$. Le diamètre de la base des prismes est de $0,04\, L$ entre $x = -L$ et $x = 2\, L$ et croît de $0,04\, L$ à $0,08\, L$ entre $x = 2\, L$ et $x = 5\, L$.

L'épaisseur de la couche de prismes est adaptée à l'épaisseur de vorticité δ_ω qui varie entre $3,7\,\text{mm} = 0,074\, L$ et $1,75\,\text{cm} = 0.35\, L$, cf. [**17**, p. 143] et [**75**, p. 119]. Il y a donc à peu près 9 points dans l'épaisseur de vorticité.

La figure 13.3.1 montre une coupe horizontale du maillage au-dessus de la cavité, c'est-à-dire la région $0 \leq x \leq L$, $-0,5\, L \leq y \leq 0,5\, L$ et $z = 0$. Le maillage produit par CENTAUR est très régulier.

13.3.2. Conditions aux limites pour le calcul. Les conditions limites à l'entrée imposent une température d'arrêt de $T_i \approx 293$ Kelvin et un profil particulier pour le débit massique. Ce profil s'obtient par interpolation de mesures physiques, cf. [**75**], et favorise le développement d'une couche de mélange turbulente. Le profil pour la vitesse et la densité de masse est

$$
\begin{aligned}
v\,(x,y,z)|_{x=-L} &= \begin{cases} v_\infty \tanh\left(6,7708\,\frac{z}{L}\right)^{0,128828} & ; \quad 0 \leq z \leq L \\ v_\infty \tanh\left(6,7708\,\left(2-\frac{z}{L}\right)\right)^{0,128828} & ; \quad L \leq z \leq 2L \end{cases} \\
\rho\,(x,y,z)|_{x=-L} &= \frac{\rho_\infty}{\left(1+0,128\left(1-\frac{u(x,y,z)}{u_\infty}\right)^2\right)} \\
v_\infty &= 258,5\frac{\text{m}}{\text{sec}} \\
\rho_\infty &= 0,88\frac{\text{kg}}{\text{m}^3}\,.
\end{aligned}
$$

Le débit massique à l'entrée est alors

$$\begin{aligned}(\rho v_x)(x,y,z)|_{x=-L} &= \rho(x,y,z)\, v(x,y,z)\\ (\rho v_y)(x,y,z)|_{x=-L} &= 0\\ (\rho v_z)(x,y,z)|_{x=-L} &= 0\end{aligned}$$

où v_x , v_y et v_z sont les composantes de la vitesse en direction x, y, et z.

Les limites latérales du canal et de la cavité sont modélisées par des parois glissantes afin de reprendre les conditions utilisées par N. Bertier. Il est utile de rappeler que le dispositif expérimental ne dispose pas de parois latérales. Larchevêque et al. ont choisi de modéliser ces conditions en imposant des limites de périodicité, ce qui se justifie par le caractère bidimensionnel de l'écoulement.

Le plafond et le plancher du canal ainsi que les parois en amont et aval de la cavité sont modélisés par des conditions limites de type paroi adiabatique.

Finalement, la sortie est modélisée par des conditions de type sortie subsonique avec une pression de $P \approx 65000$ Pascal. La manière dont Cedre impose cette condition limite permet d'éviter la réflexion d'ondes acoustiques à la sortie.

13.3.3. Présentation des méthodes de discrétisation spatiale. L'objectif principal du calcul consiste à comparer cinq options de discrétisation spatiale avec une attention particulière à la stabilité des méthodes de discrétisation. Du fait que les reconstructions d'ordre élevé décrites dans les chapitres 7 et 8 n'étaient pas encore disponibles dans Cedre, il s'agit des options de reconstruction du gradient et des options de limitation.

(0) Un premier calcul sert à vérifier que la *reconstruction du gradient sur le premier voisinage sans limiteur de gradient* est bien instable pour ce type de calcul. Les travaux du chapitre 10 montrent que cette méthode de discrétisation est instable pour la convection linéaire en maillage de tétraèdres. On constate effectivement que l'instabilité de la discrétisation spatiale se manifeste par des températures négatives dans une ou plusieurs cellules, ce qui force l'arrêt du calcul.

(1) Un calcul avec la reconstruction du gradient sur le deuxième voisinage, décrit par l'algorithme 8.4.1 de la section 8.4, mais *sans limiteur de gradient* (abrégé *Sans lim.* dans la légende). Si cette méthode, qui est stable pour la convection linéaire, ne montre pas d'instabilité, il est presque certain que les problèmes de robustesse de l'option (0) proviennent d'une instabilité linéaire du schéma comme celle décrite au chapitre 10.

(2) Un calcul avec la reconstruction du gradient sur le deuxième voisinage et le limiteur de face décrit dans la section 12.6 (abrégé *Lim. faces* dans la légende).

(3) Un calcul avec la reconstruction du gradient sur le deuxième voisinage et le limiteur directionnel basé sur le principe du maximum, défini par l'algorithme 12.4.2 dans la section 12.4 (abrégé *Lim. direct.* dans la légende).

(4) Un calcul avec la reconstruction du gradient sur le deuxième voisinage et le limiteur de gradient non directionnel, décrit dans la section 12.6 (abrégé *Lim. non dir.* dans la légende).

(5) Un calcul avec la reconstruction du gradient sur le deuxième voisinage, le limiteur de gradient non directionnel et le limiteur de face décrits dans la section 12.6 (abrégé *Lim. ndir. faces* dans la légende).

13.3.4. Intégration en temps. L'intégration en temps s'effectue par une méthode de Runge-Kutta explicite d'ordre deux, le schéma de Heun, avec un pas de temps $\Delta t = 10^{-7}\,\mathrm{s}$. Le calcul démarre à partir d'un état initial généré avec un schéma d'ordre un en espace et une méthode d'intégration implicite d'ordre un en temps.

La durée du calcul instationnaire est de $0,05\,\mathrm{s}$ dont les premières $0,01\,\mathrm{s}$ ne sont pas exploitées car elles constituent la phase de démarrage pendant laquelle les phénomènes d'oscillation s'installent. Les signaux de pression dans la cavité permettent de déterminer le moment à partir duquel l'écoulement atteint son régime asymptotique.

Expérience		(1) Sans lim.		(2) Lim. faces		(3) Lim. direct.		(4) Lim. non dir.		(5) L. ndir. faces	
f	SPL	f	SPL	f	SPL	f	SPL	f	SPL	f	SPL
1989.8	154.66	1950.	148.14	1950.	148.29	2000.	148.64	1950.	148.81	1950.	148.79
3979.5	153.36	3900.	150.00	3900.	149.89	4000.	146.77	3900.	152.28	3900.	150.99
5981.5	132.47	5800.	134.59	5800.	134.47	6050.	133.43	5850.	136.92	5850.	136.42
7910.2	132.95	7750.	139.90	7750.	137.32	8050.	134.88	7800.	139.88	7800.	138.16
9960.9	117.50	9700.	124.42	9700.	121.20	10050.	115.96	9750.	121.12	9750.	120.14
11951.	124.86	11650.	128.95	11700.	126.32	12050.	117.65	11700.	127.79	11700.	126.23
13977.	115.47	13600.	119.95	13650.	114.87	14100.	112.85	13650.	118.34	13650.	113.73

TAB. 13.4.1: Les sept premières raies du spectre : fréquences (Hz) et SPL (dB).

À la vitesse de $U_\infty \approx 260\,\frac{\mathrm{m}}{\mathrm{s}}$, l'air met à peu près $0,001\,\mathrm{s}$ pour traverser le canal d'une longueur de $5L = 30\,\mathrm{cm}$. Une durée effective du calcul de $0,04\,\mathrm{s}$ paraît donc suffisante pour observer les structures de l'écoulement et obtenir une résolution fréquentielle suffisante des signaux de pression.

13.3.5. Modèle physique. Les calculs ont été effectués avec une modélisation de la turbulence par le modèle de Smagorinsky qui est disponible dans CEDRE depuis les travaux de N. Bertier [**17**]. La section 3.3 présente les détails de ce modèle. Les paramètres utilisés sont les suivants :

(1) Constante de Prandtl : 0,09.
(2) Constante de Karman : 0,4.
(3) Nombre de Prandtl turbulent : $\mathrm{Pr}^{(\mathrm{sgs})} = 0,9$.
(4) Nombre de Schmidt turbulent : 0,9.
(5) Constante de longueur de mélange : $\mathrm{C_S} = 0,1$.

Le modèle de paroi est un modèle de Couette laminaire.

13.4. Résultats de l'étude numérique

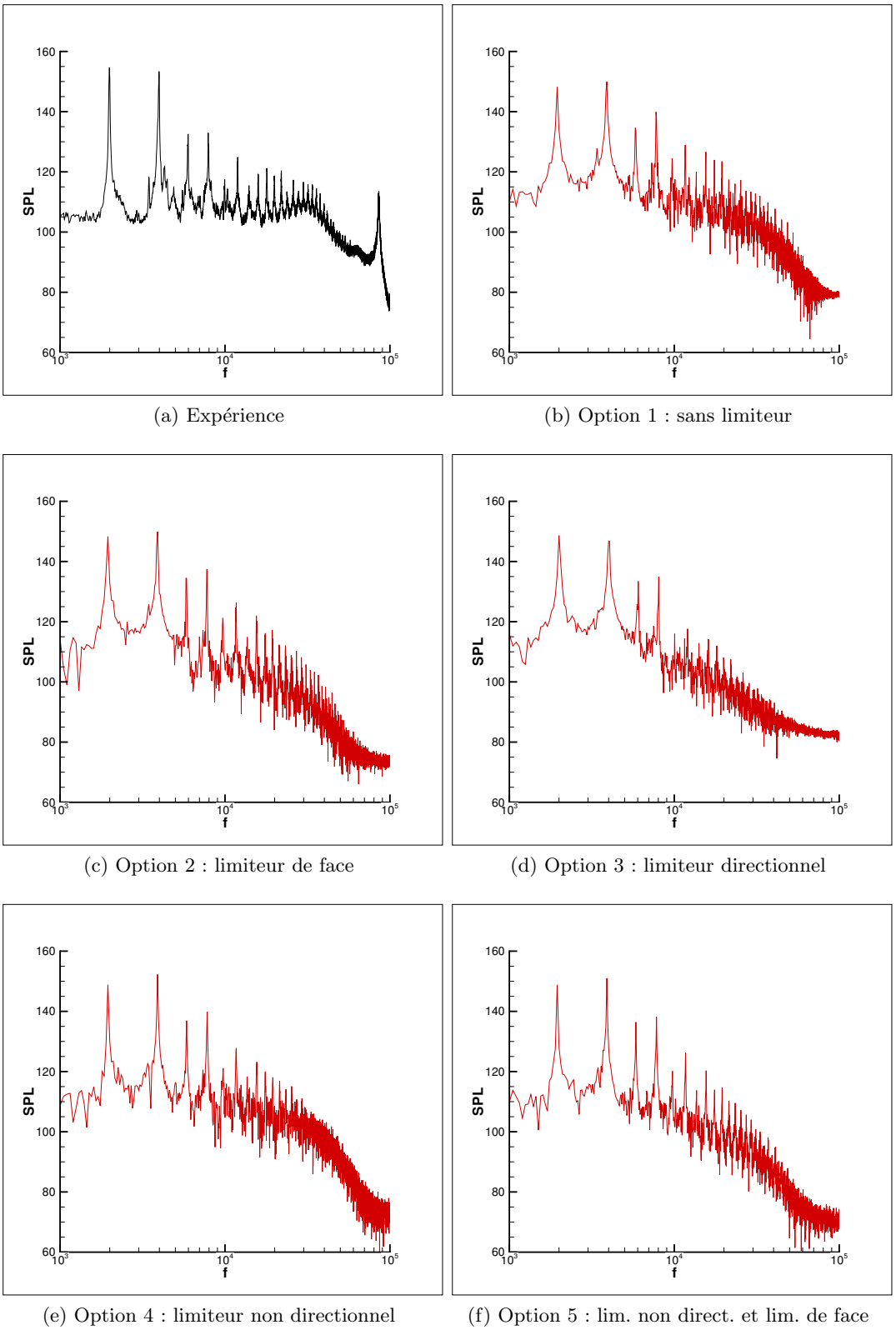
13.4.1. Étude des signaux de pression. Les signaux de pression aux cinq capteurs sont soumis à une transformation discrète de Fourier par la méthode du périodogramme qui fournit une estimation de la densité spectrale de puissance du signal. Le pas de temps du calcul est de $\Delta t = 10^{-7}\,s$ et un signal est enregistré tous les $N_e = 50$ pas de temps. La fréquence d'échantillonage est donc de $f_e = \frac{1}{N_e\,\Delta t}\,\mathrm{Hz} = 2\,10^5\,\mathrm{Hz}$ et la fréquence de Nyquist est de $f_c = \frac{1}{2} f_e = 10^5\,\mathrm{Hz}$. Avec une durée du calcul de $T = 0,04\,\mathrm{s}$, la résolution fréquentielle est a priori de $\Delta f = \frac{1}{T}\,\mathrm{Hz} = 25\,\mathrm{Hz}$. Puisqu'on utilise deux blocs pour moyenner les spectres, la résolution fréquentielle finale est de $\Delta f = 50\,\mathrm{Hz}$.

La figure 13.4.1 montre la densité spectrale de puissance du signal de pression pour l'étude expérimentale et les cinq options de discrétisation spatiale. Les niveaux sont exprimés en *sound pressure level* SPL (dB), définis par

$$\mathrm{SPL}\ (\mathrm{dB}) = 20 \log_{10} \left(\frac{\sqrt{(p - \overline{p})^2}}{p_{\mathrm{ref}}} \right)$$

où $\overline{p}$ désigne la moyenne temporelle du signal de pression et p_{ref} est la pression de référence d'une valeur de $2\,10^{-5}\,\mathrm{Pa}$.

Les figures 13.4.2 et 13.4.3 permettent de comparer les spectres pour les cinq options de discrétisation spatiale avec les données expérimentales. Le tableau 13.4.1 montre les fréquences (Hz) et SPL (dB) des sept premières raies du spectre pour les différentes options de discrétisation et les données expérimentales. La figure 13.4.4 présente graphiquement les données du tableau 13.4.1.



(a) Expérience

(b) Option 1 : sans limiteur

(c) Option 2 : limiteur de face

(d) Option 3 : limiteur directionnel

(e) Option 4 : limiteur non directionnel

(f) Option 5 : lim. non direct. et lim. de face

FIG. 13.4.1: Spectres du signal de pression.

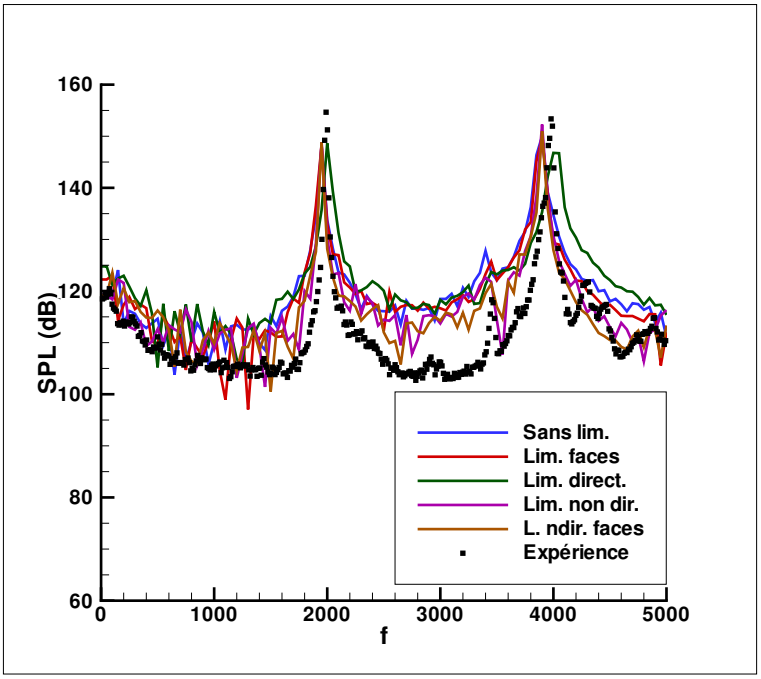


FIG. 13.4.2: Portion du spectre pour les cinq options.

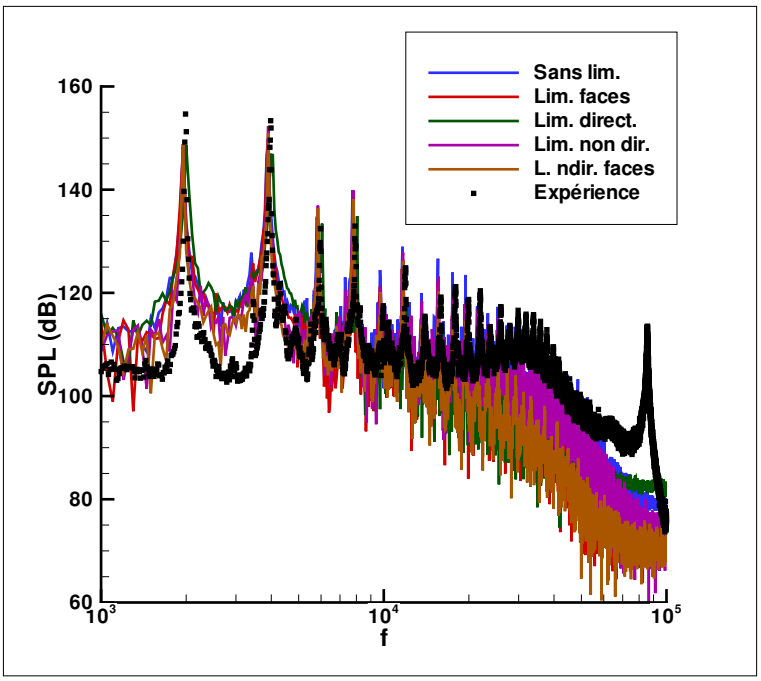


FIG. 13.4.3: Totalité du spectre pour les cinq options.

Le signal a un caractère très périodique qu'on retrouve dans les résultats numériques. On constate que toutes les options de discrétisation reproduisent assez bien les premières raies spectrales qui sont caractéristiques des cavités profondes. Il s'agit des harmoniques d'une fréquence fondamentale d'environ 2000 Hz. Les options 1, 2, 4 et 5 sous-évaluent légèrement les fréquences, alors que l'option 3, qui tend plutôt à surévaluer un peu les fréquences, donne de meilleurs résultats. Pour les options de discrétisation 1 et 2, une petite raie non harmonique est visible autour

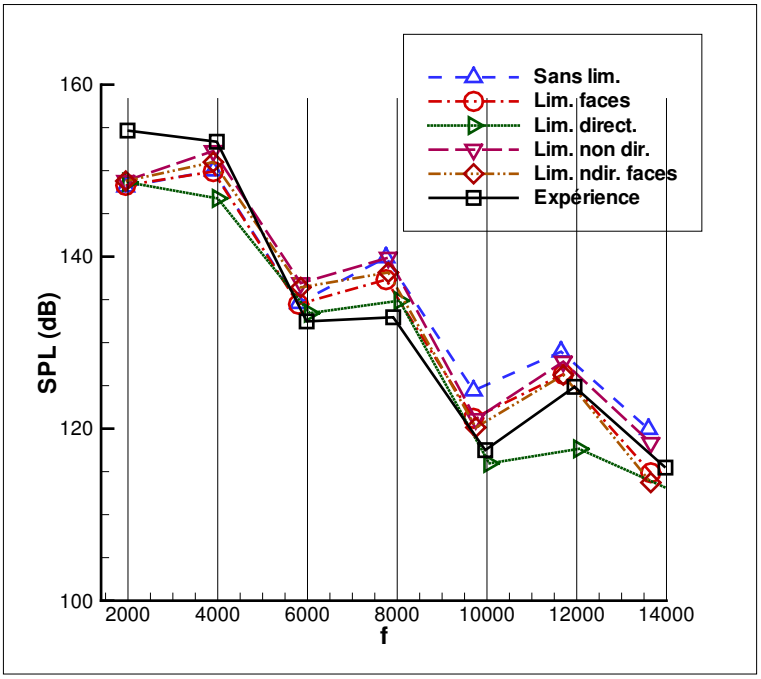


FIG. 13.4.4: Les sept premières raies du spectre : fréquences (Hz) et SPL (dB).

de 3500 Hz, à gauche de la deuxième raie principale. Cette raie se retrouve dans les données expérimentales, alors que les options 3, 4 et 5 ne permettent pas de la retrouver. Cela pourrait s'expliquer par le fait que les options 3, 4 et 5 créent plus de dissipation numérique que les options 1 et 2.

La figure 13.4.4 permet également de comparer les amplitudes des raies spectrales. On constate que l'option 3 prévoit des amplitudes plus faibles que les autres options. Pour expliquer ce fait, on peut noter que l'option 3 est la seule qui impose un critère mathématiquement rigoureux pour supprimer des oscillations. Il se peut donc que cette approche génère un peu trop de dissipation numérique.

En général, on remarque que toutes les options de discrétisation tendent à sous-évaluer les amplitudes des deux premières raies spectrales. Le niveau de la première raie est par ailleurs sensiblement le même pour les cinq options de discrétisation spatiale. Les niveaux des raies suivantes sont par contre sur-estimés, sauf dans le cas de l'option 3.

Finalement, les figures 13.4.1, 13.4.2 et 13.4.3 révèlent que le niveau entre les deux premières raies spectrales est sur-évalué par rapport aux données expérimentales, pour toutes les cinq options de discrétisation.

13.4.2. Étude des profils de vitesse. L'objectif est de comparer les profils de la vitesse moyenne horizontale $\overline{v}_x$ et de la vitesse moyenne verticale $\overline{v}_z$ avec les données expérimentales. Pour cela, $\overline{v}_x$ et $\overline{v}_z$ sont mesurées en fonction de la coordonnée verticale z dans le plan $y = 0$ aux coordonnées $x = 0,2\,L$, $x = 0,4\,L$, $x = 0,6\,L$, et $x = 0,8\,L$, c'est-à-dire juste au-dessus de la cavité. Les mesures sont prises entre les hauteurs $z = -0,2\,L$ et $z = 0,2\,L$.

La figure 13.4.5 montre le profil de la vitesse moyenne horizontale $\overline{v}_x$ au-dessus de la cavité et la figure 13.4.6 montre le profil de la vitesse moyenne verticale $\overline{v_z}$.

On constate que le profil de la vitesse moyenne horizontale $\overline{v}_x$ est correctement reproduit par toutes les options de discrétisation spatiale, avec de très légers écarts. Le profil de la vitesse moyenne $\overline{v}_z$ est encore correctement calculé pour $z \geq 0$, c'est-à-dire dans la partie supérieure de la couche de mélange. Dans la partie inférieure, c'est-à-dire pour $z \leq 0$, on commence à constater des écarts entre le calcul et l'expérience, même si les courbes ont des profils similaires.

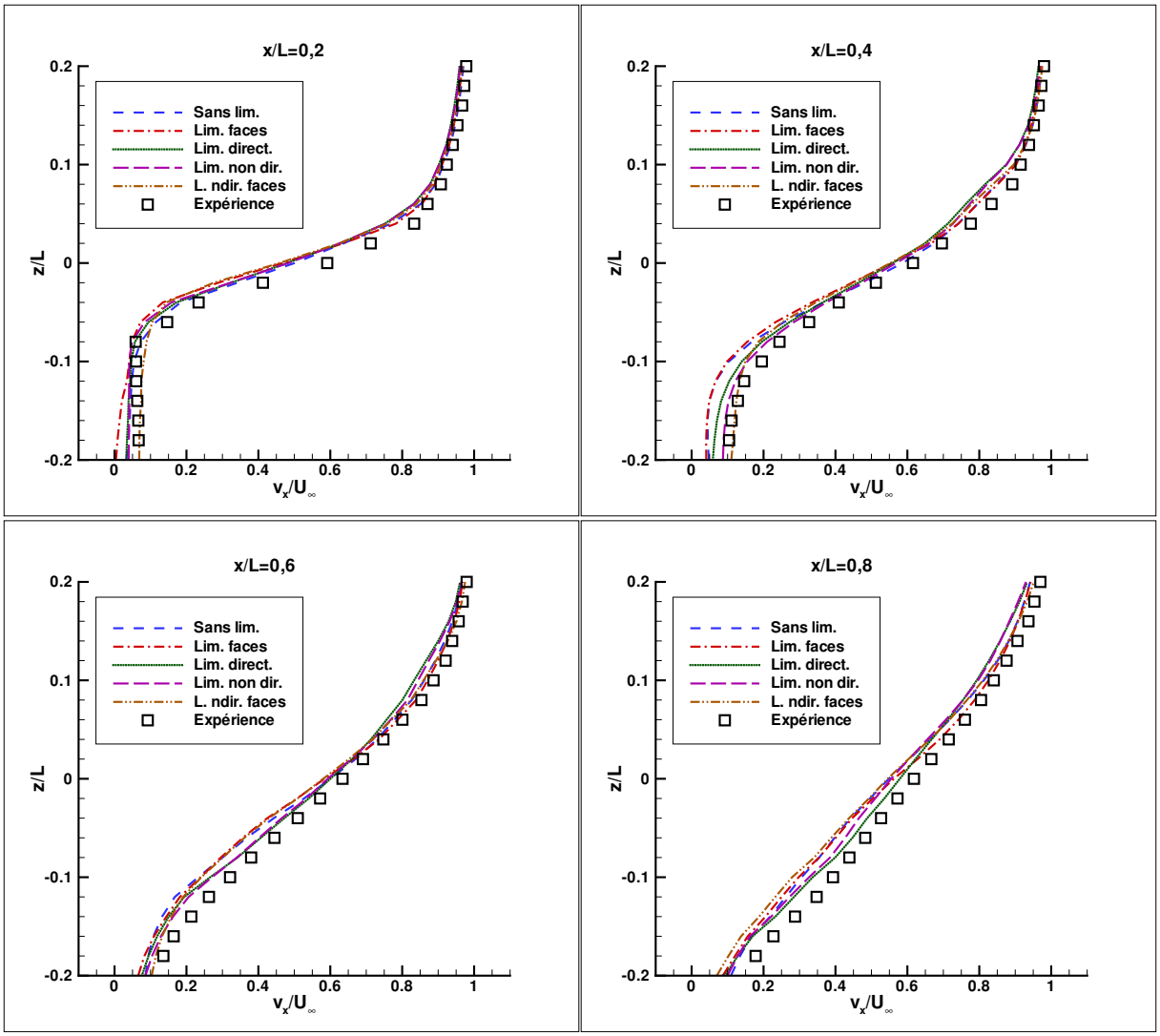


FIG. 13.4.5: Profils de la vitesse moyenne horizontale $\overline{v_x}$.

Les résultats ne permettent cependant pas de distinguer un schéma parmi les cinq qui donnerait les meilleurs résultats. Selon le cas, l'un ou l'autre des schémas capte mieux la solution.

13.4.3. Étude de la couche de mélange. L'étude expérimentale [**50**] a fourni une analyse détaillée de la couche de mélange au-dessus de la cavité : l'article présente les mesures effectuées sur deux maillages, l'un dans le plan $y = 0$ et l'autre dans le plan $z = 0$. En particulier, l'étude a permis de caractériser la croissance de l'épaisseur de quantité de mouvement δ_m au-dessus de la cavité, dans le plan $y = 0$, entre $x = 0$ et $x = L$. Dans le cas présent, l'épaisseur de quantité de mouvement est définie par

$$\delta_m(x) = \int_{z_{\min}}^{z_{\max}} \frac{\overline{v}_x(x,0,z)}{v_\infty}\left(1 - \frac{\overline{v}_x(x,0,z)}{v_\infty}\right) dz \tag{13.4.1}$$

où $z_{\min} = -2\,L$, $z_{\max} = 2\,L$ et $\overline{v}_x(x,0,z)$ est la vitesse moyenne horizontale dans le plan $y = 0$. Les bornes d'intégration $z_{\min} = -2\,L$ et $z_{\max} = 2\,L$ permettent de comparer le résultat avec les mesures expérimentales qui ont été effectuées entre ces bornes.

Les valeurs de δ_m et de x sont normalisées par rapport à l'épaisseur de quantité de mouvement initiale

$$\delta_m^0 = 0,648\,\text{mm} = 0,01296\,L\,,$$

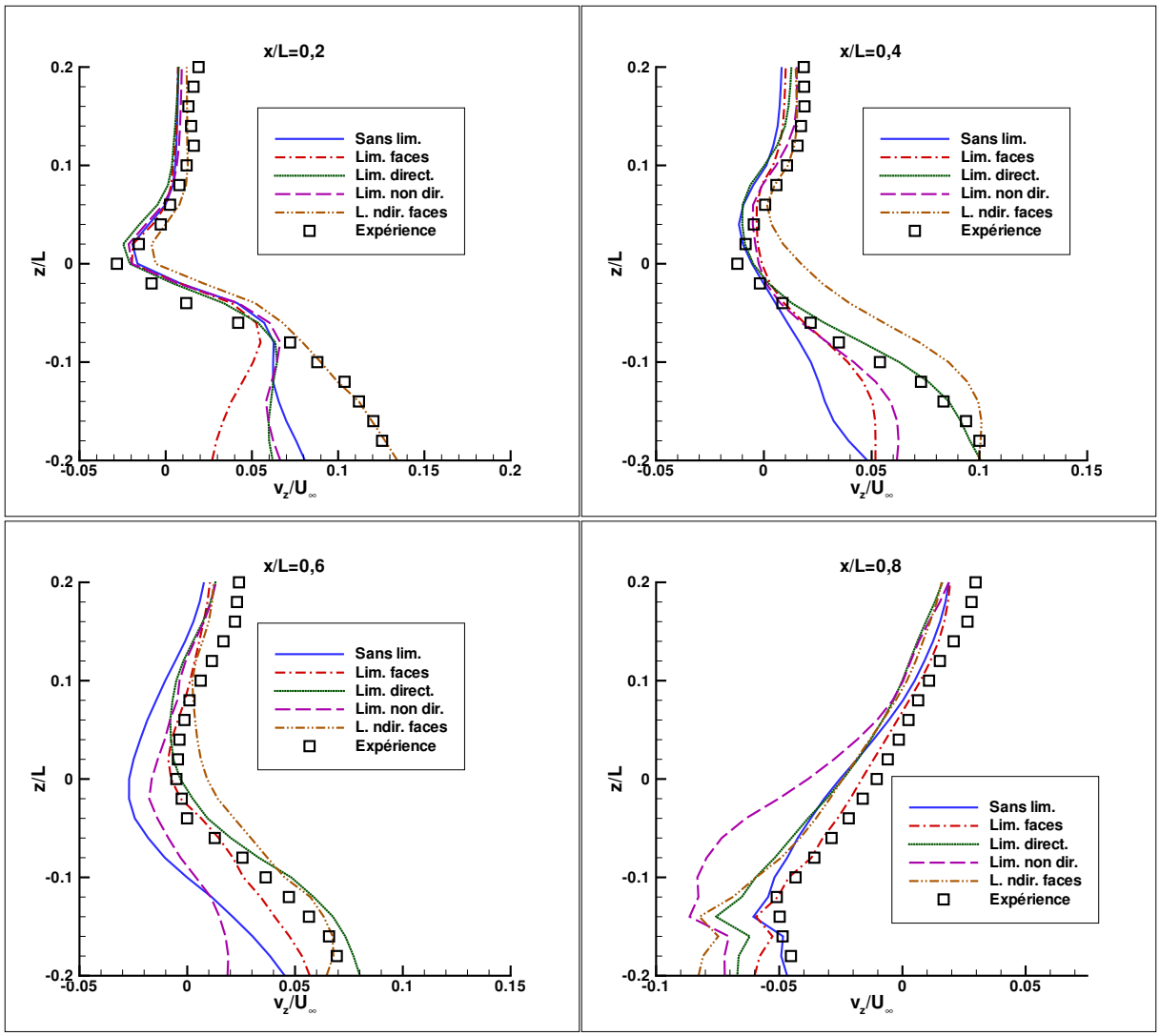


FIG. 13.4.6: Profils de la vitesse moyenne verticale $\overline{v_z}$.

mesurée à 1 mm de distance du coin amont de la cavité, c'est-à-dire à $x = -0.02\,L$, par Forestier et al. [**50**].

L'étude expérimentale met en évidence un comportement particulier de la croissance de δ_m :

(1) Dans la région $0 \leq x \leq 17\delta_m^0$, la croissance est linéaire avec une pente de $0,12$. C'est une pente trois fois plus grande que celle des couches de mélange libres.

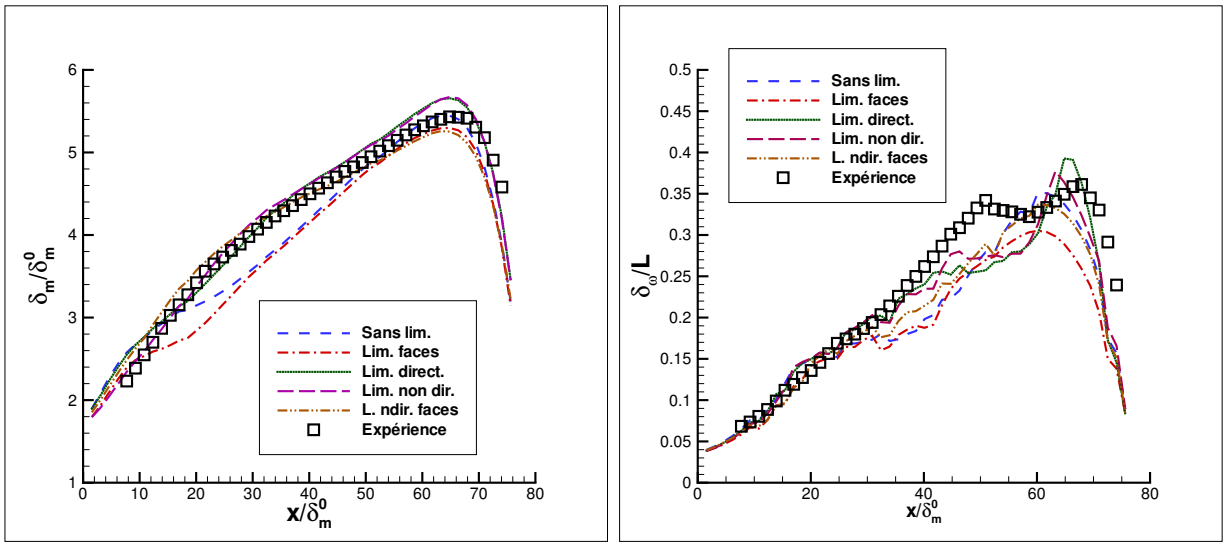
(2) Dans la région $32\delta_m^0 \leq x \leq 60\delta_m^0$, la pente revient à une valeur de $0,042$, ce qui correspond à peu près au taux de croissance des couches de mélange turbulentes libres.

(3) Dans la région $60\delta_m^0 \leq x \leq 77\delta_m^0$, l'épaisseur de quantité de mouvement diminue lorsque l'écoulement s'approche du coin aval de la cavité. La décroissance finale est cependant un artefact provoqué par le fait que les bornes d'intégration dans (13.4.1) sont trop petites dans cette région.

Une autre quantité importante est l'épaisseur de vorticité, définie dans le cas présent par

$$\delta_\omega = \frac{v_\infty}{\max\left\{ \dfrac{\partial \overline{v}_x}{\partial z} \,\middle|\, -0,2\,L \leq z \leq 0,2\,L \right\}}\,. \tag{13.4.2}$$

La figure 13.4.7 montre l'évolution de δ_m et de δ_ω en fonction de x.



(a) Epaisseur de quantité de mouvement (b) Epaisseur de vorticité

FIG. 13.4.7: Grandeurs caractéristiques de la couche de mélange.

On constate que le comportement particulier de l'épaisseur de quantité de mouvement est assez bien reproduit par les cinq options de discrétisation spatiale. Il est intéressant de noter que les deux options non monotones, l'option 1 sans limiteur et l'option 2 avec le limiteur de face, sous-estiment l'épaisseur de quantité de mouvement dans la région $10\delta_m^0 \leq x \leq 50\delta_m^0$. L'épaisseur de vorticité est bien reproduite dans la région $10\delta_m^0 \leq x \leq 30\delta_m^0$ et sous-estimée dans le reste du domaine. Ici, les options non monotones 1 et 2 fournissent également des valeurs plus faibles que les options monotones 3, 4 et 5. Il faut noter que l'analyse de l'épaisseur de vorticité 13.4.2 nécessite d'interpoler la dérivée partielle $\dfrac{\partial \overline{v}_x}{\partial z}$, ce qui pourrait expliquer pourquoi la courbe de l'épaisseur de vorticité est moins lisse que celle de l'épaisseur de quantité de mouvement.

13.4.4. Étude des tensions de Reynolds. L'objectif est de comparer les profils des tensions de Reynolds avec les données expérimentales, avec une attention particulière à la composante de cisaillement $\overline{v'_x v'_z}$ du tenseur de Reynolds. Pour cela, les composantes $\overline{v'_x v'_z}$, $\overline{v'_x v'_x}$ et $\overline{v'_z v'_z}$ sont mesurées en fonction de la coordonnée verticale z dans le plan $y = 0$ aux coordonnées $x = 0,2\,L$, $x = 0,4\,L$, $x = 0,6\,L$, et $x = 0,8\,L$, c'est-à-dire juste au-dessus de la cavité. Les mesures sont prises entre les hauteurs $z = -0,2\,L$ et $z = 0,2\,L$.

La figure 13.4.8 montre la composante de cisaillement $\overline{v'_x v'_z}$ du tenseur de Reynolds. Les résultats numériques sont proches des données expérimentales. Les écarts avec l'expérience sont comparables à ceux constatés sur des maillages structurés par Larchevêque et al. [**75, 76**] ainsi que Bertier [**17**].

L'étude de la précision des schémas, présentée dans le chapitre 11, peut expliquer ces écarts. Les schémas d'ordre deux montrent en effet une erreur dispersive, c'est-à-dire une erreur sur la vitesse de convection, relativement importante. Les schémas d'ordre trois permettent d'éliminer cette erreur, mais il n'a pas été possible de les implémenter à temps dans CEDRE pour pouvoir comparer les résultats entre la discrétisation d'ordre deux et d'ordre trois.

La figure 13.4.9 montre la composante $\overline{v'_x v'_x}$ du tenseur de Reynolds. On constate que l'accord entre les résultats numériques et l'expérience est bon aux points $x = 0,2\,L$ et $x = 0,4\,L$ et se détériore vers l'aval de la cavité. Finalement, la figure 13.4.10 montre la composante $\overline{v'_z v'_z}$ du tenseur de Reynolds, avec des écarts qui sont plus grands pour $z \leq 0$ que pour $z \geq 0$.

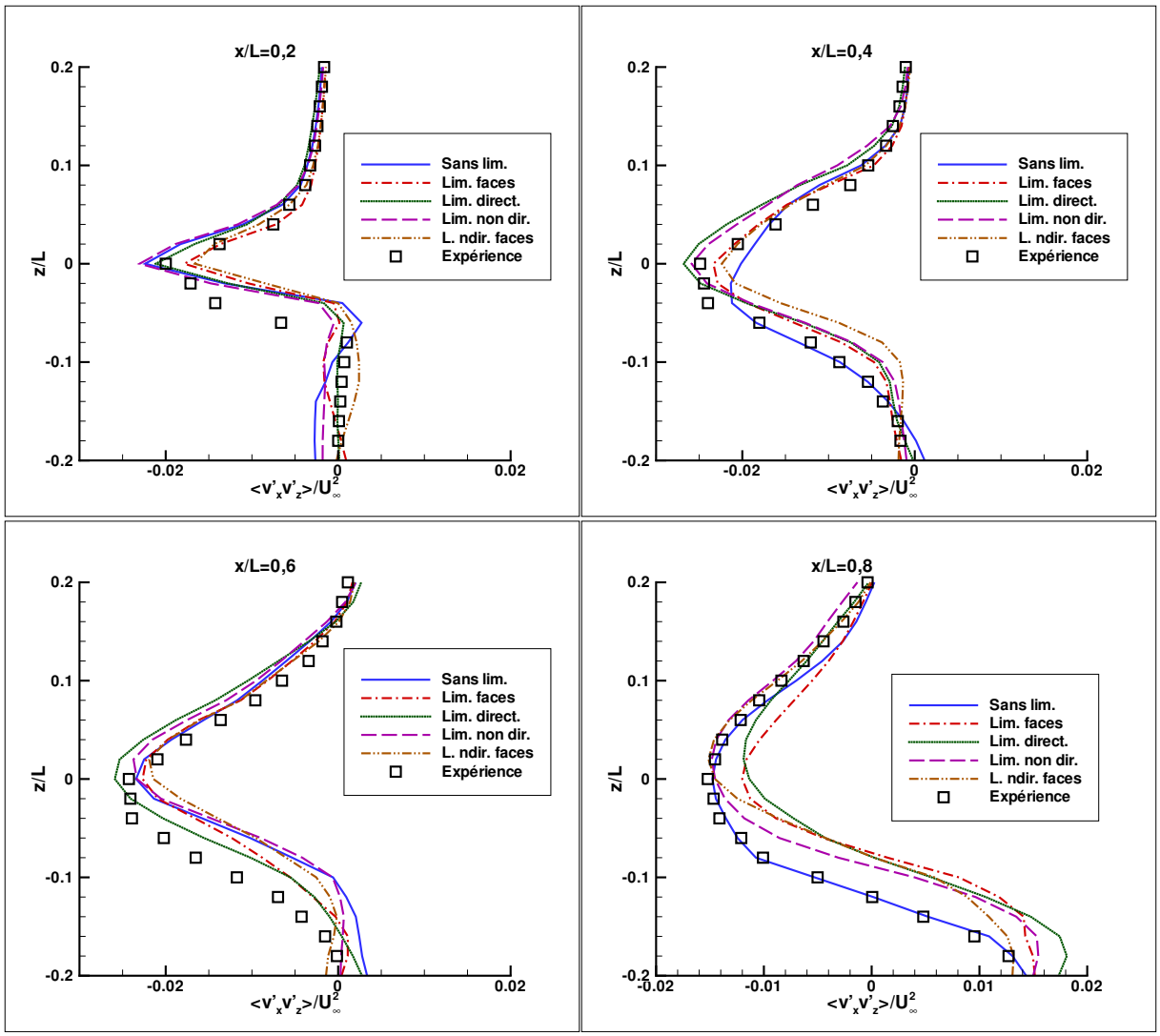


FIG. 13.4.8: Composante de cisaillement $\overline{v'_x v'_z}$ du tenseur de Reynolds.

13.5. Bilan du chapitre

Les calculs effectués permettent tout d'abord de tirer une conclusion importante pour les méthodes de discrétisation spatiale. Il s'avère que ce type de calcul peut être stable sans limiteur de gradient si la reconstruction du gradient donne une discrétisation stable de la convection linéaire. Les problèmes de stabilité, constatés avec CEDRE sur des maillages de tétraèdres, sont donc d'abord dûs à une instabilité linéaire du schéma qui a son origine dans la reconstruction du gradient. Cela montre la pertinence de l'étude de stabilité du chapitre 10.

Les simulations décrites dans ce chapitre ont également permis de comparer la précision de cinq options de discrétisation spatiale d'ordre deux à l'exemple d'une simulation des grandes échelles en maillage non structuré. Les données comparées incluent la densité spectrale de puissance d'un signal de pression dans la cavité, les profils de la vitesse moyenne et les tensions de Reynolds dans la couche de mélange au-dessus de la cavité. Finalement, l'épaisseur de quantité de mouvement et l'épaisseur de vorticité ont été calculées afin de vérifier le comportement caractéristique de la couche de mélange au-dessus de la cavité.

Les résultats numériques ont permis de retrouver la dynamique instationnaire de l'écoulement dans le spectre du signal de pression. La simulation reproduit notamment les raies spectrales qui sont des multiples d'une fréquence fondamentale $f_1 \approx 1975$. Les résultats numériques obtenues fournissent une bonne approximation des données expérimentales, notamment de la composante

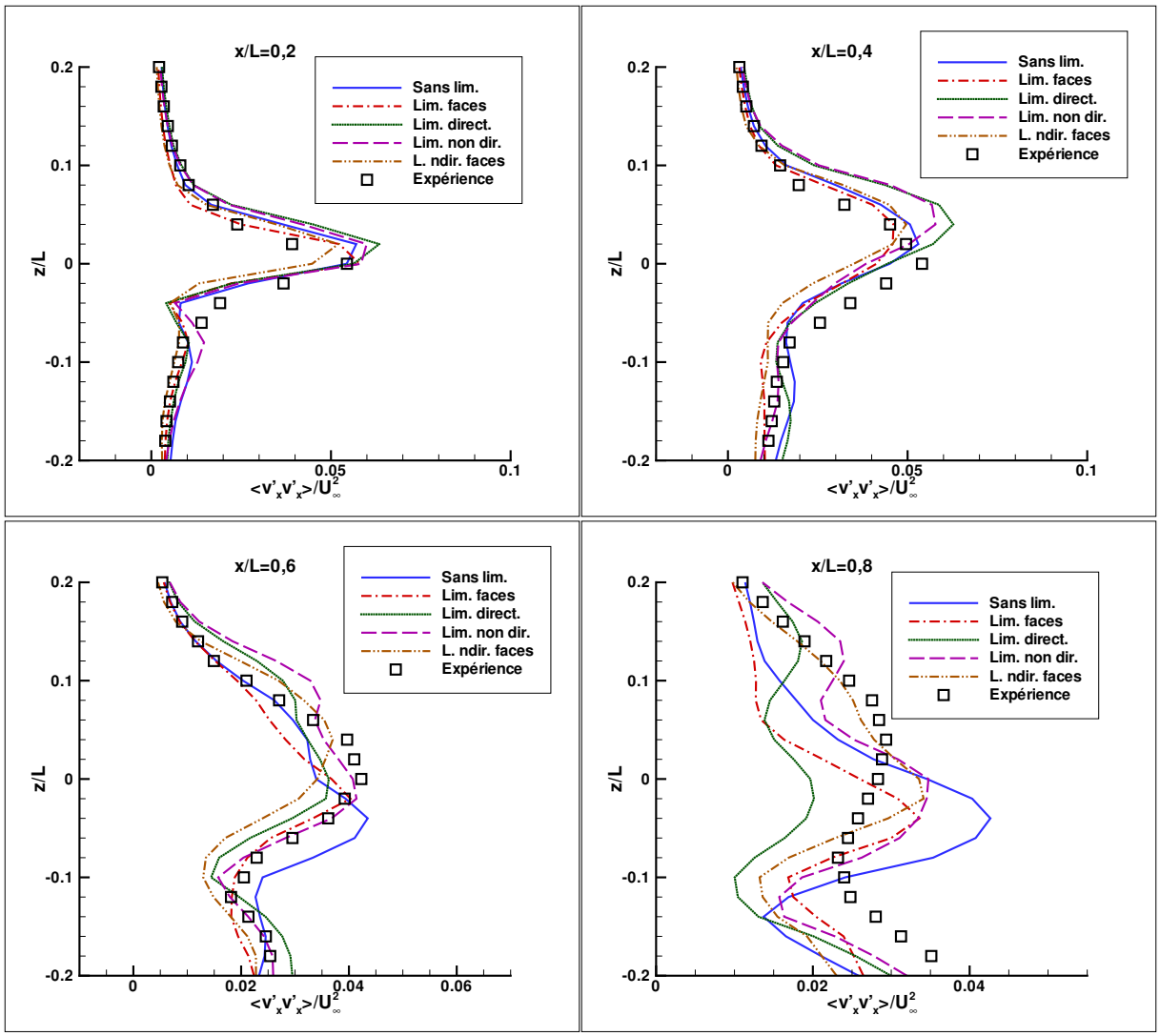


FIG. 13.4.9: Composante $\overline{v'_x v'_x}$ du tenseur de Reynolds.

de cisaillement $\overline{v'_x v'_z}$ du tenseur de Reynolds. Il a également été possible de retrouver la croissance caractéristique de l'épaisseur de quantité de mouvement dans la couche de mélange.

Finalement, on peut encore faire deux constats :

(1) Il ne semble pas y avoir de différence excessive entre les diverses options de discrétisation spatiale. Dans ce type de calcul, les limiteurs ne dégradent pas trop la précision par rapport à la discrétisation sans limiteur.

(2) Les résultats semblent être du même niveau que les résultats obtenus par N. Bertier avec CEDRE en maillage structuré.

Pour expliquer ces observations, on peut constater qu'une grande partie de la dynamique se déroule dans la couche de prismes au-dessus de la cavité. Les bases de ces prismes forment un maillage triangulaire très régulier qui est légèrement plus fin et beaucoup plus régulier que le maillage triangulaire le plus fin de l'étude de précision du chapitre 11. Lorsqu'on regarde la figure 11.3.4 sur la page 198, qui montre les résultats de convection pour le maillage triangulaire numéro neuf, on s'aperçoit que le schéma d'ordre deux fournit déjà de très bons résultats sur ces maillages très fins. Comme le maillage de prismes s'obtient par extrusion d'un maillage triangulaire très régulier et assez fin, on peut supposer que les bonnes qualités du maillage triangulaire se transmettent au maillage de prismes. Le chapitre 14 montre que les calculs sur des maillages composés exclusivement de tétraèdres sont plus difficiles.

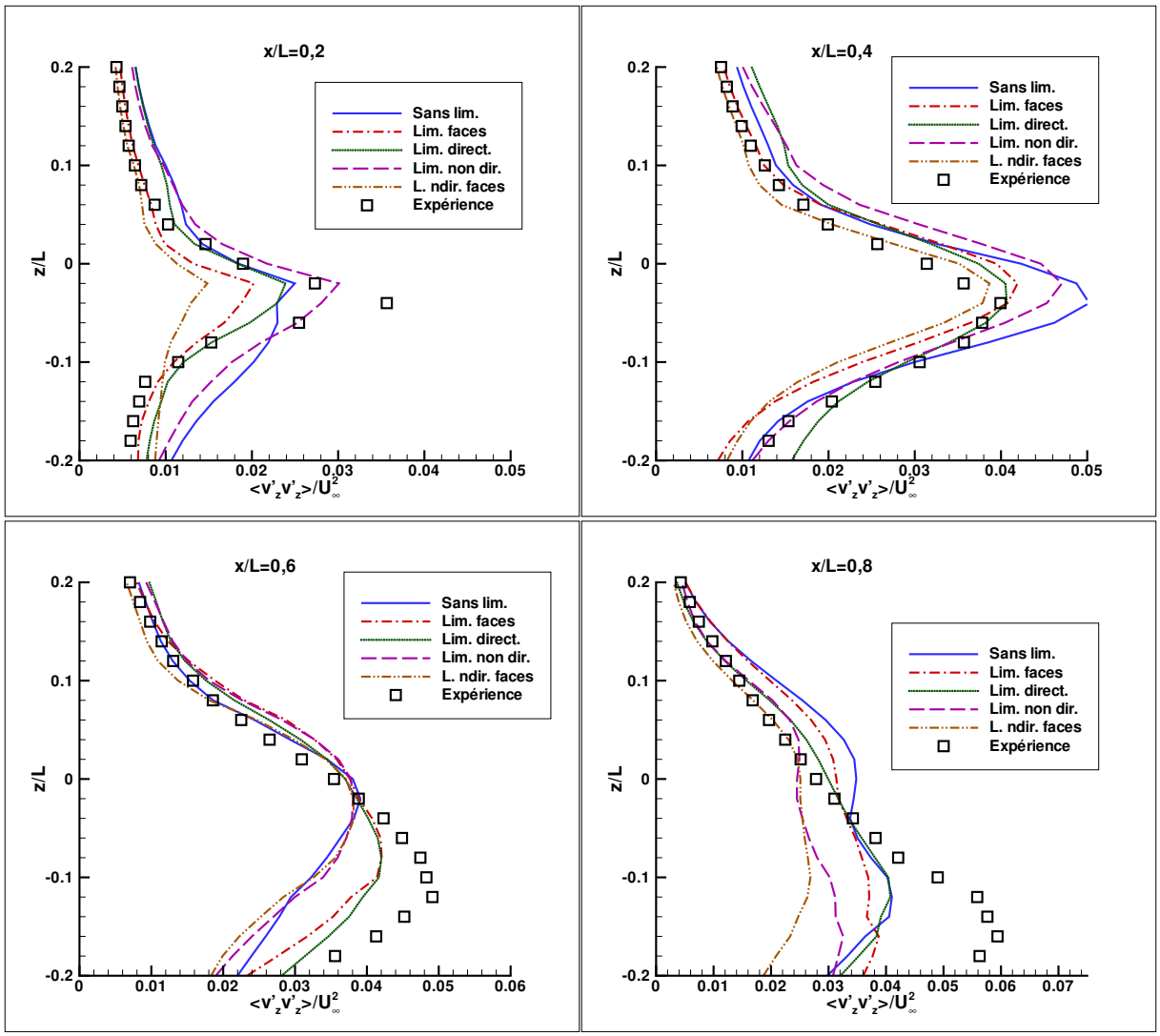


FIG. 13.4.10: Composante $\overline{v'_z v'_z}$ du tenseur de Reynolds.

Finalement, ce calcul pourra servir de cas test pour les méthodes de reconstruction d'ordre élevé développées dans le chapitre 8, et en particulier les méthodes de reconstruction quadratique. On s'attend à ce que la reconstruction quadratique permette de capter encore mieux la dynamique de la couche de mélange et de réduire les écarts entre les résultats numériques et les données expérimentales.

CHAPITRE 14

Simulation des grandes échelles d'un jet chaud supersonique

14.1. Objectif du chapitre

L'objectif est d'appliquer certaines des évolutions de la discrétisation spatiale présentées dans ce document à des simulations d'un jet supersonique. Il s'agit de la simulation instationnaire d'un jet chaud supersonique qui a fait l'objet d'une étude expérimentale par Seiner et al., cf. [**93**]. Ces simulations servent à analyser les différents aspects de la discrétisation spatiale sur des maillages non structurés en dimension trois dans le cas supersonique.

La réduction du bruit de jet est un objectif très important pour l'industrie aéronautique. Dans ce contexte, les simulations numériques peuvent aider à la conception de dispositifs de réduction du bruit. Des efforts sont actuellement entrepris pour étendre l'approche de la simulation des grandes échelles aux calculs instationnaires de jets dans des configurations industrielles proches de la réalité. Ce type de simulation peut capter les grandes échelles instationnaires de la turbulence et le champ acoustique du jet. Cela constitue un avantage par rapport aux simulations basées sur l'approche RANS qui fournissent uniquement une description statistique de la turbulence et nécessitent des modèles pour reconstruire le champ acoustique.

Les dispositifs de réduction de bruit font en général appel à des géométries complexes. Pour cette raison, il est souvent plus facile de recourir aux maillages non structurés pour simuler ce type de configuration. Ce constat a été à l'origine d'un certain nombre de simulations instationnaires du bruit de jet avec Cedre [**81**]. Ces travaux ont notamment révélé qu'une dissipation numérique excessive empêche le jet de devenir turbulent en maillage de tétraèdres, alors que la transition vers un état turbulent se fait sans problème en maillage structuré. Une amélioration satisfaisante des résultats sur des maillages de tétraèdres a été obtenue en insérant un bloc structuré dans les zones de mélange du jet avec l'air ambiant.

Ces résultats posent la question de la précision des schémas d'ordre deux en maillage non structuré. Il est intéressant de déterminer si la perte de précision est uniquement provoquée par les limiteurs où si elle provient aussi de la reconstruction du gradient. L'étude numérique de ce chapitre sert à clarifier cette question et à explorer les pistes à poursuivre pour obtenir une amélioration de la discrétisation spatiale. L'objectif n'est pas de faire une étude très détaillée de tous les phénomènes physiques du jet mais de se concentrer sur la transition vers l'état turbulent. L'indicateur de cette transition est en première ligne le profil de la vitesse axiale.

14.2. Résumé de l'étude expérimentale

L'article de Seiner et Ponton [**93**] explore la génération d'ondes acoustiques par des jets supersoniques chauds. L'étude est réalisée sur une tuyère axisymétrique refroidie à l'eau. La tuyère, dont le diamètre à la sortie est de $D_\mathrm{j} = 9,144$ cm, est conçue pour un nombre de Mach $M = 2$ à une température de 1366 K. L'étude couvre des températures allant de 313 K à 1534 K.

L'étude s'intéresse notamment à deux phénomènes physiques et leur importance relative dans la production de bruit. Le premier phénomène, appelé *eddy Mach wave emission*, repose sur la convection de structures turbulentes du jet à une vitesse supersonique par rapport à la vitesse du son dans le milieu ambiant. Le deuxième phénomène sont les ondes de type Kelvin-Helmholtz liées au cisaillement entre le jet et le milieu ambiant. Les mesures acoustiques et aérodynamiques indiquent que les instabilités de type Kelvin-Helmholtz sont le mécanisme dominant dans la production du bruit sur toute la plage des températures testées. L'étude expérimentale montre également qu'une tuyère de forme elliptique permet de réduire les ondes de Kelvin-Helmholtz.

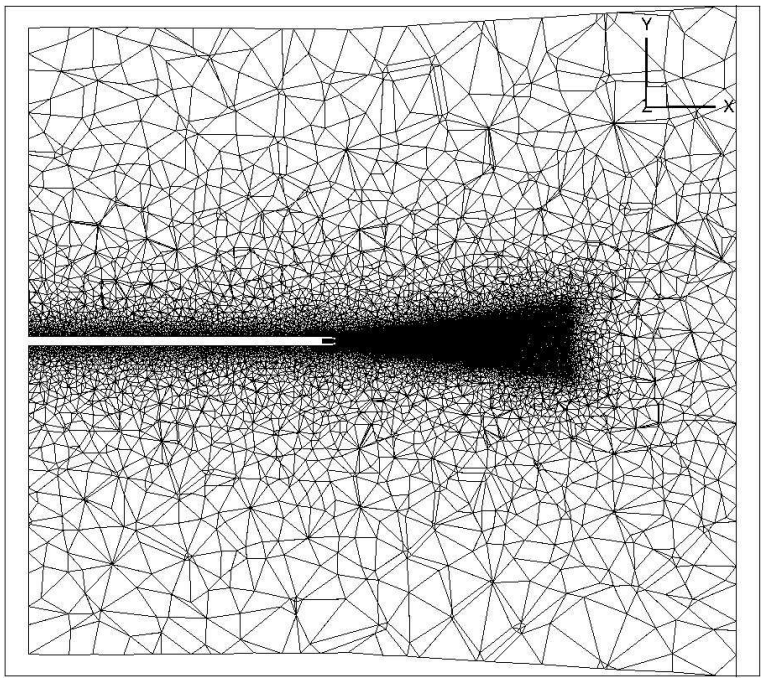


FIG. 14.3.1: Maillage du domaine cylindrique. Vue du plan $z = 0$.

Les mesures acoustiques se font à l'aide de microphones dans une chambre anéchoïque ce qui permet d'obtenir un ensemble de spectres acoustiques. Les mesures aérodynamiques concernent surtout la vitesse axiale moyenne et sa variation dans l'axe de symétrie de la tuyère. Il s'agit ici d'un résultat important que toute simulation numérique du jet doit correctement reproduire. La vitesse axiale moyenne du jet diminue dans la direction axiale en raison du développement de la turbulence. La simulation instationnaire avec CEDRE [**81**] n'a pas pu reproduire cette diminution caractéristique de la vitesse axiale sur des maillages de tétraèdres.

14.3. Description des calculs

14.3.1. Présentation de la géométrie et du maillage. La figure 14.3.1 montre le domaine de calcul en forme de cylindre d'une hauteur de $L_{\text{dom}} \approx 16,6\,\text{m}$ et d'un rayon de $R_{\text{dom}} \approx 7,3\,\text{m}$. L'axe de symétrie du cylindre coïncide avec l'axe x. La tuyère est insérée dans le domaine tel que son axe de symétrie coïncide avec celle du cylindre entourant. La figure 14.3.2 montre la forme particulière de la tuyère et le maillage autour du culot.

Le maillage, composé d'environ deux millions de tétraèdres, est raffiné dans une zone qui a la forme d'un cône. C'est dans cette zone que le jet devient turbulent. La figure 14.3.2 montre également que le maillage est encore plus fin dans la zone de mélange du jet avec l'air environnant.

Des capteurs sont positionnés derrière la sortie de la tuyère pour enregistrer les signaux des ondes acoustiques, voir la figure 14.3.3. Les spectres de ces signaux peuvent ensuite être comparés aux résultats expérimentaux. Pour le calcul parallèle sur 64 processeurs, le maillage est découpé en 256 domaines.

14.3.2. Conditions aux limites pour le calcul. La base du cylindre, c'est-à-dire le plan $x = -7,01\,\text{m}$, est une entrée subsonique qui impose une vitesse de $v_x = 10\,\frac{\text{m}}{\text{s}}$ en direction de l'axe x et une température de 280 K.

Les conditions limites aux parois extérieures du domaine cylindrique et au sommet du cylindre, c'est-à-dire au plan $x = 9,5\,\text{m}$, sont de type subsonique sortant et imposent une pression statique de $p = 101300\,\text{Pa}$. Des conditions limites particulières empêchent la réflexion d'ondes aux sorties du domaine.

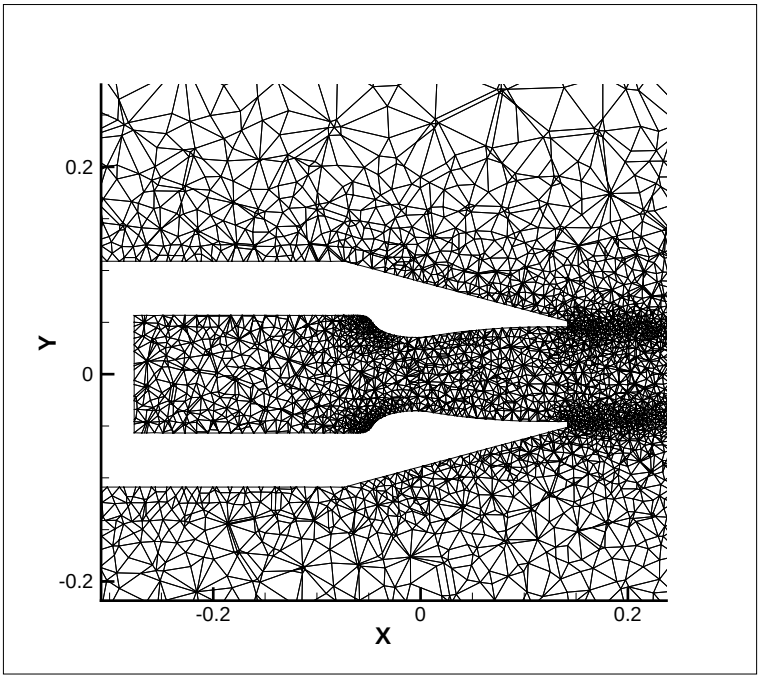


FIG. 14.3.2: Géométrie de la tuyère. Vue du plan $z = 0$.

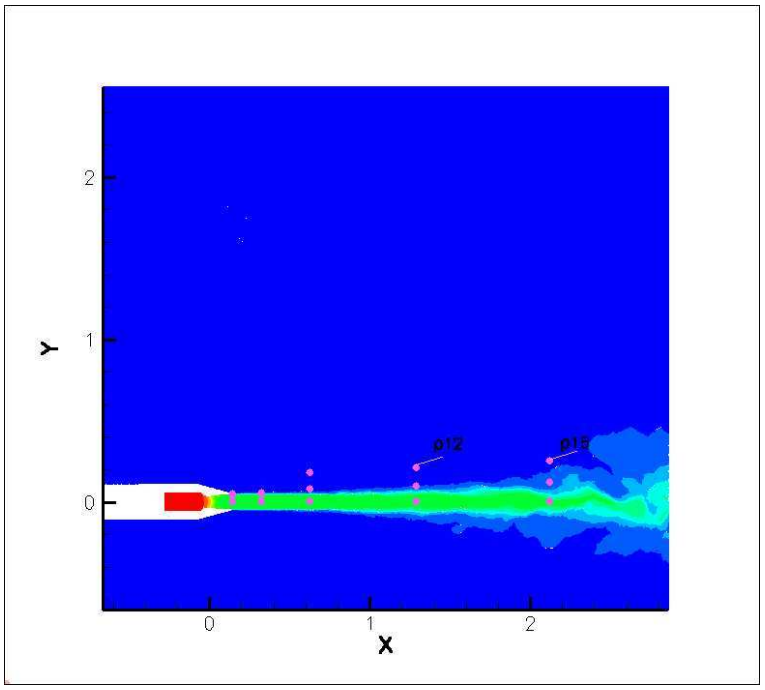


FIG. 14.3.3: Position des capteurs. Vue du plan $z = 0$.

L'entrée du jet se situe à l'intérieur de la tuyère dans le plan $x = -0,27\,\text{m}$. La condition limite de cette entrée est de type subsonique entrant et impose une pression d'arrêt de $p_i = 817950\,\text{Pa}$ ainsi qu'une température d'arrêt de $T_i = 1370\,\text{K}$.

Les parois extérieures et intérieures de la tuyère et les parois du culot sont modélisées par des conditions de type paroi.

14.3.3. Présentation des méthodes de discrétisation spatiale. L'objectif principal du calcul consiste à comparer plusieurs options de discrétisation spatiale. Puisque les reconstructions d'ordre élevé décrites dans les chapitres 7 et 8 ne sont pas encore disponibles dans Cedre, la validation concerne les options de reconstruction et de limitation du gradient. Contrairement à l'écoulement décrit au chapitre 13, l'écoulement du jet présente un choc dans la tuyère. Il est donc plausible que la simulation du jet nécessite une limitation de gradient au moins à l'intérieur de la tuyère. Cette hypothèse a été confirmée par les calculs.

Finalement, les tests couvrent les options suivantes :

(1) Un premier calcul avec le limiteur de gradient non directionnel et le limiteur de face de Cedre, décrits dans la section 12.6, sert de calcul de référence (abrégé *Lim. non dir. et faces* dans la légende). L'objectif des autres options de calcul est d'apporter une amélioration par rapport à cette option.

(2) Un deuxième calcul sert à évaluer le limiteur directionnel basé sur le principe du maximum, défini par l'algorithme 12.4.2 dans la section 12.4 (abrégé *Lim. dir.* dans la légende). On s'attend à ce que ce type de schéma génère moins de dissipation numérique et capte mieux la turbulence que l'option 1. La vitesse axiale moyenne du jet devrait donc diminuer plus rapidement avec une distance croissante à la sortie de la tuyère.

(3) Un troisième calcul sert à vérifier si les problèmes de précision proviennent des limiteurs ou de la reconstruction du gradient. Pour cela, on active les limiteurs seulement à l'intérieur de la tuyère. À l'extérieur de la tuyère, on utilise la reconstruction du gradient sur le deuxième voisinage, décrit par l'algorithme 8.4.1 de la section 8.4, sans limitation du gradient. Si le profil de vitesse est correctement reproduit, il est possible de conclure que le problème de transition vers l'état turbulent est causé par la limitation du gradient. Cette option est abrégée *Sans Lim. Reco. Stable* dans la légende.

(4) Un quatrième calcul utilise les mêmes options que le calcul 3, mais avec le limiteur de face, décrit dans la section 12.6. On rappelle que le limiteur de face ne peut pas empêcher l'apparition de nouveaux extrema locaux car il n'agit pas sur la cellule entière mais seulement sur une face à la fois. On peut donc supposer qu'il engendre moins de dissipation numérique que les limiteurs basés sur le principe du maximum. Il faut toutefois prendre en considération que le principe du maximum impose des bornes admissibles moins restrictives sur chaque face que les limiteurs de face, voir la section 12.6 pour les détails. Cette option est abrégée *Lim. faces Reco. Stable* dans la légende.

(5) Un cinquième calcul combine l'option de reconstruction du gradient sur le deuxième voisinage, décrit par l'algorithme 8.4.1, avec les limiteurs directionnels sur tout le domaine de calcul (abrégé *Lim. dir. Reco. stable* dans la légende).

14.3.4. Intégration en temps. L'intégration en temps s'effectue par une méthode de Runge-Kutta explicite d'ordre deux, le schéma de Heun, avec un pas de temps $\Delta t = 10^{-7}$ sec. Le calcul démarre à partir d'un état initial stationnaire, généré avec une méthode d'intégration implicite en temps.

La durée du calcul instationnaire est de $0,05$ s dont les premiers $0,01$ s ne sont pas exploitées car elles constituent la phase de démarrage pendant laquelle le jet devient turbulent. Le moment à partir duquel l'écoulement a atteint son régime asymptotique s'observe à l'aide des signaux de pression des capteurs.

La vitesse axiale du jet est de $U_{\mathrm{j}} \approx 1130 \frac{\mathrm{m}}{\mathrm{s}}$ et le rayon de la tuyère est de $R_{\mathrm{t}} = 0,0456\,\mathrm{m}$. Le nombre de Mach à la sortie de la tuyère est à peu près $M \approx 2$. Le temps nécessaire pour le calcul est au moins égal à $t_{\mathrm{fin}} = N_{\mathrm{fin}} \frac{D_{\mathrm{t}}}{U_{\mathrm{axi}}}$ où N_{fin} doit être supérieur à 100. La valeur de $N_{\mathrm{fin}} = 140$ donne dans le cas présent une durée de calcul minimale de $t_{\mathrm{fin}} \approx 0,01$ sec.

14.3.5. Modèle physique. Les calculs ont été effectués avec une modélisation de la turbulence par le modèle de Smagorinsky qui est disponible dans Cedre depuis les travaux de N. Bertier [**17**]. La section 3.3 présente les détails de ce modèle. Les paramètres utilisés sont les mêmes que dans la section 13.3.5 :

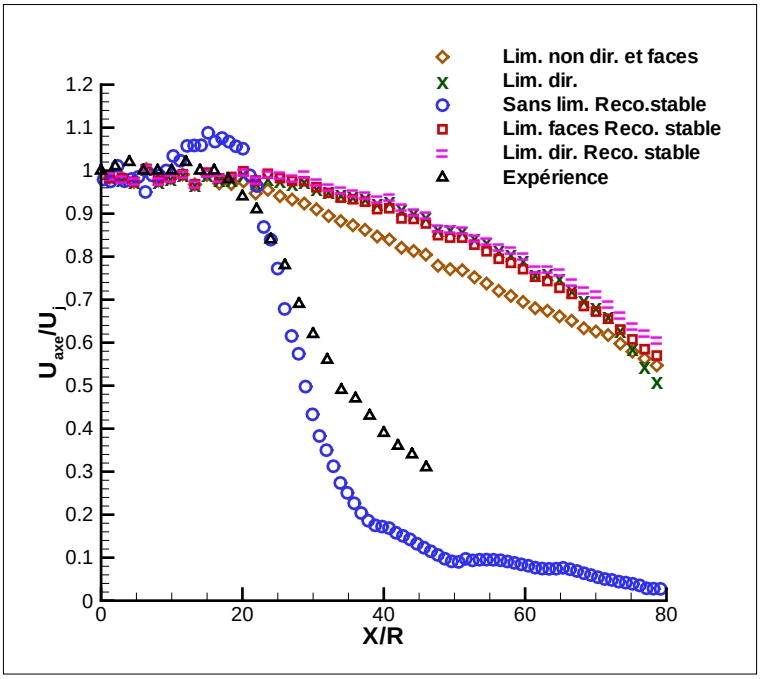


FIG. 14.4.1: Profil de la vitesse axiale pour les cinq options testées.

(1) Constante de Prandtl : 0,09.
(2) Constante de Karman : 0,4.
(3) Nombre de Prandtl turbulent : $\mathrm{Pr}^{(\mathrm{sgs})} = 0,9$.
(4) Nombre de Schmidt turbulent : 0,9.
(5) Constante de longueur de mélange : $\mathrm{C_S} = 0,1$.

Le modèle de paroi est un modèle de Couette laminaire.

14.4. Résultats de l'étude numérique

Le profil longitudinal de la vitesse axiale est un indicateur important pour évaluer la pertinence des simulations de jet. Le développement de la turbulence du jet conduit à une décroissance caractéristique de sa vitesse axiale le long de l'axe de symétrie de la tuyère. Les travaux [**81**] ont montré que la discrétisation spatiale de CEDRE empêche la transition du jet vers l'état turbulent sur des maillages de tétraèdres. Ce problème se manifeste sur le profil de la vitesse axiale moyenne qui ne décroît que faiblement à la sortie de la tuyère.

La figure 14.4.1 montre les résultats pour les cinq options de discrétisation spatiale testées. Toutes les options nécessitent une limitation du gradient à l'intérieur de la tuyère. La vitesse axiale du jet est rapportée à la vitesse $U_\mathrm{j} \approx 1130\,\frac{\mathrm{m}}{\mathrm{s}}$ et la coordonnée x est rapportée au rayon de la tuyère $R_\mathrm{t} = 0,0456\,\mathrm{m}$.

Contrairement à la simulation décrite dans le chapitre 13, ce calcul montre des grandes différences entre les différentes options de discrétisation spatiale. Les options

– gradient sur le premier voisinage , limiteur non directionnel et limiteur de face (option 1)
– gradient stable sur le deuxième voisinage et limiteur de face (option 4)
– gradient stable sur le deuxième voisinage et limiteur directionnel (option 5)

n'arrivent pas à reproduire la décroissance de la vitesse axiale.

Contrairement aux autres options de calcul, l'option 3, sans limiteur à l'extérieur de la tuyère et avec la reconstruction stable du gradient, montre une décroissance de la vitesse axiale qui est beaucoup plus forte que pour les autres options. Cette option de calcul a cependant un défaut : la vitesse axiale augmente autour de la coordonnée axiale $x \approx 20\,R_\mathrm{t}$, ce qui est certainement dû à la transition entre le domaine de calcul avec limiteur et le domaine de calcul sans limiteur.

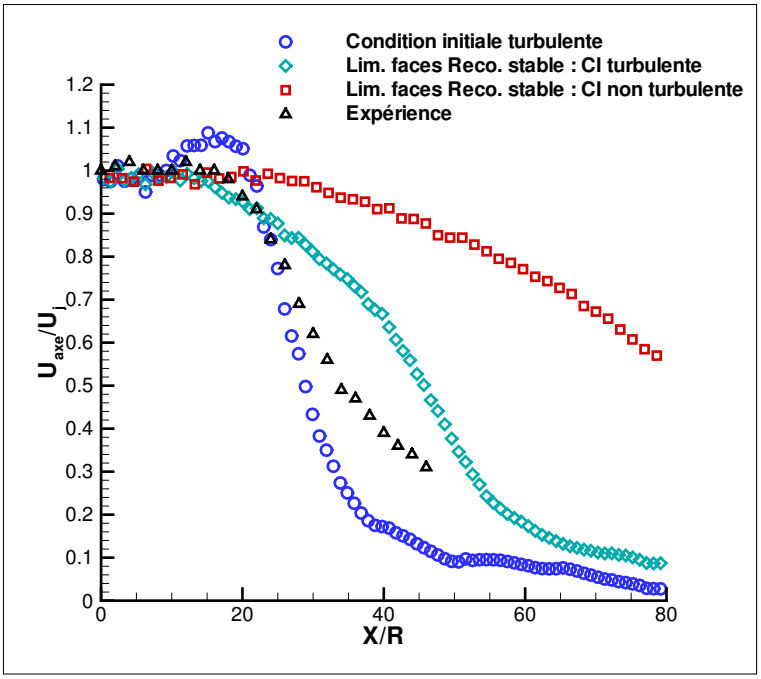


FIG. 14.4.2: Profil de la vitesse axiale moyenne obtenu pour l'option de discrétisation spatiale *Lim. faces Reco.stable* (option 4) avec la condition initiale turbulente.

Néanmoins, cette option de calcul montre clairement que c'est l'action des limiteurs qui empêche la transition de l'écoulement vers l'état turbulent, car seulement l'option de calcul sans limitation en dehors de la tuyère permet à la turbulence de se développer.

Il est alors intéressant de se poser la question suivante : les limiteurs empêchent-ils seulement la transition de l'état initial non turbulent vers l'état turbulent ou suppriment-ils complètement la turbulence, même si la condition initiale était déjà turbulente ?

Pour trouver une réponse à cette question, on fait un nouveau calcul avec l'option de calcul 4 (limiteurs de face) en utilisant le résultat final de l'option de calcul 3 comme condition initiale. Cette condition initiale présente l'avantage d'être déjà pleinement turbulente. Ce calcul permet de vérifier si l'état turbulent de l'écoulement est préservé par l'option de calcul 4.

La figure 14.4.2 montre le résultat de ce calcul pour l'option 4. On constate que le résultat avec la condition initiale turbulente est beaucoup mieux que le résultat obtenu avec la condition initiale non turbulente. Le problème semble donc plutôt résider dans le fait que les limiteurs empêchent la transition vers l'état turbulent. Une fois celui-ci établi, la turbulence n'est plus supprimée. L'utilisation des limiteurs avec la condition initiale turbulente permet également de supprimer la croissance factice de la vitesse axiale qui apparaît dans le calcul avec l'option 3.

Néanmoins, le résultat présenté dans la figure 14.4.2 peut encore être amélioré. Il est d'abord nécessaire de refaire le calcul avec la condition initiale turbulente en utilisant les nouveaux options de limitation, notamment l'option 5, pour déterminer si cela donne un résultat encore plus précis. Il faudra ensuite faire évoluer les algorithmes de limitation du chapitre 12. Il semble que l'introduction de la limitation directionnelle décrite dans le chapitre 12 ne soit pas suffisante. Il faudrait également développer un critère de monotonie encore moins restrictif que le principe du maximum du théorème 12.2.1.

Finalement, il est intéressant d'analyser des spectres acoustiques du jet pour les différentes options de discrétisation spatiale. Pour cela, on choisit le signal de pression au capteur 15 qui est le capteur le plus éloigné de la sortie de la tuyère, comme indiqué sur la figure 14.3.3. La figure 14.4.3 montre la densité spectrale de puissance du signal de pression au capteur 15. Cette figure illustre bien les différences entre les options de discrétisation : la puissance du signal est la plus forte pour le schéma sans limiteur. Les autres méthodes donnent à peu près les mêmes

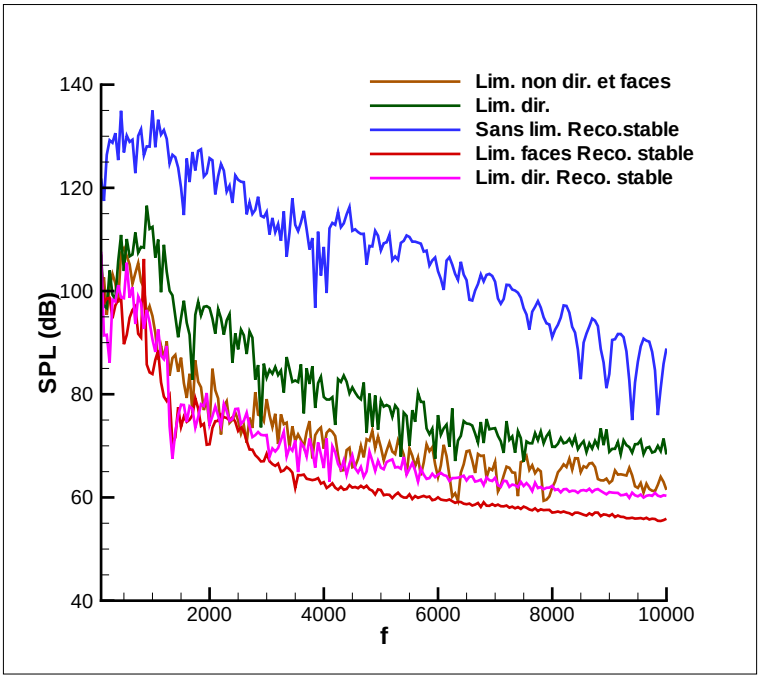


FIG. 14.4.3: Spectres du signal de pression au capteur 15.

niveaux. Il est intéressant de noter que la différence remarquable entre l'option *Lim. dir.* et l'option *Lim. dir. Reco. stable* réside seulement dans le fait que la première option utilise une reconstruction instable sur le premier voisinage. La différence entre les niveaux de la densité spectrale de puissance de ces deux méthodes est donc entièrement provoquée par l'instabilité de la discrétisation spatiale.

14.5. Bilan du chapitre

Le chapitre décrit la simulation d'un jet chaud supersonique sur un maillage de tétraèdres. Ce jet a fait l'objet d'une étude expérimentale par Seiner et Ponton [**93**] qui a entre autres permis de mesurer le profil de la vitesse axiale du jet. Des simulations instationnaires de ce jet par Lupoglazoff, Rahier et Vuillot [**81**] sur des maillages de tétraèdres avec le schéma d'ordre deux et les limiteurs de CEDRE n'ont pas permis de retrouver un état turbulent.

L'objectif de l'étude numérique a été d'analyser pourquoi ce schéma numérique ne permet pas la transition de l'écoulement vers un état turbulent en maillage de tétraèdres. La comparaison des cinq options de calcul montre que ce sont les limiteurs qui empêchent la transition vers l'état turbulent : seulement l'option de calcul sans limiteurs de gradient en dehors de la tuyère permet de retrouver une décroissance de la vitesse axiale qui correspond à un écoulement turbulent.

Le champ de vitesse final de ce calcul a alors été utilisé comme condition initiale turbulente dans un calcul avec les limiteurs de face de CEDRE. Il s'avère que ce calcul donne un résultat nettement meilleur que le calcul avec une condition initiale non turbulente. Les limiteurs empêchent donc surtout la transition vers la turbulence, mais ne semblent pas supprimer la turbulence si elle est présente dans la condition initiale. Il est encore nécessaire de tester les autres options de calcul avec la nouvelle condition turbulente.

L'une des conclusions de l'étude numérique est que le principe du maximum, exprimé par le théorème 12.2.1, est peut-être un peu trop restrictif pour la simulation des grandes échelles, au moins dans le cas du jet. Il serait important de chercher de nouveaux critères de monotonie moins restrictif car le seul principe de la limitation directionnelle ne semble pas suffisant pour permettre la transition vers l'état turbulent.

CHAPITRE 15

Conclusion

15.1. Synthèse

15.1.1. Objectifs. Les objectifs principaux de l'étude étaient
- d'améliorer la stabilité et la précision des schémas volumes finis en maillage non structuré général ;
- d'analyser des algorithmes efficaces et faciles à implémenter pour la montée en ordre du schéma volumes finis ;
- valider certaines évolutions de la discrétisation spatiale, élaborées au cours de cette thèse, sur des simulations des grandes échelles en maillage non structuré.

L'effort de l'étude s'est concentré sur la discrétisation spatiale dans le contexte de l'approche MUSCL. La motivation principale pour le présent travail était l'expérience avec le code CEDRE développé au sein du département DSNA de l'ONERA. Ce code, qui contient un solveur de type volumes finis pour les équations de Navier-Stokes compressibles, est de plus en plus utilisé pour la simulation des grandes échelles (LES), ce qui pose de nouveaux défis pour la précision et robustesse des schémas numériques.

15.1.2. Étude préliminaire. L'étude préliminaire du chapitre 5 a permis de mettre en place les outils indispensables pour les chapitres suivants : une notation efficace pour la géométrie des maillages non structurés, l'introduction de tenseurs associés au maillage et la démonstration de plusieurs identités géométriques. Les résultats de cette étude préliminaire peuvent par ailleurs servir pour d'autres méthodes numériques que les volumes finis.

Le chapitre 6 rappelle les principes de la discrétisation spatiale par la méthode des volumes finis et définit la notion d'erreur de troncature du schéma. Pour obtenir une erreur de troncature qui diminue plus rapidement avec le diamètre du maillage, on introduit l'opération de reconstruction : on appelle ainsi des méthodes capables de fournir une approximation de toute fonction u suffisamment régulière à partir des seules moyennes de cellule de u. Si l'erreur d'approximation est d'ordre $k+1$ par rapport au diamètre du maillage, on parle de reconstruction d'ordre $k+1$. La discussion dans le chapitre 6 démontre alors comment la reconstruction d'une fonction dans chaque cellule permet de calculer des flux plus précis aux interfaces des cellules : si la reconstruction est d'ordre $k+1$, l'erreur de troncature du schéma est au moins d'ordre k. Ce résultat constitue la motivation principale pour analyser des reconstructions d'ordre $k+1$ en maillage non structuré général.

15.1.3. Étude de la reconstruction des polynômes en maillage non structuré. L'étude proprement dite commence dans le chapitre 7 par le développement d'une théorie générale de la reconstruction en maillage non structuré à partir de quelques principes fondamentaux :

(1) La fonction reconstruite dépend linéairement des moyennes de cellule.

(2) Les fonctions reconstruites sont des polynômes de degré k.

(3) La reconstruction est locale, c'est-à-dire seules les moyennes de cellule dans un voisinage de la cellule contribuent à la reconstruction.

Ces trois principes déterminent la forme générale des fonctions reconstruites.

À cela, il est nécessaire d'ajouter un critère de précision qui assure que la reconstruction soit d'ordre $k+1$: dans cette étude, on exige que la reconstruction reproduise les polynômes de degré k et on appelle de telles reconstructions *consistantes de degré k*. La discussion de la section 7.2.3 prouve que le critère de consistance de degré k garantit que la reconstruction est d'ordre $k+1$ si l'hypothèse 7.2.6 est satisfaite.

À ce critère de consistance, on ajoute encore la condition de conservativité qui stipule que la reconstruction préserve la moyenne de cellule. La conservativité est un principe particulièrement important pour les applications dans l'énergétique, mais il s'agit d'une condition indépendante de l'ordre de la reconstruction.

Le choix de la base polynomiale (7.3.2) a permis de transcrire la condition de consistance de degré k en une équation d'algèbre linéaire pour les coefficients des polynômes reconstruits. Les outils du calcul tensoriel développés dans la section 5.10 permettent ensuite d'écrire ces conditions de consistance sous la forme de l'équation matricielle (7.6.3). La structure de cette équation a permis de représenter la famille des reconstructions consistantes de degré k sous deux formes mathématiques différentes.

L'étude numérique de la section 7.10 permet de tester le taux de convergence des méthodes de reconstruction de degré $k = 2$ et $k = 3$ en maillage structuré et non structuré. En raison d'une limitation des outils de test, ces tests ont été restreints à la dimension deux. Ils confirment dans le cas $k = 2$ et $k = 3$ que la reconstruction consistante de degré k permet d'obtenir des erreurs d'approximation d'ordre $k + 1$, même en maillage non structuré.

15.1.4. Étude de méthodes de reconstruction itératives. Une partie importante de cette étude est l'analyse des méthodes itératives de reconstruction dans le chapitre 8. Le but de ces méthodes est de reconstruire des polynômes de degré élevé par des algorithmes qui n'échangent des données qu'entre cellules voisines, ce qui simplifie énormément l'implémentation des algorithmes sur ordinateur. L'étude du chapitre 8 a porté sur deux méthodes particulières, la méthode des moindres carrés couplés (MCC), introduite dans la section 8.2, et la méthode des correction successives (CS), introduite dans la section 8.3. Bien que ces méthodes permettent a priori de reconstruire les polynômes de degré k, l'implémentation et les tests se restreignent à la reconstruction de degré deux dans le cadre de la présente étude. Cela donne les algorithmes suivants :

(1) La méthode des moindres carrés couplés par itération MCCI, définie par l'algorithme 8.2.2, et la méthode des moindres carrés couplés par itération sur un voisinage élargi MCCIE, définie par l'algorithme 8.4.3.

(2) La méthode des corrections successives CS, définie par l'algorithme 8.3.8, et la méthode des corrections successives sur un voisinage élargi CSE, définie par l'algorithme 8.4.4.

L'étude numérique de la section 8.5 confirme que l'erreur d'approximation de ces méthodes de reconstruction est d'ordre trois sur des maillages de triangles et des maillages hybrides en dimension deux.

15.1.5. Étude de l'intégration des flux en maillage non structuré. L'étude de l'intégration des flux du chapitre 9 a permis de présenter deux méthodes alternatives pour calculer les flux sur les interfaces entre les cellules. La discussion présente les avantages et les inconvénients de ces deux méthodes qui ont été implémentées et testées dans Cedre.

15.1.6. Étude de la stabilité des schémas volumes finis. L'étape suivante de l'étude a consisté à analyser la stabilité des schémas numériques reposant sur les méthodes de reconstruction développées dans les chapitres 7 et 8. L'objectif était de déterminer les méthodes consistantes de reconstruction qui améliorent la stabilité des schémas. Pour pouvoir évaluer l'influence de la reconstruction sur la stabilité, il était nécessaire d'étudier les schémas de discrétisation sans limiteurs et dans le contexte semi-discret sur l'exemple de l'équation de convection linéaire.

L'étude a permis de déterminer un nouveau critère local de stabilité, l'opérateur de reconstruction locale introduit par la définition 10.5.3. Il s'agit d'un opérateur linéaire qui permet d'exprimer les fluctuations entre les valeurs reconstruites aux faces et la valeur moyenne de la cellule en fonction des fluctuations entre les moyennes de cellule au voisinage. Cet opérateur est nul dans le cas du schéma basé sur une reconstruction constante par cellule. Puisque le théorème 10.4.2 démontre que ce schéma est toujours stable indépendamment du maillage, il est naturel de choisir des reconstructions consistantes qui minimisent l'opérateur de reconstruction locale dans une ou plusieurs normes. L'étude de la section 10.6 permet alors de tirer deux conclusions importantes pour le choix et la conception des méthodes de reconstruction :

(1) Le théorème 10.6.7 montre que la reconstruction des moindres carrés minimise le critère de la définition 10.6.1 pour une large famille de normes. Ce résultat suggère que si la méthode des moindres carrés donne un schéma instable, alors toute autre méthode de reconstruction consistante est également susceptible de produire un schéma instable. Ce résultat permet de voir la méthode des moindres carrés comme une méthode de stabilité optimale mais cette interprétation n'est évidemment pas totalement rigoureuse.

(2) Le résultat du théorème 10.6.9 suggère que des voisinages de reconstruction plus grands conduisent à des schémas plus robustes, au moins dans le cas de la reconstruction des moindres carrés.

Ces conclusions se généralisent aux reconstructions de degré $k > 1$.

L'étude numérique de la section 10.9 confirme ces résultats pour les degrés $k = 1, 2, 3$: les instabilités disparaissent si le voisinage de reconstruction est élargi. De plus, l'étude numérique montre que les algorithmes 8.2.2 (MCCI) et 8.3.8 (CS) conduisent à des schémas instables et doivent être remplacés par les algorithmes 8.4.3 (MCCIE) et 8.4.4 (CSE) qui s'avèrent stables sur tous les maillages.

Un autre résultat important concerne la reconstruction des polynômes de degré trois : sur des maillages de triangles en dimension deux, ces schémas présentent des modes instables avec une partie réelle faiblement positive. Ces modes s'appellent modes instables « lents » car leur croissance est plus lente. La variation spatiale de ces modes est lisse, ce qui laisse penser que les limiteurs agiront d'une manière moins efficace sur ces modes que sur les modes instables observés avec les reconstructions de degré un et deux. L'analyse de ce phénomène nécessite en tout cas une étude plus approfondie.

15.1.7. Étude de la précision des schémas volumes finis. L'étude de la précision commence par une analyse de l'équation modifiée pour l'équation de convection linéaire en maillage non structuré général. Cette étude a fourni un développement asymptotique des premiers termes d'erreur. Il a été possible de montrer que certains termes d'erreur disparaissent en maillage structuré.

L'étude numérique a permis d'analyser les taux de convergence des schémas pour les reconstructions de degré $k = 1, 2, 3$ en dimension deux. En maillage cartésien, le taux de convergence est presque égal à $k+1$. En maillage non structuré, le taux de convergence se dégrade et se situe entre k et $k + 1$ pour les cas $k = 1$ et $k = 2$. Pour la reconstruction des polynômes de degré $k = 3$ en maillage triangulaire, le taux de convergence ne dépasse pas celui de la reconstruction quadratique. Ce comportement nécessite une étude plus approfondie. Il est également nécessaire d'analyser les taux de convergence en dimension trois, ce qui n'a pas encore été fait en raison d'une limitation des outils de test.

15.1.8. Reconstruction monotone en maillage non structuré. L'étude des limiteurs dans le chapitre 12 part d'un critère de monotonie donné par le principe du maximum de Barth [**8, 10**]. L'étude met l'accent sur le développement d'algorithmes pour la limitation directionnelle. Ces algorithmes reposent sur une résolution approchée et rapide du problème géométrique de la limitation du gradient en maillage non structuré. Deux algorithmes ont été implémentés dans Cedre. Cependant, il faudrait poursuivre l'étude pour trouver un critère de monotonie moins dissipatif que le principe du maximum.

15.1.9. Implémentation des méthodes dans CEDRE. La thèse a permis d'implémenter un certain nombre d'évolutions dans Cedre :

- reconstruction stable du gradient par l'algorithme 8.4.1 ;
- formules de quadrature pour les flux numériques, adaptées à la reconstruction de degré trois. Ceci a d'ores et déjà permis d'améliorer les résultats de calcul en référentiel tournant ;
- algorithmes de limitation 12.4.2 et 12.4.4 élaborées dans le chapitre 12 ;
- structures de données en vue de l'implémentation des reconstructions de degré deux ;

L'implémentation de la reconstruction stable du gradient par l'algorithme 8.4.1 a permis de stabiliser les calculs sur tétraèdres sans les limiteurs. La retombée concrète a été de rendre possible certains calculs LES en maillage tétraédrique.

15.1.10. Calculs LES. Certaines évolutions citées ci-dessus ont été validées sur deux calculs LES en maillage non structuré. Il s'agit de calculs qui avaient déjà fait l'objet de simulations numériques sur maillage non structuré à l'ONERA et qui avaient mis en évidence certains problèmes de discrétisation spatiale.

(1) Le premier calcul est la simulation d'un écoulement subsonique au-dessus d'une cavité profonde avec le modèle de Smagorinsky. Cette configuration a fait l'objet d'une étude expérimentale à l'ONERA [**50**]. La difficulté était d'obtenir des spectres acoustiques et des profils de vitesse et de tenseur de Reynolds proches des données expérimentales. Jusqu'à présent seuls des calculs en maillage hexaédrique donnaient des résultats satisfaisants, les calculs en maillage tétraédrique étant soit instables en l'absence de limiteurs de gradient, soit trop dissipatifs avec limiteurs. Les nouveaux résultats en maillage tétraédrique ont permis d'accéder à une qualité comparable à ceux des autres maillages.

(2) Le second calcul est la simulation d'un jet chaud hypersonique qui avait fait l'objet d'une étude expérimentale de Seiner et Ponton [**93**]. La difficulté était d'obtenir la transition de l'écoulement vers un état turbulent en maillage tétraédrique. La possibilité de faire certains calculs sans limiteurs grâce à la reconstruction stable permet d'abord de démontrer que ce sont les limiteurs de gradient qui empêchent la transition vers l'état turbulent. Le calcul sans limiteur a ensuite permis d'obtenir une condition initiale turbulente pour le calcul du jet. Si le calcul avec les limiteurs part de cette condition initiale, le résultat final est nettement meilleur que les résultats obtenus jusqu'ici en maillage de tétraèdres.

15.2. Perspectives

Cette étude a permis d'analyser et clarifier plusieurs aspects de la discrétisation spatiale par la méthode des volumes finis. Cependant, l'étude a également soulevé plusieurs points qui nécessitent encore des analyses plus approfondies.

15.2.1. Stabilité. L'étude a mis en évidence une relation entre la reconstruction et la stabilité linéaire des schémas volumes finis dans l'approche MUSCL. Le principal point qui reste à analyser est le comportement des schémas basés sur des reconstructions de degré trois, c'est-à-dire les schémas qui sont formellement d'ordre quatre. Une question importante est notamment l'apparition de modes instables particuliers dont la suppression pourrait nécessiter une dissipation numérique artificielle.

15.2.2. Précision. Les études numériques des sections 7.10, 8.5 et 11.3 ont permis d'analyser les taux de convergence des schémas en dimension deux. Il est indispensable d'effectuer les mêmes tests numériques en dimension trois. En général, la discrétisation spatiale en dimension trois semble plus difficile qu'en dimension deux et un bon comportement des schémas en dimension deux ne garantit donc pas l'absence de problèmes en dimension trois.

15.2.3. Limiteurs. Il semble très important de chercher des critères de monotonie moins dissipatifs que le principe du maximum. L'introduction de la limitation directionnelle donne de bons résultats mais il est encore possible de les améliorer.

15.2.4. Perspectives pour CEDRE. La perspective à court terme pour CEDRE est l'implémentation de la méthode des corrections successives CSE et de la méthode des moindres carrés MCCIE pour la reconstruction de degré deux. Il faut ensuite continuer le travail de recherche sur les algorithmes de limitation et implémenter des méthodes de limitation encore moins dissipatives dans CEDRE.

À moyen terme, on peut envisager l'implémentation de reconstructions de degré trois si les questions autour du taux de convergence et des modes instables « lents » sont résolues.

ANNEXE A

Étude détaillée de la reconstruction des polynômes de degré un, deux et trois

A.1. Objectif du chapitre

Ce chapitre présente en détail la reconstruction des polynômes de degré un, deux et trois, en appliquant les résultats généraux du chapitre 7 aux cas particuliers $k = 1, 2, 3$. De façon à rendre indépendantes chacune des sections A.2, A.3 et A.4, les différentes sections reprennent cas par cas, pour la commodité du lecteur, la dérivation des équations traduisant la reconstruction locale des polynômes.

A.2. Étude des reconstructions linéaires

Cette section est dédiée à l'étude détaillée de la reconstruction des polynômes de degré un.

A.2.1. Établissement des conditions de reconstruction. Commençons par le cas important des reconstructions linéaires, c'est-à-dire le cas où le degré des polynômes reconstruits $w_\alpha[\mathfrak{u}]$ est égal à un. Les fonctions reconstruites (7.3.6) sont, dans ce cas, de la forme générale

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \sum_\beta \left[w^{\alpha\beta} + \sum_{j=1}^{d} w_j^{\alpha\beta} (x_j - x_{\alpha,j}) \right] \overline{u}_\beta \,. \tag{A.2.1}$$

Les composantes $w_j^{\alpha\beta}$ forment des vecteurs $\boldsymbol{w}_{\alpha\beta}^{(1)}$ et la définition $w_{\alpha\beta}^{(0)} \triangleq w^{\alpha\beta}$ permet de garder une notation uniforme. Cela permet d'écrire (A.2.1) sous la forme simple

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \sum_\beta \left[w_{\alpha\beta}^{(0)} + \boldsymbol{w}_{\alpha\beta}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) \right] \overline{u}_\beta \,. \tag{A.2.2}$$

Pour déterminer une équation générale pour les coefficients, il est nécessaire d'appliquer les conditions de consistance (7.3.8). Si v est un polynôme de degré un, la forme la plus générale de v est

$$v(\boldsymbol{x}) = v(\boldsymbol{x}_\alpha) + \sum_{j=1}^{d} v_j (x_j - x_{\alpha,j}) = v^{(0)} + \boldsymbol{v}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha)$$

où le vecteur $\boldsymbol{v}^{(1)}$ est de composantes v_j et $v^{(0)} \triangleq v_0$ par souci d'homogénéité de la notation. La moyenne $\overline{v}_\gamma$ de v sur la cellule $\mathcal{T}_\gamma$ est donnée par

$$\overline{v}_\gamma = v^{(0)} + \boldsymbol{v}^{(1)} \cdot \boldsymbol{h}_{\alpha\gamma} \tag{A.2.3}$$

où $\boldsymbol{h}_{\alpha\gamma} \triangleq \boldsymbol{x}_\gamma - \boldsymbol{x}_\alpha$, cf. (5.3.6). La moyenne (A.2.3) de v coïncide avec la valeur $v(\boldsymbol{x}_\gamma)$ de v au barycentre $\boldsymbol{x}_\gamma$ de $\mathcal{T}_\gamma$, cf. la remarque 7.3.3.

Soit $w_\alpha[\mathfrak{v}]$ le polynôme reconstruit à partir des moyennes $\mathfrak{v} \triangleq (\overline{v}_1, \ldots, \overline{v}_N)$. La reproduction des polynômes de degré un s'exprime par la condition que le polynôme $w_\alpha[\mathfrak{v}]$ doit être égal à v

$$v^{(0)} + \boldsymbol{v}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) = \sum_\beta \left[w_{\alpha\beta}^{(0)} + \boldsymbol{w}_{\alpha\beta}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) \right] \left(v^{(0)} + \boldsymbol{v}^{(1)} \cdot \boldsymbol{h}_{\alpha\beta} \right) . \tag{A.2.4}$$

La comparaison des coefficients dans le polynôme (A.2.4) donne les équations suivantes pour les $w_{\alpha\beta}^{(0)}$

$$\sum_{\beta} w_{\alpha\beta}^{(0)} = 1 \tag{A.2.5}$$

$$\sum_{\beta} w_{\alpha\beta}^{(0)} \boldsymbol{h}_{\alpha\beta} = 0\,. \tag{A.2.6}$$

Les équations pour les coefficients $\boldsymbol{w}_{\alpha\beta}^{(1)}$ sont

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} = 0 \tag{A.2.7}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \boldsymbol{h}_{\alpha\beta} = \boldsymbol{\delta}^{(2)}\,. \tag{A.2.8}$$

Le tenseur $\boldsymbol{\delta}^{(2)}$, défini par (5.2.22), est le symbole de Kronecker

$$\left(\boldsymbol{\delta}^{(2)}\right)_{ij} = \delta_{ij}\,.$$

Les deux équations (A.2.7) et (A.2.8) s'écrivent composante par composante comme

$$\sum_{\beta} w_i^{\alpha\beta} = 0\,,\ 1 \leq i \leq d$$

$$\sum_{\beta} w_i^{\alpha\beta} h_j^{\alpha\beta} = \delta_{ij}\,,\ 1 \leq i,j \leq d\,.$$

Ces conditions impliquent que les coefficients $\boldsymbol{w}_{\alpha\beta}^{(1)}$ reproduisent le gradient de tout polynôme de degré un

$$\boldsymbol{v}^{(1)} = \sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} \left(v^{(0)} + \boldsymbol{v}^{(1)} \cdot \boldsymbol{h}_{\alpha\beta}\right)\,.$$

La prochaine étape est d'imposer la contrainte sur la moyenne (6.3.8)

$$\frac{1}{|\mathcal{T}_\alpha|} \int_{\mathcal{T}_\alpha} w_\alpha\left[\mathfrak{v}\right](\boldsymbol{x})\ dx = \overline{v}_\alpha$$

au polynôme (A.2.2). Ceci donne l'équation supplémentaire

$$\overline{v}_\alpha = \sum_{\beta} w_{\alpha\beta}^{(0)} \overline{v}_\beta \tag{A.2.9}$$

car la moyenne de $\boldsymbol{x} - \boldsymbol{x}_\alpha$ sur la cellule $\mathcal{T}_\alpha$ est nulle. Comme la relation (A.2.9) doit être vraie indépendamment des valeurs $\mathfrak{v} = (\overline{v}_1, \ldots, \overline{v}_N)$, il s'ensuit que

$$w_{\alpha\beta}^{(0)} = \delta_{\alpha\beta}\,. \tag{A.2.10}$$

Il s'avère que la contrainte de la moyenne (A.2.9) détermine complètement les coefficients $w_{\alpha\beta}^{(0)}$. Il résulte de la proposition 7.5.3 que les coefficients (A.2.10) satisfont les deux équations (A.2.5) et (A.2.6). En effet, l'insertion de (A.2.10) dans (A.2.5) donne

$$\sum_{\beta} w_{\alpha\beta}^{(0)} = \sum_{\beta} \delta_{\alpha\beta} = 1$$

et l'insertion de (A.2.10) dans (A.2.6) montre que

$$\sum_{\beta} w_{\alpha\beta}^{(0)} \boldsymbol{h}_{\alpha\beta} = \sum_{\beta} \delta_{\alpha\beta} \boldsymbol{h}_{\alpha\beta} = \boldsymbol{h}_{\alpha\alpha} = 0\,.$$

L'utilisation de (A.2.10) dans (A.2.2) donne la forme générale de la reconstruction conservative de degré un

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \overline{u}_\alpha + \sum_\beta \boldsymbol{w}^{(1)}_{\alpha\beta} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha)\, \overline{u}_\beta\,. \tag{A.2.11}$$

La définition du gradient de (A.2.11)

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] \triangleq \sum_\beta \boldsymbol{w}^{(1)}_{\alpha\beta} \overline{u}_\beta$$

permet d'écrire (A.2.11) de façon plus compacte

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \overline{u}_\alpha + \boldsymbol{\sigma}_\alpha[\mathfrak{u}] \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha)\ ,\ \boldsymbol{x} \in \mathcal{T}_\alpha\,, \tag{A.2.12}$$

ce qui montre que la contrainte de la moyenne n'augmente pas seulement la précision de la reconstruction mais simplifie aussi sa forme.

Pour les raisons énoncées ci-dessus, l'étude se concentre désormais sur la reconstruction du gradient à partir des moyennes de cellule $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$. C'est pourquoi nous introduisons la notation suivante. Pour les besoins de reconstruction d'un gradient, les coefficients $\boldsymbol{w}^{(1)}_{\alpha\beta}$ s'appellent désormais $\boldsymbol{\sigma}_{\alpha\beta} \triangleq \boldsymbol{w}^{(1)}_{\alpha\beta}$

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \overline{u}_\beta\,. \tag{A.2.13}$$

Rappelons que $\boldsymbol{\sigma}_{\alpha\beta} \triangleq 0$ par définition dans (A.2.13) si la cellule $\mathcal{T}_\beta$ n'est pas dans le voisinage de reconstruction de la cellule $\mathcal{T}_\alpha$. La relation (A.2.7) permet d'éliminer le coefficient $\boldsymbol{\sigma}_{\alpha\alpha}$ dans (A.2.13) car

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta} u_\alpha = 0\,. \tag{A.2.14}$$

Il y a donc deux façons équivalentes d'écrire (A.2.13)

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} u_\beta = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}(u_\beta - u_\alpha)\ . \tag{A.2.15}$$

La première forme de (A.2.15) contient $\boldsymbol{\sigma}_{\alpha\alpha}$ mais pas la deuxième.

Le coefficient $\boldsymbol{\sigma}_{\alpha\alpha}$ n'apparaît pas dans la condition (A.2.8) qui s'écrit

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \otimes (\boldsymbol{x}_\beta - \boldsymbol{x}_\alpha) = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} = \boldsymbol{\delta}^{(2)} \tag{A.2.16}$$

car $\boldsymbol{h}_{\alpha\alpha} = \boldsymbol{x}_\alpha - \boldsymbol{x}_\alpha = 0$. La condition (A.2.16) sera désormais appelée *condition de consistance pour la reconstruction du gradient*. Elle implique que

$$\boldsymbol{\sigma} = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\ (\boldsymbol{h}_{\alpha\beta} \cdot \boldsymbol{\sigma}) \text{ pour tous les } \boldsymbol{\sigma} \in \mathbb{R}^d \tag{A.2.17}$$

ce qui signifie que les coefficients $\boldsymbol{\sigma}_{\alpha\beta}$ reproduisent le gradient de tout polynôme de degré un.

A.2.2. Calcul de l'erreur d'approximation. La condition (A.2.16) assure la reconstruction exacte de tout polynôme de degré un. Pour calculer le premier terme d'erreur sur la dérivée première d'une fonction u de classe C^3, il suffit d'insérer les coefficients $\boldsymbol{\sigma}_{\alpha\beta}$ et l'identité $\boldsymbol{z}^{(2)}_{\alpha\beta} = \boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta$ dans (7.4.1)

$$\begin{aligned}\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}(\overline{u}_\beta - \overline{u}_\alpha) = \\ = D^{(1)}u\Big|_{\boldsymbol{x}_\alpha} + \frac{1}{2!}\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\left\{ D^{(2)}u\Big|_{\boldsymbol{x}_\beta} \bullet \left[\boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta\right] + \mathrm{O}\left(h^3\right)\right\}\,.\end{aligned} \tag{A.2.18}$$

La relation (A.2.18) montre que l'erreur d'approximation du gradient est d'ordre un à condition que $\|\boldsymbol{\sigma}_{\alpha\beta}\| = \mathrm{O}\left(h^{-1}\right)$ car l'expression $\boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta$ est $\mathrm{O}\left(h^2\right)$.

A.2.3. Formulation matricielle. La prochaine étape est de mettre (A.2.16) sous une forme plus compacte. Soit m_α le nombre de cellules dans le voisinage de reconstruction de cellule α et $\mathbb{W}_\alpha \triangleq \{\beta_1, \beta_2, \ldots, \beta_{m_\alpha}\}$ les indices des cellules dans ce voisinage. Dans la cellule $\mathcal{T}_\alpha$, les vecteurs inconnus $\boldsymbol{\sigma}_{\alpha\beta}$ forment les colonnes d'une $d \times m_\alpha$ matrice S_α. De façon similaire, les vecteurs géométriques $\boldsymbol{h}_{\alpha\beta}$ forment les lignes d'une $m_\alpha \times d$ matrice H_α

$$H_\alpha^t = [\boldsymbol{h}_{\alpha\beta_1}, \boldsymbol{h}_{\alpha\beta_2}, \ldots, \boldsymbol{h}_{\alpha\beta_m}] \tag{A.2.19}$$
$$S_\alpha = [\boldsymbol{\sigma}_{\alpha\beta_1}, \boldsymbol{\sigma}_{\alpha\beta_2}, \ldots, \boldsymbol{\sigma}_{\alpha\beta_m}] \,. \tag{A.2.20}$$

Ces définitions permettent en effet d'écrire l'équation (A.2.16) comme une équation matricielle pour l'inconnue S_α

$$S_\alpha H_\alpha = \mathrm{I}_d. \tag{A.2.21}$$

L'équation (A.2.21) montre que la matrice S_α doit être un inverse à gauche de la matrice H_α. Pour que (A.2.21) admette des solutions, il faut et il suffit que la matrice H_α ait un rang égal à d. Cette condition évoquée dans la section 5.3 est satisfaite sur les maillages utilisés dans la pratique. Une condition nécessaire pour que le rang de H_α soit égal à d est que $m_\alpha \geq d$. Comme le nombre de premiers voisins du polygone le plus simple, le simplexe, est égal à $d+1$, cette condition nécessaire est, dans le cas des polynômes de degré un, toujours satisfaite.

A.3. Étude des reconstructions quadratiques

Cette section est dédiée à l'étude détaillée de la reconstruction des polynômes de degré deux.

A.3.1. Établissement des conditions de reconstruction. L'objectif de cette section est d'examiner la reconstruction par des polynômes de degré deux. A nouveau, on reprend de façon détaillée la totalité des calculs. On choisit une base constituée de polynômes centrés sur le barycentre $\boldsymbol{x}_\alpha$ de la cellule $\mathcal{T}_\alpha$. Les fonctions reconstruites (7.3.6) sont dans ce cas de la forme générale

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \sum_\beta \left[w^{\alpha\beta} + \sum_{i=1}^d w_i^{\alpha\beta}(x_i - x_{\alpha,i}) + \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d w_{ij}^{\alpha\beta}(x_i - x_{\alpha,i})(x_j - x_{\alpha,j}) \right] \overline{u}_\beta \,. \tag{A.3.1}$$

Les coefficients $w_j^{\alpha\beta}$ forment un vecteur $\boldsymbol{w}_{\alpha\beta}^{(1)}$ et les coefficients $w_{ij}^{\alpha\beta}$ constituent un tenseur symétrique d'ordre deux appelé $\boldsymbol{w}_{\alpha\beta}^{(2)}$. Pour unifier la notation, on pose $w_{\alpha\beta}^{(0)} \triangleq w^{\alpha\beta}$, ce qui permet de réécrire (A.3.1) comme

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \sum_\beta \left[w_{\alpha\beta}^{(0)} + \boldsymbol{w}_{\alpha\beta}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\alpha)^t \cdot \boldsymbol{w}_{\alpha\beta}^{(2)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) \right] \overline{u}_\beta \,. \tag{A.3.2}$$

Pour déterminer une équation générale pour les coefficients, il est de nouveau nécessaire d'appliquer les contraintes de consistance. Si v est un polynôme de degré deux, v s'écrit dans la forme la plus générale

$$v(\boldsymbol{x}) = v + \sum_{i=1}^d v_i(x_i - x_{\alpha,i}) + \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d v_{ij}(x_i - x_{\alpha,i})(x_j - x_{\alpha,j}) =$$
$$= v^{(0)} + \boldsymbol{v}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\alpha)^t \cdot \boldsymbol{v}^{(2)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) \,.$$

La formule (5.8.4) et la définition des tenseurs $\boldsymbol{z}_{\alpha\beta}^{(j)}$ par (5.7.10) permettent d'écrire la moyenne de v sur la cellule $\mathcal{T}_\gamma$ comme

$$\overline{v}_\gamma = v^{(0)} + \boldsymbol{v}^{(1)} \cdot \boldsymbol{h}_{\alpha\gamma} + \frac{1}{2}\boldsymbol{v}^{(2)} \bullet \boldsymbol{z}_{\alpha\gamma}^{(2)} \,. \tag{A.3.3}$$

La relation (A.3.3) montre qu'en général la moyenne d'une fonction polynomiale de degré deux sur une cellule n'est pas égale à sa valeur au barycentre de la cellule. Il y a donc une différence entre les reconstructions à partir de moyennes de cellule et à partir de valeurs ponctuelles. La différence entre la moyenne et la valeur au barycentre est donnée par

$$\overline{v}_\gamma - v\left(\boldsymbol{x}_\gamma\right) = \frac{1}{2}\boldsymbol{v}^{(2)} \bullet \boldsymbol{x}_\gamma^{(2)} = \mathrm{O}\left(h^2\right) .$$

L'insertion des moyennes $\mathfrak{v} = (\overline{v}_1, \ldots, \overline{v}_N)$ dans la formule (A.3.2) pour $w_\alpha[\mathfrak{v}]$ et l'application du critère de consistance de degré deux se traduisent par l'identité

$$\begin{aligned} v^{(0)} + \boldsymbol{v}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\alpha)^t \cdot \boldsymbol{v}^{(2)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) = \\ = \sum_\beta \left(w_{\alpha\beta}^{(0)} + \boldsymbol{w}_{\alpha\beta}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\alpha)^t \cdot \boldsymbol{w}_{\alpha\beta}^{(2)} \cdot (\boldsymbol{x} - \boldsymbol{x}_\alpha) \right) \times \\ \times \left(v^{(0)} + \boldsymbol{v}^{(1)} \cdot \boldsymbol{h}_{\alpha\beta} + \frac{1}{2}\boldsymbol{h}_{\alpha\beta}^t \cdot \boldsymbol{v}^{(2)} \cdot \boldsymbol{h}_{\alpha\beta} + \frac{1}{2}\boldsymbol{v}^{(2)} \bullet \boldsymbol{x}_\beta^{(2)} \right) . \end{aligned} \tag{A.3.4}$$

L'équation (A.3.4) doit être satisfaite pour toutes les valeurs de $v^{(0)}$, $\boldsymbol{v}^{(1)}$ et $\boldsymbol{v}^{(2)}$ et permet donc d'établir les conditions pour les coefficients inconnus $w_{\alpha\beta}^{(0)}$, $\boldsymbol{w}_{\alpha\beta}^{(1)}$ et $\boldsymbol{w}_{\alpha\beta}^{(2)}$. Ces conditions s'obtiennent par identification des termes linéairement indépendants dans (A.3.4). Il s'avère d'ailleurs avantageux de reformuler certaines conditions afin d'en éliminer les coefficients $\boldsymbol{w}_{\alpha\alpha}^{(1)}$ et $\boldsymbol{w}_{\alpha\alpha}^{(2)}$.

Les premières conditions concernent les coefficients $w_{\alpha\beta}^{(0)}$

$$\sum_\beta w_{\alpha\beta}^{(0)} = 1 \tag{A.3.5}$$

$$\sum_\beta w_{\alpha\beta}^{(0)} \boldsymbol{h}_{\alpha\beta} = 0 \tag{A.3.6}$$

$$\sum_\beta w_{\alpha\beta}^{(0)} \frac{1}{2}\left(\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}_\beta^{(2)} \right) = 0 . \tag{A.3.7}$$

Les équations pour les $\boldsymbol{w}_{\alpha\beta}^{(1)}$ sont :

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(1)} = 0 \tag{A.3.8}$$

$$\sum_\beta \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \boldsymbol{h}_{\alpha\beta} = \boldsymbol{\delta}^{(2)} \tag{A.3.9}$$

$$\begin{aligned} \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \frac{1}{2}\left(\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}_\beta^{(2)} \right) = \\ = \sum_\beta \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \frac{1}{2}\left(\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}_\beta^{(2)} - \boldsymbol{x}_\alpha^{(2)} \right) = 0 . \end{aligned} \tag{A.3.10}$$

Dans (A.3.10), la soustraction du produit tensoriel de (A.3.8) et $\boldsymbol{x}_\alpha^{(2)}$ a permis d'éliminer le terme contenant $\boldsymbol{w}_{\alpha\alpha}^{(1)}$. Le tenseur $\boldsymbol{\delta}^{(2)}$ défini par (5.2.22) est de composantes

$$\left(\boldsymbol{\delta}^{(2)}\right)_{ij} = \delta_{ij} .$$

Finalement, les équations pour les $\boldsymbol{w}_{\alpha\beta}^{(2)}$ s'expriment par

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} = 0 \tag{A.3.11}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} \otimes \boldsymbol{h}_{\alpha\beta} = 0 \tag{A.3.12}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} \otimes \frac{1}{2}\left(\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}_{\beta}^{(2)}\right) =$$

$$= \sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} \otimes \frac{1}{2}\left(\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}_{\beta}^{(2)} - \boldsymbol{x}_{\alpha}^{(2)}\right) = \boldsymbol{\delta}^{(4)}\,. \tag{A.3.13}$$

Dans (A.3.13), la soustraction du produit tensoriel de (A.3.11) et $\boldsymbol{x}_{\alpha}^{(2)}$ a permis de supprimer le terme contenant $\boldsymbol{w}_{\alpha\alpha}^{(2)}$. Le tenseur $\boldsymbol{\delta}^{(4)}$ défini par (5.2.22) est de composantes

$$\left(\boldsymbol{\delta}^{(4)}\right)_{ijkl} = \frac{1}{2}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)\,.$$

La condition (A.3.13) s'écrit de façon explicite

$$\sum_{\beta} w_{ij}^{\alpha\beta} \frac{1}{2}\left[h_{k}^{\alpha\beta} h_{l}^{\alpha\beta} + x_{\beta,kl}\right] = \frac{1}{2}\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}$$

où $1 \leq i, j, k, l \leq d$.

La prochaine étape est d'appliquer la contrainte de conservation (6.3.8)

$$\frac{1}{|\mathcal{T}_{\alpha}|}\int_{\mathcal{T}_{\alpha}} w_{\alpha}\left[\mathfrak{u}\right](\boldsymbol{x})\; dx = \overline{u}_{\alpha}$$

qui doit être satisfaite *même si u est une fonction arbitraire*. Ceci conduit à l'équation supplémentaire

$$\overline{u}_{\alpha} = \sum_{\beta}\left[w_{\alpha\beta}^{(0)} + \frac{1}{2}\boldsymbol{w}_{\alpha\beta}^{(2)} \bullet \boldsymbol{x}_{\alpha}^{(2)}\right]\overline{u}_{\beta}\,. \tag{A.3.14}$$

Puisque (A.3.14) doit être vraie indépendamment des valeurs $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$, il vient que

$$w_{\alpha\beta}^{(0)} = \delta_{\alpha\beta} - \frac{1}{2}\boldsymbol{w}_{\alpha\beta}^{(2)} \bullet \boldsymbol{x}_{\alpha}^{(2)} \tag{A.3.15}$$

La contrainte de la moyenne (A.3.14) détermine donc les coefficients $w_{\alpha\beta}^{(0)}$ en fonction des coefficients $\boldsymbol{w}_{\alpha\beta}^{(2)}$ et permet d'éliminer les conditions (A.3.5), (A.3.6) et (A.3.7).

Pour éviter toute contradiction, il est cependant nécessaire de vérifier que les coefficients $w_{\alpha\beta}^{(0)}$ spécifiés par (A.3.15) satisfont (A.3.5), (A.3.6) et (A.3.7). La condition (A.3.5) est satisfaite car

$$\sum_{\beta} w_{\alpha\beta}^{(0)} = 1 - \sum_{\beta}\frac{1}{2}\boldsymbol{w}_{\alpha\beta}^{(2)} \bullet \boldsymbol{x}_{\alpha}^{(2)} = 1$$

en vertu de (A.3.11). Ensuite, (A.3.12) et $\boldsymbol{h}_{\alpha\alpha} = \boldsymbol{x}_{\alpha} - \boldsymbol{x}_{\alpha} = 0$ entraînent

$$\sum_{\beta} w_{\alpha\beta}^{(0)}\boldsymbol{h}_{\alpha\beta} = \boldsymbol{h}_{\alpha\alpha} - \sum_{\beta}\frac{1}{2}\left(\boldsymbol{w}_{\alpha\beta}^{(2)} \bullet \boldsymbol{x}_{\alpha}^{(2)}\right)\boldsymbol{h}_{\alpha\beta} = 0$$

ce qui montre que (A.3.6) est vrai. Finalement, en raison de (A.3.13) et de $\boldsymbol{h}_{\alpha\alpha} = \boldsymbol{x}_{\alpha} - \boldsymbol{x}_{\alpha} = 0$ il vient

$$\sum_{\beta} w_{\alpha\beta}^{(0)}\left(\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}_{\beta}^{(2)}\right) =$$

$$= \boldsymbol{h}_{\alpha\alpha} \otimes \boldsymbol{h}_{\alpha\alpha} + \boldsymbol{x}_{\alpha}^{(2)} - \sum_{\beta}\frac{1}{2}\left(\boldsymbol{w}_{\alpha\beta}^{(2)} \bullet \boldsymbol{x}_{\alpha}^{(2)}\right)\left(\boldsymbol{h}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}_{\beta}^{(2)}\right) = \boldsymbol{x}_{\alpha}^{(2)} - \boldsymbol{x}_{\alpha}^{(2)} = 0\,,$$

ce qui prouve (A.3.7).

L'insertion de (A.3.15) dans (A.3.2) et les notations (5.2.16) et (5.2.27) simplifient la forme de (A.3.2)

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = \overline{u}_\alpha + \sum_\beta \boldsymbol{w}^{(1)}_{\alpha\beta}\cdot(\boldsymbol{x}-\boldsymbol{x}_\alpha)\,\overline{u}_\beta + \frac{1}{2}\sum_\beta \boldsymbol{w}^{(2)}_{\alpha\beta}\bullet\left[(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 - \boldsymbol{x}^{(2)}_\alpha\right]\overline{u}_\beta\,. \tag{A.3.16}$$

Le gradient et la dérivée seconde du polynôme (A.3.16) en $\boldsymbol{x}_\alpha$ sont donnés par

$$\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{w}^{(1)}_{\alpha\beta}\overline{u}_\beta \tag{A.3.17}$$

$$\boldsymbol{\theta}_\alpha\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{w}^{(2)}_{\alpha\beta}\overline{u}_\beta\,. \tag{A.3.18}$$

La dérivée seconde (A.3.18) est un tenseur symétrique d'ordre deux. L'indication du rang a été omise pour simplifier la notation. Le polynôme (A.3.16) s'écrit finalement sous forme compacte

$$w_\alpha\left[\mathfrak{u}\right](\boldsymbol{x}) = u_\alpha + \boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right]\cdot(\boldsymbol{x}-\boldsymbol{x}_\alpha) + \frac{1}{2}\boldsymbol{\theta}_\alpha\left[\mathfrak{u}\right]\bullet\left[(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 - \boldsymbol{x}^{(2)}_\alpha\right]. \tag{A.3.19}$$

Il est utile de renommer les coefficients dans (A.3.17) et (A.3.18)

$$\begin{aligned}\boldsymbol{\sigma}_{\alpha\beta} &\triangleq \boldsymbol{w}^{(1)}_{\alpha\beta}\\ \boldsymbol{\theta}_{\alpha\beta} &\triangleq \boldsymbol{w}^{(2)}_{\alpha\beta}\end{aligned}$$

afin d'indiquer qu'il s'agit des coefficients pour le gradient et la dérivée seconde.

Les équations (A.3.8) et (A.3.11) servent uniquement à déterminer les coefficients $\boldsymbol{\sigma}_{\alpha\alpha}$ et $\boldsymbol{\theta}_{\alpha\alpha}$

$$\begin{aligned}\boldsymbol{\sigma}_{\alpha\alpha} &= -\sum_{\beta\neq\alpha}\boldsymbol{\sigma}_{\alpha\beta}\\ \boldsymbol{\theta}_{\alpha\alpha} &= -\sum_{\beta\neq\alpha}\boldsymbol{\theta}_{\alpha\beta}\end{aligned}$$

en fonction des autres coefficients. Ces relations permettent d'écrire le gradient et la dérivée seconde de deux façons différentes

$$\boldsymbol{\sigma}_\alpha\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}u_\beta = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\left(u_\beta - u_\alpha\right) \tag{A.3.20}$$

$$\boldsymbol{\theta}_\alpha\left[\mathfrak{u}\right] = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}u_\beta = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}\left(u_\beta - u_\alpha\right). \tag{A.3.21}$$

La deuxième forme du gradient et de la dérivée seconde ne contient ni $\boldsymbol{\sigma}_{\alpha\alpha}$ ni $\boldsymbol{\theta}_{\alpha\alpha}$.

Les coefficients restants sont données par les quatre équations

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\otimes\boldsymbol{h}_{\alpha\beta} = \boldsymbol{\delta}^{(2)} \tag{A.3.22}$$

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\otimes\frac{1}{2}\left(\boldsymbol{h}_{\alpha\beta}\otimes\boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta - \boldsymbol{x}^{(2)}_\alpha\right) = 0 \tag{A.3.23}$$

$$\sum_\beta \boldsymbol{\theta}_{\alpha\beta}\otimes\boldsymbol{h}_{\alpha\beta} = 0 \tag{A.3.24}$$

$$\sum_\beta \boldsymbol{\theta}_{\alpha\beta}\otimes\frac{1}{2}\left(\boldsymbol{h}_{\alpha\beta}\otimes\boldsymbol{h}_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta - \boldsymbol{x}^{(2)}_\alpha\right) = \boldsymbol{\delta}^{(4)}\,. \tag{A.3.25}$$

Remarque A.3.1. Les quatre équations (A.3.22-A.3.25) montrent qu'il n'y a aucune différence entre la reconstruction à partir de moyennes de cellule et à partir de valeurs ponctuelles *si toutes les cellules ont le même moment d'ordre deux* $\boldsymbol{x}^{(2)}_\beta$. Ceci est en particulier le cas des maillages réguliers.

A.3.2. Calcul de l'erreur d'approximation. Si p est un polynôme de degré deux, les équations (A.3.22) et (A.3.23) montrent que les paramètres $\boldsymbol{\sigma}_{\alpha\beta}$ reconstruisent le gradient de p en $\boldsymbol{x}_\alpha$. Les conditions (A.3.24) et (A.3.25) signifient que les paramètres $\boldsymbol{\theta}_{\alpha\beta}$ reconstruisent la dérivée seconde de p en $\boldsymbol{x}_\alpha$. Si v est une fonction générale mais suffisamment régulière, la formule (5.8.3) permet d'écrire

$$\overline{v}_\beta - \overline{v}_\alpha = D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} \cdot \boldsymbol{h}_{\alpha\beta} + \frac{1}{2!}\, D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} \bullet \left[(\boldsymbol{h}_{\alpha\beta})^2 + \boldsymbol{x}_\beta^{(2)}\right] + \frac{1}{3!}\, D^{(3)}v\Big|_{\boldsymbol{x}_\alpha} \bullet \boldsymbol{z}_{\alpha\beta}^{(3)} + \mathrm{O}\left(h^4\right) . \quad \text{(A.3.26)}$$

L'insertion de (A.3.26) dans (A.3.20) et les équations (A.3.22) et (A.3.23) entraînent que

$$\boldsymbol{\sigma}_\alpha[\mathfrak{p}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\left(\overline{v}_\beta - \overline{v}_\alpha\right) = D^{(1)}v\Big|_{\boldsymbol{x}_\alpha} + \frac{1}{3!}\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\left\{ D^{(3)}v\Big|_{\boldsymbol{x}_\alpha} \bullet \left[\boldsymbol{z}_{\alpha\beta}^{(3)} + \mathrm{O}\left(h^4\right)\right]\right\} . \quad \text{(A.3.27)}$$

L'expression $\boldsymbol{z}_{\alpha\beta}^{(3)}$ dans (A.3.27) est $\mathrm{O}\left(h^3\right)$. Le premier terme d'erreur dans (A.3.27) est $\mathrm{O}\left(h^2\right)$ à condition que $\|\boldsymbol{\sigma}_{\alpha\beta}\| = \mathrm{O}\left(h^{-1}\right)$.

L'insertion de (A.3.26) dans (A.3.21) et les équations (A.3.24) et (A.3.25) impliquent que

$$\boldsymbol{\theta}_\alpha[\mathfrak{p}] = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}\left(\overline{v}_\beta - \overline{v}_\alpha\right) = D^{(2)}v\Big|_{\boldsymbol{x}_\alpha} + \frac{1}{3!}\sum_\beta \boldsymbol{\theta}_{\alpha\beta}\left\{ D^{(3)}v\Big|_{\boldsymbol{x}_\alpha} \bullet \left[\boldsymbol{z}_{\alpha\beta}^{(3)} + \mathrm{O}\left(h^4\right)\right]\right\} . \quad \text{(A.3.28)}$$

L'erreur dans (A.3.28) est $\mathrm{O}\left(h\right)$ à condition que $\|\boldsymbol{\theta}_{\alpha\beta}\| = \mathrm{O}\left(h^{-2}\right)$.

A.3.3. Formulation matricielle. Afin de mettre les équations (A.3.22-A.3.25) sous une forme d'équation matricielle, il est nécessaire de les écrire de façon explicite afin de déterminer les inconnues et les équations indépendantes. Ceci donne :

$$\sum_\beta \sigma_i^{\alpha\beta} h_k^{\alpha\beta} = \delta_{ik} \quad \text{(A.3.29)}$$

$$\sum_\beta \sigma_i^{\alpha\beta} \frac{1}{2}\left[h_k^{\alpha\beta} h_l^{\alpha\beta} + x_{\beta,kl} - x_{\alpha,kl}\right] = 0 \quad \text{(A.3.30)}$$

$$\sum_\beta \theta_{ij}^{\alpha\beta} h_k^{\alpha\beta} = 0 \quad \text{(A.3.31)}$$

$$\sum_\beta \theta_{ij}^{\alpha\beta} \frac{1}{2}\left[h_k^{\alpha\beta} h_l^{\alpha\beta} + x_{\beta,kl} - x_{\alpha,kl}\right] = \frac{1}{2}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) \quad \text{(A.3.32)}$$

où $1 \leq i, j, k, l \leq d$. On rappelle que m_α désigne le nombre de cellules dans le voisinage de reconstruction $\mathbb{W}_\alpha$, où $\alpha \notin \mathbb{W}_\alpha$, cf. la définition (7.2.2). Il existe $m_\alpha d$ coefficients $\sigma_i^{\alpha\beta}$ pour $\beta \neq \alpha$ et $m_\alpha \frac{1}{2} d\,(d+1)$ coefficients indépendants $\theta_{ij}^{\alpha\beta}$ pour $\beta \neq \alpha$ en raison de la symétrie du tenseur $\boldsymbol{\theta}_{\alpha\beta}$. Les conditions (A.3.29) et (A.3.30) représentent respectivement d^2 et $\frac{1}{2}d^2\,(d+1)$ équations indépendantes pour les coefficients $\sigma_i^{\alpha\beta}$. De façon analogue, (A.3.31) et (A.3.32) constituent respectivement $\frac{1}{2}d^2\,(d+1)$ et $\frac{1}{4}d^2\,(d+1)^2$ équations pour les coefficients $\theta_{ij}^{\alpha\beta}$. Une condition nécessaire pour l'existence de solutions de (A.3.29) et (A.3.30) est donc

$$d^2 + \frac{1}{2}d^2\,(d+1) = \frac{1}{2}d^2\,(d+3) \leq m_\alpha d\,. \quad \text{(A.3.33)}$$

Une condition nécessaire pour l'existence de solutions de (A.3.31) et (A.3.32) est

$$\frac{1}{2}d^2\,(d+1) + \frac{1}{4}d^2\,(d+1)^2 = \frac{1}{4}d^2\,(d+1)\,(d+3) \leq m_\alpha \frac{1}{2} d\,(d+1)\,. \quad \text{(A.3.34)}$$

Les inégalités (A.3.33) et (A.3.34) donnent *la même condition nécessaire* pour m_α

$$\frac{1}{2}d\,(d+3) \leq m_\alpha\,. \quad \text{(A.3.35)}$$

La condition (A.3.35) s'obtient également par la comparaison entre la dimension de l'espace des polynômes de degré deux $\mathbb{P}_2\left(\mathbb{R}^d\right)$ et le nombre de cellules utilisées pour la reconstruction,

c'est-à-dire $m_\alpha + 1$ en comptant la cellule $\mathcal{T}_\alpha$ elle-même,

$$\frac{1}{2}(d+1)(d+2) = \frac{1}{2}d(d+3) + 1 \leq m_\alpha + 1\,, \tag{A.3.36}$$

ce qui est exactement la même condition que (A.3.35).

La prochaine étape est de reformuler les conditions (A.3.29-A.3.32) sous forme d'équations matricielles. Il est possible de définir dans chaque cellule une matrice géométrique H_α avec m_α lignes et $\frac{1}{2}d(d+3)$ colonnes. La transposée de H_α a la forme suivante

$$\widetilde{H}_\alpha^t = \begin{pmatrix} h_1^{\alpha\beta_1} & \cdots & h_1^{\alpha\beta_m} \\ \vdots & \vdots & \vdots \\ h_d^{\alpha\beta_1} & \cdots & h_d^{\alpha\beta_m} \\ \frac{1}{2}\left[\left(h_1^{\alpha\beta_1}\right)^2 + x_{\beta_1,11} - x_{\alpha,11}\right] & \cdots & \frac{1}{2}\left[\left(h_1^{\alpha\beta_m}\right)^2 + x_{\beta_m,11} - x_{\alpha,11}\right] \\ h_1^{\alpha\beta_1}h_2^{\alpha\beta_1} + x_{\beta_1,12} - x_{\alpha,12} & \cdots & h_1^{\alpha\beta_m}h_2^{\alpha\beta_m} + x_{\beta_m,12} - x_{\alpha,12} \\ \vdots & \vdots & \vdots \\ \frac{1}{2}\left[\left(h_d^{\alpha\beta_1}\right)^2 + x_{\beta_1,dd} - x_{\alpha,dd}\right] & \cdots & \frac{1}{2}\left[\left(h_d^{\alpha\beta_m}\right)^2 + x_{\beta_m,dd} - x_{\alpha,dd}\right] \end{pmatrix}. \tag{A.3.37}$$

Les coefficients $\sigma_i^{\alpha\beta}$ et $\theta_{ij}^{\alpha\beta}$ forment de la même façon une matrice S_α avec m_α colonnes et $\frac{1}{2}d(d+3)$ lignes.

$$S_\alpha = \begin{pmatrix} \sigma_1^{\alpha\beta_1} & \sigma_1^{\alpha\beta_2} & \sigma_1^{\alpha\beta_3} & \cdots & \sigma_1^{\alpha\beta_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_d^{\alpha\beta_1} & \sigma_d^{\alpha\beta_2} & \sigma_d^{\alpha\beta_3} & \cdots & \sigma_d^{\alpha\beta_m} \\ \theta_{11}^{\alpha\beta_1} & \theta_{11}^{\alpha\beta_2} & \theta_{11}^{\alpha\beta_3} & \cdots & \theta_{11}^{\alpha\beta_m} \\ \theta_{12}^{\alpha\beta_1} & \theta_{12}^{\alpha\beta_2} & \theta_{12}^{\alpha\beta_3} & \cdots & \theta_{12}^{\alpha\beta_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{dd}^{\alpha\beta_1} & \theta_{dd}^{\alpha\beta_2} & \theta_{dd}^{\alpha\beta_3} & \cdots & \theta_{dd}^{\alpha\beta_m} \end{pmatrix}. \tag{A.3.38}$$

Les équations (A.3.29) à (A.3.32) s'expriment alors par la simple identité

$$S_\alpha H_\alpha = \mathrm{I}_{\frac{1}{2}d(d+3)} \tag{A.3.39}$$

qui est de la même forme que l'équation (A.2.21), dans le cas des polynômes de degré un.

C'est pourquoi la discussion de la section A.2 s'applique également à la reconstruction par des polynômes de degré deux. En particulier, le rang de H_α doit être égal à $\frac{1}{2}d(d+3)$. Si $\mathcal{W}$ est un espace supplémentaire de $\mathrm{Im}\,(H_\alpha)$ dans $\mathbb{R}^{\frac{1}{2}d(d+3)}$ et si W est une matrice dont les colonnes forment une base de l'espace orthogonal $\mathcal{W}^\perp$, alors la forme générale de la solution de (A.3.39) est

$$S_{\mathcal{W},\alpha} = \left(W^t H_\alpha\right)^{-1} W^t\,.$$

A.4. Étude des reconstructions cubiques

Cette section est dédiée à l'étude détaillée de la reconstruction des polynômes de degré trois.

A.4.1. Établissement des conditions de reconstruction. L'objectif de cette section est d'analyser la reconstruction par des polynômes de degré trois. On reprend ici de façon détaillée la totalité des calculs, comme dans les deux sections précédentes. Les fonctions reconstruites (7.3.6) sont, dans ce cas, de la forme générale

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \sum_\beta \left[w^{\alpha\beta} + \sum_{i=1}^d w_i^{\alpha\beta}(x_i - x_{\alpha,i}) + \frac{1}{2!}\sum_{i=1}^d\sum_{j=1}^d w_{ij}^{\alpha\beta}(x_i - x_{\alpha,i})(x_j - x_{\alpha,j}) + \right.$$
$$\left. + \frac{1}{3!}\sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d w_{ijk}^{\alpha\beta}(x_i - x_{\alpha,i})(x_j - x_{\alpha,j})(x_k - x_{\alpha,k}) \right] \overline{u}_\beta\,. \quad \text{(A.4.1)}$$

Les coefficients $w_i^{\alpha\beta}$ forment un vecteur $\boldsymbol{w}_{\alpha\beta}^{(1)}$, les coefficients $w_{ij}^{\alpha\beta}$ un tenseur symétrique d'ordre deux noté $\boldsymbol{w}_{\alpha\beta}^{(2)}$ et les coefficients $w_{ijk}^{\alpha\beta}$ un tenseur symétrique d'ordre trois noté $\boldsymbol{w}_{\alpha\beta}^{(3)}$. La définition $w_{\alpha\beta}^{(0)} = w^{\alpha\beta}$ sert à unifier les notations.

Cette notation simplifie l'écriture de (A.4.1) qui devient

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = \sum_\beta \left\{ w_{\alpha\beta}^{(0)} + \boldsymbol{w}_{\alpha\beta}^{(1)}\cdot(\boldsymbol{x}-\boldsymbol{x}_\alpha) + \frac{1}{2!}\boldsymbol{w}_{\alpha\beta}^{(2)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 + \frac{1}{3!}\boldsymbol{w}_{\alpha\beta}^{(3)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^3 \right\}\overline{u}_\beta\,. \quad \text{(A.4.2)}$$

Les conditions de consistance déterminent les équations pour les coefficients de reconstruction. Une fonction polynomiale v de degré trois s'écrit sous la forme

$$v(\boldsymbol{x}) = v^{(0)} + \boldsymbol{v}^{(1)}\cdot(\boldsymbol{x}-\boldsymbol{x}_\alpha) + \frac{1}{2!}\boldsymbol{v}^{(2)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 + \frac{1}{3!}\boldsymbol{v}^{(3)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^3\,.$$

On rappelle la définition des tenseurs $\boldsymbol{z}_{\alpha\gamma}^{(j)}$ par (5.7.10) et celle des tenseurs $\boldsymbol{y}_{\alpha\gamma}^{(j|l)}$ par (5.7.9). La moyenne $\overline{v}_\gamma$ sur la cellule $\mathcal{T}_\gamma$ est, selon la formule (5.8.4),

$$\overline{v}_\gamma = v^{(0)} + \boldsymbol{v}^{(1)}\cdot\boldsymbol{h}_{\alpha\gamma} + \frac{1}{2!}\boldsymbol{v}^{(2)}\bullet\left[\boldsymbol{h}_{\alpha\gamma}^2 + \boldsymbol{x}_\gamma^{(2)}\right] + \frac{1}{3!}\boldsymbol{v}^{(3)}\bullet\left[\boldsymbol{h}_{\alpha\gamma}^3 + 3\boldsymbol{y}_{\alpha\gamma}^{(1|2)} + \boldsymbol{x}_\gamma^{(3)}\right]\,.$$

L'insertion des moyennes $\mathfrak{v} = (\overline{v}_1, \ldots, \overline{v}_N)$ dans la formule (A.3.2) pour $w_\alpha[\mathfrak{v}]$ et l'application des conditions de consistance de degré trois a pour résultat

$$v^{(0)} + \boldsymbol{v}^{(1)}\cdot(\boldsymbol{x}-\boldsymbol{x}_\alpha) + \frac{1}{2!}\boldsymbol{v}^{(2)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 + \frac{1}{3!}\boldsymbol{v}^{(3)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^3 =$$
$$= \sum_\beta \left(w_{\alpha\beta}^{(0)} + \boldsymbol{w}_{\alpha\beta}^{(1)}\cdot(\boldsymbol{x}-\boldsymbol{x}_\alpha) + \frac{1}{2!}\boldsymbol{w}_{\alpha\beta}^{(2)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 + \frac{1}{3!}\boldsymbol{w}_{\alpha\beta}^{(3)}\bullet(\boldsymbol{x}-\boldsymbol{x}_\alpha)^3 \right) \times$$
$$\times \left(v^{(0)} + \boldsymbol{v}^{(1)}\cdot\boldsymbol{h}_{\alpha\beta} + \frac{1}{2!}\boldsymbol{v}^{(2)}\bullet\left[\boldsymbol{h}_{\alpha\beta}^2 + \boldsymbol{x}_\beta^{(2)}\right] + \frac{1}{3!}\boldsymbol{v}^{(3)}\bullet\left[\boldsymbol{h}_{\alpha\beta}^3 + 3\boldsymbol{y}_{\alpha\gamma}^{(1|2)} + \boldsymbol{x}_\beta^{(3)}\right] \right)\,. \quad \text{(A.4.3)}$$

Les tenseurs $\boldsymbol{w}_{\alpha\beta}^{(l)}$ doivent satisfaire la condition (A.4.3) quelles que soient les valeurs de $v^{(0)}$, $\boldsymbol{v}^{(1)}$, $\boldsymbol{v}^{(2)}$ et $\boldsymbol{v}^{(3)}$. Les équations pour les coefficients s'obtiennent par identification des termes indépendants dans (A.4.3). Certaines équations sont reformulées afin d'éliminer des termes contenant $\boldsymbol{w}_{\alpha\alpha}^{(1)}$, $\boldsymbol{w}_{\alpha\alpha}^{(2)}$ et $\boldsymbol{w}_{\alpha\alpha}^{(3)}$. Ceci donne les équations suivantes pour $w_{\alpha\beta}^{(0)}$

$$\sum_\beta w_{\alpha\beta}^{(0)} = 1 \quad \text{(A.4.4)}$$

$$\sum_\beta w_{\alpha\beta}^{(0)}\boldsymbol{h}_{\alpha\beta} = 0 \quad \text{(A.4.5)}$$

$$\sum_\beta w_{\alpha\beta}^{(0)}\frac{1}{2!}\left(\boldsymbol{h}_{\alpha\beta}^2 + \boldsymbol{x}_\beta^{(2)}\right) = 0 \quad \text{(A.4.6)}$$

$$\sum_\beta w_{\alpha\beta}^{(0)}\frac{1}{3!}\left(\boldsymbol{h}_{\alpha\beta}^3 + 3\boldsymbol{y}_{\alpha\gamma}^{(1|2)} + \boldsymbol{x}_\beta^{(3)}\right) = 0\,. \quad \text{(A.4.7)}$$

Les équations pour les $\boldsymbol{w}_{\alpha\beta}^{(1)}$ sont

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} = 0 \tag{A.4.8}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \boldsymbol{h}_{\alpha\beta} = \boldsymbol{\delta}^{(2)} \tag{A.4.9}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \frac{1}{2!}\left(\boldsymbol{h}_{\alpha\beta}^{2} + \boldsymbol{x}_{\beta}^{(2)}\right) =$$

$$= \sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \frac{1}{2!}\left(\boldsymbol{h}_{\alpha\beta}^{2} + \boldsymbol{x}_{\beta}^{(2)} - \boldsymbol{x}_{\alpha}^{(2)}\right) = 0 \tag{A.4.10}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \frac{1}{3!}\left(\boldsymbol{h}_{\alpha\beta}^{3} + 3\boldsymbol{y}_{\alpha\gamma}^{(1|2)} + \boldsymbol{x}_{\beta}^{(3)}\right) =$$

$$= \sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(1)} \otimes \frac{1}{3!}\left(\boldsymbol{h}_{\alpha\beta}^{3} + 3\boldsymbol{y}_{\alpha\gamma}^{(1|2)} + \boldsymbol{x}_{\beta}^{(3)} - \boldsymbol{x}_{\alpha}^{(3)}\right) = 0\,. \tag{A.4.11}$$

où $\boldsymbol{\delta}^{(2)}$ est le tenseur de composantes

$$\left(\boldsymbol{\delta}^{(2)}\right)_{ik} = \delta_{ik}\,.$$

Les équations pour les tenseurs $\boldsymbol{w}_{\alpha\beta}^{(2)}$ s'écrivent

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} = 0 \tag{A.4.12}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} \otimes \boldsymbol{h}_{\alpha\beta} = 0 \tag{A.4.13}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} \otimes \frac{1}{2!}\left(\boldsymbol{h}_{\alpha\beta}^{2} + \boldsymbol{x}_{\beta}^{(2)}\right) =$$

$$= \sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} \otimes \frac{1}{2!}\left(\boldsymbol{h}_{\alpha\beta}^{2} + \boldsymbol{x}_{\beta}^{(2)} - \boldsymbol{x}_{\alpha}^{(2)}\right) = \boldsymbol{\delta}^{(4)} \tag{A.4.14}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} \otimes \frac{1}{3!}\left(\boldsymbol{h}_{\alpha\beta}^{3} + 3\boldsymbol{y}_{\alpha\gamma}^{(1|2)} + \boldsymbol{x}_{\beta}^{(3)}\right) =$$

$$= \sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(2)} \otimes \frac{1}{3!}\left(\boldsymbol{h}_{\alpha\beta}^{3} + 3\boldsymbol{y}_{\alpha\gamma}^{(1|2)} + \boldsymbol{x}_{\beta}^{(3)} - \boldsymbol{x}_{\alpha}^{(3)}\right) = 0 \tag{A.4.15}$$

où $\boldsymbol{\delta}^{(4)}$ est le tenseur d'ordre quatre défini par (5.2.22), de composantes

$$\left(\boldsymbol{\delta}^{(4)}\right)_{ijkl} \triangleq \frac{1}{2!}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)\,.$$

Finalement, les équations pour les tenseurs $\boldsymbol{w}_{\alpha\beta}^{(3)}$ deviennent

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(3)} = 0 \tag{A.4.16}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(3)} \otimes \boldsymbol{h}_{\alpha\beta} = 0 \tag{A.4.17}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(3)} \otimes \frac{1}{2!}\left(\boldsymbol{h}_{\alpha\beta}^{2} + \boldsymbol{x}_{\beta}^{(2)}\right) =$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(3)} \otimes \frac{1}{2!}\left(\boldsymbol{h}_{\alpha\beta}^{2} + \boldsymbol{x}_{\beta}^{(2)} - \boldsymbol{x}_{\alpha}^{(2)}\right) = 0 \tag{A.4.18}$$

$$\sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(3)} \otimes \frac{1}{3!}\left(\boldsymbol{h}_{\alpha\beta}^{3} + \boldsymbol{y}_{\alpha\gamma}^{(3)} + \boldsymbol{x}_{\beta}^{(3)}\right) =$$

$$= \sum_{\beta} \boldsymbol{w}_{\alpha\beta}^{(3)} \otimes \frac{1}{3!}\left(\boldsymbol{h}_{\alpha\beta}^{3} + 3\boldsymbol{y}_{\alpha\gamma}^{(1|2)} + \boldsymbol{x}_{\beta}^{(3)} - \boldsymbol{x}_{\alpha}^{(3)}\right) = \boldsymbol{\delta}^{(6)}\,. \tag{A.4.19}$$

D'après la définition (5.2.22), les composantes du tenseur $\boldsymbol{\delta}^{(6)}$ dans (A.4.19) sont

$$\left(\boldsymbol{\delta}^{(6)}\right)_{ijmkln} \triangleq \frac{1}{3!}\left(\delta_{ik}\delta_{jl}\delta_{mn} + \delta_{ik}\delta_{jn}\delta_{ml} + \delta_{il}\delta_{jk}\delta_{mn} + \delta_{il}\delta_{jn}\delta_{mk} + \delta_{in}\delta_{jk}\delta_{ml} + \delta_{in}\delta_{jl}\delta_{mk}\right)\,.$$

La condition (A.4.19) s'explicite sous la forme

$$\sum_{\beta} w_{ijm}^{\alpha\beta}\left[h_{k}^{\alpha\beta}h_{l}^{\alpha\beta}h_{n}^{\alpha\beta} + h_{k}^{\alpha\beta}x_{\beta,ln} + h_{l}^{\alpha\beta}x_{\beta,nk} + h_{n}^{\alpha\beta}x_{\beta,kl} + x_{\beta,kln} - x_{\alpha,kln}\right] =$$

$$= \delta_{ik}\delta_{jl}\delta_{mn} + \delta_{ik}\delta_{jn}\delta_{ml} + \delta_{il}\delta_{jk}\delta_{mn} + \delta_{il}\delta_{jn}\delta_{mk} + \delta_{in}\delta_{jk}\delta_{ml} + \delta_{in}\delta_{jl}\delta_{mk}\,.$$

Il est important de noter que les coefficients $\boldsymbol{w}_{\alpha\alpha}^{(1)}$, $\boldsymbol{w}_{\alpha\alpha}^{(2)}$ et $\boldsymbol{w}_{\alpha\alpha}^{(3)}$ apparaissent uniquement dans les équations (A.4.8), (A.4.12) et (A.4.16).

La contrainte de conservativité (6.3.8)

$$\frac{1}{|\mathcal{I}_{\alpha}|}\int_{\mathcal{I}_{\alpha}} w_{\alpha}\left[\mathfrak{u}\right](\boldsymbol{x})\,dx = \overline{u}_{\alpha}$$

donne pour les polynômes de degré trois l'équation supplémentaire

$$\overline{u}_{\alpha} = \sum_{\beta}\left[w_{\alpha\beta}^{(0)} + \frac{1}{2!}\boldsymbol{w}_{\alpha\beta}^{(2)} \bullet \boldsymbol{x}_{\alpha}^{(2)} + \frac{1}{3!}\boldsymbol{w}_{\alpha\beta}^{(3)} \bullet \boldsymbol{x}_{\alpha}^{(3)}\right]\overline{u}_{\beta}\,. \tag{A.4.20}$$

Comme (A.4.3) doit être vraie quelles que soient les valeurs $\mathfrak{u} = (\overline{u}_1, \ldots, \overline{u}_N)$, on exprime $w_{\alpha\beta}^{(0)}$ en fonction des tenseurs $\boldsymbol{w}_{\alpha\beta}^{(2)}$ et $\boldsymbol{w}_{\alpha\beta}^{(3)}$ par

$$w_{\alpha\beta}^{(0)} = \delta_{\alpha\beta} - \frac{1}{2}\boldsymbol{w}_{\alpha\beta}^{(2)} \bullet \boldsymbol{x}_{\alpha}^{(2)} - \frac{1}{3!}\boldsymbol{w}_{\alpha\beta}^{(3)} \bullet \boldsymbol{x}_{\alpha}^{(3)}\,. \tag{A.4.21}$$

La contrainte sur la moyenne (A.4.20) détermine donc les coefficients $w_{\alpha\beta}^{(0)}$ en fonction des coefficients $\boldsymbol{w}_{\alpha\beta}^{(2)}$ et $\boldsymbol{w}_{\alpha\beta}^{(3)}$. Elle permet de cette façon d'éliminer les conditions (A.4.4) à (A.4.7).

L'insertion de (A.4.21) dans (A.4.2) a pour résultat

$$w_{\alpha}\left[\mathfrak{u}\right](\boldsymbol{x}) = u_{\alpha} + \sum_{\beta}\left\{\boldsymbol{w}_{\alpha\beta}^{(1)} \cdot (\boldsymbol{x} - \boldsymbol{x}_{\alpha}) + \frac{1}{2!}\boldsymbol{w}_{\alpha\beta}^{(2)} \bullet \left[(\boldsymbol{x} - \boldsymbol{x}_{\alpha})^{2} - \boldsymbol{x}_{\alpha}^{(2)}\right] + \right.$$

$$\left. + \frac{1}{3!}\boldsymbol{w}_{\alpha\beta}^{(3)} \bullet \left[(\boldsymbol{x} - \boldsymbol{x}_{\alpha})^{3} - \boldsymbol{x}_{\alpha}^{(3)}\right]\right\}\overline{u}_{\beta}\,. \tag{A.4.22}$$

Les trois dérivées successives du polynôme (A.4.22) en $\boldsymbol{x}_\alpha$ sont données par

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{w}^{(1)}_{\alpha\beta}\overline{u}_\beta \tag{A.4.23}$$

$$\boldsymbol{\theta}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{w}^{(2)}_{\alpha\beta}\overline{u}_\beta \tag{A.4.24}$$

$$\boldsymbol{\omega}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{w}^{(3)}_{\alpha\beta}\overline{u}_\beta\,. \tag{A.4.25}$$

L'indication du rang a été omise pour alléger la notation. Le polynôme (A.4.22) devient finalement

$$w_\alpha[\mathfrak{u}](\boldsymbol{x}) = u_\alpha + \boldsymbol{\sigma}_\alpha[\mathfrak{u}]\cdot(\boldsymbol{x}-\boldsymbol{x}_\alpha) + \frac{1}{2!}\boldsymbol{\theta}_\alpha[\mathfrak{u}]\bullet\left[(\boldsymbol{x}-\boldsymbol{x}_\alpha)^2 - \boldsymbol{x}^{(2)}_\alpha\right] + \\ + \frac{1}{3!}\boldsymbol{\omega}_\alpha[\mathfrak{u}]\bullet\left[(\boldsymbol{x}-\boldsymbol{x}_\alpha)^3 - \boldsymbol{x}^{(3)}_\alpha\right]\,. \tag{A.4.26}$$

Pour plus de clarté, il convient de renommer les coefficients dans (A.4.23), (A.4.24) et (A.4.25)

$$\begin{aligned}\boldsymbol{\sigma}_{\alpha\beta} &\triangleq \boldsymbol{w}^{(1)}_{\alpha\beta}\\ \boldsymbol{\theta}_{\alpha\beta} &\triangleq \boldsymbol{w}^{(2)}_{\alpha\beta}\\ \boldsymbol{\omega}_{\alpha\beta} &\triangleq \boldsymbol{w}^{(3)}_{\alpha\beta}\,.\end{aligned}$$

Les équations (A.4.8), (A.4.12) et (A.4.16) servent uniquement à déterminer les coefficients $\boldsymbol{\sigma}_{\alpha\alpha}$, $\boldsymbol{\theta}_{\alpha\alpha}$ et $\boldsymbol{\omega}_{\alpha\alpha}$

$$\begin{aligned}\boldsymbol{\sigma}_{\alpha\alpha} &= -\sum_{\beta\neq\alpha}\boldsymbol{\sigma}_{\alpha\beta}\\ \boldsymbol{\theta}_{\alpha\alpha} &= -\sum_{\beta\neq\alpha}\boldsymbol{\theta}_{\alpha\beta}\\ \boldsymbol{\omega}_{\alpha\alpha} &= -\sum_{\beta\neq\alpha}\boldsymbol{\omega}_{\alpha\beta}\end{aligned}$$

en fonction des autres coefficients. Ces relations permettent de réécrire les dérivées de deux façons différentes de la manière suivante

$$\boldsymbol{\sigma}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}u_\beta = \sum_\beta \boldsymbol{\sigma}_{\alpha\beta}(u_\beta - u_\alpha) \tag{A.4.27}$$

$$\boldsymbol{\theta}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}u_\beta = \sum_\beta \boldsymbol{\theta}_{\alpha\beta}(u_\beta - u_\alpha) \tag{A.4.28}$$

$$\boldsymbol{\omega}_\alpha[\mathfrak{u}] = \sum_\beta \boldsymbol{\omega}_{\alpha\beta}u_\beta = \sum_\beta \boldsymbol{\omega}_{\alpha\beta}(u_\beta - u_\alpha)\,. \tag{A.4.29}$$

Les coefficients restants sont donnés par les équations

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\otimes\boldsymbol{h}_{\alpha\beta} = \boldsymbol{\delta}^{(2)} \tag{A.4.30}$$

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\otimes\frac{1}{2!}\left(\boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_\beta - \boldsymbol{x}^{(2)}_\alpha\right) = 0 \tag{A.4.31}$$

$$\sum_\beta \boldsymbol{\sigma}_{\alpha\beta}\otimes\frac{1}{3!}\left(\boldsymbol{h}^3_{\alpha\beta} + 3\boldsymbol{y}^{(1|2)}_{\alpha\beta} + \boldsymbol{x}^{(3)}_\beta - \boldsymbol{x}^{(3)}_\alpha\right) = 0 \tag{A.4.32}$$

$$\sum_{\beta} \boldsymbol{\theta}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} = 0 \tag{A.4.33}$$

$$\sum_{\beta} \boldsymbol{\theta}_{\alpha\beta} \otimes \frac{1}{2!}\left(\boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_{\beta} - \boldsymbol{x}^{(2)}_{\alpha}\right) = \boldsymbol{\delta}^{(4)}\,. \tag{A.4.34}$$

$$\sum_{\beta} \boldsymbol{\theta}_{\alpha\beta} \otimes \frac{1}{3!}\left(\boldsymbol{h}^3_{\alpha\beta} + 3\boldsymbol{y}^{(1|2)}_{\alpha\beta} + \boldsymbol{x}^{(3)}_{\beta} - \boldsymbol{x}^{(3)}_{\alpha}\right) = 0 \tag{A.4.35}$$

$$\sum_{\beta} \boldsymbol{\omega}_{\alpha\beta} \otimes \boldsymbol{h}_{\alpha\beta} = 0 \tag{A.4.36}$$

$$\sum_{\beta} \boldsymbol{\omega}_{\alpha\beta} \otimes \frac{1}{2!}\left(\boldsymbol{h}^2_{\alpha\beta} + \boldsymbol{x}^{(2)}_{\beta} - \boldsymbol{x}^{(2)}_{\alpha}\right) = 0 \tag{A.4.37}$$

$$\sum_{\beta} \boldsymbol{\omega}_{\alpha\beta} \otimes \frac{1}{3!}\left(\boldsymbol{h}^3_{\alpha\beta} + 3\boldsymbol{y}^{(1|2)}_{\alpha\beta} + \boldsymbol{x}^{(3)}_{\beta} - \boldsymbol{x}^{(3)}_{\alpha}\right) = \boldsymbol{\delta}^{(6)}\,. \tag{A.4.38}$$

Remarque A.4.1. Les 9 équations (A.4.30) à (A.4.38) montrent qu'il peut y avoir une différence entre la reconstruction à partir de moyennes de cellule et à partir de valeurs ponctuelles même si toutes les cellules ont les mêmes moments d'ordre deux $\boldsymbol{x}^{(2)}_{\beta}$ et trois $\boldsymbol{x}^{(3)}_{\beta}$. Ceci vient du fait qu'en général

$$\boldsymbol{y}^{(1|2)}_{\alpha\beta} \triangleq \boldsymbol{h}_{\alpha\beta} \odot \boldsymbol{x}^{(2)}_{\beta} \neq 0\,.$$

A.4.2. Calcul de l'erreur d'approximation. Si p est un polynôme de degré trois, les équations (A.4.30), (A.4.31) et (A.4.32) montrent que les $\boldsymbol{\sigma}_{\alpha\beta}$ reconstruisent le gradient de p en $\boldsymbol{x}_{\alpha}$. De façon analogue, les conditions (A.4.33), (A.4.34) et (A.4.35) signifient que les $\boldsymbol{\theta}_{\alpha\beta}$ reconstruisent la dérivée seconde de p en $\boldsymbol{x}_{\alpha}$. Les conditions (A.4.36), (A.4.37) et (A.4.38) entraînent que les $\boldsymbol{\omega}_{\alpha\beta}$ reconstruisent la dérivée troisième de p en $\boldsymbol{x}_{\alpha}$.

Si v est une fonction suffisamment régulière, la formule (5.8.3) permet d'écrire

$$\overline{v}_{\beta} - \overline{v}_{\alpha} = \sum_{k=0}^{4} \frac{1}{k!}\, D^{(k)}v\Big|_{\boldsymbol{x}_{\alpha}} \bullet \left[\boldsymbol{z}^{(k)}_{\alpha\beta} - \boldsymbol{x}^{(k)}_{\alpha}\right] + \mathrm{O}\left(h^5\right)\,. \tag{A.4.39}$$

L'insertion de (A.4.39) dans (A.4.27), ainsi que les équations (A.4.30), (A.4.31) et (A.4.32) entraînent que

$$\boldsymbol{\sigma}_{\alpha}\left[\mathfrak{v}\right] = \sum_{\beta} \boldsymbol{\sigma}_{\alpha\beta}\left(\overline{v}_{\beta} - \overline{v}_{\alpha}\right) = D^{(1)}v\Big|_{\boldsymbol{x}_{\alpha}} + \frac{1}{4!}\sum_{\beta} \boldsymbol{\sigma}_{\alpha\beta}\left\{D^{(4)}v\Big|_{\boldsymbol{x}_{\alpha}} \bullet \boldsymbol{z}^{(4)}_{\alpha\beta} + \mathrm{O}\left(h^5\right)\right\}\,. \tag{A.4.40}$$

L'expression $\boldsymbol{z}^{(4)}_{\alpha\beta}$ dans (A.4.40) est $\mathrm{O}\left(h^4\right)$. L'erreur dans (A.4.40) est $\mathrm{O}\left(h^3\right)$ à condition que $\|\boldsymbol{\sigma}_{\alpha\beta}\| = \mathrm{O}\left(h^{-1}\right)$. De façon analogue, l'erreur sur la dérivée seconde approchée

$$\boldsymbol{\theta}_{\alpha}\left[\mathfrak{v}\right] = \sum_{\beta} \boldsymbol{\theta}_{\alpha\beta}\left(\overline{v}_{\beta} - \overline{v}_{\alpha}\right) = D^{(2)}v\Big|_{\boldsymbol{x}_{\alpha}} + \frac{1}{4!}\sum_{\beta} \boldsymbol{\theta}_{\alpha\beta}\left\{D^{(4)}v\Big|_{\boldsymbol{x}_{\alpha}} \bullet \boldsymbol{z}^{(4)}_{\alpha\beta} + \mathrm{O}\left(h^5\right)\right\} \tag{A.4.41}$$

est $\mathrm{O}\left(h^2\right)$ à condition que $\|\boldsymbol{\theta}_{\alpha\beta}\| = \mathrm{O}\left(h^{-2}\right)$. L'approximation de la dérivée troisième est

$$\boldsymbol{\omega}_{\alpha}\left[\mathfrak{v}\right] = \sum_{\beta} \boldsymbol{\omega}_{\alpha\beta}\left(\overline{v}_{\beta} - \overline{v}_{\alpha}\right) = D^{(3)}v\Big|_{\boldsymbol{x}_{\alpha}} + \frac{1}{4!}\sum_{\beta} \boldsymbol{\omega}_{\alpha\beta}\left\{D^{(4)}v\Big|_{\boldsymbol{x}_{\alpha}} \bullet \boldsymbol{z}^{(4)}_{\alpha\beta} + \mathrm{O}\left(h^5\right)\right\} \tag{A.4.42}$$

et l'erreur dans (A.4.42) est $\mathrm{O}\left(h\right)$ si $\|\boldsymbol{\omega}_{\alpha\beta}\| = \mathrm{O}\left(h^{-3}\right)$.

A.4.3. Formulation matricielle. La prochaine étape est de reformuler les conditions (A.4.30-A.4.38) sous la forme d'une équation matricielle. Les équations s'écrivent de façon explicite

$$\sum_\beta \sigma_i^{\alpha\beta} h_k^{\alpha\beta} = \left(\boldsymbol{\delta}^{(2)}\right)_{ik} \tag{A.4.43}$$

$$\sum_\beta \sigma_i^{\alpha\beta} \frac{1}{2!}\left[h_k^{\alpha\beta} h_l^{\alpha\beta} + x_{\beta,kl} - x_{\alpha,kl}\right] = 0 \tag{A.4.44}$$

$$\sum_\beta \sigma_i^{\alpha\beta} \frac{1}{3!}\left[h_k^{\alpha\beta} h_l^{\alpha\beta} h_n^{\alpha\beta} + 3y_{kln}^{\alpha\beta} + x_{\beta,kln} - x_{\alpha,kln}\right] = 0 \tag{A.4.45}$$

$$\sum_\beta \theta_{ij}^{\alpha\beta} h_k^{\alpha\beta} = 0 \tag{A.4.46}$$

$$\sum_\beta \theta_{ij}^{\alpha\beta} \frac{1}{2!}\left[h_k^{\alpha\beta} h_l^{\alpha\beta} + x_{\beta,kl} - x_{\alpha,kl}\right] = \left(\boldsymbol{\delta}^{(4)}\right)_{ijkl} \tag{A.4.47}$$

$$\sum_\beta \theta_{ij}^{\alpha\beta} \frac{1}{3!}\left[h_k^{\alpha\beta} h_l^{\alpha\beta} h_n^{\alpha\beta} + 3y_{kln}^{\alpha\beta} + x_{\beta,kln} - x_{\alpha,kln}\right] = 0 \tag{A.4.48}$$

$$\sum_\beta \omega_{ijm}^{\alpha\beta} h_k^{\alpha\beta} = 0 \tag{A.4.49}$$

$$\sum_\beta \omega_{ijm}^{\alpha\beta} \frac{1}{2!}\left[h_k^{\alpha\beta} h_l^{\alpha\beta} + x_{\beta,kl} - x_{\alpha,kl}\right] = 0 \tag{A.4.50}$$

$$\sum_\beta \omega_{ijm}^{\alpha\beta} \frac{1}{3!}\left[h_k^{\alpha\beta} h_l^{\alpha\beta} h_n^{\alpha\beta} + 3y_{kln}^{\alpha\beta} + x_{\beta,kln} - x_{\alpha,kln}\right] = \left(\boldsymbol{\delta}^{(6)}\right)_{ijmkln} \tag{A.4.51}$$

où $1 \leq i, j, m, k, l, n \leq d$.

Le nombre d'inconnues dans (A.4.30) à (A.4.38) est

$$N_{\text{inc}} = m_\alpha \left[d + \frac{1}{2}d\,(d+1) + \frac{1}{6}d\,(d+1)\,(d+2)\right] = m_\alpha \frac{1}{6}d\left[11 + 6d + d^2\right]$$

et le nombre d'équations est

$$N_{\text{equ}} = \frac{1}{36}d^2\left[11 + 6d + d^2\right]^2 .$$

Il est possible de définir dans chaque cellule une matrice géométrique H_α avec m_α lignes et $\frac{1}{6}d\left[11 + 6d + d^2\right]$ colonnes. La transposée de H_α a la forme (A.4.53). Les coefficients $\sigma_i^{\alpha\beta}$, $\theta_{ij}^{\alpha\beta}$ et $\omega_{ijk}^{\alpha\beta}$ forment de la même façon la matrice S_α de (A.4.53) avec m_α colonnes et $\frac{1}{6}d\left[11 + 6d + d^2\right]$ lignes. Les équations (A.4.43) à (A.4.51) s'expriment alors par la simple relation

$$S_\alpha H_\alpha = \mathrm{I}_{\frac{1}{6}d[11+6d+d^2]} \tag{A.4.52}$$

qui est de la même forme que l'équation (A.2.21) dans le cas des polynômes de degré un.

C'est pourquoi la discussion de la section A.2 s'applique également à la reconstruction par des polynômes de degré trois. En particulier, le rang de H_α doit être égal à $\frac{1}{6}d\left[11 + 6d + d^2\right]$. Si $\mathcal{W}$ est un espace supplémentaire de $\mathrm{Im}\,(H_\alpha)$ dans $\mathbb{R}^{\frac{1}{6}d\left[11+6d+d^2\right]}$ et si W est une matrice dont les colonnes forment une base de l'espace orthogonal $\mathcal{W}^\perp$, alors la forme générale de la solution de (A.4.52) est

$$S_{\mathcal{W},\alpha} = \left(W^t H_\alpha\right)^{-1} W^t .$$

$$H_\alpha^t = \begin{pmatrix} h_1^{\alpha\beta_1} & \cdots & h_1^{\alpha\beta_m} \\ \vdots & \cdots & \vdots \\ h_d^{\alpha\beta_1} & \cdots & \vdots \\ \frac{1}{2!}\left[\left(h_1^{\alpha\beta_1}\right)^2 + x_{\beta_1,11} - x_{\alpha,11}\right] & \cdots & \vdots \\ h_1^{\alpha\beta_1} h_2^{\alpha\beta_1} + x_{\beta_1,12} - x_{\alpha,12} & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \frac{1}{2!}\left[\left(h_d^{\alpha\beta_1}\right)^2 + x_{\beta_1,dd} - x_{\alpha,dd}\right] & \cdots & \vdots \\ \frac{1}{3!}\left[\left(h_1^{\alpha\beta_1}\right)^3 + 3y_{111}^{\alpha\beta_1} + x_{\beta_1,111} - x_{\alpha,111}\right] & \cdots & \vdots \\ \frac{1}{2!}\left[\left(h_1^{\alpha\beta_1}\right)^2 h_2^{\alpha\beta_1} + 3y_{112}^{\alpha\beta_1} + x_{\beta_1,112} - x_{\alpha,112}\right] & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ h_1^{\alpha\beta_1} h_1^{\alpha\beta_1} h_3^{\alpha\beta_1} + 3y_{123}^{\alpha\beta_1} + x_{\beta_1,123} - x_{\alpha,123} & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \frac{1}{3!}\left[\left(h_d^{\alpha\beta_1}\right)^3 + 3y_{ddd}^{\alpha\beta_1} + x_{\beta_1,ddd} - x_{\alpha,ddd}\right] & \cdots & \vdots \end{pmatrix} \tag{A.4.53}$$

$$S_\alpha = \begin{pmatrix} \sigma_1^{\alpha\beta_1} & \sigma_1^{\alpha\beta_2} & \sigma_1^{\alpha\beta_3} & \cdots & \sigma_1^{\alpha\beta_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_d^{\alpha\beta_1} & \sigma_d^{\alpha\beta_2} & \sigma_d^{\alpha\beta_3} & \cdots & \sigma_d^{\alpha\beta_m} \\ \theta_{11}^{\alpha\beta_1} & \theta_{11}^{\alpha\beta_2} & \theta_{11}^{\alpha\beta_3} & \cdots & \theta_{11}^{\alpha\beta_m} \\ \theta_{12}^{\alpha\beta_1} & \theta_{12}^{\alpha\beta_2} & \theta_{12}^{\alpha\beta_3} & \cdots & \theta_{12}^{\alpha\beta_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{dd}^{\alpha\beta_1} & \theta_{dd}^{\alpha\beta_2} & \theta_{dd}^{\alpha\beta_3} & \cdots & \theta_{dd}^{\alpha\beta_m} \\ \omega_{111}^{\alpha\beta_1} & \omega_{111}^{\alpha\beta_2} & \omega_{111}^{\alpha\beta_3} & \cdots & \omega_{111}^{\alpha\beta_m} \\ \omega_{112}^{\alpha\beta_1} & \omega_{112}^{\alpha\beta_2} & \omega_{112}^{\alpha\beta_3} & \cdots & \omega_{112}^{\alpha\beta_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \omega_{123}^{\alpha\beta_1} & \omega_{123}^{\alpha\beta_2} & \omega_{123}^{\alpha\beta_3} & \cdots & \omega_{123}^{\alpha\beta_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \omega_{ddd}^{\alpha\beta_1} & \omega_{ddd}^{\alpha\beta_2} & \omega_{ddd}^{\alpha\beta_3} & \cdots & \omega_{ddd}^{\alpha\beta_m} \end{pmatrix} \tag{A.4.54}$$

ANNEXE B

Étude détaillée de la stabilité en dimension un

B.1. Objectif du chapitre

L'objectif de ce chapitre est de présenter de façon détaillée l'étude de la stabilité du chapitre 10 dans le cas particulier de la reconstruction linéaire en dimension un.

B.2. Construction du schéma semi-discret en dimension un

Le sujet de cette section est la construction du schéma MUSCL semi-discret (10.2.13) en dimension un. Dans ce cas particulier, l'équation de convection linéaire pour une fonction u : $\mathbb{R} \times [0,T] \to \mathbb{R}$ est

$$\partial_t u + c\partial_x u = 0 \tag{B.2.1}$$

La discrétisation de cette équation se fait sur un maillage irrégulier de taille N. Les cellules sont indexées de 1 à N et la définition $N+1 \triangleq 1$ permet de tenir compte de la périodicité du maillage. En dimension un, la plupart des vecteurs géométriques du maillage s'expriment en fonction des volumes. Les écarts entre les barycentres des cellules sont donnés par

$$h_{\alpha,\alpha+1} = \frac{1}{2}\left(|\mathcal{T}_\alpha| + |\mathcal{T}_{\alpha+1}|\right) > 0 \qquad , \qquad h_{\alpha,\alpha-1} = -\frac{1}{2}\left(|\mathcal{T}_\alpha| + |\mathcal{T}_{\alpha-1}|\right) < 0\,.$$

Les autres grandeurs géométriques sont données par

$$k_{\alpha,\alpha+1} = \frac{|\mathcal{T}_\alpha|}{2} > 0 \qquad k_{\alpha,\alpha-1} = -\frac{|\mathcal{T}_\alpha|}{2} < 0 \tag{B.2.2}$$

$$j_{\alpha,\alpha+1} = \frac{|\mathcal{T}_\alpha|}{2} > 0 \qquad j_{\alpha,\alpha-1} = -\frac{|\mathcal{T}_\alpha|}{2} < 0 \tag{B.2.3}$$

$$a_{\alpha,\alpha+1} = 1 \qquad a_{\alpha,\alpha-1} = -1\,. \tag{B.2.4}$$

Dans cette section, l'étude se restreint à la reconstruction sur le premier voisinage qui se trouve en dimension un réduit à $\mathbb{V}_\alpha = \{\alpha-1, \alpha+1\}$. La matrice géométrique H_α, définie par (A.2.19), est dans ce cas donnée par

$$H_\alpha \triangleq \begin{pmatrix} h_{\alpha,\alpha+1} \\ h_{\alpha,\alpha-1} \end{pmatrix}. \tag{B.2.5}$$

La matrice de reconstruction du gradient S_α, définie par (A.2.20), s'écrit sous la forme

$$S_\alpha \triangleq \begin{pmatrix} s_{\alpha,\alpha+1} & s_{\alpha,\alpha-1} \end{pmatrix}. \tag{B.2.6}$$

Avec les définitions (B.2.5) et (B.2.6), la condition de consistance (A.2.21) pour la matrice de reconstruction S_α est

$$S_\alpha H_\alpha = s_{\alpha,\alpha+1}h_{\alpha,\alpha+1} + s_{\alpha,\alpha-1}h_{\alpha,\alpha-1} = 1\,. \tag{B.2.7}$$

D'après le théorème 7.8.1, la forme générale des solutions de (B.2.7) est

$$S_\alpha = \widetilde{S}_\alpha + \Lambda_\alpha B_\alpha$$

où $\widetilde{S}_\alpha \in \mathbb{M}_{1,2}(\mathbb{R})$ est une solution particulière de (B.2.7) et $B_\alpha \in \mathbb{M}_{1,2}(\mathbb{R})$ est une solution de rang un de l'équation homogène

$$B_\alpha H_\alpha = b_{\alpha,\alpha+1}h_{\alpha,\alpha+1} + b_{\alpha,\alpha-1}h_{\alpha,\alpha-1} = 0\,. \tag{B.2.8}$$

La matrice $\Lambda_\alpha \in \mathbb{M}_{1,1}(\mathbb{R})$ est dans le cas présent un nombre réel arbitraire. Une solution particulière de (B.2.7) est donnée par la matrice de la méthode des moindres carrés (7.9.4)

$$S_\alpha = \left(H_\alpha^t H_\alpha\right)^{-1} H_\alpha^t\,.$$

Les coefficients de la matrice (7.9.4) sont les paramètres de la reconstruction du gradient par la méthode des moindres carrés

$$s_{\alpha,\alpha+1} = \frac{h_{\alpha,\alpha+1}}{h_{\alpha,\alpha+1}^2 + h_{\alpha,\alpha-1}^2}\,,\; s_{\alpha,\alpha-1} = \frac{h_{\alpha,\alpha-1}}{h_{\alpha,\alpha+1}^2 + h_{\alpha,\alpha-1}^2}\,. \tag{B.2.9}$$

Une solution B_α de rang un de (B.2.8) est donnée par

$$B_\alpha \triangleq \begin{pmatrix} \dfrac{-h_{\alpha,\alpha-1}}{h_{\alpha,\alpha+1}^2 + h_{\alpha,\alpha-1}^2} & \dfrac{h_{\alpha,\alpha+1}}{h_{\alpha,\alpha+1}^2 + h_{\alpha,\alpha-1}^2} \end{pmatrix}. \tag{B.2.10}$$

Comme le symbole λ est déjà utilisé pour les valeurs propres, on note ξ_α le seul élément de la matrice $\Lambda_\alpha \in \mathbb{M}_{1,1}(\mathbb{R})$. La forme générale de la reconstruction est alors donnée par la matrice

$$S_\alpha = \begin{pmatrix} \dfrac{h_{\alpha,\alpha+1} - \xi_\alpha h_{\alpha,\alpha-1}}{h_{\alpha,\alpha+1}^2 + h_{\alpha,\alpha-1}^2} & \dfrac{h_{\alpha,\alpha-1} + \xi_\alpha h_{\alpha,\alpha+1}}{h_{\alpha,\alpha+1}^2 + h_{\alpha,\alpha-1}^2} \end{pmatrix}.$$

Le théorème 7.8.2 fournit une autre méthode pour écrire la solution générale de (B.2.7). D'après le théorème 7.8.2, le choix d'une solution S_α de (B.2.7) est équivalent au choix d'un sous-espace $\mathcal{W}$ qui est un supplémentaire de $\mathrm{Im}\,(H_\alpha)$ dans $\mathbb{R}^2$. Introduisons pour cela le vecteur $\mathfrak{h}_\alpha = \begin{pmatrix} h_{\alpha,\alpha+1} & h_{\alpha,\alpha-1} \end{pmatrix}$ qui engendre le sous-espace $\mathrm{Im}\,(H_\alpha)$. Comme la dimension de $\mathrm{Im}\,(H_\alpha)$ est un, la dimension de $\mathcal{W}$ est également un. Pour fixer $\mathcal{W}$ il suffit alors de choisir un vecteur unitaire non nul $\mathfrak{v}_\alpha = \begin{pmatrix} v_{\alpha,\alpha+1} & v_{\alpha,\alpha-1} \end{pmatrix}$ qui est linéairement indépendant de $\mathfrak{h}_\alpha$. Cela signifie que le déterminant de $\mathfrak{h}_\alpha$ et de $\mathfrak{v}_\alpha$ satisfait

$$\begin{vmatrix} h_{\alpha,\alpha+1} & h_{\alpha,\alpha-1} \\ v_{\alpha,\alpha+1} & v_{\alpha,\alpha-1} \end{vmatrix} = h_{\alpha,\alpha+1}v_{\alpha,\alpha-1} - h_{\alpha,\alpha-1}v_{\alpha,\alpha+1} \neq 0\,. \tag{B.2.11}$$

D'après le théorème (7.8.2), il faut ensuite choisir une matrice dont les colonnes forment une base de $\mathcal{W}^\perp$. Un choix possible pour cette matrice est

$$W_\alpha \triangleq \begin{pmatrix} w_{\alpha,\alpha+1} \\ w_{\alpha,\alpha-1} \end{pmatrix} \triangleq \begin{pmatrix} v_{\alpha,\alpha-1} \\ -v_{\alpha,\alpha+1} \end{pmatrix}.$$

La relation (B.2.11) équivaut en fait à

$$W_\alpha^t H_\alpha = h_{\alpha,\alpha+1}w_{\alpha,\alpha+1} + h_{\alpha,\alpha-1}w_{\alpha,\alpha-1} \neq 0\,.$$

La forme générale de la reconstruction s'écrit alors sous la forme

$$S_\alpha \triangleq \begin{pmatrix} \dfrac{w_{\alpha,\alpha+1}}{w_{\alpha,\alpha+1}h_{\alpha,\alpha+1} + w_{\alpha,\alpha-1}h_{\alpha,\alpha-1}} & \dfrac{w_{\alpha,\alpha-1}}{w_{\alpha,\alpha+1}h_{\alpha,\alpha+1} + w_{\alpha,\alpha-1}h_{\alpha,\alpha-1}} \end{pmatrix}$$

qui montre que la matrice S_α ne dépend que de la droite engendrée par W_α. En effet, la multiplication de W_α par un nombre réel non nul produit la même matrice de reconstruction S_α. La forme des matrices montre également que S_α devient la reconstruction des moindres carrés (7.9.4) si $W_\alpha = H_\alpha$.

Pour la reconstruction de Green, la matrice W_α est identique à la matrice N_α définie par (7.9.11) et la matrice de reconstruction est

$$S_\alpha = \left(N_\alpha^t H_\alpha\right)^{-1} N_\alpha^t\,.$$

Avec les identités pour les vecteurs surface (B.2.4) et les vecteurs (B.2.3) ainsi que les inégalités $h_{\alpha,\alpha+1} > 0$ et $h_{\alpha,\alpha-1} < 0$, la matrice N_α est donnée par

$$N_\alpha \triangleq \begin{pmatrix} \dfrac{|j_{\alpha,\alpha+1}|}{|h_{\alpha,\alpha+1}|} \\ \\ -\dfrac{|j_{\alpha,\alpha-1}|}{|h_{\alpha,\alpha-1}|} \end{pmatrix} = \begin{pmatrix} \dfrac{|\mathcal{T}_\alpha|}{2h_{\alpha,\alpha+1}} \\ \\ \dfrac{|\mathcal{T}_\alpha|}{2h_{\alpha,\alpha-1}} \end{pmatrix}. \tag{B.2.12}$$

La formule (B.2.12) implique

$$N_\alpha^t H_\alpha = |j_{\alpha,\alpha+1}| + |j_{\alpha,\alpha-1}| = |\mathcal{T}_\alpha|\,.$$

La matrice de reconstruction S_α pour la méthode Green est alors donnée par

$$S_\alpha = \left(\begin{array}{cc} \dfrac{1}{2h_{\alpha,\alpha+1}} & \dfrac{1}{2h_{\alpha,\alpha-1}} \end{array} \right) . \tag{B.2.13}$$

La formulation du schéma MUSCL semi-discret (10.2.13) en dimension un consiste dans un premier temps à interpoler des valeurs aux interfaces des cellules à l'aide de la reconstruction du gradient

$$\left. \begin{array}{l} u_{\alpha,\alpha+1} = u_\alpha + k_{\alpha,\alpha+1}\sigma_\alpha = u_\alpha + \dfrac{|\mathcal{T}_\alpha|}{2}\sigma_\alpha \\ u_{\alpha,\alpha-1} = u_\alpha + k_{\alpha,\alpha-1}\sigma_\alpha = u_\alpha - \dfrac{|\mathcal{T}_\alpha|}{2}\sigma_\alpha \end{array} \right\} . \tag{B.2.14}$$

Il faut ensuite tenir compte des identités

$$\left. \begin{array}{l} (c \cdot a_{\alpha,\alpha+1})_+ = -(c \cdot a_{\alpha,\alpha-1})_- = (c)_+ = \begin{cases} c & \text{si } c > 0 \\ 0 & \text{si } c \leq 0 \end{cases} \\ (c \cdot a_{\alpha,\alpha+1})_- = -(c \cdot a_{\alpha,\alpha-1})_+ = (c)_- = \begin{cases} 0 & \text{si } c \geq 0 \\ c & \text{si } c < 0 \end{cases} \end{array} \right\} . \tag{B.2.15}$$

L'insertion des formules (B.2.14) et (B.2.15) dans le schéma (10.2.13) donne l'équation différentielle ordinaire

$$\begin{aligned} \frac{du_\alpha}{dt} &= -\frac{c}{|\mathcal{T}_\alpha|}\left(u_{\alpha,\alpha+1} - u_{\alpha-1,\alpha}\right) = \\ &= -\frac{c}{|\mathcal{T}_\alpha|}\left(u_\alpha + \frac{|\mathcal{T}_\alpha|}{2}\sigma_\alpha - u_{\alpha-1} - \frac{|\mathcal{T}_{\alpha-1}|}{2}\sigma_{\alpha-1}\right), 1 \leq \alpha \leq N \end{aligned} \tag{B.2.16}$$

qui s'écrit de façon explicite

$$\begin{aligned} \frac{du_\alpha}{dt} = -\frac{c}{|\mathcal{T}_\alpha|}\{&u_\alpha + k_{\alpha,\alpha+1}\left[s_{\alpha,\alpha+1}\left(u_{\alpha+1} - u_\alpha\right) + s_{\alpha,\alpha-1}\left(u_{\alpha-1} - u_\alpha\right)\right] - \\ &-u_{\alpha-1} - k_{\alpha-1,\alpha}\left[s_{\alpha-1,\alpha}\left(u_\alpha - u_{\alpha-1}\right) + s_{\alpha-1,\alpha-2}\left(u_{\alpha-2} - u_{\alpha-1}\right)\right]\} \end{aligned} \tag{B.2.17}$$

où l'indice α varie de 1 à N.

B.3. Analyse de stabilité du schéma d'ordre deux en dimension un

Le sujet de cette section est l'étude de la stabilité asymptotique du schéma MUSCL (10.2.13) en dimension un qui est alors donné par l'équation différentielle (B.2.17). Dans ce cas particulier, la matrice R_α de l'opérateur de reconstruction locale

$$R_\alpha = \left[\begin{array}{cc} k_{\alpha,\alpha+1}s_{\alpha,\alpha+1} & k_{\alpha,\alpha+1}s_{\alpha,\alpha-1} \\ k_{\alpha,\alpha-1}s_{\alpha,\alpha+1} & k_{\alpha,\alpha-1}s_{\alpha,\alpha-1} \end{array} \right] \tag{B.3.1}$$

n'a que deux éléments indépendants. Les définitions

$$\left. \begin{array}{l} r_\alpha^{(+)} \triangleq k_{\alpha,\alpha+1}s_{\alpha,\alpha+1} = -k_{\alpha,\alpha-1}s_{\alpha,\alpha+1} = \dfrac{|\mathcal{T}_\alpha|}{2}s_{\alpha,\alpha+1} \\ r_\alpha^{(-)} \triangleq k_{\alpha,\alpha-1}s_{\alpha,\alpha-1} = -k_{\alpha,\alpha+1}s_{\alpha,\alpha-1} = -\dfrac{|\mathcal{T}_\alpha|}{2}s_{\alpha,\alpha-1} \end{array} \right\} \tag{B.3.2}$$

permettent d'écrire la matrice (B.3.1) comme

$$R_\alpha = \left[\begin{array}{cc} r_\alpha^{(+)} & -r_\alpha^{(-)} \\ -r_\alpha^{(+)} & r_\alpha^{(-)} \end{array} \right] . \tag{B.3.3}$$

La définition (B.3.2) permet de simplifier (B.2.17)

$$\begin{aligned} \frac{du_\alpha}{dt} = -\frac{c}{|\mathcal{T}_\alpha|}\Big\{&u_\alpha + r_\alpha^{(+)}\left(u_{\alpha+1} - u_\alpha\right) - r_\alpha^{(-)}\left(u_{\alpha-1} - u_\alpha\right) - \\ &-u_{\alpha-1} - r_{\alpha-1}^{(+)}\left(u_\alpha - u_{\alpha-1}\right) + r_{\alpha-1}^{(-)}\left(u_{\alpha-2} - u_{\alpha-1}\right)\Big\}, 1 \leq \alpha \leq N . \end{aligned} \tag{B.3.4}$$

La forme quadratique (10.5.11) associée à (B.3.4) devient en dimension un

$$\Phi(\delta\mathfrak{u}) = -c\sum_{\alpha} |u_{\alpha+1} - u_\alpha|^2 + \\ + 2c\sum_{\alpha} \Re\left\{ \left(u^*_{\alpha+1} - u^*_\alpha\right) \left(r^{(+)}_\alpha \left(u_{\alpha+1} - u_\alpha\right) + r^{(-)}_\alpha \left(u_\alpha - u_{\alpha-1}\right)\right)\right\} . \quad \text{(B.3.5)}$$

Une condition suffisante mais non nécessaire pour la stabilité asymptotique de (10.2.14) en dimension un est donnée par

$$\Phi(\delta\mathfrak{u}) \leq 0 . \quad \text{(B.3.6)}$$

Même en dimension un, la question de la stabilité de l'opérateur associé à (B.3.4) s'avère trop difficile pour trouver une réponse générale, cf. [**57**]. Pour cette raison, on énonce dans cette section uniquement un résultat pour le cas des maillages uniformes. Ce résultat s'obtient par un calcul explicite des valeurs propres de l'opérateur MUSCL.

Dans le cas d'un maillage régulier, les entités géométriques se simplifient pour donner

$$h_{\alpha,\alpha+1} = h \qquad k_{\alpha,\alpha+1} = \frac{h}{2} \qquad j_{\alpha,\alpha+1} = \frac{h}{2} \qquad |\mathcal{T}_\alpha| = h$$
$$h_{\alpha,\alpha-1} = -h \qquad k_{\alpha,\alpha-1} = -\frac{h}{2} \qquad j_{\alpha,\alpha-1} = -\frac{h}{2} .$$

L'uniformité du maillage suggère d'utiliser la même reconstruction du gradient dans toutes les cellules. Pour cela, on choisit le même paramètre $\xi_\alpha = \xi$ dans chaque cellule, ce qui donne

$$S_\alpha \triangleq \begin{pmatrix} \dfrac{1+\xi}{2h} & \dfrac{-1+\xi}{2h} \end{pmatrix} .$$

Le gradient σ_α dans la cellule $\mathcal{T}_\alpha$ est dans ce cas

$$\sigma_\alpha = \frac{1+\xi}{2h}\left(u_{\alpha+1} - u_\alpha\right) + \frac{-1+\xi}{2h}\left(u_{\alpha-1} - u_\alpha\right) .$$

Certains choix pour ξ donnent les formules centrées et décentrées classiques pour l'approximation du gradient

$$\sigma_\alpha = \begin{cases} \dfrac{u_{\alpha+1} - u_{\alpha-1}}{2h} & \text{si } \xi = 0 \\ \dfrac{u_{\alpha+1} - u_\alpha}{h} & \text{si } \xi = 1 \\ \dfrac{u_\alpha - u_{\alpha-1}}{h} & \text{si } \xi = -1 \end{cases} .$$

Le schéma semi-discret (B.2.16) devient, dans le cas d'un maillage uniforme

$$\frac{du_\alpha}{dt} \;=\; -\frac{c}{h}\left(u_\alpha + \frac{h}{2}\sigma_\alpha - u_{\alpha-1} - \frac{h}{2}\sigma_{\alpha-1}\right) . \quad \text{(B.3.7)}$$

Pour calculer les valeurs propres de l'opérateur semi-discret associé à (B.3.7), il est utile de définir une matrice de permutation P_N qui agit sur le vecteur de la solution semi-discrète $\mathfrak{u} = (u_1, \ldots, u_N)$ par

$$\left(P_N\mathfrak{u}\right)_\alpha = u_{\alpha+1} .$$

La définition $N+1 \triangleq 1$ permet de tenir compte de la périodicité du maillage. L'inverse P_N^{-1} de P_N est définie par

$$\left(P_N^{-1}\mathfrak{u}\right)_\alpha = u_{\alpha-1} .$$

La matrice de permutation P_N est orthogonale car

$$\left(P_N\mathfrak{u}, P_N\mathfrak{v}\right) = \sum_{\alpha=1}^{N} u^*_{\alpha+1} v_{\alpha+1} = \sum_{\alpha=1}^{N} u^*_\alpha v_\alpha = \left(\mathfrak{u}, \mathfrak{v}\right) \text{ pour tout } \mathfrak{u}, \mathfrak{v} \in \mathbb{C}^N ,$$

ce qui entraîne $P_N^{-1} = P_N^t$. La matrice P_N est donc normale, c'est-à-dire qu'elle satisfait

$$P_N P_N^t = P_N^t P_N .$$

Les matrices normales sont exactement les matrices qui sont diagonalisables par une transformation unitaire. Il existe par conséquent une matrice unitaire U telle que

$$UP_NU^* = \mathrm{diag}\,(\mu_1, \ldots, \mu_N)$$

où les μ_k, $1 \leq k \leq N$, sont les valeurs propres de P_N. L'identité

$$P_N^N = \mathrm{I}_N$$

entraîne que les μ_k, $1 \leq k \leq N$ sont des racines N-ièmes de l'unité. Un calcul explicite révèle que les valeurs propres de P_N sont données par

$$\mu_k = \exp\left(\frac{2\pi i}{N}k\right)\,,\ 1 \leq k \leq N\,. \tag{B.3.8}$$

En dimension un, le gradient σ_α, pour $1 \leq \alpha \leq N$, est un nombre réel, ce qui permet de définir le vecteur $\mathfrak{s} \triangleq (\sigma_1, \ldots, \sigma_N)$ des gradients de cellule. Les définitions de $\mathfrak{s}$ et de la matrice P_N permettent d'écrire le schéma semi-discret (B.3.7) sous la forme compacte

$$\frac{d\mathfrak{u}}{dt} = -\frac{c}{h}\left(\mathfrak{u} + \frac{h}{2}\mathfrak{s} - P_N^{-1}\mathfrak{u} - \frac{h}{2}P_N^{-1}\mathfrak{s}\right) = -\frac{c}{h}\left(\mathrm{I}_N - P_N^{-1}\right)\left(\mathfrak{u} + \frac{h}{2}\mathfrak{s}\right) \tag{B.3.9}$$

où le vecteur des gradients de cellule $\mathfrak{s}$ est donné par la formule

$$\mathfrak{s} = \left(\frac{1+\xi}{2h}\left(P_N - \mathrm{I}_N\right) + \frac{-1+\xi}{2h}\left(P_N^{-1} - \mathrm{I}_N\right)\right)\mathfrak{u}\,.$$

La matrice J de l'opérateur de (B.3.9) s'écrit alors comme un polynôme de degré un dans les matrices P_N et P_N^{-1}

$$J = -\frac{c}{h}\left(\mathrm{I}_N - P_N^{-1}\right)\left(\mathrm{I}_N + \frac{1+\xi}{4}\left(P_N - \mathrm{I}_N\right) + \frac{-1+\xi}{4}\left(P_N^{-1} - \mathrm{I}_N\right)\right)\,. \tag{B.3.10}$$

Cela permet d'énoncer la

Proposition B.3.1 (Valeurs propres de l'opérateur Muscl sur maillage uniforme en dimension un). *Considérons le système linéaire (B.3.7) avec une vitesse de convection positive $c > 0$ sur un maillage régulier avec la même méthode de reconstruction dans chaque cellule. Dans ces conditions, l'opérateur J donné par (B.3.10) est diagonalisable. Les vecteurs propres de J sont exactement les vecteurs propres de P_N et ils forment une base orthogonale de $\mathbb{C}^N$. Les valeurs propres de l'opérateur J sont explicitement données par*

$$\begin{aligned} \Re\left(\lambda_k\right) &= (1-\xi)\,\frac{c}{4h}\left(1 - \cos\left(\frac{2\pi k}{N}\right)\right)^2 \\ \Im\left(\lambda_k\right) &= -\frac{c}{2h}\sin\left(\frac{2\pi k}{N}\right)\left[2 + (1-\xi)\left(1 - \cos\left(\frac{2\pi k}{N}\right)\right)\right] \end{aligned} \tag{B.3.11}$$

où $-\frac{N}{2} < k \leq \left[\frac{N}{2}\right]$. Cela montre que sur un maillage uniforme et une vitesse $c > 0$, le schéma Muscl *est stable si et seulement si $\xi \geq 1$.*

Démonstration. Les matrices P_N et $P_N^{-1} = P_N^t$ sont normales et ont les mêmes vecteurs propres qui forment une base orthogonale de $\mathbb{C}^N$. Puisque la matrice J définie par (B.3.10) est un polynôme dans les matrices P_N et P_N^{-1}, J est elle-même diagonalisable par la même transformation de similitude qui diagonalise P_N et P_N^{-1}. Les vecteurs propres de J sont alors exactement ceux de P_N. L'insertion des valeurs propres (B.3.8) dans l'expression (B.3.10) de J donne les valeurs propres de J. □

Bibliographie

[1] R. Abgrall : An extension of Roe's upwind scheme to algebraic equilibrium real gas models. *Computers & Fluids*, 19(2):171 – 182, 1991.

[2] R. Abgrall : Design of an essentially non-oscillatory reconstruction procedure on finite-element type meshes. Rapport technique 1584, INRIA, 1992.

[3] R. Abgrall : Essentially non-oscillatory Residual Distribution schemes for hyperbolic problems. *Journal of Computational Physics*, 214(2):773 – 808, 2006.

[4] R. Abgrall : Residual distribution schemes : Current status and future trends. *Computers & Fluids*, 35(7):641 – 669, 2006. Special Issue Dedicated to Professor Stanley G. Rubin on the Occasion of his 65th Birthday.

[5] R. Abgrall et H. Deconinck : Special issue on Residual Distribution schemes, Discontinuous Galerkin schemes multidimensional schemes and mesh adaptation. *Computers & Fluids*, 34(4-5):399 – 400, 2005. Residual Distribution Schemes, Discontinuous Galerkin Schemes and Adaptation.

[6] R. Abgrall et M. Mezine : Construction of second order accurate monotone and stable residual distribution schemes for unsteady flow problems. *Journal of Computational Physics*, 188(1):16 – 55, 2003.

[7] R. Abgrall et M. Mezine : Construction of second-order accurate monotone and stable residual distribution schemes for steady problems. *Journal of Computational Physics*, 195(2):474 – 507, 2004.

[8] T. Barth et M. Ohlberger : Finite volume methods : Foundation and analysis. *Dans* E. Stein, R. de Borst et T.J.R. Hughes, éditeurs : *Encyclopedia of Computational Mechanics*. John Wiley & Sons, 2004.

[9] T. J. Barth et P. O. Frederickson : Higher order solution of the Euler equation on unstructured grids using quadratic reconstruction. *Dans AIAA 90*, numéro AIAA-90-0013, pages 1–12, Reno Nevada, January 1990. AIAA.

[10] T.J. Barth : Numerical methods for conservation laws on structured and unstructured meshes. Rapport technique, VKI Lectures Series, 2003.

[11] G. K. Batchelor : *An introduction in fluid dynamics*. Cambridge University Press, 2000.

[12] M. Ben-Artzi et J. Falcovitz : *Generalized Riemann problems in Computational Fluid Dynamics*. Cambridge Univ. Press, 2003.

[13] M. Berger, M. J. Aftosmis et S. M. Murman : Analysis of slope limiters on irregular grids. Rapport technique NAS-05-007, NASA, Jan. 2005. AIAA Paper 2005-0490, Reno, NV.

[14] C. Berthon : Stability of the muscl schemes for the euler equations. *Communications in Mathematical Sciences*, 3(2):133–157, 2005.

[15] C. Berthon : Why the muscl-hancock scheme is l1-stable. *Numerische Mathematik*, 104:27–46, 2006.

[16] Christophe Berthon : Robustness of muscl schemes for 2d unstructured meshes. *Journal of Computational Physics*, 218(2):495 – 509, 2006.

[17] N. Bertier : *Simulation des grandes échelles en aérothermique sur des maillages non structurés généraux*. Thèse de doctorat, Université Pierre et Marie Curie Paris VI, 2006.

[18] F. Bouchut : *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well balanced schemes for sources*. Frontiers in Mathematics. Birkhäuser, 2004.

[19] F. Bouchut, Ch. Bourdarias et B. Perthame : A muscl method satisfying all the numerical entropy inequalities. *Mathematics of Computation*, 65:1439–1461, 1996.

[20] P. Brenner : Rapport de recherche non publié. 2008.

[21] F. Brezzi, B. Cockburn, L.D. Marini et E. Süli : Stabilization mechanisms in Discontinuous Galerkin finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 195(25-28):3293 – 3310, 2006. Discontinuous Galerkin Methods.

[22] T. Buffard et S. Clain : Multi-slope MUSCL methods for unstructured meshes. *Preprint Univ. Clermont-Ferrand*, 2008.

[23] M.D. Buhmann et R. Fletcher : M.J.D. Powell's work in univariate and multivariate approximation theory and his contribution to optimization. Rapport technique 96-16, Seminar für Angewandte Mathematik, ETH Zurich, 1996.

[24] S. Camarri, M.V. Salvetti, B. Koobus et A. Dervieux : A low-diffusion MUSCL scheme for LES on unstructured grids. *Comput. and Fluids*, 33:1101–1129, 2004.

[25] S. Candel : *Mécanique des fluides*. Bordas Paris, 1990.

[26] L.A. Catalano : A new reconstruction scheme for the computation of inviscid compressible flows on 3D unstructured grids. *Int. J. Numer. Meth. Fluids*, 40:273–279, 2002.

[27] C. Chainais-Hillairet : Second order finite volume schemes for a nonlinear hyperbolic equation : error estimate. *M2AS*, 23:467–490, 2000.

[28] B. Cockburn : Devising Discontinuous Galerkin methods for non-linear hyperbolic conservation laws. *Journal of Computational and Applied Mathematics*, 128(1-2):187 – 204, 2001.

[29] B. Cockburn, G. Kanschat et D. Schötzau : The local Discontinuous Galerkin method for linearized incompressible fluid flow : a review. *Computers & Fluids*, 34(4-5):491 – 506, 2005.

[30] B. Cockburn, S.-Y. Lin et C.W. Shu : TVB Runge-Kutta local projection Discontinuous Galerkin finite element method for conservation laws III : One-dimensional systems. *Journal of Computational Physics*, 84(1):90 – 113, 1989.

[31] B. Cockburn et C.W. Shu : The local Discontinuous Galerkin method for time-dependent convection diffusion systems. *SIAM J. Numer. Anal.*, 35(6):2440–2463, 1998.

[32] B. Cockburn et C.W. Shu : The Runge-Kutta Discontinuous Galerkin method for conservation laws V : Multidimensional systems. *Journal of Computational Physics*, 141(2):199 – 224, 1998.

[33] B. Cockburn et C.W. Shu : Runge-Kutta Discontinuous Galerkin methods for convection-dominated problems. *J. Sci. Comput.*, 16(3):173–261, 2001.

[34] I. M. Cohen et P. K. Kundu : *Fluid mechanics*. Elsevier Academic Press, 3rd édition, 2004.

[35] P.M. Cohn : *Classic Algebra*. John Wiley & Sons, 2000.

[36] P. Colella : Glimm's method for gas dynamics. *SIAM J. Sci. Comput.*, 3 (1):76–110, 1982.

[37] F. Coquel et P.G. LeFloch : An entropy satisfying muscl scheme for systems of conservation laws. *Numerische Mathematik*, 74:1–33, 1996.

[38] F. Coquel et M.S. Liu : Stable and low diffusive hybrid upwind splitting methods. *Dans Computational Fluid Dynamics 92*, numéro A95-95357, pages 9 – 16. Proceedings of the European Computational Fluid Dynamics Conference Brussels, 1992.

[39] F. Coquel et B. Perthame : Relaxation of energy and approximate riemann solvers for general pressure laws in fluid dynamics. *SIAM J. Numer. Anal.*, 35(6):2223–2249, 1998.

[40] B. Courbet : Rapport de recherche Onera non publié. 2008.

[41] J-P. Croisille et P. Villedieu : A kinetic flux-splitting scheme for hypersonic flows. *Dans Proceedings of the Thirteenth International Conference on Numerical Methods in Fluid Dynamics*, volume 414, pages 310 – 314, Consiglio Nazionale delle Ricerche Rome, Italy, 1993. Thirteenth International Conference on Numerical Methods in Fluid Dynamics.

[42] L. Cueto-Felgueroso, I. Colominas, J. Fe, F. Navarrina et M. Casteleiro : High-order finite volume schemes on unstructured grids using moving least-squares reconstruction. Application to shallow water dynamics. *Int. J. Numer. Meth. Eng.*, 65:295–331, 2006.

[43] C. de Boor : *Multivariate Approximation*. Oxford Univ. Press, 1987.

[44] M. Delanaye : *Polynomial Reconstruction Finite Volume Schemes for the Compressible Euler and Navier-Stokes Equations on Unstructured Adaptive Grids*. Thèse de doctorat, Université Liège, Faculté des Sciences Appliquées, 1996.

[45] K. Van den Abeele, C. Lacor et Z.J. Wang : On the connection between the spectral volume and the spectral difference method. *Journal of Computational Physics*, 227(2):877 – 885, 2007.

[46] B. Despres : An explicit a priori estimate for a finite volume approximation of linear advection on non-cartesian grids. *SIAM J. Numer. Anal.*, 42(2):484–504, 2004.

[47] P. Deuflhard et F. Bornemann : *Scientific Computing with Ordinary Differential Equations*. Springer-Verlag, 2002.

[48] R. Eymard, T. Gallouët et R. Herbin : The finite volume method. *Dans* P.G. Ciarlet et J.L. Lions, éditeurs : *Handbook of Numerical Analysis*, volume 7, pages 715–1022. North-Holland, 2000.

[49] A. Favre, M. Coantic, R. Dumas, J. Gaviglio, L. S. G. Kovasznay et E. A. Brun : *La Turbulence en mécanique des fluides : bases théoriques et expérimentales, méthodes statistiques*. Gauthier-Villars, Paris, 1976.

[50] N. Forestier, L. Jacquin et P. Geffroy : The mixing layer over a deep cavity at high-subsonic speed. *J. Fluid Mech.*, 475:101–145, 2003.

[51] G.P. Galdi : *An Introduction to the Mathematical Theory of the Navier-Stokes Equations : Volume 2 : Nonlinear Steady Problems*. Springer, 2001.

[52] E. Godlewski et P-A. Raviart : *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Springer-Verlag, 2002.

[53] S. K. Godunov : A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb. (N.S.)*, 47(89)(3):271–306, 1959.

[54] G.H. Golub et C.F. Van Loan : *Matrix computations*. John Hopkins Univ. Press, 3rd édition, 1996.

[55] J. B. Goodman et R. Leveque : On the accuracy of stable schemes for 2D scalar conservation laws. *Math. Comp.*, 45:955–970, 1985.

[56] S. Gottlieb, C.W. Shu et E. Tadmor : Strong stability-preserving high-order time discretization methods. *SIAM Review*, 43(1):89–112, 2001.

[57] F. Haider, B. Courbet et J-P. Croisille : Stabilité du schéma volumes finis MUSCL sur maillage quelconque. Rapport Technique, ONERA, 2008.

[58] R. Harris, Z.J. Wang et Y. Liu : Efficient quadrature-free high-order spectral volume method on unstructured grids : Theory and 2D implementation. *Journal of Computational Physics*, 227(3):1620 – 1642, 2008.

[59] A. Harten : High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357 – 393, 1983.

[60] A. Harten et S. Chakravarthy : Multidimensional ENO schemes for general geometries. Rapport technique 91-76, NASA-ICASE,Langley Research Center, Hampton., 1991.

[61] A. Harten et S. Osher : Uniformly high-order accurate nonoscillatory schemes. I. *SIAM Journal on Numerical Analysis*, 24(2):279–309, 1987.

[62] C. Hirsch : *Numerical Computation of Internal and External Flows*, volume 1. John Wiley & Sons, 1988.

[63] C. Hirsch : *Numerical Computation of Internal and External Flows*, volume 2. John Wiley & Sons, 1988.

[64] M. Holt : Review of Godunov methods. Rapport technique 96-525, NASA-ICASE, 1996.

[65] R.A. Horn et C.R. Johnson : *Matrix Analysis*. Cambridge Univ. Press, 1985.

[66] R.A. Horn et C.R. Johnson : *Topics in Matrix Analysis*. Cambridge Univ. Press, 1991.

[67] A.S. Householder : *The Theory of Matrices in Numerical Analysis*. Dover Publications, New-York, 1975.

[68] M.E. Hubbard : Multidimensional slope limiters for MUSCL-type finite volume schemes on unstructured grids. *J. Comput. Phys.*, 155:54–74, 1999.

[69] A. Iserles : *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Univ. Press, 1996.

[70] A. Iske et T. Sonar : On the structure of function spaces in optimal recovery of point data for ENO-schemes by radial basis functions. *Numerische Mathematik*, 74:177–201, 1996.

[71] S. Khosla, P.J. Dionne, M.E. Lee et C.E. Smith : Using fourth order spatial integration on unstructured meshes to reduce LES run time. Numéro AIAA 2008-782. 46th AIAA Aerospace Sciences Meeting and Exhibit, AIAA, January 2008.

[72] A. N. Kolmogorov : Dissipation of energy in the locally isotropic turbulence. *Royal Society of London Proceedings Series A*, 434:15–17, juillet 1991.

[73] D. Kröner : *Numerical Schemes for Conservation Laws*. John Wiley and Sons, Chichester, and B.G. Teubner, Stuttgart, 1997.

[74] F. Lafon et S. Osher : High order filtering methods for approximating hyperbolic systems of conservation laws. *Journal of Computational Physics*, 96(1):110 – 142, 1991.

[75] L. Larchevêque : *Simulation des grandes échelles de l'écoulement au-dessus d'une cavité*. Thèse de doctorat, Université Pierre et Marie Curie Paris VI, 2003.

[76] L. Larchevêque, P. Sagaut, I. Mary et O. Labbéo : Large-eddy simulation of a compressible flow past à deep cavity. *Phys. Fluid*, 15(1):193–210, 2003.

[77] A. Leonard : Energy cascade in large-eddy simulations of turbulent fluid flows. *Adv. Geophys.*, 18(A):237–248, 1974.

[78] N. Leterrier : *Discrétisation spatiale en maillage non structuré général*. Thèse de doctorat, Paris 6 et ONERA, 2003.

[79] Randall J. LeVeque : *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[80] Y. Liu, M. Vinokur et Z.J. Wang : Spectral (finite) volume method for conservation laws on unstructured grids V : Extension to three-dimensional systems. *Journal of Computational Physics*, 212(2):454 – 472, 2006.

[81] N. Lupoglazoff, G. Rahier et F. Vuillot : Application of the Cedre unstructured flow solver to jet noise computations. First European Conference for Aerospace Sciences (EUCASS), 2005.

[82] D. Mavriplis : Revisiting the least-squares procedure for gradient reconstruction on unstructured meshes. NIA Report 2003-06, NASA, 2003.

[83] J-L. Montagné : Etude de schémas numeriques décentrés en dynamique des gaz bidimensionnelle. *La Recherche Aerospatiale*, 5:323–338, 1984.

[84] J-L. Montagné, H. Yee et M. Vinokur : Comparative study of high-resolution shock capturing schemes for a real gas. *Dans Proceeding of the 7.th GAMM Conf.,*, Louvain la Neuve, Sept. 1987.

[85] J. Von Neumann et R. D. Richtmyer : A method for the numerical calculation of hydrodynamic shocks. *Journal of Applied Physics*, 21(3):232–237, 1950.

[86] P. J. O'Rourke et M. S. Sahota : A variable explicit/implicit numerical method for calculating advection on unstructured meshes. *J. Comput. Phys.*, 143(2):312–345, 1998.

[87] S. Osher et F. Solomon : Upwind difference schemes for hyperbolic systems of conservation laws. *Mathematics of Computation*, 38(158):339–374, Apr. 1982.

[88] B. Perthame et Y. Qiu : A variant of Van Leer's method for multidimensional systems of conservation laws. *Journal of Computational Physics*, 112:370–381, 1994.

[89] W.H. Reed et T.R. Hill : Triangular mesh methods for the neutron transport equation. Rapport technique LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, New Mexico, 1973.

[90] M. Ricchiuto, N. Villedieu, R. Abgrall et H. Deconinck : On uniformly high-order accurate residual distribution schemes for advection-diffusion. *Journal of Computational and Applied Mathematics*, 215(2):547 – 556, 2008. Proceedings of the Third International Conference on Advanced Computational Methods in Engineering (ACOMEN 2005)., Proceedings of the Third International Conference on Advanced Computational Methods in Engineering (ACOMEN 2005).

[91] P.L. Roe : Approximate Riemann solvers, parameter vectors and difference schemes. *Journal of Computational Physics*, 43:357–372, 1981.

[92] P. Sagaut : *Large Eddy Simulation for Incompressible Flows*. Springer, 2nd ed. édition, 2002.

[93] J. M. Seiner, M.K. Ponton, B. J. Jansen et N. T. Lagen : The effects of temperature on supersonic jet noise emission. *DGLR/AIAA*, 92-02-046:295–307, 1992.

[94] C.W. Shu, G. Erlebacher, T.A. Zang, D. Whitaker et S. Osher : High-order ENO schemes applied to two- and three-dimensional compressible flows. *Appl. Num. Math.*, 9:45–71, 1992.

[95] C.W. Shu et S. Osher : Efficent implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, 77:439–471, 1988.

[96] C.W. Shu et S. Osher : Efficent implementation of essentially non-oscillatory shock-capturing schemes, II. *Journal of Computational Physics*, 83:32–78, 1989.

[97] T. Sonar : Optimal recovery using thin plate splines in finite volume methods for the numerical solution of hyperbolic conservation laws. *IMA Journal of Numerical Analysis*, 16:549–581, 1996.

[98] T. Sonar : On the construction of esentially non-oscillatory finite volume approximations to hyperbolic conservation laws on general triangulations : polynomial recovery, accuracy and stencil selection. *Computer Methods in Applied Mechanics and Engineering*, 140:157–181, 1997.

[99] T. Sonar : On families of pointwise optimal finite volume ENO approximations. *SIAM J. Numer. Anal.*, 35:2350–2369, 1998.

[100] Y. Sun, Z.J. Wang et Y. Liu : Spectral (finite) volume method for conservation laws on unstructured grids VI : Extension to viscous flow. *Journal of Computational Physics*, 215(1):41 – 58, 2006.

[101] Ambady Suresh : Positivity-preserving schemes in multidimensions. *SIAM Journal on Scientific Computing*, 22(4):1184–1198, 2000.

[102] P. K. Sweby : High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM Journal on Numerical Analysis*, 21(5):995–1011, 1984.

[103] E. Tadmor : Approximate solutions of nonlinear conservation laws. *Dans* A. Quarteroni, éditeur : *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, Lecture Notes in Mathematics*, numéro 1697. Springer Verlag, 1997.

[104] R. Temam : *Navier-Stokes Equations : Theory and Numerical Analysis*. AMS Chelsea, 3rd édition, 2001.

[105] B. van Leer : Towards the ultimate conservative difference scheme. I. The quest of monotonicity. *Dans Lecture Notes in Physics*, volume 18, pages 163–168. Third International Conference on Numerical Methods in Fluid Mechanics, Springer, 1973.

[106] B. van Leer : Towards the ultimate conservative difference scheme. II. Monotonicity and conservation combined in a second-order scheme. *Journal of Computational Physics*, 14(4):361 – 370, 1974.

[107] B. van Leer : Towards the ultimate conservative difference scheme III. Upstream-centered finite-difference schemes for ideal compressible flow. *Journal of Computational Physics*, 23(3):263 – 275, 1977.

[108] B. van Leer : Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. *Journal of Computational Physics*, 23(3):276 – 299, 1977.

[109] B. van Leer : Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. *Journal of Computational Physics*, 32(1):101 – 136, 1979.

[110] R. S. Varga : *Geršgorin and His Circles*. Springer Verlag, 2004.

[111] B. Vreman : *Direct and large-eddy simulation of the compressible turbulent mixing layer*. Thèse de doctorat, University of Twente, 1995.

[112] B. Vreman, B. Geurts et H. Kuerten : A priori tests of large-eddy simulation of compressible plane mixing layer. *J. Eng. Math.*, 29:299–327, 1995.

[113] H. Wackernagel : *Multivariate geostistics*. Springer, 2003.

[114] N.J. Walkington : Quadrature on simplices of arbitrary dimension. Rapport technique 00-CNA-023, Carnegie Mellon University, Pittsburgh, 2000.

[115] Z. J. Wang : Spectral (finite) volume method for conservation laws on unstructured grids : Basic formulation. *Journal of Computational Physics*, 178(1):210 – 251, 2002.

[116] Z. J. Wang et Y. Liu : Spectral (finite) volume method for conservation laws on unstructured grids : II. extension to two-dimensional scalar equation. *Journal of Computational Physics*, 179(2):665 – 697, 2002.

[117] Z. J. Wang, L. Zhang et Y. Liu : Spectral (finite) volume method for conservation laws on unstructured grids IV : extension to two-dimensional systems. *Journal of Computational Physics*, 194(2):716 – 741, 2004.

[118] Z.J. Wang et Y. Liu : Extension of the spectral volume method to high-order boundary representation. *Journal of Computational Physics*, 211(1):154 – 178, 2006.

[119] R. F. Warming et B. J. Hyett : The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of Computational Physics*, 14(2):159 – 179, 1974.

[120] H. Wendland : *Scattered Data Approximation*. Cambridge Univ. Press, 2004.

[121] H. Wendland : On the convergence of a general class of finite volume methods. *SIAM J. Numer. Anal.*, 43(3):987–1002, 2005.

[122] P. Woodward et P. Colella : The numerical simulation of two-dimensional fluid flow with strong shocks. *Journal of Computational Physics*, 54(1):115 – 173, 1984.

[123] H. C. Yee, R. F. Warming et A. Harten : On a class of TVD schemes for gas dynamic calculations. *Dans Proc. of the sixth international symposium on computing methods in applied sciences and engineering, VI*, pages 491–492, Amsterdam, The Netherlands, 1985. North-Holland Publishing Co.

[124] H.C. Yee, R.F. Warming et A. Harten : Application of TVD schemes for the Euler and gas dynamics. *Dans* R. L. Lee, R. L. Sani, T. M. Shih et P. M. Gresho, éditeurs : *Large-scale computations in fluid mechanics ; Fifteenth Summer Seminar on Applied Mathematics, La Jolla, CA, June 27-July 8, 1983.*, numéro A85-48201 23-34, pages 357–377. American Mathematical Society, 1985.

[125] H.Q. Yee : Upwind and symmetric shock capturing schemes. Rapport technique 89464, NASA-technical Memorandum, 1987.

Index

Discrétisation en maillage non structuré général et applications LES

L'objectif est d'améliorer la stabilité et la précision de la discrétisation spatiale de type volumes finis sur des maillages non structurés. L'intérêt réside dans l'application croissante des volumes finis à la simulation des grandes échelles (LES) qui exige une discrétisation précise. Un autre objectif est le développement d'algorithmes permettant de reconstruire les polynômes de degré élevé en maillage non structuré sur de petits voisinages (stencils).

L'étude commence par une analyse générale de la reconstruction des polynômes de degré k en maillage non structuré et une étude numérique de la convergence en maillage pour la reconstruction des polynômes de degré 2 et 3. L'étude présente plusieurs algorithmes permettant de reconstruire des polynômes sur de petits voisinages. Des expériences numériques confirment l'ordre d'approximation de ces méthodes pour les polynômes de degré 2 en dimension 2.

L'étude théorique de la stabilité dégage des principes généraux pour concevoir des méthodes de reconstruction stables. L'étude théorique de la précision caractérise les erreurs induites par le maillage non structuré à l'aide de l'approche de l'équation modifiée. Ces études sont complétées par des expériences numériques.

L'étude formule également des algorithmes de limitation en maillage non structuré basés sur une approche géométrique.

Les calculs LES d'un écoulement subsonique au-dessus d'une cavité et d'un jet supersonique permettent de valider et comparer plusieurs options de discrétisation spatiale implémentées dans le code Cedre de l'Onera. Les résultats de l'étude de stabilité permettent d'améliorer le calcul du jet en maillage de tétraèdres.

Mots-clés : discrétisation spatiale, méthode de volumes finis, Muscl, maillage non structuré général, lois de conservation, mécanique des fluides compressibles, simulation des grandes échelles, SGE, turbulence ;

Discretization on general unstructured grids and applications to LES

The objective is to improve the stability and accuracy of finite volume spatial discretization on unstructured grids. The interest lies in the growing use of finite volumes for large eddy simulation (LES) that requires accurate discretization methods. Another goal is the design of algorithms capable of reconstructing polynomials of higher degree on unstructured grids using only small and compact stencils.

The study starts with a general analysis of the reconstruction of polynomials of degree k on unstructured grids, completed by numerical measurements of the convergence rate of the reconstruction error for polynomials of degree 2 and 3. The study presents algorithms for the reconstruction of polynomials on small stencils. Numerical experiments confirm the order of the approximation of these reconstruction methods for quadratic polynomials in 2 dimensions.

A theoretical stability analysis exhibits general principles for the design of stable reconstruction methods. A theoretical accuracy analysis, based on the modified equation approach, highlights the errors induced by unstructured grids. The theoretical investigations are completed and confirmed by numerical experiments.

The study of slope limiters on unstructured grids formulates algorithms based on a geometric approach.

Large eddy simulations of a subsonic flow over a cavity and of a supersonic jet allow the validation and comparison of several discretization features implemented in the code Cedre of Onera.The results of the theoretical stability analysis make it possible to obtain better results for the jet computation on tetrahedral grids.

Keywords: spatial discretization, finite volume method, Muscl, general unstructured grids, conservation laws, compressible fluid dynamics, large eddy simulation, LES, turbulence;

Les travaux de recherche ont été effectués au
Département de Simulation Numérique et Aéroacoustique de l'Onera
BP72 – 29 avenue de la Division Leclerc 92322 Châtillon Cedex
Tel +33 1 46 73 40 40 Fax +33 1 46 73 41 41
http://www.onera.fr/dsna/index.php