Asymptotic-preserving and well-balanced schemes for transport models using Trefftz discontinuous Galerkin method

Guillaume Morel

Sorbonne Université

École doctorale ED386

Laboratoire Jacques-Louis Lions

Asymptotic-preserving and well-balanced schemes for transport models using Trefftz discontinuous Galerkin method

Présentée pour l'obtention du grade de DOCTEUR
DE SORBONNE UNIVERSITÉ
par

Guillaume MOREL

Thèse de doctorat de Mathématiques appliquées

Dirigée par **Bruno DESPRÉS**
et **Christophe BUET**

Présentée et soutenue publiquement le 26 septembre 2018

Devant un jury composé de :

Président du jury	**Grégoire Allaire**	Professeur	École polytechnique
Rapporteur	**Christophe Berthon**	Professeur	Université de Nantes
Examinateur	**Alexandre Ern**	Professeur	Université Paris-Est
Examinateur	**Frédéric Hecht**	Professeur	Sorbonne Université
Directeur de thèse	**Bruno Després**	Professeur	Sorbonne Université
Co-directeur de thèse	**Christophe Buet**	Ingénieur chercheur	CEA

Après avis favorables des rapporteurs: Christophe Berthon et Ilaria Perugia.

Thèse effectuée au sein du **Laboratoire Jacques-Louis Lions**
de Sorbonne Université
4, place Jussieu, 75005 Paris, France

ainsi qu'au sein du **CEA, DAM, DIF**
F-91297 Arpajon, France

Asymptotic-preserving and well-balanced schemes for transport models using Trefftz discontinuous Galerkin method

Résumé

Cette thèse traite de l'étude et de l'analyse d'un schéma de type Trefftz Galerkin discontinu (TDG) pour un problème modèle de transport avec relaxation linéaire. Nous montrons que la méthode TDG fournie naturellement des discrétisations bien équilibrées et *asymptotic-preserving* puisque des solutions exactes, éventuellement non polynomiales, sont utilisées localement dans les fonctions de base. En particulier, la formulation de la méthode du TDG est donnée dans le cas général des systèmes de Friedrichs. En pratique, une attention particulière est consacrée à l'approximation P_N de l'équation de transport. Pour ce modèle bidimensionnel, des fonctions de base polynomiales et exponentielles sont construites et la convergence du schéma est étudiée. Les exemples numériques sur les modèles P_1 et P_3 montrent que la méthode TDG surpasse la méthode Galerkin discontinue standard pour certains tests avec termes source raides. En particulier, la méthode TDG permet d'obtenir des schémas efficaces pour capturer les couches limites et la limite de diffusion de l'équation de transport.

Mots-clés: Schémas *asymptotic-preserving* et bien équilibré, Méthode de Trefftz Galerkin discontinue, équation de transport, modèles P_N, couches limites, limite de diffusion.

Abstract

This thesis deals with the study and analysis of a Trefftz Discontinuous Galerkin (TDG) scheme for a model problem of transport with linear relaxation. We show that natural well-balanced and asymptotic-preserving discretization are provided by the TDG method since exact solutions, possibly non-polynomials, are used locally in the basis functions. In particular, the formulation of the TDG method for the general case of Friedrichs systems is given. For the practical examples, a special attention is devoted to the P_N approximation of the transport equation. For this two dimensional model, polynomial and exponential basis functions are constructed and the convergence of the scheme is studied. Numerical examples on the P_1 and P_3 models show that the TDG method outperforms the standard discontinuous Galerkin method when considering stiff coefficients. In particular, the TDG method leads to efficient schemes to capture boundary layers and the diffusion limit of the transport equation.

Keywords: Asymptotic-preserving and well-balanced schemes, Trefftz discontinuous Galerkin method, transport equation, P_N model, boundary layers, diffusion limit.

Remerciements

Tout d'abord, je tiens à remercier Bruno Després et Christophe Buet pour m'avoir accordé leur confiance en acceptant d'encadrer ce travail. Je les remercie, entre autres, pour leurs visites régulières à Teratec, leur investissement et disponibilité de tous les instants, leurs nombreux conseils et surtout leur bonne humeur permanente qui ont rendu ces trois années extrêmement agréables et enrichissantes.

Je suis très reconnaissant à Christophe Berthon et Ilaria Perugia pour le temps qu'ils ont consacré à rapporter ce document. Merci également à Grégoire Allaire pour avoir accepté de présider le jury ainsi qu'à Alexandre Ern et Frederic Hecht en avoir été les examinateurs.

Un remerciement particulier à Stéphane Del Pino et Emmanuel Labourasse pour m'avoir fait découvrir le monde de la recherche lors de mon stage de fin d'études. C'est notamment grâce à ce stage que j'ai souhaité poursuivre en thèse et j'en garde d'excellents souvenirs.

Je souhaite aussi remercier Hervé Jourdren, Cédric Enaux, Gérald Samba, Alexandra Claisse et Patricia Cargo pour les échanges enrichissants au cours de ces trois dernières années. Je remercie également Isabelle Visotto, Brigitte Sadoule et Céline Poussin pour leur aide dans les différentes tâches administratives ainsi que Thao, Denis et Yves qui ont toujours été disponibles en cas de besoin.

Je passe maintenant à Teratec et je commence par mes collègues d'open space à qui je souhaite bon courage pour la fin de leurs thèses respectives même si, avec mon départ et bientôt celui d'Hoby, l'ambiance risque d'en prendre un sacré coup. Merci à Hoby donc qui a prouvé que les meilleurs ne partent pas toujours en premier, à Éloise pour les dessins d'animaux bizarres que l'on retrouve parfois sur son bureau, à Christina, notre encadrante préférée, pour le ravitaillement régulier de l'open space, à Nestor pour avoir le courage de faire toutes ces thèses en même temps et à Théo (S.) pour ses discussions toujours passionnées. Merci aussi au shérif Hugo (B.) pour s'être occupé des bandits de Teratec, à Hugo (T.) et Arthur non présent dans l'open space mais dont les éclats de rire ne semblaient jamais vraiment très loin et à Théo (C.) un spécimen rare de post-doc exilé parmi les stagiaires. J'en profite pour remercier les nombreux stagiaires rencontrés pendant la thèse et notamment Fabien, Simon, Ludovic, Clément, Ewan, Elyès, Antony et Jérémy. Merci également à Alexis et Rémi avec qui j'ai découvert Teratec du temps où la climatisation n'existait pas encore, à Gautier (l'école doctorale le déteste, ce jeune chercheur simplifie la vie des doctorants) pour m'avoir prêté son graveur CD lors de la fin de la thèse ainsi qu'aux anciens Sébastien, Xavier, Tony et Christelle.

Je continue avec les Nantais, connaissance de (très) longue date. Merci notamment à Thomas qui attend toujours le retour de ses chaussettes avec l'impatience d'un elfe de maison, à Benjamin pour ses inspirations *insane* ainsi qu'à Audrey, Fabien, Guillaume (C.) et Axel que j'ai eu le plaisir de voir (plus ou moins) régulièrement au cours de ces trois dernières années.

Merci à Christine, Michel, Martin et Anouk pour leur présence à la soutenance, à Hélène pour son aide dans la préparation du pot ainsi qu'à Corentin ("Corentin il a fait une thèse intéressante lui", *Antoine le 26/09/2018*) pour son aide lors de la soutenance.

Je finis en remerciant ma famille qui m'a soutenu et encouragé. Merci notamment à Antoine et Marine pour les quelques gaufres dégustées cet été à la paillote (au nutella évidemment) et bien sûr à mes parents pour leur soutien inconditionnel durant les 26 dernières années.

Contents

List of Figures

Introduction

This document deals with the study and the analysis of a Trefftz Galerkin discontinuous (TDG) method applied to transport models.

Transport equations have many practical applications in biology, radiotherapy, radiative transfer or more generally astrophysics. In this work, we are interested in the transport of particles such as neutrons or photons. Such physical phenomena often involve absorption and scattering processes which may lead to complex behaviors where no analytic solutions are known. In these cases numerical methods are required.

However, the numerical approximation of the transport equation remains challenging because of the two spatially dependent absorption and scattering coefficients. In highly scattering regimes, the transport equation has a limit where the behavior of the solution is govern by a diffusion equation. It is known that naive schemes fail to give a good approximation of the diffusion limit on coarse meshes. Another potential issue comes from boundary layers which may occur in the solutions to the transport equation. In both cases, one often needs to consider very fine meshes to get a correct approximation of the solution which can drastically increase the computational time.

To address these issues, a possibility is to construct numerical methods which satisfy some particular properties. On one hand, schemes which preserve exactly stationary solutions are called well-balanced and they may be very efficient to capture boundary layers. On the other hand, the so called asymptotic-preserving schemes are able to capture the diffusion limit with reasonable computational time. Generally speaking, well-balanced and asymptotic-preserving schemes are two related concepts and it is often desirable to satisfy simultaneously these two properties.

Objective and main results

The goal of this work is to derive and analyze an asymptotic-preserving and well-balanced scheme for transport models using Trefftz discontinuous Galerkin (TDG) method. The principle of the TDG method is to use the standard discontinuous Galerkin (DG) framework but with a change of basis functions: the basis functions of the TDG scheme are solutions to the equation and therefore not necessarily polynomials.

Several original results are produced in this document. In particular, the well-balanced property of the TDG method is given in Proposition 2.13 of Chapter 2. The construction of exponential and polynomial basis functions for the general two dimensional P_N model is given in Theorems 4.25 and 4.34 of Chapter 4. Moreover, a proof of high order convergence of the TDG method applied to the stationary P_N model is given in Theorem 4.75. Finally, the asymptotic-preserving and well-balanced properties are illustrated through various numerical examples in Chapter 5.

A first article with application to the P_1 model has been published [BDM18] and others are in preparation.

Plan of the thesis

Chapter 1

In the first chapter the physical and mathematical context is given. In particular, the transport equation is introduced together with the popular S_N and P_N reduced models. Then, the two notions of asymptotic-preserving and well-balanced schemes are recalled and a definition for two dimensional well-balanced schemes is proposed. Finally, a brief bibliographical review on Trefftz method is made with particular interest in the Trefftz discontinuous Galerkin method.

Chapter 2

In Chapter 2, the Trefftz discontinuous Galerkin method is presented in the context of general Friedrichs systems. The well-balanced property of the scheme is studied and some estimates in various norms are provided. Such estimates will be useful in Chapter 4 to study the convergence of the scheme. Note however that the procedure given in this chapter does not cover the construction of the basis functions which will be treated in Chapters 3 and 4 for some particular transport models.

Chapter 3

In Chapter 3, the TDG method is applied to the one dimensional P_1 and Su-Olson model. The basis functions are constructed for these two systems with the possibility to get high order scheme in space and time. An asymptotic study of the scheme in the diffusion regime is made for the one dimensional P_1 model. Finally, the properties of the TDG method are illustrated with some numerical examples.

Chapter 4

The Chapter 4 is the central chapter of this document. It deals with the analysis of the TDG method for the general two dimensional P_N model. As a first step, the derivation and some properties of the P_N model are recalled. Then, exponential and polynomial spatial solutions are constructed together with some time dependent solutions. Finally, high order convergence of the scheme, in particular through the study of the approximation properties of the basis functions, is provided for the stationary case.

Chapter 5

In Chapter 5, the TDG method is applied to the two dimensional P_1 and P_3 models. In particular, the basis functions are explicitly calculated for these two models using the results of Chapter 4. Additionally, numerical results are provided to illustrate some properties such as the convergence, the well-balanced property (through numerical tests with boundary layers) and the asymptotic behavior of the scheme.

Chapter 6

The Chapter 6 is an independent part devoted to the study and analysis of an asymptotic-preserving multidimensional ALE method for a system of two compressible flows coupled with friction. This chapter, taken from a published article [PLM18], proposes a multidimensional

scheme to approximate solutions to a particular kind of bi-fluid system which depends on a friction parameter. Properties such as conservation, stability, consistency and asymptotic-preserving (with respect to the friction parameter) are studied. Various numerical results are also provided.

Chapter 1

Physical and mathematical context

Contents

In this Chapter, the physical and mathematical context is given. First, the transport equation is deduced from the radiative transfer equations and the popular P_N and S_N approximations are presented. Then, the main motivations behind the construction of asymptotic-preserving and well-balanced schemes are recalled. Finally, a brief bibliographical review on Trefftz methods is provided with a particular interest in the Trefftz discontinuous Galerkin method.

1-1 The transport equation

The study of the evolution of a population through transport equations has many practical applications in astrophysics, optics, atmospheric science, population dynamics or, in our case, radiative transfer. Radiative transfer is the branch of physics describing the transport of energy by electromagnetic radiation through a material medium. We consider a population of particles such as photons or neutrons and study how they travel through the material.

In this section, we briefly recall the equations of the radiative transfer [Cha50]. Seven variables are required to describe the evolution of a particle: one time variable t, three space variables $\mathbf{x} = (x, y, z)^T = (x_1, x_2, x_3)^T \in \mathbb{R}^3$, two for the direction $\mathbf{\Omega} = (\Omega_1, \Omega_2, \Omega_3)^T \in \mathbb{R}^3$ and one for the frequency $\nu \in \mathbb{R}_+$. The function of distribution of the particles reads $f := f(t, \mathbf{x}, \mathbf{\Omega}, \nu)$. A useful quantity when studying a population of photons is the radiative intensity

$$\mathcal{I}(t, \mathbf{x}, \mathbf{\Omega}, \nu) := ch\nu f(t, \mathbf{x}, \mathbf{\Omega}, \nu),$$

where c is the speed of light and h is the Planck constant. To simplify the model, we consider physical quantities averaged with respect to the frequency. We define the grey moment as

$$\mathcal{I}(t, \mathbf{x}, \mathbf{\Omega}) := \int_0^{+\infty} \mathcal{I}(t, \mathbf{x}, \mathbf{\Omega}, \nu) d\nu.$$

At the local thermal equilibrium the radiative intensity is governed by the Planck function

$$B\big(T(t,\mathbf{x})\big) = \int_0^{+\infty} \frac{2h\nu^3}{c^2[e^{\frac{h\nu}{kT}}-1]} d\nu,$$

where k is the Boltzmann constant. The interaction between the particles and their environment can be described using three mechanisms:

- The absorption,
- The scattering,
- The emission.

A particle can be absorbed by the material through the absorption coefficient $\sigma_a\big(T(t,\mathbf{x})\big) \geq 0$ and the particles interact with each other through the scattering coefficient $\sigma_s\big(T(t,\mathbf{x})\big) \geq 0$. Finally, the emission of particles depends on the Planck function B and can be written $\sigma_a\big(T(t,\mathbf{x})\big)B\big(T(t,\mathbf{x})\big)$. We can now introduce the grey (i.e. average in frequency) radiative transfer system

$$\begin{cases} \frac{1}{c}\partial_t \mathcal{I}(t,\mathbf{x},\mathbf{\Omega}) + \mathbf{\Omega}\cdot\nabla\mathcal{I}(t,\mathbf{x},\mathbf{\Omega}) = \sigma_a(\mathbf{x})\Big(B\big(T(t,\mathbf{x})\big) - \mathcal{I}(t,\mathbf{x},\mathbf{\Omega})\Big) \\ \qquad\qquad +\sigma_s(\mathbf{x})\Big(\frac{1}{4\pi}\int_{S^2} p(\Omega,\Omega')\mathcal{I}(t,\mathbf{x},\Omega')d\Omega' - \mathcal{I}(t,\mathbf{x},\mathbf{\Omega})\Big), \\ \frac{1}{c}\partial_t \mathcal{E}\big(T(t,\mathbf{x})\big) = \frac{1}{4\pi}\int_{S^2} \sigma_a\big(T(t,\mathbf{x})\big)\Big(B\big(T(t,\mathbf{x})\big) - \mathcal{I}(t,\mathbf{x},\mathbf{\Omega})\Big)d\Omega, \end{cases}$$

where $p(\Omega,\Omega')$ is an angular distribution function which defines the anisotropy of the scattering and $\mathcal{E}(T)$ is the internal energy density of the material at the temperature T. For example, when considering perfect gazes one has

$$\mathcal{E}\big(T(t,\mathbf{x})\big) = c_V T(t,\mathbf{x}),$$

where c_V is the heat capacity at constant volume of the medium. For simplicity we assume that the temperature $T(t,\mathbf{x})$ is given, there is no emission of particles (that is $B=0$) and the scattering is isotropic (that is $p(\Omega,\Omega')=1$). The grey radiative transfer equation now reads

$$\frac{1}{c}\partial_t \mathcal{I}(t,\mathbf{x},\mathbf{\Omega}) + \mathbf{\Omega}\cdot\nabla\mathcal{I}(t,\mathbf{x},\mathbf{\Omega}) = -\sigma_a(\mathbf{x})\mathcal{I}(t,\mathbf{x},\mathbf{\Omega}) + \sigma_s(\mathbf{x})\Big(<\mathcal{I}>(t,\mathbf{x}) - \mathcal{I}(t,\mathbf{x},\mathbf{\Omega})\Big), \quad (1.1)$$

where we use the following notation

$$<\cdot>(t,\mathbf{x}) = \frac{1}{4\pi}\int_{S^2} d\mathbf{\Omega},$$

with S^2 the unit sphere in $\mathbb{R}^3$.

In the following, we may refer to the equation (1.1) as the transport equation. Interesting physical phenomena depend on the coefficients σ_a and σ_s. For example, when they vary significantly boundary layers may occur. Also, in the asymptotic regime $t >> 1$ and $\sigma_s >> 1$, the transport equation (1.1) tends to a diffusion limit (see for example [ABDG15] in french)

$$\partial_t <\mathcal{I}>(t,\mathbf{x}) - \operatorname{div}\Big(\frac{1}{3\sigma_s}\nabla <\mathcal{I}>(t,\mathbf{x})\Big) + \sigma_a <\mathcal{I}>(t,\mathbf{x}) = 0.$$

1-2 Approximate models of the transport equation

In practice, equation (1.1) is difficult to solve numerically because of the large number of variables (up to three space variables, two for the direction and one time variable). Probabilistic methods

such has Monte-Carlo schemes can be used directly on the transport equation. However, for deterministic schemes one often needs to consider approximate models. In the following, we present two popular approximations to discretize the angular variable in equation (1.1). More general reviews of approximate models for the transport equation can also be found in [Bru02] or [Fra12] (in french).

1-2.1 The discrete ordinate method

The discrete ordinates method [Cha50] (or S_N method) assumes that the particles can only travel through some particular directions. As pointed in [Bru02] this is equivalent to write the density as a sum of Dirac mass

$$\mathcal{I}(t,\mathbf{x},\mathbf{\Omega}) = \sum_{i=1}^{m} \mathcal{I}_i(t,\mathbf{x})\delta(\mathbf{\Omega}-\mathbf{\Omega}_i).$$

The transport equation (1.1) then reads

$$\frac{1}{c}\partial_t \mathcal{I}_i(t,\mathbf{x}) + \mathbf{\Omega}_i\cdot\nabla\mathcal{I}_i(t,\mathbf{x}) = -\Big(\sigma_a(\mathbf{x})+\sigma_s(\mathbf{x})\Big)\mathcal{I}_i(t,\mathbf{x}) + \sigma_s(\mathbf{x})\sum_{j=1}^{m} w_j\mathcal{I}_j(t,\mathbf{x}), \quad i=1,...,m, \tag{1.2}$$

where $\mathcal{I}_i$ are the unknown and w_i the integration weights. One often choose a symmetric quadrature that is

$$\sum_{i=1}^{m} w_i = 1, \quad \sum_{i=1}^{m} w_i\mathbf{\Omega}_i = 0.$$

Moreover, to recover the correct diffusion coefficient one needs to impose [LMM87]

$$\sum_{i=1}^{m} w_i\mathbf{\Omega}_i\otimes\mathbf{\Omega}_i = \frac{1}{3}I_m,$$

where $I_m \in \mathbb{R}^{m\times m}$ is the identity matrix.

The main advantages of the S_N method is that the system (1.2) is diagonal and the positivity is preserved for each unknown. However, this system is not invariant under rotation and has a well-known defect called ray effects. These effects come from the choice of the discrete values $\mathbf{\Omega}_i$ for the directions which are then favored in the numerical simulation [Bru02].

1-2.2 Spherical harmonics approximation

The idea behind the spherical harmonic approximation (or P_N model) is to decompose the solution to (1.1) on the spherical harmonics basis

$$\mathcal{I}(t,\mathbf{x},\mathbf{\Omega}) = \sum_{k\geq 0}\sum_{|l|\leq k} Y_{k,l}(\mathbf{\Omega})u_k^l(t,x),$$

where $Y_{k,l}$ are the real or complex spherical harmonics. The P_N approximation assumes that if $k>N$ then the moments satisfy $u_k^l = 0$. Multiplying by $Y_{k,l}$, integrating over the direction and using the recursion relations of the spherical harmonics one finally gets a system of the form

$$\partial_t\mathbf{u}(t,\mathbf{x}) + \sum_{i=1}^{3} A_i\partial_{x_i}\mathbf{u}(t,\mathbf{x}) = -R\mathbf{u}(t,\mathbf{x}),$$

where $\mathbf{u}(t,\mathbf{x})$ is the unknown, R is a diagonal positive matrix and the matrices A_i have the following block structure [Her16]

$$A_1 = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & C \\ C^T & 0 \end{pmatrix}.$$

A more detailed construction of the P_N model and some properties of this system will be given in Chapter 4.

In particular, a nice property of the P_N model is that its solutions are invariant under rotation. However, a well-known defect of the P_N method is that it can lead to negative density when considering time dependent case [Bru02]. There have been several attempts to address this problem [BH01, HM10, MH10, Ols12]. Among them, a popular approach is the so-called filtered P_N (FP_N) method [FHK16, MH10, RARO13].

1-3 Asymptotic-preserving and well-balanced schemes

1-3.1 Asymptotic-preserving schemes

The introduction of a small parameter ε in an equation is often a good way to model a particular physical behavior. For example, the parameter ε may represent some scaling of the coefficients or different time scales of the physical quantities. In practice, the parameter ε may vary in the domain and it is therefore mandatory to derive a numerical scheme which behaves well whatever the parameter value is. Naive schemes may fail to capture the limit $\varepsilon \to 0$ on coarse meshes typically because the error behaves as $O(\frac{\Delta x}{\varepsilon})$. On the contrary, schemes which are able to capture efficiently the limit $\varepsilon \to 0$ have been called asymptotic-preserving (AP) schemes.

Definition 1.1 (Asymptotic-preserving schemes)**.** A scheme is said to be asymptotic-preserving (AP) if its consistency error does not depend on ε in the limit $\varepsilon \to 0$.

AP schemes have been applied to a wide range of kinetic and hyperbolic equations, see, for example, the review [Jin10]. For the transport equation, it is known that under the correct scaling it tends to a diffusion limit. To capture the diffusion limit with reasonable computational time, asymptotic-preserving schemes have been introduced [JL91, JL96] and applied to transport problems [BT11, BDF15, Fra12, Gos13, GT02, Jin10, RGK12]. Such schemes are usually obtained with a modification of the fluxes by including a dependence in ε in the new fluxes.

A typical example is the hyperbolic heat equation in dimension one

$$\begin{cases} \partial_t p + \dfrac{1}{\varepsilon}\partial_x v = 0, \\ \partial_t u + \dfrac{1}{\varepsilon}\partial_x p = -\dfrac{\sigma_s}{\varepsilon^2} v. \end{cases}$$

Here the unknown are (p, v), $\sigma_s \in \mathbb{R}^+$ and $0 < \varepsilon \leq 1$. In particular, when $\varepsilon \to 0$ the variable p follows a diffusion equation (see Chapter 3 for details)

$$\partial_t p - \partial_x\Big(\frac{1}{\sigma_s}\partial_x p\Big) = 0.$$

To show why it can be challenging for numerical method to capture the diffusion limit, consider a standard finite volume scheme written under the form

$$\begin{cases} \dfrac{p_j^{n+1} - p_j^n}{\Delta t} + \dfrac{v_{j+\frac{1}{2}}^n - v_{j-\frac{1}{2}}^n}{\varepsilon \Delta x} = 0, \\ \dfrac{v_j^{n+1} - v_j^n}{\Delta t} + \dfrac{p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n}{\varepsilon \Delta x} = -\dfrac{\sigma_s}{\varepsilon^2} v_j, \end{cases} \quad , \quad j = 0, ..., N, \tag{1.3}$$

where Δt, Δx are the time and space step, p_j and v_j the approximations of p and v in the cell j. To get the fluxes $p_{j\pm\frac{1}{2}}$ and $v_{j\pm\frac{1}{2}}$, a first possible choice is to solve the associated Riemann

problem at the interface. One gets after simplification

$$\begin{cases} p^n_{j+\frac{1}{2}} = \frac{1}{2}\Big(p^n_j + p^n_{j+1} + v^n_j - v^n_{j+1}\Big), \\ v^n_{j+\frac{1}{2}} = \frac{1}{2}\Big(v^n_j + v^n_{j+1} + p^n_j - p^n_{j+1}\Big). \end{cases} \tag{1.4}$$

However with this particular choice of fluxes the scheme is not AP.

Proposition 1.2. *The consistency error of the scheme* (1.3) *with the fluxes* (1.4) *is* $O(\frac{\Delta x}{\varepsilon} + \Delta t)$. *Therefore the scheme* (1.3)-(1.4) *is not AP in the limit* $\varepsilon \to 0$.

Proof. See for example [BDF12]. ■

A possible way to get an AP scheme is to modify the source term in the scheme (1.3)

$$\begin{cases} \dfrac{p^{n+1}_j - p^n_j}{\Delta t} + \dfrac{v^n_{j+\frac{1}{2}} - v^n_{j-\frac{1}{2}}}{\varepsilon \Delta x} = 0, \\ \dfrac{v^{n+1}_j - v^n_j}{\Delta t} + \dfrac{p^n_{j+\frac{1}{2}} - p^n_{j-\frac{1}{2}}}{\varepsilon \Delta x} = -\dfrac{\sigma_s}{2\varepsilon^2}\Big(v_{j+\frac{1}{2}} + v_{j-\frac{1}{2}}\Big), \end{cases} \tag{1.5}$$

and consider the fluxes

$$\begin{cases} p^n_{j+\frac{1}{2}} = \frac{1}{2}\Big(p^n_j + p^n_{j+1} + v^n_j - v^n_{j+1}\Big), \\ v^n_{j+\frac{1}{2}} = \dfrac{1}{2(1+a)}\Big(v^n_j + v^n_{j+1} + p^n_j - p^n_{j+1}\Big), \end{cases} \tag{1.6}$$

with $a = \frac{\sigma_s \Delta}{2\varepsilon}$. The scheme (1.5)-(1.6) has been proposed by Gosse and Toscani [GT02]. This scheme is AP.

Proposition 1.3. *The consistency error of the scheme* (1.5) *with the fluxes* (1.6) *is* $O(\Delta x + \Delta t)$. *Therefore the scheme* (1.5)-(1.6) *is AP in the limit* $\varepsilon \to 0$.

Proof. See [BDF12, GT02]. ■

We compare the behavior of an AP and a non AP scheme applied to the hyperbolic heat equation in Figure 1.1. The Figure 1.1 shows that, in the diffusive regime ($\varepsilon << 1$), naive schemes needs lots of degrees of freedom to approximate correctly the limit solution. On the contrary, even with few degrees of freedom, the AP scheme captures the numerical solution very well.

1-3.2 Well-balanced schemes

A concept which is strongly related to the asymptotic-preserving property is the notion of well-balanced (WB) schemes. The common definition for a well-balanced scheme is a scheme which preserves stationary solutions to the model. Such schemes are related to AP schemes since the stationary states can be seen as limit solutions when $t >> 1$. Well-balanced schemes have been introduced in [GL96] and, since then, have been widely used [BPV03, DB16, Gos13, GT02, Jin04, JTH09, LeV98, MDBCF16]. They have several advantages:

- they can improve the numerical calculation when considering stiff source terms,
- they increase the accuracy of the scheme around the steady states,
- they can be a good starting point to derive efficient AP schemes.

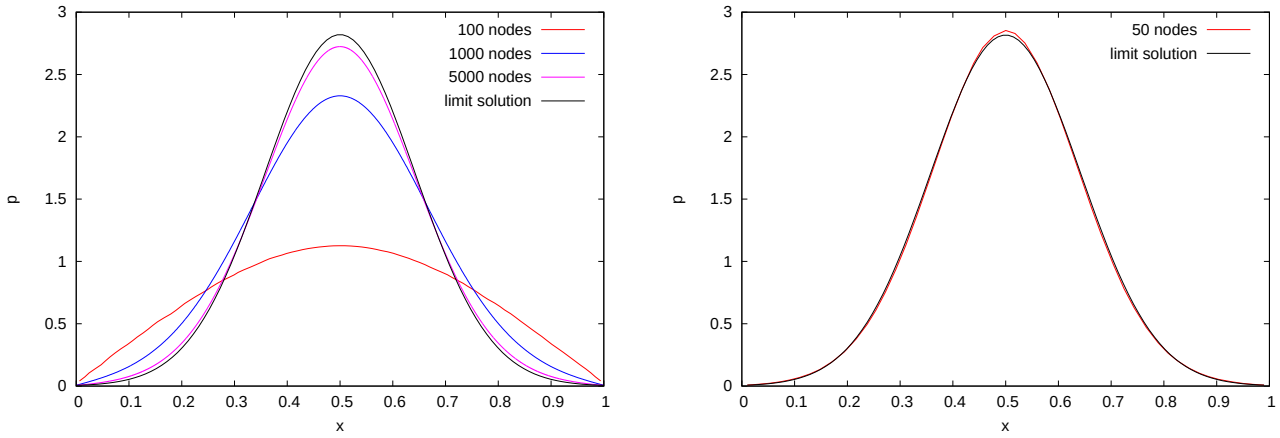

Figure 1.1 – Representation of the variable p of the hyperbolic heat equation when $\varepsilon = 10^{-3}$ and comparison with the limit solution. On the left: numerical solution obtained with a non-AP scheme. On the right: numerical solution obtained with an AP scheme.

For our applications in particular, we expect well-balanced schemes to be able to capture efficiently boundary layers.

Schemes which are both asymptotic-preserving and well-balanced have been designed and studied in one dimension [GT02, JTH09]. However, direct extensions in higher dimensions may fail to capture boundary layers [Tan09]. In general, and except in some particular cases, two dimensional asymptotic-preserving schemes are not well-balanced. It comes from the particular definition of a well-balanced scheme in one dimension.

Definition 1.4 (Well-balanced scheme in 1D)**.** A one dimensional scheme is said to be well-balanced if it preserves all the stationary states.

There is an important difference between the one-dimensional case and higher dimensions. In one dimension, a scheme is well-balanced if it captures all the stationary states of a hyperbolic system. This is possible because, in one dimension, the number of linearly independent stationary solutions is finite.

However, in two dimensions, the space of stationary solutions becomes infinite. It has a huge impact on what is a well-balanced scheme in space dimensions higher than one. For a two dimensional well-balanced scheme, one chooses a finite subset of solutions for which the scheme will be exact.

Definition 1.5 (Well-balanced scheme in 2D)**.** A two dimensional scheme will be said to be well-balanced for some solutions to the model if it is exact for any linear combinations of these solutions.

Note that since it could also be a good idea to preserve time dependent states, we do not restrict the Definition 1.5 to stationary states.

We illustrate why the concept of well-balanced schemes plays an important role in the numerical approximation of solutions to the transport equation. We consider the P_1 model in one dimension which is a very simple approximation of the transport equation (see Chapters 3 and 4 for details). The Figure 1.2 shows that well-balanced schemes may be very efficient to capture boundary layers (in this case because the stationary states are exponential solutions) which is not the case for some other naive schemes.

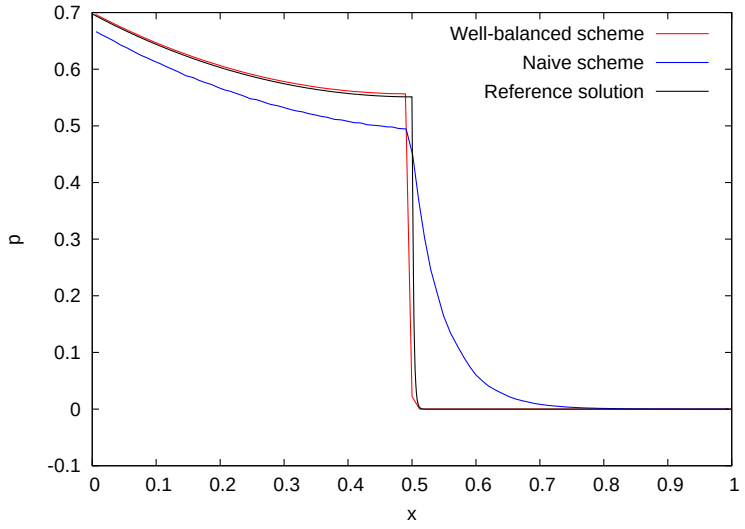

Figure 1.2 – Representation of the first variable of the P_1 model. Comparison between a well-balanced and a naive scheme for a test case with a boundary layer at $x = 0.5$.

1-4 Trefftz methods

1-4.1 Trefftz and related methods

The idea of adding information about the solution in the basis functions is known, in some cases, to greatly improve the quality of the numerical solutions. The so-called enrichment methods include for example the partition of unity method (PUM) [MB96, BM96, GS00], the Generalized finite element method (GFEM) [SBC00, SCB00], and the extended finite element method (XFEM) [BB99, MDB99]. The enrichment methods are based on a standard polynomial basis which may come from the finite element method or the discontinuous Galerkin method. The idea is then to add locally some special basis functions to the approximation space. These special basis functions may be, for example, solutions or asymptotic solutions to the model. This can be very useful when considering physical problems which involve discontinuities, singularities or high gradients. For detailed reviews of these methods, see [AH08, FB10] and reference therein. In this document, we are interested in the Trefftz discontinuous Galerkin (TDG) method which uses *only* solutions to the equation as basis functions.

The name of the Trefftz methods comes from the seminal paper of E. Trefftz which has recently been translated in English [Mau03]. In his paper, Trefftz proposed the new concept of using trial functions which satisfy the governing differential equations (for Trefftz it was the 2D Laplace problem). At the time, the benefits of using such trial functions was to obtained a lower bound of the error. This lower bound combined with the upper bound given by the Ritz method allowed Trefftz to give a general bound of the error.

Since then, Trefftz methods gained in popularity and have been applied on various problems. In particular, Trefftz methods have been widely used for the Helmholtz equation, see for example the reviews [Luo13, Chapter 3], [Moi11, Chapter 1], [PvHVD07, HMP16b] and reference therein. For review of Trefftz methods applied on other type of equations see [KK95, Li08, Qin05, CZ97].

1-4.2 The Trefftz discontinuous Galerkin method

The discontinuous Galerkin (DG) methods have been introduced by Reed and Hill in 1973 for solving the steady neutron transport equation [RH73] and the first mathematical analysis was performed by LeSaint and Raviart in 1974 [LR74]. Since then, the DG methods have been successfully applied to a large range of problems, see for example [CKS99, DPE11, HW07] and reference therein. The DG methods combine feature of the finite element and finite volume methods. In particular, they depend on a weak formulation of the problem and on the choice of finite dimensional trial and test spaces of piecewise continuous functions. In the following, we will call basis functions, the functions which belong to the trial and test spaces. A classical choice is to consider polynomials basis functions.

In this document, we are interested in the Trefftz discontinuous Galerkin (TDG) method which combined the discontinuous Galerkin framework with Trefftz's original idea. This method has been somehow rediscovered by Cessenat and Despres [CD98] with the ultra weak variational formulation (UWVF), see also [HMK02, BM08, IG13, IGD14, IG15b]. Later, it has been noticed that the UWVF is in fact equivalent to a DG method with a special choice of basis functions [BM08, GA07, HMM07]. Since the basis functions of this formulation were solutions to the equation, it has taken the name of Trefftz discontinuous Galerkin method. The reformulation of the UWVF into the DG formalism allows to use all the techniques of analysis developed in the DG framework. In particular, for the TDG method applied to the Helmholtz equation, it has been used to study h-convergence [KMPS16], p-convergence [HMP11] and even hp-convergence [HMP16a].

TDG methods have their pros and cons.

- **Pros**:
  - — Incorporate a priori knowledge in the basis functions which are therefore well adapted to multiscale problems.
  - — Often need less degrees of freedom to reach a given accuracy. A typical example is the 2D version of the P_1 model in the dominant absorption regime $\sigma_a > 0$ illustrated in the table below where we compare the number p of basis functions needed to achieve a given fractional order. The first line is for the TDG method. One gets $p_{\text{TDG}} = 2(\text{order} + 1)$ which is a rephrasing of the result of Theorem 4.75 given in Chapter 4 for the case $N = 1$. The second line is the optimal number of basis function for a general DG method $p_{\text{DG}} = \frac{3}{2}(\text{order} + \frac{1}{2})(\text{order} + \frac{3}{2})$.

order	1/2	3/2	5/2	7/2	9/2
p_{TDG}	3	5	7	9	11
p_{DG}	3	9	18	30	45

    In particular, the number of basis functions is the same to get order $= 1/2$ and one always gets $p_{\text{TDG}} \leq p_{\text{DG}}$.
  - — Is easy to incorporate in DG codes since one only needs to change the basis functions.
- **Cons**:
  - — May suffer ill-conditioning due to poor linear independence of the basis functions [CD98, HMK02]. For wave problems, some remedies exist in the literature [GHP09].
  - — The practical calculation of the basis functions adds to the computational burden. If one can calculate the basis functions analytically, the computational burden is moderate. If it is not the case, the computational burden is heavier: several options could be considered such as computing numerically the basis functions or relying on a general procedure [IGD14, IG15a, IG15b].

In the following, we briefly recall the idea behind the TDG method. Consider the following model problem with Dirichlet boundary condition

$$\begin{aligned} &\mathcal{L}\mathbf{u} = \mathbf{0}, \qquad \text{on } \Omega, \\ &\mathbf{u} = \mathbf{u}_{ex}, \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.7}$$

Here, $\mathbf{u}$ is the unknown, Ω is an open bounded set of $\mathbb{R}^2$ or $\mathbb{R}^3$, $\mathbf{u}_{ex}$ is the exact solution and $\mathcal{L}$ is a linear differential operator. Typically, one could think of $\mathcal{L}$ as $\mathcal{L} = \Delta$. Moreover, we assume that $\mathcal{T}_h$ is a mesh of the domain Ω and we will denote Ω_k a cell of $\mathcal{T}_h$. All these notations will be made rigorous in the Chapter 2.

The only difference between the DG and the TDG method is the choice of basis functions in the approximation space V_h. For the DG method, the basis functions are simple polynomials and the approximation space V_{DG} reads

$$V_{DG}(\mathcal{T}_h) = \Big\{ \mathbf{v} \in H^1(\mathcal{T}_h),\ \mathbf{v}_k \in \mathbb{P}_h \quad \forall \Omega_k \in \mathcal{T}_h \Big\},$$

where $\mathbb{P}_h$ is the polynomials space.

For the TDG method, the basis functions are exact solutions to the equation and the approximation space reads

$$V_{TDG}(\mathcal{T}_h) = \Big\{ \mathbf{v} \in H^1(\mathcal{T}_h), \quad \mathcal{L}\mathbf{v}_k = \mathbf{0} \quad \forall \Omega_k \in \mathcal{T}_h \Big\}.$$

Now, we assume one can apply the standard discontinuous Galerkin method to the model problem (1.7) and we denote $a_{DG}(\cdot,\cdot)$ and $l(\cdot)$ respectively the bilinear and linear form obtained with the DG method [DF15, DPE11, HW07], see also Chapter 2.

Definition 1.6. Assume $V_h(\mathcal{T}_h)$ is a finite subspace of $V_{DG}(\mathcal{T}_h)$ or $V_{TDG}(\mathcal{T}_h)$. The standard DG/TDG method reads

$$\begin{cases} \text{find } \mathbf{u}_h \in V_h(\mathcal{T}_h) \text{ such that} \\ a_{DG}(\mathbf{u}_h, \mathbf{w}_h) = l(\mathbf{w}_h), \quad \forall \mathbf{w}_h \in V_h(\mathcal{T}_h), \end{cases} \tag{1.8}$$

where $a_{DG}(\cdot,\cdot)$ and $l(\cdot)$ are respectively the bilinear and linear form obtained with the DG method.

Solving the formulation (1.8) is equivalent to find the solution of a linear system. Indeed, since V_h is finite-dimensional space, there exists functions $\mathbf{v}_i(t,\mathbf{x})$ such that

$$V_h = \text{Span}\Big\{ \mathbf{v}_1(t,\mathbf{x}), ..., \mathbf{v}_n(t,\mathbf{x}) \Big\},$$

for some $n \in \mathbb{N}$. The functions $\mathbf{v}_i$ are called the basis functions. Since $\mathbf{u}_h \in V_h$, one can write $\mathbf{u}_h = \sum_{i=1}^n a_i \mathbf{v}_i$, $a_i \in \mathbb{R}$. Therefore, the formulation (1.8) can be written

$$\begin{cases} \text{find } a_i \in \mathbb{R},\ i = 1, ..., n, \text{ such that} \\ a_{DG}(\sum_{i=1}^n a_i \mathbf{v}_i, \mathbf{v}_j) = l(\mathbf{v}_j), \quad j = 1, ..., n. \end{cases}$$

Using the linearity of $a_{DG}(\cdot,\cdot)$, one gets the following linear system: find $\mathbf{a} = (a_1, ..., a_n) \in \mathbb{R}^n$ such that

$$M\mathbf{a} = \mathbf{b},$$

where $M = (M_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ and $\mathbf{b} = (b_1, ..., b_n)^T \in \mathbb{R}^n$, with

$$M_{ij} = a_{DG}(\mathbf{v}_i, \mathbf{v}_j), \quad b_i = l(\mathbf{v}_i).$$

With the standard DG method, the approximation space V_h is typically made of polynomials (for example one can consider simple monomials such as $1, x, y...$). This is not the case anymore for the TDG method. In the following, we give some examples of subspace V_h when considering the TDG method. For simplicity, we consider the same basis functions $\mathbf{v}_1, ..., \mathbf{v}_k$, $k \in \mathbb{N}$, in all the cells and make a slight abuse of notation by denoting $V_h = \text{Span}(\mathbf{v}_1, ..., \mathbf{v}_k)$. The basis functions have compact support in the cell and the total number of basis functions n is then k multiplied by the number of cells. The first example that we consider is the Helmholtz equation which was the model problem used for the UWVF [CD98].

Example 1.7. The two dimensional Helmholtz equation reads

$$\Delta u = -\omega^2 u,$$

where $\omega \in \mathbb{R}$. A typical choice for V_h is then

$$V_h = \text{Span}\left\{e^{i\omega(\mathbf{d}_1,\mathbf{x})}, ..., e^{i\omega(\mathbf{d}_k,\mathbf{x})}\right\},$$

where $\mathbf{x} = (x, y)^T \in \mathbb{R}^2$, $k \in \mathbb{N}$ is the number of basis function and $\mathbf{d}_i \in \mathbb{R}^2$ are directions on the unit circle $\mathbf{d}_i = (\cos\theta_i, \sin\theta_i)^T$. ●

For a more transport related model we give the example of the P_1 model in one dimension.

Example 1.8. The P_1 model in one dimension reads

$$\begin{cases} \partial_t p + \partial_x v = -\sigma_a p, \\ \partial_t v + \partial_x p = -\sigma_t v. \end{cases}$$

The unknown is $\mathbf{u} = (p, v)^T$ and $\sigma_a, \sigma_s \in \mathbb{R}^+$, $\sigma_t = \sigma_a + \frac{\sigma_s}{\varepsilon^2}$. A possible choice for V_h is

$$V_h = \text{Span}\left\{ \begin{pmatrix} -\sqrt{\sigma_t} \\ \sqrt{\sigma_a} \end{pmatrix} e^{\sqrt{\sigma_a\sigma_t}x}, \begin{pmatrix} \sqrt{\sigma_t} \\ \sqrt{\sigma_a} \end{pmatrix} e^{-\sqrt{\sigma_a\sigma_t}x} \right\}.$$

In some regimes, it can be very interesting to consider non polynomial basis functions. When $\sigma_a\sigma_t >> 1$ for example, the two basis functions of V_h are stiff exponential functions and may therefore be very well adapted to capture boundary layers. ●

However, the basis functions are not always exponentials. Consider for example the hyperbolic heat equation in two dimensions.

Example 1.9. The hyperbolic heat equation in two dimensions reads

$$\begin{cases} \partial_t p + \text{div}\,\mathbf{v} = 0, \\ \partial_t \mathbf{v} + \nabla p = -\sigma_s \mathbf{v}, \end{cases}$$

the unknown is $\mathbf{u} = (p, \mathbf{v})^T \in \mathbb{R}^3$ and $\sigma_s \in \mathbb{R}^+$. For simplicity, we consider stationary solutions. Deriving the second equation and inserting in the first equation, one gets $\Delta p = 0$. Therefore, a possible choice for V_h is

$$V_h = \text{Span}\left\{ \begin{pmatrix} \sigma_s q_1(\mathbf{x}) \\ -\nabla q_1(\mathbf{x}) \end{pmatrix}, ..., \begin{pmatrix} \sigma_s q_k(\mathbf{x}) \\ -\nabla q_k(\mathbf{x}) \end{pmatrix} \right\},$$

where $\mathbf{x} = (x, y)^T \in \mathbb{R}^2$, $k \in \mathbb{N}$ is the number of basis functions and the functions $q_i(\mathbf{x})$ denote the two dimensional harmonic polynomials. ●

Chapter 2

Trefftz discontinuous Galerkin method for Friedrichs systems with linear relaxation

Contents

In this chapter, the TDG method for Friedrichs systems with linear relaxation is presented. After introducing such systems, the standard discontinuous Galerkin method for Friedrichs systems [DPE11, EG06, MR05] is recalled. This is particularly useful since the TDG method is simply a DG method with a special choice of basis functions. The derivation of the TDG method is then given. Finally, some error estimates are provided and the well-balanced property of the scheme is deduced from the quasi-optimality result. Note however that this chapter does not cover the general construction of the basis functions which may be a difficult point. A first look at the basis functions in some particular cases is given in the previous chapter in Examples 1.7, 1.8 and 1.9.

2-1 Friedrichs systems with linear relaxation

In this section, we present the general systems (2.1) which are considered in this document.

The method is presented in a general framework to consider both stationary and time dependent problems. Let Ω_S be a bounded polygonal/polyhedral Lipschitz space domain in $\mathbb{R}^d$ and consider a time interval $[0,T]$, $T>0$. We denote $\Omega=\Omega_S$ for stationary problems and $\Omega=\Omega_S\times[0,T]$

for time dependent problems. We consider Friedrichs systems with linear relaxation [Fri58]

$$\begin{cases} \sum_{i=0}^{d} A_i \partial_i \mathbf{u}(t,\mathbf{x}) = -R(\mathbf{x})\mathbf{u}(t,\mathbf{x}), & \text{in } \Omega, \\ M^-\mathbf{u}(t,\mathbf{x}) = M^-\mathbf{g}(t,\mathbf{x}), & \text{in } \partial\Omega. \end{cases} \tag{2.1}$$

The space variable is $\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$ and the time variable is t. The unknown is $\mathbf{u} \in \mathbb{R}^m$. Moreover the matrices

$$A_i \in \mathbb{R}^{m\times m}, \quad R(\mathbf{x}) \in \mathbb{R}^{m\times m},$$

are symmetric and we assume $R(\mathbf{x})$ is a non negative matrix, that is

$$(R(\mathbf{x})\mathbf{v}, \mathbf{v}) \geq 0, \quad \text{for all } \mathbf{v} \in \mathbb{R}^m, \ \mathbf{x} \in \mathbb{R}^d.$$

We use the notation $\partial_0 = \partial_t$, $\partial_i = \partial_{x_i}$ for $i = 1, ..., d$. For time dependent problem $\mathbf{u} = \mathbf{u}(t,\mathbf{x})$ and the matrix A_0 is a non negative matrix (and often $A_0 = I_m$). For stationary problem $\mathbf{u} = \mathbf{u}(\mathbf{x})$ and therefore $\partial_t \mathbf{u} = \mathbf{0}$. The outward normal unit vector is $\mathbf{n}(t,\mathbf{x}) = (n_t, n_{x_1}, ..., n_{x_d})$ for $(t,\mathbf{x}) \in \partial\Omega$ and of course for stationary problems $n_t = 0$ for all $\mathbf{x} \in \partial\Omega$. We set

$$M(\mathbf{n}) = A_0 n_t + \sum_{i=1}^{d} A_i n_{x_i}, \quad \text{on } \partial\Omega. \tag{2.2}$$

Since the matrices A_i are symmetric, M is also symmetric and one has the standard decomposition $M(\mathbf{n}) = M^+(\mathbf{n}) + M^-(\mathbf{n})$ where M^+ is a non negative matrix and M^- is a non positive matrix. More precisely denoting λ_i the eigenvalues of the matrix M associated with the eigenvectors $\mathbf{r}_i$ one can take

$$M^+(\mathbf{n}) = \sum_{\lambda_i>0} \lambda_i \mathbf{r}_i \mathbf{r}_i^T, \quad M^-(\mathbf{n}) = \sum_{\lambda_i<0} \lambda_i \mathbf{r}_i \mathbf{r}_i^T. \tag{2.3}$$

Finally we use the matrix M^- to write the boundary conditions with $\mathbf{g} \in L^2(\partial\Omega)$ and assume the problem (2.1) admits a unique solution [EG06].

2-2 Presentation of the method

2-2.1 Mesh notation and generic discontinuous Galerkin formulation

The partition or mesh of the space domain $\Omega = \Omega_S \subset \mathbb{R}^d$ is denoted as $\mathcal{T}_h$. It is made of polyhedral non overlapping subdomains $\Omega_{S,r}$, that is

$$\mathcal{T}_h = \cup_r \Omega_{S,r}.$$

For a space time problem, we first split the time interval into smaller time intervals (t_n, t_{n+1}) with $0 = t_0 < t_1 < ... < t_N = T$. Making an abuse of notation, the mesh of the space-time domain $\Omega = \Omega_S \times [0,T] \subset \mathbb{R}^{d+1}$ is still denoted as

$$\mathcal{T}_h = \cup_{r,n} \Omega_{S,r} \times (t_n, t_{n+1}).$$

One must therefore be careful that $\mathcal{T}_h$ denotes either a purely spatial mesh for stationary models or a space-time mesh for time dependent models. Moreover the cells or subdomains will be referred to with the same notation, that is

$$\Omega_k = \Omega_{S,r} \quad \text{or} \quad \Omega_k = \Omega_{S,r} \times (t_n, t_{n+1}).$$

In summary, one can write in both cases $\mathcal{T}_h = \cup_k \Omega_k$ and the context makes these notations non ambiguous. They are several advantages to consider space-time meshes

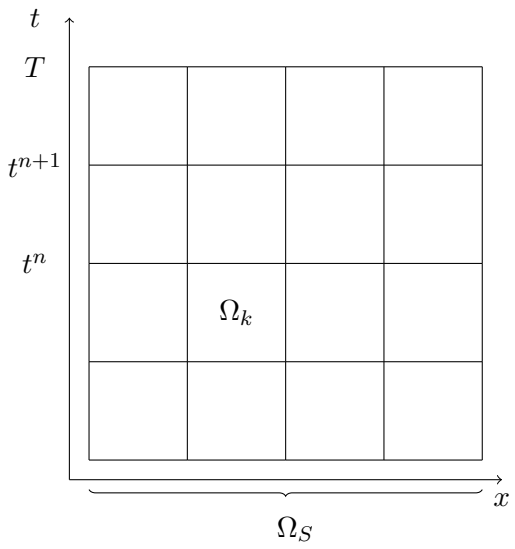

Figure 2.1 – Illustration of the partition $\mathcal{T}_h$ for a time dependent problem.

- No distinction is made between stationary and time dependent problems. Therefore, the formulation and analysis of the method are the same in both cases.
- Time dependent basis functions can be used. This can be particularly useful for Trefftz methods since it allows to be well-balanced for time dependent solutions.
- With the right choice of flux, the method can be solved iteratively in time (time step after time step) like any standard scheme.

Finally, note that other kind of space-time meshes could also be considered [FR00, LRv95, MR05].

The broken Sobolev space is

$$H^1(\mathcal{T}_h) := \Big\{ \mathbf{v} \in L^2(\Omega),\ \mathbf{v}_{|\Omega_k} \in H^1(\Omega_k)\ \forall \Omega_k \in \mathcal{T}_h \Big\}.$$

In the following we assume $\mathbf{u} \in H^1(\mathcal{T}_h)$. For convenience, we may rewrite the system (2.1) under the form $L\mathbf{u} = \mathbf{0}$ and consider the adjoint operator

$$L := \sum_i A_i \partial_i + R, \quad L^* := -\sum_i A_i \partial_i + R = -L + 2R.$$

The matrices A_i are constant and we assume that the matrix $R(\mathbf{x})$ is constant in each cell. Multiplying the system (2.1) by $\mathbf{v} \in H^1(\mathcal{T}_h)$ and integrating on Ω gives

$$\sum_k \int_{\Omega_k} \mathbf{v}_k^T L \mathbf{u}_k = 0, \tag{2.4}$$

where $\mathbf{v}_k = \mathbf{v}_{|\Omega_k}$, $\mathbf{u}_k = \mathbf{u}_{|\Omega_k}$. Integrating by parts one gets

$$\sum_k \int_{\Omega_k} \left(L^* \mathbf{v}_k\right)^T \mathbf{u}_k + \sum_k \int_{\partial\Omega_k} \mathbf{v}_k^T M_k \mathbf{u}_k = 0,$$

where $\partial\Omega_k$ is the contour of the element Ω_k. Here, we have generalized the notation (2.2) on $\partial\Omega_k$ where $\mathbf{n}_k = (n_t, n_{x_1}, ..., n_{x_d})^T$ is the outward unit normal and $M_k = M(\mathbf{n}_k) = A_0 n_t + \sum_i A_i n_i$. Denoting Σ_{kj} the edge oriented from Ω_k to Ω_j when $k \neq j$ and Σ_{kk} the edges belonging to $\Omega_k \cap \partial\Omega$ (for simplicity we use the same notation even if there is more than one edge in $\Omega_k \cap \partial\Omega$),

one can write

$$\sum_k \int_{\Omega_k} \left(L^* \mathbf{v}_k\right)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}^T M \mathbf{u})_k + (\mathbf{v}^T M \mathbf{u})_j + \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g}.$$

For $\mathbf{u}$ satisfying the equation (2.1), the normal flux is

$$f_{kj}(\mathbf{u}_k, \mathbf{u}_j) := M_k \mathbf{u}_k = -M_j \mathbf{u}_j, \quad \text{on } \Sigma_{kj}. \tag{2.5}$$

Note that $f_{kj}(\mathbf{u}_k, \mathbf{u}_j) + f_{jk}(\mathbf{u}_j, \mathbf{u}_k) = \mathbf{0}$. One has

$$\sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}^T M \mathbf{u})_k + (\mathbf{v}^T M \mathbf{u})_j = \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T f_{kj}(\mathbf{u}_k, \mathbf{u}_j).$$

Because M is symmetric, one can decompose M under the form $M = M^+ + M^-$ where M^+ is a non negative matrix and M^- is a non positive matrix, see (2.3). In the following we consider the upwind flux

$$f_{kj}(\mathbf{u}_k, \mathbf{u}_j) = M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j,$$

where $M_{kj} = M_{k|\Sigma_{kj}}$. Finally one gets

$$\sum_k \int_{\Omega_k} \left(L^* \mathbf{v}_k\right)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) + \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g}. \tag{2.6}$$

We define the bilinear form $a_{DG} : H^1(\mathcal{T}_h) \times H^1(\mathcal{T}_h) \to \mathbb{R}$ and the linear form $l : H^1(\mathcal{T}_h) \to \mathbb{R}$ as

$$\begin{aligned} a_{DG}(\mathbf{u}, \mathbf{v}) = &\sum_k \int_{\Omega_k} (L^* \mathbf{v}_k)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) \\ &+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k, \quad \mathbf{u}, \mathbf{v} \in H^1(\mathcal{T}_h), \\ l(\mathbf{v}) = &-\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g}, \quad \mathbf{v} \in H^1(\mathcal{T}_h). \end{aligned} \tag{2.7}$$

One can rewrite (2.6) as $a_{DG}(\mathbf{u}, \mathbf{v}) = l(\mathbf{v})$, $\forall \mathbf{v} \in H^1(\mathcal{T}_h)$. We can now define the classic discontinuous Galerkin method for Friedrichs systems with polynomial basis functions [DPE11, EG06, FR00, MR05]. We define $\mathbb{P}_q^d$ the space of polynomials of d variables, of total degree at most q and the broken polynomial space

$$\mathbb{P}_q^d(\mathcal{T}_h) := \left\{ \mathbf{v} \in L^2(\Omega), \mathbf{v}_{|\Omega_k} \in \mathbb{P}_q^d \ \forall \Omega_k \in \mathcal{T}_h \right\} \subset H^1(\mathcal{T}_h).$$

Now we can introduce the DG method.

Definition 2.1 (DG method)**.** Assume $P_h(\mathcal{T}_h)$ is a finite subspace of $\mathbb{P}_q^d(\mathcal{T}_h)$. The standard upwind discontinuous Galerkin method for Friedrichs systems is formulated as follows

$$\begin{cases} \text{find } \mathbf{u}_h \in P_h(\mathcal{T}_h) \text{ such that} \\ a_{DG}(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in P_h(\mathcal{T}_h). \end{cases} \tag{2.8}$$

Note that, because of the conservation equation (2.5), the exact solution to (2.1) also satisfies

$$a_{DG}(\mathbf{u}, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in H^1(\mathcal{T}_h). \tag{2.9}$$

2-2.2 Trefftz Discontinuous Galerkin formulation

Since our goal is to use Trefftz method we take discontinuous basis functions which are solutions to (2.1) in each cell

$$V(\mathcal{T}_h) = \Big\{ \mathbf{v} \in H^1(\mathcal{T}_h), L\mathbf{v}_k = \mathbf{0} \quad \forall \Omega_k \in \mathcal{T}_h \Big\} \subset H^1(\mathcal{T}_h). \tag{2.10}$$

The space $V(\mathcal{T}_h)$ is a genuine subspace of $H^1(\mathcal{T}_h)$ except in the case $L = 0$. Starting from the bilinear form a_{DG} from (2.7), one notices that the volume term can be written for all functions in $V(\mathcal{T}_h)$ as

$$\int_{\Omega_k} \big(L^*\mathbf{v}_k\big)^T \mathbf{u}_k = \int_{\Omega_k} \Big((-L+2R)\mathbf{v}_k\Big)^T \mathbf{u}_k = 2\int_{\Omega_k} \mathbf{v}_k^T R\mathbf{u}_k, \quad \forall \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h). \tag{2.11}$$

One can therefore define a bilinear form $a_T : V(\mathcal{T}_h) \times V(\mathcal{T}_h) \to \mathbb{R}$ as

$$\begin{aligned} a_T(\mathbf{u}, \mathbf{v}) = &\sum_k 2\int_{\Omega_k} \mathbf{v}_k^T R\mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) \\ &+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k, \quad \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h). \end{aligned} \tag{2.12}$$

Of course, one has $a_{DG}(\mathbf{u}, \mathbf{v}) = a_T(\mathbf{u}, \mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h)$. We give an equivalent formulation of the bilinear form $a_T(\cdot, \cdot)$. Thanks to an integration by part one has for $\mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h)$

$$\begin{aligned} a_T(\mathbf{u}, \mathbf{v}) &= \sum_k \int_{\Omega_k} (L^*\mathbf{v}_k)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j), \\ &= \sum_k \int_{\Omega_k} \mathbf{v}_k^T L\mathbf{u}_k - \sum_k \int_{\partial\Omega_k} \mathbf{v}_k^T M_k \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j). \end{aligned} \tag{2.13}$$

Since the functions $\mathbf{u}_k \in V(\mathcal{T}_h)$ are piecewise homogeneous solutions of the equation, that is $L\mathbf{u}_k = \mathbf{0}$, one gets

$$a_T(\mathbf{u}, \mathbf{v}) = -\sum_k \sum_{j<k} \int_{\Sigma_{kj}} (M_{kj}^- \mathbf{v}_k + M_{kj}^+ \mathbf{v}_j)^T (\mathbf{u}_k - \mathbf{u}_j) - \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{u}_k, \quad \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h). \tag{2.14}$$

The relaxation term R completely disappeared in the formulation (2.14). It might seem a paradox at first sight but it is not because, for a Trefftz method, some information about R is encoded in the basis functions. Since there is no volume term in the formulation (2.14) compared to (2.12) it may be easier to implement. The related linear form $l : V(\mathcal{T}_h) \to \mathbb{R}$ is unchanged with respect to (2.7), that is $l(\mathbf{v}) = -\sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g}$ for all $\mathbf{v} \in V(\mathcal{T}_h)$.

Definition 2.2 (TDG method)**.** Assume $V_h(\mathcal{T}_h)$ is a finite subspace of $V(\mathcal{T}_h)$. The upwind Trefftz discontinuous Galerkin method for the model problem (2.1) is formulated as follows

$$\begin{cases} \text{find } \mathbf{u}_h \in V_h(\mathcal{T}_h) \text{ such that} \\ a_T(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in V_h(\mathcal{T}_h). \end{cases} \tag{2.15}$$

Remark 2.3 (Iterative scheme in time)**.** In case of a time dependent problem, even if the classic upwind discontinuous Galerkin formulation (2.8) and the upwind Trefftz discontinuous Galerkin formulation (2.15) are posed on the whole space-time domain Ω, they still can be decoupled time

step after time step. It comes from the fact that the matrix A_0 is non negative and therefore $M^-(\mathbf{n}) = 0$ if $\mathbf{n} = (1, 0, ..., 0)$.

Define $a_T^n : V(\mathcal{T}_h) \times V(\mathcal{T}_h) \to \mathbb{R}$ (related to the general bilinear form (2.14)) and $l^n : V(\mathcal{T}_h) \to \mathbb{R}$ as the "space part" of the previous bilinear and linear form

$$\begin{aligned} a_T^n(\mathbf{u}, \mathbf{v}) = &- \sum_k \sum_{j<k} \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n} \mathbf{v}_k^n + M^+_{k^n j^n} \mathbf{v}_j^n)^T (\mathbf{u}_k^n - \mathbf{u}_j^n) - \sum_k \int_{\partial\Omega_S \cap \partial\Omega_{k^n}} (\mathbf{v}_k^n)^T M^-_{k^n} \mathbf{u}_k^n \\ &- \sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{v}_k^n)^T M^-_{k^n k^{n-1}} \mathbf{u}_k^n, \quad \mathbf{u}, \mathbf{v} \in V(\mathcal{T}_h), \\ l^n(\mathbf{v}) = &- \sum_k \int_{\partial\Omega_S \cap \partial\Omega_{k^n}} (\mathbf{v}_k^n)^T M^-_{k^n} \mathbf{g} - \sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{v}_k^n)^T M^-_{k^n k^{n-1}} \mathbf{u}_k^{n-1}, \quad \mathbf{v} \in V(\mathcal{T}_h). \end{aligned} \tag{2.16}$$

Here the index k^n is a notation for an element of the space-time mesh: it denotes the element with index k in the spacial mesh at the time step n. We also used the convention $\Sigma_{k^1 k^0} = \partial\Omega_{k^1} \cap (\partial\Omega \times \{0\})$ (*i.e.* $\Sigma_{k^1 k^0}$ is a cell of the spacial mesh at time $t = 0$) and $\Sigma_{k^{N+1} k^N} = \partial\Omega_{k^N} \cap (\partial\Omega \times \{T\})$ (*i.e.* $\Sigma_{k^{N+1} k^N}$ is a cell of the spacial mesh at the final time). The formulation (2.15) is equivalent to the series of space problems

$$\begin{cases} \text{find } \mathbf{u}_h^n, \ n = 1, ..., N, \ \text{such that} \\ a_T^n(\mathbf{u}_h^n, \mathbf{v}_h^n) = l^n(\mathbf{v}_h^n), \quad \forall \mathbf{v}_h^n \in V_h(\mathcal{T}_h). \end{cases} \tag{2.17}$$

The scheme obtained with the formulation (2.17) is implicit. ●

Remark 2.4 (Exact integration of the basis functions)**.** In this document, the basis functions that we consider are products of polynomials and exponentials. To calculate the contributions of the basis functions in the bilinear and linear form (2.15), one therefore needs to integrate products of polynomials and exponentials on the edges/faces of the mesh. Even if it is always possible to use quadrature formulas, it may be desirable to calculate exactly such integrals. We refer the reader to [Gab09] for a convenient way to integrate products of polynomials and exponentials in two and three dimensions. In our numerical tests, the integrals will be calculated exactly. ●

Remark 2.5 (Adjoint basis functions)**.** A fully different choice of basis functions is also possible using the adjoint operator L^* instead of L in (2.10). Define $V^*(\mathcal{T}_h) = \{\mathbf{v} \in H^1(\mathcal{T}_h), \ L^*\mathbf{v}_k = \mathbf{0} \ \forall \Omega_k \in \mathcal{T}_h\} \subset H^1(\mathcal{T}_h)$. With this choice of basis functions one has $L^*\mathbf{v}_k = \mathbf{0}$ in (2.7) and we therefore define $a_{AT} : V^*(\mathcal{T}_h) \times V^*(\mathcal{T}_h) \to \mathbb{R}$ as

$$a_{AT}(\mathbf{u}, \mathbf{v}) = \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M^+_{kj} \mathbf{u}_k + M^-_{kj} \mathbf{u}_j) + \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M^+_k \mathbf{u}_k, \tag{2.18}$$

and consider $V_h^*(\mathcal{T}_h)$ a finite subspace of $V^*(\mathcal{T}_h)$. The upwind adjoint Trefftz discontinuous Galerkin method for the model problem (2.1) reads

$$\begin{cases} \text{find } \mathbf{u}_h \in V_h^*(\mathcal{T}_h) \text{ such that} \\ a_{AT}(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in V_h^*(\mathcal{T}_h), \end{cases} \tag{2.19}$$

with l a linear form as in (2.7). Even if when $R = 0$ these two approaches coincide, the problems we are interested in are such that $R = R^T \neq 0$, so these two methods are different in our case. The numerical solution is by construction in the space $V^* \neq V$ and it is not clear if a finite subspace of V^* can give a good approximation of V using standard norms. Some numerical examples are given in appendix C.

Another possibility is to adopt a Petrov-Galerkin approach choosing trial functions in $V(\mathcal{T}_h)$ and test functions in $V^*(\mathcal{T}_h)$ [Gab06, Gab07]. However, we have tested this approach and noticed some stability issue for time dependent problems. Therefore these methods will not be studied further. ●

2-2.3 Trefftz discontinuous Galerkin formulation for systems with a source term

In this section, we show how to derive the TDG method for a model problem with a source term. The only difference with the system (2.1) is the addition of a source term $\mathbf{f}$ in the right hand side

$$\begin{cases} \sum_{i=0}^{d} A_i \partial_i \mathbf{u} = -R(\mathbf{x})\mathbf{u} + \mathbf{f}(t,\mathbf{x}), & \text{in } \Omega, \\ M^- \mathbf{u} = M^- \mathbf{g}, & \text{in } \partial\Omega, \end{cases} \tag{2.20}$$

where $\mathbf{f}(t,\mathbf{x}) \in L^2(\Omega)$ is constant in each cell and all the hypothesis made for the model (2.1) hold.

With the same approximation space V_h as before, the TDG method may give a bad approximation near the source. If we want the basis functions to "see" the source term $\mathbf{f}$, the approximation space (2.10) must be changed. We introduce the following space

$$V_h^{\mathbf{f}}(\mathcal{T}_h) = \left\{ \mathbf{v} \in H^1(\mathcal{T}_h),\ \alpha_k \in \mathbb{R},\ L\mathbf{v}_k = \alpha_k \mathbf{f}, \quad \forall \Omega_k \in \mathcal{T}_h \right\} \subset H^1(\mathcal{T}_h). \tag{2.21}$$

To get the new bilinear form $a_T(\cdot,\cdot)$ one can use the equality $L\mathbf{u}_k = \alpha_k \mathbf{f}$ in (2.13). For the linear form $l(\cdot)$ one needs to add the contribution $\sum_k \int_{\Omega_k} \mathbf{v}_k^T \mathbf{f}$. The bilinear form $a_T(\cdot,\cdot)$ and the linear form $l(\cdot)$ now read

$$\begin{aligned} a_T(\mathbf{u},\mathbf{v}) = &\sum_k \alpha_k \int_{\Omega_k} \mathbf{v}_k^T \mathbf{f} - \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (M_{kj}^- \mathbf{v}_k + M_{kj}^+ \mathbf{v}_j)^T (\mathbf{u}_k - \mathbf{u}_j) \\ &- \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{u}_k, \quad \mathbf{u},\mathbf{v} \in V(\mathcal{T}_h), \\ l(\mathbf{v}) = &- \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{g} + \sum_k \int_{\Omega_k} \mathbf{v}_k^T \mathbf{f}, \quad \mathbf{v} \in H^1(\mathcal{T}_h). \end{aligned} \tag{2.22}$$

Definition 2.6. Assume $V_h(\mathcal{T}_h)$ is a finite subspace of $V_h^{\mathbf{f}}(\mathcal{T}_h)$. The upwind Trefftz discontinuous Galerkin method for the model problem (2.20) is formulated as follows

$$\begin{cases} \text{find } \mathbf{u}_h \in V_h^{\mathbf{f}}(\mathcal{T}_h) \text{ such that} \\ a_T(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in V_h^{\mathbf{f}}(\mathcal{T}_h), \end{cases} \tag{2.23}$$

where $a_T(\cdot,\cdot)$ and $l(\cdot)$ are given in (2.22).

In practice, the formulation (2.23) is not so different from the formulation (2.15). Typically, a possible choice will be first to consider some basis functions which are solutions to the homogeneous problem (that is take $\alpha = 0$ in (2.21)). After that, one could add basis functions which depend on $\mathbf{f}$. For example, if the matrix R is invertible, one could consider the basis function $R^{-1}\mathbf{f}$ (that is take $\alpha = 1$ in (2.21)). With this procedure, the only difference between the formulations (2.23) and (2.15) is the addition of the one single basis function $R^{-1}\mathbf{f}$. The additional computational cost of the formulation (2.15) compare to the formulation (2.23) is therefore limited. Some numerical examples of the TDG method applied to a system with a source term will be given in Chapter 5.

2-3 Analysis of the Trefftz Discontinuous Galerkin method

2-3.1 Quasi-optimality

In this section, we study the TDG method applied to the model problem (2.1). In particular, we show a quasi-optimality bound in mesh-dependent norms. Our analysis follows some results of [KMPS16] where a special case with $R = 0$ was studied, see also [FR00, MR05] for the general case. We define two semi-norms on $H^1(\mathcal{T}_h)$

$$\begin{aligned}\|\mathbf{u}\|_{DG}^2 &= \sum_k \int_{\Omega_k} \mathbf{u}_k^T R \mathbf{u}_k + \sum_k \sum_{j<k} \frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{u}_k - \mathbf{u}_j)^T |M_{kj}| (\mathbf{u}_k - \mathbf{u}_j) + \sum_k \frac{1}{2} \int_{\Sigma_{kk}} \mathbf{u}_k^T |M_k| \mathbf{u}_k, \\ \|\mathbf{u}\|_{DG^*}^2 &= \sum_k \int_{\partial\Omega_k} -\mathbf{u}_k^T M_k^- \mathbf{u}_k,\end{aligned} \tag{2.24}$$

with $|M_{kj}| = |M_{jk}| = M_{kj}^+ - M_{kj}^-$. First, we show that these two semi-norms are in fact norms on the Trefftz space. We will need the following lemmas.

Lemma 2.7. *One has the inequality* $\|\mathbf{v}\|_{DG} \leq C\|\mathbf{v}\|_{DG^*}$ *for all* $\mathbf{v} \in V(\mathcal{T}_h)$*, with* $C = \sqrt{\frac{5}{2}}$.

Proof. Assume $\mathbf{v} \in V(\mathcal{T}_h)$ then $L\mathbf{v}_k = \mathbf{0}$, $\forall \Omega_k \in \mathcal{T}_h$. Multiply by $\mathbf{v}_k^T$ and integrate over Ω_k one has $\int_{\Omega_k} \mathbf{v}_k^T L \mathbf{v}_k = 0$. Integrating by parts one finds

$$\int_{\Omega_k} (L^* \mathbf{v}_k)^T \mathbf{v}_k + \int_{\partial\Omega_k} \mathbf{v}_k M_k \mathbf{v}_k = 0.$$

Using $L^* = -L + 2R$ and $L\mathbf{v}_k = \mathbf{0}$ one gets

$$\frac{1}{2} \int_{\partial\Omega_k} \mathbf{v}_k^T M_k \mathbf{v}_k + 2 \int_{\Omega_k} \mathbf{v}_k^T R \mathbf{v}_k = 0. \tag{2.25}$$

Therefore one has

$$\sum_k \int_{\Omega_k} \mathbf{v}_k^T R \mathbf{v}_k \leq -\frac{1}{2} \sum_k \int_{\partial\Omega_k} \mathbf{v}_k^T M_k^- \mathbf{v}_k = \frac{1}{2} \|\mathbf{v}\|_{DG^*}^2, \tag{2.26}$$

which is a bound for the first term in the definition of the DG norm (2.24). Moreover, because R is non negative and using (2.25), one also finds $\int_{\partial\Omega_k} \mathbf{v}_k^T M_k \mathbf{v}_k \leq 0$ that is $\int_{\partial\Omega_k} \mathbf{v}_k^T M_k^+ \mathbf{v}_k \leq -\int_{\partial\Omega_k} \mathbf{v}_k^T M_k^- \mathbf{v}_k$ and consequently

$$\int_{\partial\Omega_k} \mathbf{v}_k^T |M_k| \mathbf{v}_k \leq -2 \int_{\partial\Omega_k} \mathbf{v}_k^T M_k^- \mathbf{v}_k. \tag{2.27}$$

An elementary inequality gives $\frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T |M_{kj}| (\mathbf{v}_k - \mathbf{v}_j) \leq \int_{\Sigma_{kj}} \mathbf{v}_k^T |M_{kj}| \mathbf{v}_k + \mathbf{v}_j^T |M_{kj}| \mathbf{v}_j$ thus

$$\sum_k \sum_{j<k} \frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T |M_{kj}| (\mathbf{v}_k - \mathbf{v}_j) + \sum_k \frac{1}{2} \int_{\Sigma_{kk}} \mathbf{v}_k^T |M_k| \mathbf{v}_k \leq \sum_k \int_{\partial\Omega_k} \mathbf{v}_k^T |M_k| \mathbf{v}_k,$$

and therefore using (2.27)

$$\sum_k \sum_{j<k} \frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T |M_{kj}| (\mathbf{v}_k - \mathbf{v}_j) + \sum_k \frac{1}{2} \int_{\Sigma_{kk}} \mathbf{v}_k^T |M_k| \mathbf{v}_k \leq -2 \sum_k \int_{\partial\Omega_k} \mathbf{v}_k M_k^- \mathbf{v}_k = 2\|\mathbf{v}\|_{DG^*}^2, \tag{2.28}$$

which is a bound for the second and third terms in the definition of the DG norm (2.24). Finally combining (2.26) and (2.28) with the definition of the DG norm (2.24) one gets $\|\mathbf{v}\|_{DG}^2 \leq \frac{5}{2} \|\mathbf{v}\|_{DG^*}^2$. ∎

Lemma 2.8. *Assume $M \in \mathbb{R}^{m\times m}$ is a symmetric matrix. Then one has*

$$\mathbf{z}^T M^2 \mathbf{z} \leq C \mathbf{z}^T |M| \mathbf{z}, \quad \forall \mathbf{z} \in \mathbb{R}^m,$$

where we have used the decomposition of $M = M^+ + M^-$, M^+ is a non negative matrix, M^- is a non positive matrix and $|M| = M^+ - M^-$.

Proof. First, we notice that $\mathbf{z}^T|M|\mathbf{z} = \mathbf{z}^T M^+ \mathbf{z} - \mathbf{z}^T M^- \mathbf{z}$ and $\mathbf{z}^T M^2 \mathbf{z} = \mathbf{z}^T (M^+)^2 \mathbf{z} + \mathbf{z}^T (M^-)^2 \mathbf{z}$. Let λ^+ be the maximum eigenvalue of M^+. Denoting λ_i and $\mathbf{r}_i$ the eigenvalue and eigenvector of M^+ one has $\sum_{\lambda_i \geq 0} \mathbf{z}^T (M^+)^2 \mathbf{z} = \lambda_i^2 (\mathbf{z}, \mathbf{r}_i)^2 \leq \lambda^+ \sum_{\lambda_i \geq 0} (\mathbf{z}, \mathbf{r}_i)^2 = \lambda^+ \mathbf{z}^T M^+ \mathbf{z}$. A similar inequality applies to the matrix M^- gives finally $\mathbf{z}^T M^2 \mathbf{z} \leq \rho(M) \mathbf{z}^T |M| \mathbf{z}$, $\forall \mathbf{z} \in \mathbb{R}^m$. This completes the proof. ∎

We can now show that the two semi-norms $\|\cdot\|_{DG}$ and $\|\cdot\|_{DG^*}$ are in fact norms on the Trefftz space $V(\mathcal{T}_h)$.

Proposition 2.9. *The semi-norms $\|\cdot\|_{DG}$ and $\|\cdot\|_{DG^*}$ are norms on the Trefftz space $V(\mathcal{T}_h)$ defined in (2.10).*

Proof. Assume $\mathbf{u} \in V(\mathcal{T}_h)$ and $\|\mathbf{u}\|_{DG} = \mathbf{0}$. Since $\|\mathbf{u}\|_{DG} = \mathbf{0}$ one has $\sum_k \sum_{j<k} \frac{1}{2} \int_{\Sigma_{kj}} (\mathbf{u}_k - \mathbf{u}_j)^T |M_{kj}| (\mathbf{u}_k - \mathbf{u}_j) = \mathbf{0}$ and Lemma 2.8 implies that $M\mathbf{u}$ has vanishing jump across each edge of $\mathcal{T}_h$. Thus $\mathbf{u}$ is a solution to the general problem $L\mathbf{u} = \mathbf{0}$ in Ω. Moreover $\int_{\partial\Omega} \mathbf{u}^T |M| \mathbf{u} = 0$. Therefore $\mathbf{u}$ is solution of

$$\begin{cases} L\mathbf{u} = \mathbf{0}, & \text{in } \Omega, \\ M^- \mathbf{u} = \mathbf{0}, & \text{on } \partial\Omega. \end{cases}$$

We conclude $\mathbf{u} = \mathbf{0}$ in Ω using the uniqueness of the solution. Thus $\|\cdot\|_{DG}$ is a norm on $V(\mathcal{T}_h)$. Thanks to Lemma 2.7, we also conclude that $\|\cdot\|_{DG^*}$ is also a norm on $V(\mathcal{T}_h)$. This completes the proof. ∎

Next, we study the coercivity and the continuity of the bilinear form $a(\cdot,\cdot)$ regarding the norms $\|\cdot\|_{DG}$ and $\|\cdot\|_{DG^*}$.

Proposition 2.10 (Coercivity)**.** *For all $\mathbf{u} \in H^1(\mathcal{T}_h)$ one has $a_{DG}(\mathbf{u},\mathbf{u}) = \|\mathbf{u}\|_{DG}^2$. Therefore, one gets $a_T(\mathbf{u},\mathbf{u}) = \|\mathbf{u}\|_{DG}^2$ for all $\mathbf{u} \in V(\mathcal{T}_h)$.*

Proof. The proof is taken from [MR05]. Let $\mathbf{u}, \mathbf{v} \in H^1(\mathcal{T}_h)$. The bilinear form (2.7) reads

$$\begin{aligned} a_{DG}(\mathbf{u},\mathbf{v}) = &\sum_k \int_{\Omega_k} \big([-\sum_i A_i \partial_i + R]\mathbf{v}_k\big)^T \mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) \\ &+ \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k. \end{aligned}$$

Integrating by part and using $M_{kj} = -M_{jk}$ one has

$$\begin{aligned} a_{DG}(\mathbf{u},\mathbf{v}) = &\sum_k \int_{\Omega_k} \mathbf{v}_k^T (\sum_i A_i \partial_i + R)\mathbf{u}_k + \sum_k \sum_{j<k} \int_{\Sigma_{kj}} -\mathbf{v}_k^T M_{kj} \mathbf{u}_k + \mathbf{v}_j^T M_{kj} \mathbf{u}_j \\ &+ (\mathbf{v}_k - \mathbf{v}_j)^T (M_{kj}^+ \mathbf{u}_k + M_{kj}^- \mathbf{u}_j) + \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^+ \mathbf{u}_k - \mathbf{v}_k^T M_{kj} \mathbf{u}_k. \end{aligned}$$

Using $M = M^+ + M^-$ one finds

$$a_{DG}(\mathbf{u},\mathbf{v}) = \sum_k \int_{\Omega_k} \mathbf{v}_k^T L \mathbf{u}_k - \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (M_{kj}^- \mathbf{v}_k + M_{kj}^+ \mathbf{v}_j)^T (\mathbf{u}_k - \mathbf{u}_j) - \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{u}_k.$$

Since $L = -L^* + 2R$ one gets

$$\begin{aligned} a_{DG}(\mathbf{u},\mathbf{v}) = &- \sum_k \int_{\Omega_k} \mathbf{v}_k^T L^* \mathbf{u}_k + \sum_k 2 \int_{\Omega_k} \mathbf{v}_k^T R \mathbf{u}_k \\ &- \sum_k \sum_{j<k} \int_{\Sigma_{kj}} (M_{kj}^- \mathbf{v}_k + M_{kj}^+ \mathbf{v}_j)^T (\mathbf{u}_k - \mathbf{u}_j) - \sum_k \int_{\Sigma_{kk}} \mathbf{v}_k^T M_k^- \mathbf{u}_k. \end{aligned}$$

Summing the above expression of $a(\cdot,\cdot)$ and the one in (2.7) one gets with $\mathbf{v} = \mathbf{u}$ the equality $2a_{DG}(\mathbf{u},\mathbf{u}) = 2\|\mathbf{u}\|_{DG}^2$. Moreover, from (2.11) one deduces $a_{DG}(\mathbf{u},\mathbf{u}) = a_T(\mathbf{u},\mathbf{u})$, $\forall \mathbf{u} \in V(\mathcal{T}_h)$. This completes the proof. ■

Proposition 2.11 (Continuity). *The continuity bound $a_T(\mathbf{u},\mathbf{v}) \leq \sqrt{2}\|\mathbf{u}\|_{DG}\|\mathbf{v}\|_{DG^*}$ holds for all $\mathbf{u},\mathbf{v} \in V(\mathcal{T}_h)$.*

Proof. Using $-M_{jk}^- = M_{kj}^+$, the norm DG^* can be recast into the form

$$\|\mathbf{u}\|_{DG^*}^2 = \sum_k \sum_{j<k} \int_{\Sigma_{kj}} -\mathbf{u}_k^T M_{kj}^- \mathbf{u}_k + \mathbf{u}_j^T M_{kj}^+ \mathbf{u}_j - \sum_k \int_{\Sigma_{kk}} \mathbf{u}_k^T M_k^- \mathbf{u}_k. \tag{2.29}$$

Since $|M^-| = -M^-$ and M^+, M^- are respectively non negative and non positive symmetric matrices, the bilinear form a_T (2.14) can be written as

$$\begin{aligned} a_T(\mathbf{u},\mathbf{v}) = \sqrt{2}\Big[&\sum_k \sum_{j<k} \int_{\Sigma_{kj}} \Big(\sqrt{|M_{kj}^-|}\mathbf{v}_k\Big)^T \sqrt{|M_{kj}^-|}\Big(\frac{\mathbf{u}_k - \mathbf{u}_j}{\sqrt{2}}\Big) + \Big(-\sqrt{M_{kj}^+}\mathbf{v}_j\Big)^T \sqrt{M_{kj}^+}\Big(\frac{\mathbf{u}_k - \mathbf{u}_j}{\sqrt{2}}\Big) \\ &+ \sum_k \int_{\Sigma_{kk}} \Big(\sqrt{|M_k^-|}\mathbf{v}_k\Big)^T \Big(\sqrt{|M_k^-|}\frac{\mathbf{u}_k}{\sqrt{2}}\Big)\Big]. \end{aligned}$$

Using the Cauchy-Schwartz inequality, one sees that the first term of each scalar product is bounded by $\|\mathbf{v}\|_{DG^*}$ and the second term by $\|\mathbf{u}\|_{DG}$. This completes the proof. ■

We can now give the following classical quasi-optimality result.

Proposition 2.12 (Quasi-optimality). *For any finite dimensional space $V_h(\mathcal{T}_h) \subset V(\mathcal{T}_h)$, the TDG formulation (2.15) admits a unique solution $\mathbf{u}_h \in V_h(\mathcal{T}_h)$. Moreover, the quasi-optimality bound holds*

$$\|\mathbf{u} - \mathbf{u}_h\|_{DG} \leq \sqrt{2} \inf_{\mathbf{v}_h \in V_h(\mathcal{T}_h)} \|\mathbf{u} - \mathbf{v}_h\|_{DG^*},$$

where $\mathbf{u}$ stands for the exact solution to (2.1).

Proof. From Propositions 2.9 and 2.10, one deduces the uniqueness of the discrete solution $\mathbf{u}_h$. Existence of $\mathbf{u}_h$ follows from uniqueness. Moreover $\forall \mathbf{v}_h \in V_h(\mathcal{T}_h)$ one has

$$\|\mathbf{u} - \mathbf{u}_h\|_{DG}^2 = a_T(\mathbf{u} - \mathbf{u}_h, \mathbf{u} - \mathbf{u}_h) = a_T(\mathbf{u} - \mathbf{u}_h, \mathbf{u} - \mathbf{v}_h) \leq \sqrt{2}\|\mathbf{u} - \mathbf{u}_h\|_{DG}\|\mathbf{u} - \mathbf{v}_h\|_{DG^*},$$

thanks to propositions 2.10 and 2.11, to the consistency equality (2.9) and to (2.15). ■

2-3.2 Well-balanced property

Using the quasi-optimality proposition one has the well-balanced property of the scheme in the sense of the Definition 1.5. Of course, a standard DG scheme has the same quasi-optimality result but it can be well-balanced only for some particular polynomial functions. On the contrary, the TDG method can be well-balanced for more general solutions which contain for example exponential factors as in Example 1.8.

Proposition 2.13 (Well-balanced scheme)**.** *The TDG method is well-balanced for its basis functions.*

Proof. Assume $\mathbf{u}$ is a linear combination of the basis functions in each cell. One can take $\mathbf{v}_h = \mathbf{u}$ in Proposition 2.12. Therefore one has $\|\mathbf{u} - \mathbf{u}_h\|_{DG} = 0$. Since $\mathbf{u} - \mathbf{u}_h \in V(\mathcal{T}_h)$ one concludes using Proposition 2.9. ■

2-3.3 Estimate in standard norms

In the previous section, the error is bounded in terms of DG-norm. It is of course desirable to have estimates in a more standard norm. In this section, we present some elementary L^2 lower bounds of the DG norm which take advantage of the relaxation matrix R and an L^2 upper bound of the DG^* norm.

Proposition 2.14. *Assume $\Omega_k \in \mathcal{T}_h$, $R_k = R(\mathbf{x})_{|\Omega_k}$, and $\forall k$ R_k is definite positive. One has*

$$\frac{1}{\sup_{k\in\mathcal{T}_h} \|\sqrt{R_k}^{-1}\|^2}\|\mathbf{u}\|_{L^2(\Omega)} \leq \|\mathbf{u}\|_{DG}, \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h).$$

Proof. A basic inequality is $\mathbf{v}^2 \leq \|\sqrt{R_k}^{-1}\|^2\Big(\mathbf{v}^T R_k \mathbf{v}\Big)$. Let $\mathbf{v} \in H^1(\mathcal{T}_h)$. Integrating over Ω_k, summing over all cells and using the definition of the DG-norm (2.24), one gets the assertion. ■

This inequality holds when R is definite positive but degenerates when $R \to 0$. For non stationary problems, one can give a L^2 lower bound at the final time that does not depend on R.

Proposition 2.15. *Assume A_0 is non singular. For time dependent problems one has*

$$\|\mathbf{u}\|_{L^2(\Omega_S\times\{T\})} \leq C\|\mathbf{u}\|_{DG}, \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h).$$

where the constant C depends on the eigenvalues of A_0. In particular if $A_0 = I_m$ then $C = 1$.

Proof. Consider $\mathbf{n}(t,\mathbf{x})$ on $\partial\Omega$ with $\mathbf{n}(t,\mathbf{x}) = (n_t, n_{x_1}, ..., n_{x_d})^T = (1,0,...,0)^T$ one has $|M|((1,0,...,0)^T) = A_0$. Therefore, since A_0 is non singular and positive

$$\sum_k \int_{\Omega_{S,k}\times\{T\}} \mathbf{u}_k^2 \leq C \sum_k \frac{1}{2}\int_{\Omega_{S,k}\times\{T\}} \mathbf{u}_k^T A_0 \mathbf{u}_k \leq C \sum_k \frac{1}{2}\int_{\Omega_{S,k}\times\{T\}} \mathbf{u}_k^T |M_{kj}| \mathbf{u}_k, \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h).$$

The notation $\Omega_{S,k} \times \{T\}$ represents the edges on the top of the space-time mesh and therefore $\cup_k \Omega_{S,k} \times \{T\} \subset \cup_k \Sigma_{kk}$. One finally has

$$\sum_k \int_{\Omega_{S,k}\times\{T\}} \mathbf{u}_k^2 \leq C \sum_k \frac{1}{2}\int_{\Sigma_{kk}} \mathbf{u}_k^T |M_{kj}| \mathbf{u}_k, \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h),$$

and the assertion follows from the definition of the DG-norm. ■

Let us define the semi-norm

$$|\mathbf{u}|_{1,\Omega}^2 := \int_\Omega \sum_{i=1}^{n}\sum_{j=1}^{d}(\partial_j \mathbf{u}_i)^2.$$

The previous propositions have given lower bounds of the DG norm. The following proposition gives an upper bound of the DG^* norm.

Proposition 2.16. *One has*

$$\|\mathbf{u}\|_{DG^*}^2 \leq C\sum_k \|\mathbf{u}\|_{L^2(\Omega_k)}\Big(\frac{1}{h_k}\|\mathbf{u}\|_{L^2(\Omega_k)} + |\mathbf{u}|_{1,\Omega_k}\Big), \quad \forall \mathbf{u} \in H^1(\mathcal{T}_h), \tag{2.30}$$

where $h_k = \operatorname{diam}(\Omega_k)$ *and the constant* C *depends on the* A_i.

More precisely, if one A_i is $O(\frac{1}{\varepsilon})$ with respect to ε, the constant C scales like $\frac{1}{\varepsilon}$.

Proof. Let $\mathbf{u} \in \mathcal{T}_h$ one has $\|\mathbf{u}\|_{DG^*}^2 = \sum_k \int_{\partial\Omega_k} -\mathbf{u}_k^T M_{kj}^- \mathbf{u}_k$ and therefore

$$\|\mathbf{u}\|_{DG^*}^2 \leq C\sum_k \int_{\partial\Omega_k} \mathbf{u}_k^2.$$

We now use the trace inequality from [DPE11, Lemma 1.49] in each cell Ω_k on each component of the vector $\mathbf{u}$

$$\|\mathbf{u}\|_{L^2(\partial\Omega_k)}^2 \leq C\|\mathbf{u}\|_{L^2(\Omega_k)}\Big(\frac{1}{h_k}\|\mathbf{u}\|_{L^2(\Omega_k)} + |\mathbf{u}|_{1,\Omega_k}\Big), \quad \forall \mathbf{u} \in H^1(\Omega_k).$$

Summing over all cells one finally gets the estimate (2.30). This completes the proof. ∎

Chapter 3

Application to transport models in 1D

Contents

In this chapter, the TDG method is applied to one dimensional transport models. More precisely, two models are considered

— The first one is the P_1 model. As a first step, stationary and time dependent solutions are constructed. Then the asymptotic-preserving property of the scheme is proven by means of Hilbert expansion. Finally, the convergence and the asymptotic behavior of the TDG scheme in the diffusive regime are numerically illustrated.

— The second one is the Su-Olson model [SO96]. Compare to the P_1 model, the particularity of the Su-Olson model comes from the degenerate matrices A_0 and A_1. Stationary and time dependent solutions are constructed and the convergence of the scheme is studied. The results obtained on the numerical test given in [SO96] are very similar to those obtained with the standard DG method.

3-1 The P_1 model

The P_1 model is a first simple approximation of the transport equation using spherical harmonic expansion of the solution. An interesting property of the P_1 model is that, like the transport equation, it admits a diffusive limit when $\varepsilon \to 0$. The time dependent version of the P_1 model in one dimension reads

$$\begin{cases} \varepsilon \partial_t p + \dfrac{c}{\sqrt{3}} \partial_x v = -\varepsilon \sigma_a(x) p, \\ \varepsilon \partial_t v + \dfrac{c}{\sqrt{3}} \partial_x p = -\sigma_t(x) v. \end{cases} \tag{3.1}$$

The unknown is $\mathbf{u} = (p, v)^T$, $c, \sigma_a, \sigma_s \in \mathbb{R}^+$, $\varepsilon \in \mathbb{R}_*^+$ and

$$\sigma_t := \sigma_t^\varepsilon := \varepsilon\sigma_a + \frac{\sigma_s}{\varepsilon}.$$

The reader should be aware that σ_t depends on ε and behave as $\frac{1}{\varepsilon}$ when $\sigma_s > 0$ and $\varepsilon \to 0$. When $\varepsilon \to 0$, the variable p of the system (3.1) follows a diffusion equation.

Proposition 3.1. *When $\varepsilon \to 0$, the variable p and v of (3.1) behave formally as*

$$\left|\begin{array}{l} \partial_t p - \dfrac{c^2}{3\sigma_s}\partial_{xx} p = -\sigma_a p, \\ v = -\dfrac{c\varepsilon}{\sqrt{3}\sigma_s}\partial_x p. \end{array}\right. \tag{3.2}$$

Proof. Multiplying the second equation of (3.12) by ε and neglecting the term in ε^2, one gets $v = -\frac{c\varepsilon}{\sqrt{3}\sigma_s}\partial_x p$. Inserting this expression in the first equation of (3.12) one finds $\partial_t p - \frac{c^2}{3\sigma_s}\partial_{xx} p = -\sigma_a p$. ∎

One challenge for numerical methods is to capture the diffusion limit (3.2) on coarse meshes. The asymptotic behavior when $\varepsilon \to 0$ of the scheme is studied in section 3-1.2.

3-1.1 Construction of the basis functions for high order time dependent scheme

In order to use the Trefftz method (2.15), one needs to find solutions to the model (3.1). In particular, we would like to give a general procedure to increase the number of basis functions in order to get high order of convergence if needed. In the following, we search for particular solutions to (3.1) under the form

$$\mathbf{u}(t, x) = \mathbf{q}(t, x)e^{\lambda x},$$

where $\mathbf{q}(t, x)$ is a polynomial in space and time, $\lambda \in \mathbb{R}$. For simplicity, we consider a polynomial of degree at most one in space and time. There are other ways to construct time dependent solutions to the P_N model, see Section 4-2.4 for 2D examples. We recast the one dimensional P_1 model (3.1) under the form of a Friedrichs system (2.1) with $d = 1$, $m = 2$. It reads

$$\varepsilon\partial_t \mathbf{u} + A_1\partial_x \mathbf{u} = -R\mathbf{u}, \tag{3.3}$$

with

$$A_1 = \frac{c}{\sqrt{3}}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} \varepsilon\sigma_a & 0 \\ 0 & \sigma_t \end{pmatrix}.$$

We can now give some solutions to the one dimensional P_1 model and use them as basis functions when $\sigma_a > 0$.

Proposition 3.2 (Solution to the P_1 model when $\sigma_a > 0$)**.** *The P_1 model (3.3) admits the following four solutions*

$$\begin{aligned} \mathbf{v}_1^\pm(x) &= \begin{pmatrix} \mp\sqrt{\sigma_t} \\ \sqrt{\varepsilon\sigma_a} \end{pmatrix} e^{\pm\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}x}, \\ \mathbf{v}_2^\pm(t, x) &= \begin{pmatrix} -\frac{c}{\varepsilon}(\varepsilon\sigma_a - \sigma_t) \mp \sqrt{\frac{3\sigma_a\sigma_t}{\varepsilon}}(\varepsilon\sigma_a + \sigma_t)x - 2\frac{c}{\varepsilon}\sigma_a\sigma_t t \\ \sqrt{3}\sigma_a(\varepsilon\sigma_a + \sigma_t)x \pm 2c\sigma_a\sqrt{\frac{\sigma_a\sigma_t}{\varepsilon}}t \end{pmatrix} e^{\pm\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}x}. \end{aligned} \tag{3.4}$$

Proof. We search for particular solutions to (3.3) under the form

$$\mathbf{u}(t,x) = \mathbf{q}(x,t)e^{\lambda x}, \tag{3.5}$$

with $\lambda \in \mathbb{R}$ and where $\mathbf{q} \in \mathbb{R}^2$ is a polynomial in x and t. We consider

$$\mathbf{q}(t,x) = \mathbf{q}_0 + \mathbf{q}_1 x + \mathbf{q}_2 t. \tag{3.6}$$

Using (3.5) in (3.3) and dropping the exponential terms, one has $[\varepsilon\partial_t + A_1\partial_x + (A_1\lambda + R)]\mathbf{q}(x,t) = \mathbf{0}$. Extending $\mathbf{q}$ one finds

$$\Big((A_1\lambda + R)\mathbf{q}_0 + A_1\mathbf{q}_1 + \varepsilon\mathbf{q}_2\Big) + \Big((A_1\lambda + R)\mathbf{q}_1\Big)x + \Big((A_1\lambda + R)\mathbf{q}_2\Big)t = \mathbf{0}.$$

This equality holds for all x and t, thus one gets the following system

$$\begin{cases} (A_1\lambda + R)\mathbf{q}_2 = \mathbf{0}, \\ (A_1\lambda + R)\mathbf{q}_1 = \mathbf{0}, \\ (A_1\lambda + R)\mathbf{q}_0 = -A_1\mathbf{q}_1 - \varepsilon\mathbf{q}_2. \end{cases} \tag{3.7}$$

Therefore the solutions to (3.3) under the form (3.5) with $\mathbf{q}$ given by (3.6) satisfy the system (3.7). A necessary condition for the system (3.7) to admits a non zero solution is $\det(A_1\lambda + R) = 0$. Since

$$A_1\lambda + R = \begin{pmatrix} \varepsilon\sigma_a & \frac{c}{\sqrt{3}}\lambda \\ \frac{c}{\sqrt{3}}\lambda & \sigma_t \end{pmatrix},$$

one deduces

$$\lambda = \pm\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}.$$

We define $\mathbf{w}$ a vector which belongs to the kernel of $A_1\lambda + R$. With $\lambda = \pm\sqrt{3\sigma_a\sigma_t}$ one notices $\ker(A_1\lambda + R) = \text{Span}\left((\mp\sqrt{\sigma_t}, \sqrt{\sigma_a})^T\right)$ and one can take

$$\mathbf{w} = (\mp\sqrt{\sigma_t}, \sqrt{\varepsilon\sigma_a})^T.$$

Using the relations (3.7) one has

$$\mathbf{q}_1 = \alpha\mathbf{w}, \quad \mathbf{q}_2 = \beta\mathbf{w}, \quad \alpha, \beta \in \mathbb{R}.$$

From the last equality of (3.7), one sees that $-A_1\mathbf{q}_1 - \varepsilon\mathbf{q}_2 \in \text{Im}(A_1\lambda + R)$ which implies

$$-A_1\mathbf{q}_1 - \varepsilon\mathbf{q}_2 \in \ker\left((A_1\lambda + R)^T\right)^{\perp}.$$

Since the matrices A_1 and R are symmetric, $\ker(A_1\lambda + R)^T = \ker(A_1\lambda + R) = \text{Span}(\mathbf{w})$. A necessary condition is then $\mathbf{w}^T(-A_1\mathbf{q}_1 - \mathbf{q}_2) = 0$ which is equivalent to

$$\beta = \pm\frac{2c\sqrt{\sigma_a\sigma_t}}{\sqrt{3\varepsilon}(\varepsilon\sigma_a + \sigma_t)}\alpha.$$

With $\alpha = 0$ one finds the solutions $\mathbf{v}_1^{\pm}(x)$. With $\alpha = 1$ one gets from the fourth equation of (3.7)

$$\mathbf{q}_0 = \left(\frac{c(-\varepsilon\sigma_a + \sigma_t)}{\sqrt{3\varepsilon\sigma_a}(\varepsilon\sigma_a + \sigma_t)}, 0\right)^T + \gamma\mathbf{w},$$

with $\gamma \in \mathbb{R}$. To sum up, one has the following relations

$$\left|\begin{array}{l} \mathbf{q}_2 = \Big(-\dfrac{2c\sqrt{\sigma_a}\sigma_t}{\sqrt{3\varepsilon}(\varepsilon\sigma_a + \sigma_t)}, \pm\dfrac{2c\sigma_a\sqrt{\sigma_t}}{\sqrt{3}(\varepsilon\sigma_a + \sigma_t)}\Big)^T, \\ \mathbf{q}_1 = \Big(\mp\sqrt{\sigma_t}, \sqrt{\varepsilon\sigma_a}\Big)^T, \\ \mathbf{q}_0 = \left(\dfrac{c(-\varepsilon\sigma_a + \sigma_t)}{\sqrt{3\varepsilon\sigma_a}(\varepsilon\sigma_a + \sigma_t)}, 0\right)^T + \gamma\mathbf{w}. \end{array}\right. \tag{3.8}$$

In the following we take $\gamma = 0$. The case $\alpha = 0$ gives the solutions $\mathbf{v}_1^{\pm}(x)$. With $\alpha = 1$, multiplying the solutions by $\sqrt{\frac{3\sigma_a}{\varepsilon}}(\varepsilon\sigma_a + \sigma_t)$ and using $\mathbf{u}(x,t) = (\mathbf{q}_0 + \mathbf{q}_1 x + \mathbf{q}_2 t)e^{\lambda x}$ with $\lambda = \pm\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}$, one finds the solutions $\mathbf{v}_2^{\pm}(t,x)$. This completes the proof. ■

The solutions (3.4) can be used as basis functions when $\sigma_a > 0$. However, when $\sigma_a \to 0$, all these functions converge toward the same limit and one can no longer use them as basis functions. To address this problem, we construct some linear combinations of the solutions (3.4) which remain stable in the case $\sigma_a \to 0$. The limit solutions give the basis functions when $\sigma_a = 0$.

Proposition 3.3 (Solution to the P_1 model when $\sigma_a = 0$)**.** *The following functions are solutions to the P_1 model when $\sigma_a = 0$*

$$\begin{aligned}
\mathbf{v}_1(t,x) &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
\mathbf{v}_2(t,x) &= \begin{pmatrix} \frac{\sqrt{3}\sigma_t}{c}x \\ -1 \end{pmatrix}, \\
\mathbf{v}_3(t,x) &= \begin{pmatrix} -\frac{3\sigma_t}{c}x^2 - 2\frac{c}{\varepsilon}t \\ 2\sqrt{3}x \end{pmatrix}, \\
\mathbf{v}_4(t,x) &= \begin{pmatrix} -\frac{\sqrt{3}\sigma_t^2}{c^2}x^3 - \frac{2\sqrt{3}\sigma_t}{\varepsilon}xt - 2\sqrt{3}x \\ \frac{3\sigma_t}{c}x^2 + 2\frac{c}{\varepsilon}t \end{pmatrix}.
\end{aligned} \tag{3.9}$$

Before proving Proposition 3.3 we begin with the a lemma. To make the solutions more convenient to read, we use the notations $z_x = \frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}x$ and $\cosh(x) = \frac{e^x+e^{-x}}{2}$, $\sinh(x) = \frac{e^x-e^{-x}}{2}$.

Lemma 3.4. *The following four functions are linear combinations of the solutions* (3.4)

$$\begin{aligned}
\tilde{\mathbf{v}}_1(x) &= \begin{pmatrix} \cosh(z_x) \\ -\sqrt{\frac{\varepsilon\sigma_a}{\sigma_t}}\sinh(z_x) \end{pmatrix}, \\
\tilde{\mathbf{v}}_2(x) &= \begin{pmatrix} \sqrt{\frac{\sigma_t}{\varepsilon\sigma_a}}\sinh(z_x) \\ -\cosh(z_x) \end{pmatrix}, \\
\tilde{\mathbf{v}}_3(t,x) &= \begin{pmatrix} -\sqrt{3}\frac{\sigma_t+\varepsilon\sigma_a}{\sqrt{\varepsilon\sigma_a\sigma_t}}x\sinh(z_x) - 2\frac{c}{\varepsilon}t\cosh(z_x) \\ c\frac{\sigma_t-\varepsilon\sigma_a}{\sigma_t\sqrt{\varepsilon\sigma_a\sigma_t}}\sinh(z_x) + \sqrt{3}\frac{\sigma_t+\varepsilon\sigma_a}{\sigma_t}x\cosh(z_x) + 2c\sqrt{\frac{\sigma_a}{\varepsilon\sigma_t}}t\sinh(z_x) \end{pmatrix}, \\
\tilde{\mathbf{v}}_4(t,x) &= \begin{pmatrix} \frac{c}{\varepsilon}\frac{\sigma_t-\varepsilon\sigma_a}{\sigma_a\sqrt{\varepsilon\sigma_a\sigma_t}}\sinh(z_x) - \sqrt{3}\frac{\sigma_t+\varepsilon\sigma_a}{\varepsilon\sigma_a}x\cosh(z_x) - 2\frac{c}{\varepsilon}\sqrt{\frac{\sigma_t}{\varepsilon\sigma_a}}t\sinh(z_x) \\ \sqrt{3}\frac{\sigma_t+\varepsilon\sigma_a}{\sqrt{\varepsilon\sigma_a\sigma_t}}x\sinh(z_x) + 2\frac{c}{\varepsilon}t\cosh(z_x) \end{pmatrix}.
\end{aligned} \tag{3.10}$$

Proof. One defines the following linear combinations of the functions (3.4)

$$\mathbf{l}_1^{\pm}(t,x) = \mathbf{v}_2^{\pm} \mp \frac{c}{\varepsilon}\mathbf{v}_1^{\pm}.$$

Then defining the four solutions

$$\begin{aligned}
\tilde{\mathbf{v}}_1(t,x) &= \frac{1}{2\sqrt{\sigma_t}}\left(\mathbf{v}_1^-(t,x) - \mathbf{v}_1^+(t,x)\right), \\
\tilde{\mathbf{v}}_2(t,x) &= \frac{-1}{2\varepsilon\sqrt{\sigma_a}}\left(\mathbf{v}_1^+(t,x) + \mathbf{v}_1^-(t,x)\right), \\
\tilde{\mathbf{v}}_3(t,x) &= \frac{1}{2\sigma_a\sigma_t}\left(\mathbf{l}_1^+(t,x) + \mathbf{l}_1^-(t,x)\right), \\
\tilde{\mathbf{v}}_4(t,x) &= \frac{1}{2\sigma_a\sqrt{\varepsilon\sigma_a\sigma_t}}\left(\mathbf{v}_2^+(t,x) - \mathbf{v}_2^-(t,x)\right),
\end{aligned}$$

one gets the functions (3.10). ■

We show that these solutions degenerate toward polynomials in the limit case $\sigma_a \to 0$ and that their limit are the functions given in Proposition 3.3.

Proposition 3.5. *When $\sigma_a \to 0$, ($\sigma_t \to \frac{\sigma_s}{\varepsilon^2}$), the solutions (3.10) tend to the following functions*

$$\begin{aligned}
\tilde{\mathbf{v}}_1(t,x) &\underset{\sigma_a\to 0}{\to} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
\tilde{\mathbf{v}}_2(t,x) &\underset{\sigma_a\to 0}{\to} \begin{pmatrix} \frac{\sqrt{3}\sigma_t}{c}x \\ -1 \end{pmatrix}, \\
\tilde{\mathbf{v}}_3(t,x) &\underset{\sigma_a\to 0}{\to} \begin{pmatrix} -\frac{3\sigma_t}{c}x^2 - 2\frac{c}{\varepsilon}t \\ 2\sqrt{3}x \end{pmatrix}, \\
\tilde{\mathbf{v}}_4(t,x) &\underset{\sigma_a\to 0}{\to} \begin{pmatrix} -\frac{\sqrt{3}\sigma_t^2}{c^2}x^3 - \frac{2\sqrt{3}\sigma_t}{\varepsilon}xt - 2\sqrt{3}x \\ \frac{3\sigma_t}{c}x^2 + 2\frac{c}{\varepsilon}t \end{pmatrix}.
\end{aligned}$$

Proof. One notices that

$$\cosh(z_x) \underset{\sigma_a\to 0}{\to} 1, \quad \frac{\sinh(z_x)}{\sqrt{\varepsilon\sigma_a\sigma_t}} \underset{\sigma_a\to 0}{\to} \frac{\sqrt{3}}{c}x. \tag{3.11}$$

The limit of $\tilde{\mathbf{v}}_1(t,x)$, $\tilde{\mathbf{v}}_2(t,x)$ and $\tilde{\mathbf{v}}_3(t,x)$ are simply obtained by using the expressions (3.11) in (3.10). The limit of the second component of $\tilde{\mathbf{v}}_4(t,x)$ can be obtained in a similar way. It remains to study the first component of $\tilde{\mathbf{v}}_4(t,x)$. One has

$$\begin{aligned}
\frac{c}{\varepsilon}\frac{\sigma_t - \varepsilon\sigma_a}{\sigma_a\sqrt{\varepsilon\sigma_a\sigma_t}}\sinh(z_x) - \sqrt{3}\frac{\sigma_t + \varepsilon\sigma_a}{\varepsilon\sigma_a}x\cosh(z_x) &= \frac{c}{\varepsilon}\frac{\sigma_t - \varepsilon\sigma_a}{\sigma_a}\Big(\frac{\sqrt{3}}{c}x + \frac{3\sqrt{3}\varepsilon\sigma_a\sigma_t}{3!c^3}x^3 + o(\sigma_a^2)\Big) \\
&\quad - \sqrt{3}\frac{\sigma_t + \varepsilon\sigma_a}{\varepsilon\sigma_a}x\Big(1 + \frac{3\varepsilon\sigma_a\sigma_t x^2}{2!c^2} + o(\sigma_a^2)\Big), \\
&= -2\sqrt{3}\varepsilon x + \frac{3\sqrt{3}\sigma_t^2}{c^2}\Big(\frac{1}{6} - \frac{1}{2}\Big)x^3 + o(\sigma_a), \\
&= -2\sqrt{3}\varepsilon x - \frac{\sqrt{3}\sigma_t^2}{c^2}x^3 + o(\sigma_a),
\end{aligned}$$

Because $-2\frac{c}{\varepsilon}\sqrt{\frac{\sigma_t}{\varepsilon\sigma_a}}t\sinh(z_x) \underset{\sigma_a\to 0}{\to} -\frac{2\sqrt{3}}{\varepsilon}\sigma_t t x$, one gets the expression of the limit of $\tilde{\mathbf{v}}_4(t,x)$. This completes the proof. ■

Remark 3.6. Note that the solutions (3.4) used when $\sigma_a > 0$ are only defined in the case $c \neq 0$. However, up to a multiplication by c or c^2 if needed, the solutions (3.9) used when $\sigma_a = 0$ can also be defined when $c = 0$. ●

3-1.2 Asymptotic behavior when $\varepsilon << 1$

In this section, we study the behavior of the TDG scheme when $\varepsilon \to 0$. The main result is Proposition 3.14 which gives the AP property of the scheme for a particular choice of basis functions. Here we choose to interpret the TDG scheme (2.15) as a finite difference scheme. This has several advantages

- Under this form one observes that the scheme is new compared to other popular one dimensional asymptotic-preserving and well-balanced finite difference schemes [BDF12, GT02].

- One can study, at least formally, the asymptotic behavior of a finite difference scheme by means of Hilbert expansions.

We consider the P_1 model with no absorption

$$\begin{cases} \varepsilon \partial_t p + \dfrac{c}{\sqrt{3}} \partial_x v = 0, \\ \varepsilon \partial_t v + \dfrac{c}{\sqrt{3}} \partial_x p = -\dfrac{\sigma_s}{\varepsilon} v, \end{cases} \tag{3.12}$$

with $\varepsilon \in \mathbb{R}_*^+$, $\sigma_s, c \in \mathbb{R}^+$. We consider the stationary basis functions e_1 and e_2 from (3.9) defined in each cell as

$$\mathbf{e}_{k,1}(t,x) = \begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & \text{if } (t,x) \in \Omega_k, \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & \text{else,} \end{cases} \qquad \mathbf{e}_{k,2}(t,x) = \begin{cases} \begin{pmatrix} -\frac{\sqrt{3}\sigma_s}{c\varepsilon}(x - x_k) \\ 1 \end{pmatrix}, & \text{if } (t,x) \in \Omega_k, \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & \text{else,} \end{cases} \tag{3.13}$$

where x_k is the abscissa of the center of the cell k.

3-1.2.1 Finite difference scheme

Proposition 3.7. *The TDG scheme (2.15) with periodic boundary conditions and the basis functions (3.13) can be recast as the following finite difference scheme*

$$\begin{cases} \varepsilon \dfrac{p_k^{n+1} - p_k^n}{\Delta t} + \dfrac{c}{2\sqrt{3}h} \Big[- p_{k+1} + 2p_k - p_{k-1} + (1-a)(v_{k+1} - v_{k-1}) \Big]^{n+1} = 0, \\ \varepsilon (1 + \dfrac{a^2}{3}) \dfrac{v_k^{n+1} - v_k^n}{\Delta t} + \dfrac{c}{2\sqrt{3}h} \Big[a^2 (v_{k+1} + 2v_k + v_{k-1}) + (-v_{k+1} + 2v_k - v_{k-1}) \\ \quad + (1+a)(p_{k+1} - p_{k-1}) \Big]^{n+1} = -\dfrac{\sigma_s}{\varepsilon} v_k^{n+1}, \end{cases} \tag{3.14}$$

with $a = \frac{\sqrt{3}\sigma_s h}{2c\varepsilon}$.

Remark 3.8. One can interpret the first component of the basis function $\mathbf{e}_{k,2}(t,x)$ in (3.13) as a correction compared to the standard finite volume method. Indeed, the standard finite volume method is equivalent to consider the formulation (2.8) with the two basis functions $\mathbf{e}_{k,1} = (1,0)^T$ and $\mathbf{e}_{k,2} = (0,1)^T$. The scheme is then (3.14) with $a = 0$. As illustrated in Section 3-1.3.2 this finite volume scheme is not asymptotic-preserving when $\varepsilon \to 0$. ●

To get the scheme (3.14) we first recast the model (3.12) into the form of a Friedrichs systems (2.1) with

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_1 = \frac{c}{\sqrt{3}\varepsilon} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 \\ 0 & -\frac{\sigma_s}{\varepsilon^2} \end{pmatrix}.$$

For the sake of simplicity we assume that σ_s is constant in the domain and that the step space $h = x_{k+1} - x_k$ is constant for all k. We consider basis functions $\mathbf{e}_{i,l}$ (3.13) where i is the global number of the cell and l the local number of the basis function in the cell i. We denote by $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$ the edges of the spatial cell $\Omega_{S,i}$, i.e. $\Omega_{S,i} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and x_i the midpoint. Finally, we use the notation $\mathbf{e}_{i,1}^n, \mathbf{e}_{i,2}^n$ when designing the basis functions from the spatial cell $\Omega_{S,i}$ at the time step n.

Consider the bilinear and linear forms obtained from the decoupled formulation (2.17)

$$\begin{aligned} a_T^n(\mathbf{u},\mathbf{v}) = &-\sum_k \sum_{j<k} \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n}\mathbf{v}_k^n + M^+_{k^n j^n}\mathbf{v}_j^n)^T(\mathbf{u}_k^n - \mathbf{u}_j^n) - \sum_k \int_{\partial\Omega_S\cap\partial\Omega_{k^n}} (\mathbf{v}_k^n)^T M^-_{k^n}\mathbf{u}_k^n \\ &- \sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{v}_k^n)^T M^-_{k^n k^{n-1}}\mathbf{u}_k^n, \quad \mathbf{u},\mathbf{v}\in V(\mathcal{T}_h), \\ l^n(\mathbf{v}) = &-\sum_k \int_{\partial\Omega_S\cap\partial\Omega_{k^n}} (\mathbf{v}_k^n)^T \mathbf{g}_S - \sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{v}_k^n)^T M^-_{k^n k^{n-1}}\mathbf{u}_k^{n-1}, \quad \mathbf{v}\in V(\mathcal{T}_h). \end{aligned} \tag{3.15}$$

In the following, we explicitly write the equality

$$a_T^n(\mathbf{u},\mathbf{e}_{l,i}^n) = l^n(\mathbf{e}_{l,i}^n), \quad l=1,2, \tag{3.16}$$

for any time step n and any spatial cell $\Omega_{S,i}$. For simplicity, we will consider periodic boundary conditions, a uniform space step h and a uniform time step Δt. We introduce some notations.

Definition 3.9. We define $C^n_{S,i,l}$, $C^{n-1}_{T,i,l}$ and $C^n_{T,i,l}$ as

$$C^n_{S,i,l} = -\sum_k \sum_j \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n}\mathbf{e}_{i,l}^n)^T(\mathbf{u}_k^n - \mathbf{u}_j^n), \tag{3.17}$$

$$C^{n-1}_{T,i,l} = -\sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{e}_{i,l}^n)^T M^-_{k^n k^{n-1}}\mathbf{u}_k^{n-1}, \tag{3.18}$$

$$C^n_{T,i,l} = -\sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{e}_{i,l}^n)^T M^-_{k^n k^{n-1}}\mathbf{u}_k^n. \tag{3.19}$$

Since $\mathbf{u}_k$ is a combination of the basis functions in each cell, one can make the following assumption.

Assumption 3.10. *We assume that $\mathbf{u}_k$ admits the following decomposition in each cell Ω_k*

$$\mathbf{u}_k = \alpha_k \mathbf{e}_{k,1} + \beta_k \mathbf{e}_{k,2}, \quad \alpha_k,\beta_k\in\mathbb{R},$$

or, in an identical way, when considering the time step n and the spatial cell $\Omega_{S,i}$

$$\mathbf{u}_i^n = \alpha_i^n \mathbf{e}_{i,1}^n + \beta_i^n \mathbf{e}_{i,2}^n, \quad \alpha_i^n,\beta_i^n\in\mathbb{R}. \tag{3.20}$$

Before proving Proposition 3.7, we need some lemmas. First, we write the equality (3.16) with the notations introduced in the Definition 3.9.

Lemma 3.11. *Consider the TDG method applied to the model (3.12) with the basis functions (3.13). The TDG formulation (3.16) with periodic boundary conditions at the time step n in any spatial cell $\Omega_{S,i}$ reads*

$$\begin{aligned} C^n_{T,i,1} - C^{n-1}_{T,i,1} + C^n_{S,i,1} = 0, \\ C^n_{T,i,2} - C^{n-1}_{T,i,2} + C^n_{S,i,2} = 0. \end{aligned} \tag{3.21}$$

Proof. Since we consider periodic boundary conditions, the term $\int_{\partial\Omega_S\cap\partial\Omega_{k^n}} (\mathbf{v}_k^n)^T M^-_{k^n}\mathbf{u}_k^n$ in the bilinear form and the term $\int_{\partial\Omega_S\cap\partial\Omega_{k^n}} (\mathbf{v}_k^n)^T \mathbf{g}_S$ in the linear form of (3.15) are equal to zero. Moreover one notices that

$$-\sum_k \sum_{j<k} \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n}\mathbf{v}_k^n - M^+_{k^n j^n}\mathbf{v}_j^n)^T(\mathbf{u}_k^n - \mathbf{u}_j^n) = -\sum_k \sum_j \int_{\Sigma_{k^n j^n}} (M^-_{k^n j^n}\mathbf{v}_k^n)^T(\mathbf{u}_k^n - \mathbf{u}_j^n). \tag{3.22}$$

Therefore one has

$$a_T(\mathbf{u}, \mathbf{e}^n_{l,i}) = C^n_{T,i,l} + C^n_{S,i,l}, \quad l(\mathbf{e}^n_{l,i}) = C^{n-1}_{T,i,l}.$$

The equality (3.16) gives for $l = 1$ and $l = 2$ respectively the first and second equations of (3.21). This completes the proof. ■

Now we can study the values of the coefficients $C_{S,i,l}$ and $C_{T,i,l}$.

Lemma 3.12. *One has*

$$C^n_{S,i,1} = \frac{c\Delta t}{2\sqrt{3}}\Big(- \alpha^n_{i-1} + 2\alpha^n_i - \alpha^n_{i+1} + (1 - \frac{\sqrt{3}\sigma_s h}{2c\varepsilon})(\beta^n_{i+1} - \beta^n_{i-1})\Big)^n, \tag{3.23}$$

and

$$\begin{aligned} C^n_{S,i,2} =& \frac{c\Delta t}{2\sqrt{3}}\Big((\frac{\sqrt{3}\sigma_s h}{2c\varepsilon})^2(\beta^n_{i+1} + 2\beta^n_i + \beta^n_{i-1}) + \frac{\sqrt{3}\sigma_s h}{2c\varepsilon}\beta^n_i \\ &+ (-\beta^n_{i-1} + 2\beta^n_i - \beta^n_{i+1}) + (1 + \frac{\sqrt{3}\sigma_s h}{2c\varepsilon})(\alpha^n_{i+1} - \alpha^n_{i-1})\Big)^n. \end{aligned} \tag{3.24}$$

Proof. For simplicity, we will use the notation $M^-_{\pm 1} = M^-((0, \pm 1)^T)$, $M^+_{\pm 1} = M^+((0, \pm 1)^T)$ and $(\lambda^{m,l}_{k,j})^\pm = (M^+_{\pm 1}\mathbf{e}_{j,l})^T\mathbf{e}_{k,m}$. Since the function $\mathbf{e}_{i,l}$ is only non-zero in the cell Ω_i one can write $C_{S,i,l}$ from (3.17) as

$$C_{S,i,l} = \int_{t^{n-1}}^{t^n} \Big(- (M^-_{-1}\mathbf{e}_{i,l})^T(\mathbf{u}_i - \mathbf{u}_{i-1})(x_{i-\frac{1}{2}}) - (M^-_1\mathbf{e}_{i,l})^T(\mathbf{u}_i - \mathbf{u}_{i+1})(x_{i+\frac{1}{2}})\Big). \tag{3.25}$$

Using $M^-_{\pm 1} = -M^+_{\mp 1}$, the decomposition of $\mathbf{u}^n_i$ (3.20) and the fact that the basis (3.13) does not depend on time, the equality (3.25) reads

$$\begin{aligned} C^n_{S,i,l} =& \Delta t\Big(\alpha^n_i(\lambda^{1l}_{i,i})^+(x_{i-\frac{1}{2}}) + \beta_i(\lambda^{2l}_{i,i})^+(x_{i-\frac{1}{2}}) - \alpha^n_{i-1}(\lambda^{l1}_{i,i-1})^+(x_{i-\frac{1}{2}}) - \beta^n_{i-1}(\lambda^{l2}_{i,i-1})^+(x_{i-\frac{1}{2}}) \\ &+ \alpha^n_i(\lambda^{1l}_{i,i})^-(x_{i+\frac{1}{2}}) + \beta^n_i(\lambda^{2l}_{i,i})^+(x_{i+\frac{1}{2}}) - \alpha^n_{i+1}(\lambda^{l1}_{i,i+1})^-(x_{i+\frac{1}{2}}) - \beta^n_{i+1}(\lambda^{l2}_{i,i+1})^-(x_{i+\frac{1}{2}})\Big)^n. \end{aligned} \tag{3.26}$$

For $n_t = 0$, one has

$$M(\mathbf{n}) = M(0, n_x) = \frac{c}{\sqrt{3}}\begin{pmatrix} 0 & n_x \\ n_x & 0 \end{pmatrix},$$

$$M^+(0, n_x) = \frac{c}{2\sqrt{3}}\begin{pmatrix} 1 & n_x \\ n_x & 1 \end{pmatrix}, \quad M^-(0, n_x) = \frac{c}{2\sqrt{3}}\begin{pmatrix} -1 & n_x \\ n_x & -1 \end{pmatrix},$$

and one notices that

$$\begin{aligned} (\lambda^{11}_{ji})^\pm(x) &= \frac{c}{2\sqrt{3}}, \\ (\lambda^{12}_{ji})^\pm(x) &= \frac{c}{2\sqrt{3}}\Big(- \frac{\sqrt{3}\sigma_s}{c\varepsilon}(x - x_i) \pm 1\Big), \\ (\lambda^{22}_{ji})^\pm(x) &= \frac{c}{2\sqrt{3}}\Big(1 \mp \frac{\sqrt{3}\sigma_s}{c\varepsilon}\big((x - x_i) + (x - x_j)\big) + (\frac{\sqrt{3}\sigma_s}{c\varepsilon})^2(x - x_i)(x - x_j)\Big). \end{aligned} \tag{3.27}$$

Recalling that $h = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ for all i and inserting (3.27) in (3.26) one finds for $l = 1$

$$C^n_{S,i,1} = \frac{c\Delta t}{2\sqrt{3}}\Big(- \alpha^n_{i-1} + 2\alpha^n_i - \alpha^n_{i+1} + (1 - \frac{\sqrt{3}\sigma_s h}{2c\varepsilon})(\beta^n_{i+1} - \beta^n_{i-1})\Big)^n,$$

and for $l = 2$

$$C^n_{S,i,2} = \frac{c\Delta t}{2\sqrt{3}}\Big((\frac{\sqrt{3}\sigma_s h}{2c\varepsilon})^2(\beta^n_{i+1}+2\beta^n_i+\beta^n_{i-1})+\frac{\sqrt{3}\sigma_s h}{2c\varepsilon}\beta^n_i+(-\beta^n_{i-1}+2\beta^n_i-\beta^n_{i+1})+(1+\frac{\sqrt{3}\sigma_s h}{2c\varepsilon})(\alpha^n_{i+1}-\alpha^n_{i-1})\Big)^n.$$

This completes the proof. ■

Lemma 3.13. *One has*

$$C^n_{T,i,1} = \varepsilon h \alpha^n_i \tag{3.28}$$

$$C^n_{T,i,2} = \varepsilon h(1 + \frac{3\sigma_s^2 h^2}{48c^2\varepsilon^2})\beta^n_i. \tag{3.29}$$

Proof. Since $-M^-_{k^n k^{n-1}} = \varepsilon I_m$, $C^n_{T,i,l}$ reads

$$C^n_{T,i,l} = -\sum_k \int_{\Sigma_{k^n k^{n-1}}} (\mathbf{e}^n_{i,l})^T M^-_{k^n k^{n-1}} \mathbf{u}^n_k = \varepsilon \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\mathbf{e}^n_{i,l})^T \mathbf{u}^n_i.$$

One notices that $\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\mathbf{e}^n_{i,1})^T \mathbf{e}^n_{i,2} = 0$. Therefore, using the decomposition of $\mathbf{u}^n_i$ (3.20) one finds

$$C^n_{T,i,1} = \varepsilon\alpha^n_i \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\mathbf{e}^n_{i,1})^T \mathbf{e}^n_{i,1} = \varepsilon h \alpha^n_i,$$

$$C^n_{T,i,2} = \varepsilon\beta^n_i \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (\mathbf{e}^n_{i,2})^T \mathbf{e}^n_{i,2} = \varepsilon h(1 + \frac{3\sigma_s^2 h^2}{48c^2\varepsilon^2})\beta^n_i.$$

This completes the proof. ■

We can now find the scheme (3.14) and prove Proposition 3.7.

Proof of Proposition 3.7. Starting from (3.21) one has

$$C^n_{T,i,1} - C^{n-1}_{T,i,1} + C^n_{S,i,1} = 0,$$
$$C^n_{T,i,2} - C^{n-1}_{T,i,2} + C^n_{S,i,2} = 0.$$

We recall the decomposition (3.20) which is $\mathbf{u}^n_i(x) = \alpha^n_i e^n_{i,1}(x) + \beta^n_i e_{i,2}(x) = (p^n_i, v^n_i)^T(x)$. In particular, considering the center of the cell one finds $\alpha^n_i = p^n_i(x_i)$ and $\beta^n_i = v^n_i(x_i)$. Therefore using (3.23), (3.24), (3.28) and (3.29) in (3.21) and making the simplification $\alpha^n_i = p^n_i$ and $\beta^n_i = v^n_i$, one finally gets the scheme (3.14). This completes the proof. ■

3-1.2.2 Asymptotic-preserving property

Using Hilbert expansion, we can now formally show that the scheme (3.14) is asymptotic-preserving.

Proposition 3.14 (Asymptotic-preserving property of the scheme (3.14))**.** *When $\varepsilon \to 0$ the scheme (3.14) admits the formal limit*

$$\begin{cases} (v^0_{k+1} + v^0_k)^{n+1} = 0, \\ \Big(\dfrac{v^1_{k+1} + 2v^1_k + v^1_{k-1}}{4}\Big)^{n+1} = -\dfrac{c}{\sqrt{3}\sigma_s}\Big(\dfrac{p^0_{k+1} - p^0_{k-1}}{2h}\Big)^{n+1}, \\ \dfrac{(\bar{p}^0_k)^{n+1} - (\bar{p}^0_k)^n}{\Delta t} - \dfrac{c^2}{3\sigma_s}\Big(\dfrac{p^0_{k+2} - 2p^0_k + p^0_{k-2}}{4h^2}\Big)^{n+1} = 0, \end{cases} \tag{3.30}$$

with $\bar{p}_k^0 = (\frac{2}{3}p_{k+2}^0 + 4p_{k+1}^0 + \frac{20}{3}p_k^0 + 4p_{k-1}^0 + \frac{2}{3}p_{k-2}^0)/16$ *a local mean value of* p_k^0.

The limit scheme (3.30) *is consistent with the limit model* (3.2) *and therefore the scheme is asymptotic-preserving.*

Proof. For convenience, we divide the two equalities of (3.14) by ε. Moreover, we adopt the notations $\{\!\{f\}\!\}_{k+\frac{1}{2}} = \frac{f_{k+1}+f_k}{2}$, $[\![f]\!]_{k+\frac{1}{2}} = \frac{f_{k+1}-f_k}{2}$ and $\delta_t f = \frac{f^{n+1}-f^n}{\Delta t}$. With these notations the scheme (3.14) can be written under the form

$$\delta_t p_k + \frac{c}{\sqrt{3}\varepsilon h}\Big[-([\![p]\!]_{k+\frac{1}{2}} - [\![p]\!]_{k-\frac{1}{2}}) + (1-a)(\{\!\{v\}\!\}_{k+\frac{1}{2}} - \{\!\{v\}\!\}_{k-\frac{1}{2}})\Big]^{n+1} = 0, \tag{3.31}$$

$$\begin{aligned}(1+\frac{a^2}{3})\delta_t v_k + \frac{c}{\sqrt{3}\varepsilon h}\Big[a^2(\{\!\{v\}\!\}_{k+\frac{1}{2}} + \{\!\{v\}\!\}_{k-\frac{1}{2}}) + 2av_k - ([\![v]\!]_{k+\frac{1}{2}} - [\![v]\!]_{k-\frac{1}{2}}) \\ +(1+a)([\![p]\!]_{k+\frac{1}{2}} + [\![p]\!]_{k-\frac{1}{2}})\Big]^{n+1} = 0.\end{aligned} \tag{3.32}$$

We assume that the variables p and v can be written under the form

$$p = \sum_{i\geq 0} p^i \varepsilon^i, \quad v = \sum_{i\geq 0} v^i \varepsilon^i.$$

We inject these expressions in (3.31) and (3.32) and expand all coefficients and variables with respect to ε. In particular one needs to expand a with respect to ε using the definition $a = \frac{\sqrt{3}\sigma_s h}{2c\varepsilon}$. The terms $O(\frac{1}{\varepsilon^2})$ in (3.31) and $O(\frac{1}{\varepsilon^3})$ in (3.32) are

$$\{\!\{v\}\!\}_{k+\frac{1}{2}}^0 - \{\!\{v\}\!\}_{k-\frac{1}{2}}^0 = 0,$$

$$\{\!\{v\}\!\}_{k+\frac{1}{2}}^0 + \{\!\{v\}\!\}_{k-\frac{1}{2}}^0 = 0.$$

These two equations together give

$$\{\!\{v\}\!\}_{k+\frac{1}{2}}^0 = 0, \forall k. \tag{3.33}$$

Now, we study the terms in $O(\frac{1}{\varepsilon})$ in (3.31) and in $O(\frac{1}{\varepsilon^2})$ in (3.32). Using (3.33) one gets

$$-([\![p]\!]_{k+\frac{1}{2}}^0 - [\![p]\!]_{k-\frac{1}{2}}^0) - \frac{\sqrt{3}\sigma_s h}{2c}(\{\!\{v\}\!\}_{k+\frac{1}{2}}^1 - \{\!\{v\}\!\}_{k-\frac{1}{2}}^1) = 0,$$

$$\frac{\sqrt{3}\sigma_s h}{6c}\delta_t v_k^0 + \frac{c}{\sqrt{3}h}\Big[[\![p]\!]_{k+\frac{1}{2}}^0 + [\![p]\!]_{k-\frac{1}{2}}^0 + \frac{\sqrt{3}\sigma_s h}{2c}(\{\!\{v\}\!\}_{k+\frac{1}{2}}^1 + \{\!\{v\}\!\}_{k-\frac{1}{2}}^1) + 2v_k^0\Big] = 0.$$

Therefore, multiplying the first equation by $\frac{2c}{\sqrt{3}\sigma_s h}$, the second by $\frac{2}{\sigma_s}$ and subtracting these two equations one finds

$$\frac{\sqrt{3}h}{3c}\delta_t v_k^0 + \{\!\{v\}\!\}_{k+\frac{1}{2}}^1 + \frac{4c}{\sqrt{3}\sigma_s h}v_k^0 = -\frac{2c}{\sqrt{3}\sigma_s h}[\![p]\!]_{k+\frac{1}{2}}^0, \forall k.$$

Adding this equality for k and $k-1$ and using (3.33) one deduces

$$\{\!\{v\}\!\}_{k+\frac{1}{2}}^1 + \{\!\{v\}\!\}_{k-\frac{1}{2}}^1 = -\frac{2c}{\sqrt{3}\sigma_s h}([\![p]\!]_{k+\frac{1}{2}}^0 + [\![p]\!]_{k-\frac{1}{2}}^0), \forall k. \tag{3.34}$$

Finally, with the terms in $O(1)$ for (3.31) and in $O(\frac{1}{\varepsilon})$ for (3.32)

$$\delta_t p_k^0 + \frac{c}{\sqrt{3}h}\left[-([\![p]\!]_{k+\frac{1}{2}}^1 - [\![p]\!]_{k-\frac{1}{2}}^1) + (\{\!\{v\}\!\}_{k+\frac{1}{2}}^1 - \{\!\{v\}\!\}_{k-\frac{1}{2}}^1) - \frac{\sqrt{3}\sigma_s h}{2c}(\{\!\{v\}\!\}_{k+\frac{1}{2}}^2 - \{\!\{v\}\!\}_{k-\frac{1}{2}}^2)\right]^{n+1} = 0,$$

$$\frac{3\sigma_s^2h^2}{12c^2}\delta_t v_k^1+\frac{c}{\sqrt{3}h}\left[\frac{\sqrt{3}\sigma_s h}{2c}(2v_k^1+[\![p]\!]_{k+\frac12}^1+[\![p]\!]_{k-\frac12}^1)+[\![p]\!]_{k+\frac12}^0+[\![p]\!]_{k-\frac12}^0-([\![v]\!]_{k+\frac12}^0-[\![v]\!]_{k-\frac12}^0)\right.$$
$$\left.+\frac{3\sigma_s^2h^2}{4c^2}(\{\!\{v\}\!\}_{k+\frac12}^2+\{\!\{v\}\!\}_{k-\frac12}^2)\right]^{n+1}=0.$$

Dividing the first equation by σ_s, using (3.33), (3.34) and multiplying by $\frac{2c}{\sqrt{3}\sigma_s^2h}$ the second equation, one gets

$$\frac{1}{\sigma_s}\delta_t p_k^0+\left[\frac{c}{\sqrt{3}\sigma_s h}\left(-([\![p]\!]_{k+\frac12}^1-[\![p]\!]_{k-\frac12}^1)+\{\!\{v\}\!\}_{k+\frac12}^1-\{\!\{v\}\!\}_{k-\frac12}^1\right)-\frac{\{\!\{v\}\!\}_{k+\frac12}^2-\{\!\{v\}\!\}_{k-\frac12}^2}{2}\right]^{n+1}=0,$$
$$\frac{\sqrt{3}h}{6c}\delta_t v_k^1+\left[\frac{c}{\sqrt{3}\sigma_s h}\left(-\{\!\{v\}\!\}_{k+\frac12}^1+2v_k^1-\{\!\{v\}\!\}_{k-\frac12}^1+[\![p]\!]_{k+\frac12}^1+[\![p]\!]_{k-\frac12}^1+\frac{4c}{\sqrt{3}\sigma_s h}v_k^0\right)\right.$$
$$\left.+\frac{\{\!\{v\}\!\}_{k+\frac12}^2+\{\!\{v\}\!\}_{k-\frac12}^2}{2}\right]^{n+1}=0.$$

Adding and subtracting these two equations one finds

$$\{\!\{v\}\!\}_{k-\frac12}^2+\frac{2c}{\sqrt{3}\sigma_s h}[\![p]\!]_{k-\frac12}^1+\frac{4c^2}{3\sigma_s^2h^2}v_k^0=-\frac{1}{\sigma_s}\delta_t p_k^0-\frac{\sqrt{3}h}{6c}\delta_t v_k^1-\frac{2c}{\sqrt{3}\sigma_s h}(v_k^1-\{\!\{v\}\!\}_{k-\frac12}^1)^{n+1},\quad(3.35)$$

and

$$\{\!\{v\}\!\}_{k+\frac12}^2+\frac{2c}{\sqrt{3}\sigma_s h}[\![p]\!]_{k+\frac12}^1+\frac{4c^2}{3\sigma_s^2h^2}v_k^0=\frac{1}{\sigma_s}\delta_t p_k^0-\frac{\sqrt{3}h}{6c}\delta_t v_k^1-\frac{2c}{\sqrt{3}\sigma_s h}(v_k^1-\{\!\{v\}\!\}_{k+\frac12}^1)^{n+1}.\quad(3.36)$$

Using (3.35) in $k+1$ and subtracting (3.36) to (3.35) one gets

$$\frac{1}{\sigma_s}\delta_t(p_{k+1}^0+p_k^0)+\frac{\sqrt{3}h}{6c}\delta_t(v_{k+1}^1-v_k^1)+\frac{2c}{\sqrt{3}h\sigma_s}(v_{k+1}^1-v_k^1)^{n+1}=\frac{4c^2}{3\sigma_s^2h^2}(v_k^0-v_{k+1}^0).$$

Adding this equation for k and $k-1$ and using (3.33) one has

$$\frac{1}{\sigma_s}\delta_t(p_{k+1}^0+2p_k^0+p_{k-1}^0)+\frac{\sqrt{3}h}{3c}\delta_t(\{\!\{v\}\!\}_{k+\frac12}^1-\{\!\{v\}\!\}_{k-\frac12}^1)-\frac{4c}{\sqrt{3}\sigma_s h}(\{\!\{v\}\!\}_{k+\frac12}^1-\{\!\{v\}\!\}_{k-\frac12}^1)^{n+1}=0.$$

Summing this equation for k and $k+1$ one gets

$$\frac{1}{\sigma_s}\delta_t(p_{k+2}^0+3p_{k+1}^0+3p_k^0+p_{k-1}^0)+\frac{\sqrt{3}h}{3c}\delta_t(\{\!\{v\}\!\}_{k+\frac32}^1-\{\!\{v\}\!\}_{k-\frac12}^1)-\frac{4c}{\sqrt{3}\sigma_s h}(\{\!\{v\}\!\}_{k+\frac32}^1-\{\!\{v\}\!\}_{k-\frac12}^1)^{n+1}=0.$$

Summing this equation for k and $k-1$ one finally finds

$$\frac{1}{\sigma_s}\delta_t(p_{k+2}^0+4p_{k+1}^0+6p_k^0+4p_{k-1}^0+p_{k-2}^0)+\frac{\sqrt{3}h}{3c}\delta_t(\{\!\{v\}\!\}_{k+\frac32}^1+\{\!\{v\}\!\}_{k+\frac12}^1-\{\!\{v\}\!\}_{k-\frac12}^1-\{\!\{v\}\!\}_{k-\frac32}^1)$$
$$-\frac{4c}{\sqrt{3}\sigma_s h}(\{\!\{v\}\!\}_{k+\frac32}^1+\{\!\{v\}\!\}_{k+\frac12}^1-\{\!\{v\}\!\}_{k-\frac12}^1-\{\!\{v\}\!\}_{k-\frac32}^1)^{n+1}=0.$$

Using (3.34) one deduces

$$(\{\!\{v\}\!\}_{k+\frac32}^1+\{\!\{v\}\!\}_{k+\frac12}^1)-(\{\!\{v\}\!\}_{k-\frac12}^1+\{\!\{v\}\!\}_{k-\frac32}^1)=-\frac{c}{\sqrt{3}\sigma_s h}(p_{k+2}^0-2p_k+p_{k-2}^0).$$

Therefore one finally has

$$\delta_t\left(\frac{2}{3}p_{k+2}^0+4p_{k+1}^0+\frac{20}{3}p_k^0+4p_{k-1}^0+\frac{2}{3}p_{k-2}^0\right)-\frac{4c^2}{3\sigma_s}\left(\frac{p_{k+2}^0-2p_k^0+p_{k-2}^0}{h^2}\right)^{n+1}=0.$$

This equality is consistent with the first equation of the limit model (3.2). Moreover, the equality (3.34) is consistent with the second equation of (3.2) and the equality (3.33) with the first equation of (3.2). This completes the proof. ■

3-1.3 Numerical results

In the following we use random meshes made of N nodes and constructed as follow: we start from a uniform mesh and then moved the vertices randomly around their initial position by a factor of at most 33%.

3-1.3.1 Study of the order

For the time dependent P_1 model in one dimension (3.1) we consider the case $\Omega_S = [0,1]$, $\varepsilon = 1, c = \sqrt{3}, \sigma_a = 1, \sigma_s = 1$, $h = 1/N$ for $N = 20, 40, 60, 80, 100$, $T = 0.024$ and $dt = T/N$. The exact solution is $\mathbf{u}_{ex} = (e^{-t}, e^{-2t})$ and we set $M^-\mathbf{u} = M^-\mathbf{u}_{ex}$ on the boundary.

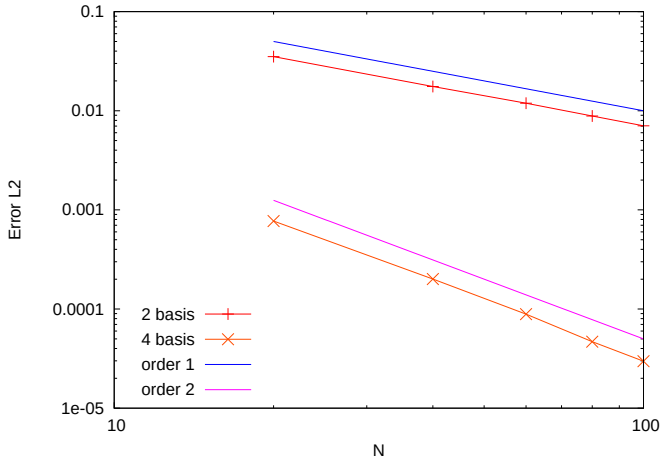

Figure 3.1 – Study of the L^2 error on the final time step in logarithmic scale for temporal one dimensional model. Error with the two stationary basis functions and the four basis functions. Random meshes.

The functions (3.4) are used as basis functions. We study two cases: a first one with only the two stationary basis functions $\mathbf{v}_1^-, \mathbf{v}_1^+$ and a second one with four basis functions $\mathbf{v}_1^-, \mathbf{v}_1^+, \mathbf{v}_2^-, \mathbf{v}_2^+$. Figure 3.1 shows that the scheme is convergent with the two basis functions $\mathbf{v}_1^-, \mathbf{v}_1^+$ and that one increases the order by adding the basis functions $\mathbf{v}_2^-, \mathbf{v}_2^+$. More precisely, order 1 is achieved with the two basis functions $\mathbf{v}_1^-, \mathbf{v}_1^+$ whereas order 2 is achieved with the four basis functions $\mathbf{v}_1^-, \mathbf{v}_1^+, \mathbf{v}_2^-, \mathbf{v}_2^+$.

3-1.3.2 Asymptotic regime when $\varepsilon << 1$

We test the asymptotic behavior of the scheme (3.14) for the P_1 model (3.12). We have shown previously that the TDG method leads to a new asymptotic-preserving scheme and we can now illustrate this property. We consider a case where $\Omega_S = [0,1]$, $\varepsilon = 0.001$, $\sigma_s = 1$, $c = \sqrt{3}$ and $T = 0.01$. For the limit solution, we consider p_0 the fundamental solution to the heat equation and the variable v_0 associated in the limit $\varepsilon \to 0$

$$p_0(t,x) = \frac{1}{2\sqrt{\pi(t+10^{-4})}} e^{\frac{-(x-0.5)^2}{4(t+10^{-4})}}, \quad v_0(t,x) = -\varepsilon \partial_x p_0(t,x).$$

The limit solution is imposed on the boundary that is $M^-(p,v)^T = M^-(p_0,v_0)^T$.

We compare the numerical solution with $(p_0, v_0)^T$. For the TDG method we use the basis functions (3.13) that is

$$\mathbf{e}_1(x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2(x) = \begin{pmatrix} -\frac{\sqrt{3}\sigma_s}{\varepsilon} x \\ 1 \end{pmatrix}. \tag{3.37}$$

And we compare the result obtained with the DG method which uses the same number of basis functions. That is the following constant functions

$$\mathbf{e}_1(x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2(x) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{3.38}$$

Note that the only difference between the basis functions (3.37) and (3.38) is the first component of $\mathbf{e}_2$.

The Figure 3.2 shows that even with few degrees of freedom the limit solution is correctly approximated by the TDG method with the basis functions (3.37). It illustrates the asymptotic-preserving property of the scheme. On the contrary, the Figure 3.3 shows that the standard DG scheme with the two basis functions (3.38) is not AP.

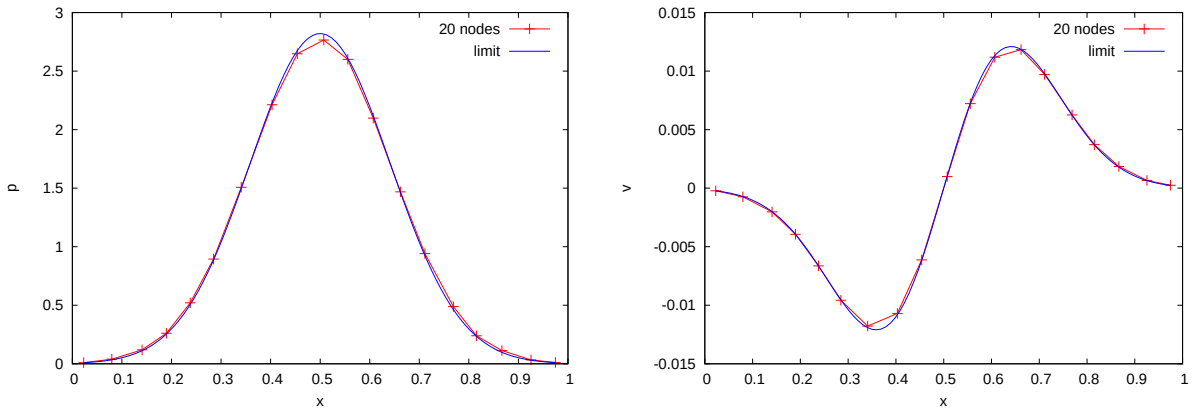

Figure 3.2 – Numerical solution obtained for the variable p (on the left) and v (on the right) with the TDG scheme (3.14) with $\varepsilon = 0.001$. Random mesh with 20 nodes and $dt = 0.01/20$. Good accuracy illustrates the AP properties of the TDG scheme.

3-2 The Su-Olson model

In this section we consider the Su-Olson model [SO96]. Compare to the P_1 model, the main difficulties when trying to solve numerically the Su-Olson model come from the degeneracy of the matrices A_0 and A_1. Standard well-balanced schemes may give bad approximation for this kind of models [GT02]. For the TDG method however, this is not a problem as soon as one can calculate the basis functions.

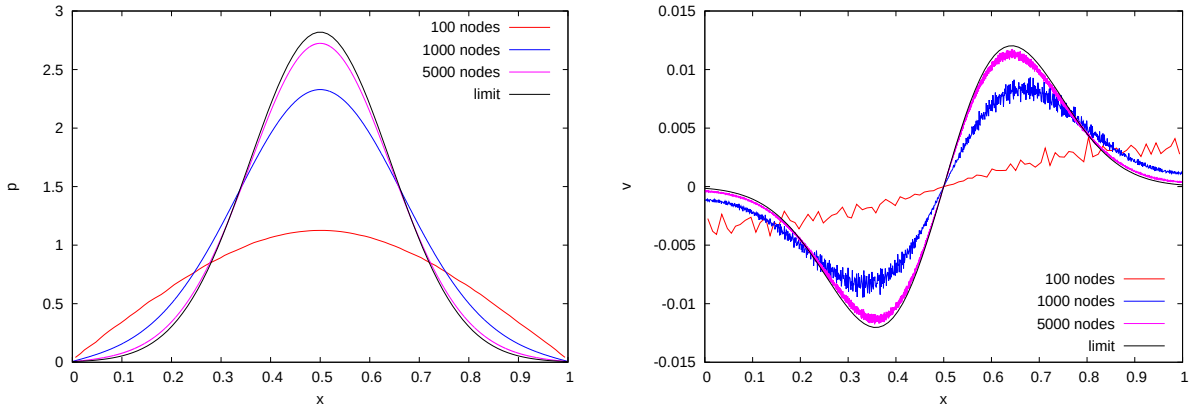

Figure 3.3 – Numerical solution obtained for the variable p (on the left) and v (on the right) with the standard DG method with two constant basis functions and different number of nodes. Bad accuracy on coarse meshes illustrates that this DG scheme is not AP.

3-2.1 Construction of the basis functions

The Su-Olson model reads [SO96]

$$\begin{cases} \varepsilon \partial_t E + \dfrac{1}{2}\partial_x F = -(E-\theta), \\ \dfrac{1}{2}\partial_x E = -\dfrac{3}{4}F, \\ \partial_t \theta = -(\theta - E), \end{cases} \tag{3.39}$$

where the unknown is $\mathbf{u} = (E, F, \theta)^T \in \mathbb{R}^3$ and $\varepsilon \in \mathbb{R}^+$. The model (3.39) can be recast under the form of a Friedrichs system (2.1) with

$$A_0 = \begin{pmatrix} \varepsilon & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 0 & -1 \\ 0 & \frac{3}{4} & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

Note that both A_0 and A_1 are degenerated matrices in the sense that they admit a row which is zero.

Proposition 3.15. *The following functions are solutions to the system* (3.39)

$$\begin{gathered} \mathbf{v}_1(t,\mathbf{x}) = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2(t,\mathbf{x}) = \begin{pmatrix} x \\ -\frac{2}{3} \\ x \end{pmatrix}, \quad \mathbf{v}_3(t,\mathbf{x}) = \begin{pmatrix} 1 \\ 0 \\ -\varepsilon \end{pmatrix} e^{-\frac{\varepsilon+1}{\varepsilon}t}, \\ \mathbf{v}_4(t,\mathbf{x}) = \begin{pmatrix} x \\ 0 \\ -\varepsilon x \end{pmatrix} e^{-\frac{\varepsilon+1}{\varepsilon}t}, \quad \mathbf{v}_5^{\pm}(t,\mathbf{x}) = \begin{pmatrix} 1+\lambda \\ \mp\frac{2}{\sqrt{3}}(1+\lambda)\sqrt{\lambda\frac{\varepsilon\lambda+\varepsilon+1}{1+\lambda}} \\ 1 \end{pmatrix} e^{\lambda t \pm \sqrt{3\lambda\frac{\varepsilon\lambda+\varepsilon+1}{1+\lambda}}x}, \end{gathered} \tag{3.40}$$

with $\lambda \in \mathbb{R}$, $\lambda \neq -1$.

Proof. Injecting F in the first equation we recast the model (3.39) under the form of a second order system

$$\begin{aligned} \varepsilon \partial_t E - \frac{1}{3}\partial_x^2 E &= -(E-\theta), \\ \partial_t \theta &= -(\theta - E). \end{aligned} \tag{3.41}$$

We search for solutions under the form

$$\mathbf{v}(t,x) = \begin{pmatrix} v_1(x) \\ v_2(x) \end{pmatrix} e^{\lambda t}.$$

Injecting $\mathbf{v}$ in the system (3.39), one finds after removing the exponentials

$$\begin{aligned} \Big(\varepsilon\lambda - \frac{1}{3}\partial_x^2\Big) v_1(x) &= -(v_1(x) - v_2(x)), \\ \lambda v_2(x) &= -(v_2(x) - v_1(x)). \end{aligned} \tag{3.42}$$

From the second equation one gets

$$v_2 = v_1/(1+\lambda), \quad \lambda \neq -1. \tag{3.43}$$

Using this equality in the first equation of (3.42) gives $(\varepsilon\lambda - \frac{1}{3}\partial_x^2)v_1(x) = -\lambda v_1(x)/(1+\lambda)$. That is

$$\partial_x^2 v_1(x) = 3\lambda \frac{\varepsilon\lambda + \varepsilon + 1}{1+\lambda} v_1(x), \quad \lambda \neq -1.$$

They are several cases:

- If $\lambda = 0$ or $\varepsilon\lambda + \varepsilon + 1 = 0$ then one has $\partial_x^2 v_1(x) = 0$.
  - When $\lambda = 0$ one finds from (3.43) the equality $v_1(x) = v_2(x)$. Therefore one has the following solutions
  $$\mathbf{v}_1(t,\mathbf{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2(t,\mathbf{x}) = \begin{pmatrix} x \\ x \end{pmatrix}. \tag{3.44}$$
  - When $\varepsilon\lambda + \varepsilon + 1 = 0$ one has $\lambda = -(\varepsilon+1)/\varepsilon$. Using (3.43) one finds $v_2(x) = -\varepsilon v_1(x)$ and one gets the following solutions
  $$\mathbf{v}_3(t,\mathbf{x}) = \begin{pmatrix} 1 \\ -\varepsilon \end{pmatrix} e^{-\frac{\varepsilon+1}{\varepsilon}t}, \quad \mathbf{v}_4(t,\mathbf{x}) = \begin{pmatrix} x \\ -\varepsilon x \end{pmatrix} e^{-\frac{\varepsilon+1}{\varepsilon}t}. \tag{3.45}$$
- If $\lambda \neq -1$ then, using (3.43), one finds the following solutions
$$\mathbf{v}_4(t,x) = \begin{pmatrix} 1+\lambda \\ 1 \end{pmatrix} e^{\lambda t \pm \sqrt{3\lambda\frac{\varepsilon\lambda+\varepsilon+1}{1+\lambda}}x}. \tag{3.46}$$

From the solutions (3.44)-(3.45)-(3.46) to the system (3.41) one deduces solutions to the system (3.39) using $F = -(2/3)\partial_x E$. The proof is complete. ■

3-2.2 Numerical results

We apply the TDG method on the numerical test given in [SO96]. Consider the domain $\Omega_S = [0, 15]$ and a total of 200 nodes with $dt = T/100$ where T is the final time. We take $\varepsilon = 0.1$ and for the boundary condition $\mathbf{u}_{|\partial\Omega}(t,x) = (\delta_0(x), 0, 0)^T$ where δ is the Kronecker symbol. In the following, we use random meshes constructed as follow: we start from a uniform mesh and moved the vertices randomly around their initial position by a factor of at most 33%.

For the basis functions, we consider the solutions (3.40) from $\mathbf{v}_1$ to $\mathbf{v}_4$ and take $\lambda = -\sqrt{\frac{\varepsilon+1}{\varepsilon}}$ for $\mathbf{v}_5^+$ and $\mathbf{v}_5^-$. After multiplying $\mathbf{v}_5^+$ and $\mathbf{v}_5^-$ by $\sqrt{\varepsilon}$ one gets

$$\begin{aligned} \mathbf{v}_5^+(t,x) &= \begin{pmatrix} \sqrt{\varepsilon} - \sqrt{\varepsilon+1} \\ -\frac{2}{\sqrt{3}}\Big(\sqrt{\varepsilon(\varepsilon+1)} - (\varepsilon+1)\Big) \\ \sqrt{\varepsilon} \end{pmatrix} e^{-\sqrt{\frac{\varepsilon+1}{\varepsilon}}t + \sqrt{3(\varepsilon+1)}x}, \\ \mathbf{v}_5^-(t,x) &= \begin{pmatrix} \sqrt{\varepsilon} - \sqrt{\varepsilon+1} \\ \frac{2}{\sqrt{3}}\Big(\sqrt{\varepsilon(\varepsilon+1)} - (\varepsilon+1)\Big) \\ \sqrt{\varepsilon} \end{pmatrix} e^{-\sqrt{\frac{\varepsilon+1}{\varepsilon}}t - \sqrt{3(\varepsilon+1)}x}. \end{aligned} \tag{3.47}$$

At first, we take the 6 basis functions $\mathbf{v}_1, ..., \mathbf{v}_5^+, \mathbf{v}_5^-$ and represent the numerical results obtained with the TDG method on the left of Figure 3.4 for the final times $T = 10^{-3}, 10^{-2}, 10^{-1}, 1, 10$. These results are consistent with those obtained in [SO96].

On the right of Figure 3.4 we compare the DG and TDG method by varying the number of basis functions. More precisely, we study the following cases

- The DG method with 3 basis functions per cell (constant basis functions only).
- The DG method with 6 basis functions per cell (affine basis functions).
- The TDG method with the 3 basis functions per cell $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$.
- The TDG method with the 5 basis functions per cell $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, $\mathbf{v}_5^+$ and $\mathbf{v}_5^-$.
- The TDG method with the 6 basis functions per cell $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, $\mathbf{v}_4$ $\mathbf{v}_5^+$ and $\mathbf{v}_5^-$.

To get the most accurate approximation one needs to take 6 basis functions for the DG and TDG method. Thus, the TDG and DG method give a similar result on this test. Moreover, note that we use logarithmic scale and the comparison is therefore not representative of the error.

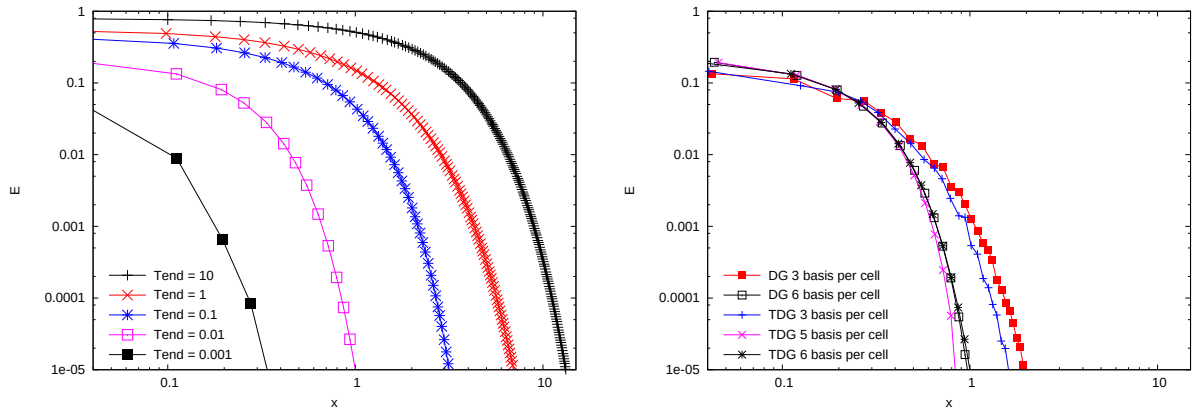

Figure 3.4 – On the left: representation of the variable E for the TDG method with 6 basis functions. On the right: comparison between the DG and TDG method at $T = 10^{-2}$ for different number of basis functions. Logarithmic scale.

Chapter 4

Analysis of the Trefftz discontinuous Galerkin method for the P_N model in 2D

Contents

In this chapter, the TDG method applied to the general P_N model is studied and analyzed. First, the P_N model is derived and some of its properties, including the rotational invariance, are given. Then, polynomial and exponential solutions are constructed. In particular, to deal with boundary layers, stationary exponential solutions are derived. Finally, the approximation properties of the stationary solutions and the convergence of the scheme are studied. A nice property of the TDG method is recovered: the number of additional basis functions to gain one order from k to $k+1$ does not depend on k. Therefore, to get high order schemes the TDG method uses, at least asymptotically, less basis functions than the standard DG method.

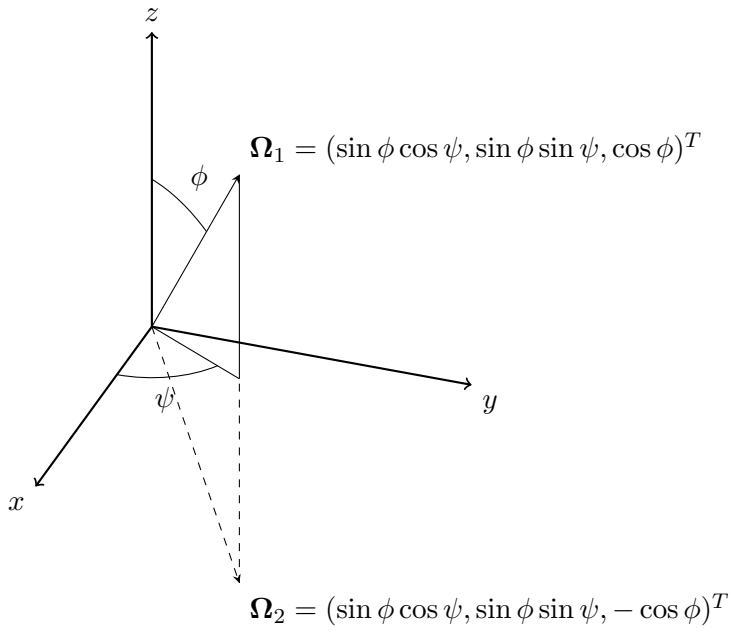

Figure 4.1 – Representation of directions $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$. If $\mathbf{u}$ is an even function of $\cos\phi$ then $\mathbf{u}(t, \mathbf{x}, \mathbf{\Omega}_1) = \mathbf{u}(t, \mathbf{x}, \mathbf{\Omega}_2)$.

4-1 The P_N model

The derivation of our model respect several principles

- Consider the P_N model with N odd. Indeed, even if the analysis can be carried out for the case N even too, the distinction odd/even has to be made in various cases which lengthens the presentation. In practice, the P_N model is rarely applied for even values of N (see for example [GH16, Section 2] for a discussion on the benefits of considering N odd) and it is therefore natural to consider only the case N odd in our analysis.
- Use the block structure given in [Her16]. This will be useful to simplify the structure, study some properties and calculate solutions to the model.
- Consider the two dimensional P_N model. In two dimensions, the size of the system is reduced with two assumptions
  (i) The solution does not depends on the variable z that is $\partial_z \mathbf{u} = 0$.
  (ii) The solution $\mathbf{u}$ is an even function of $\cos\phi$. This is equivalent to assume that the solution is symmetric with respect to the plan xy see Figure 4.1. In three dimensions, it can be interpreted as pure reflective conditions at the boundaries of the domain.
- Consider the P_N model in the plan xy. The plan xz may also be a possible choice [BH01, BDF15], however the rotation matrix associated with the spherical harmonics is more difficult to calculate in the plan xz [BFB97, IR96, PH07]. In practice, the rotation matrix will be useful to
  (i) Deduce two dimensional special solutions from the one dimensional case using the rotational invariance of the P_N model.
  (ii) Simplify the calculation of the matrices $M(\mathbf{n})$, $M^+(\mathbf{n})$, $M^-(\mathbf{n})$ for the numerical simulations, see Remark 4.14 below.

Such configurations are studied in the literature but the notations or the symmetry assumptions vary from one author to another. In the following, the presentation is unified.

4-1.1 Derivation from the transport equation

Let $\psi \in [0, 2\pi)$ and $\phi \in [0, \pi)$ be the polar and azimutal angles on the sphere, so that in Cartesian coordinate with usual notations

$$\mathbf{\Omega} := (\Omega_1, \Omega_2, \Omega_3)^T = (\sin\phi\cos\psi, \sin\phi\sin\psi, \cos\phi)^T \in \mathbb{R}^3.$$

To be consistent with the standard notation of the spherical harmonics, the uppercase letter $Y_{k,l}$ is used to denoted the real spherical harmonics. We make a slight abuse of notation by denoting indifferently

$$Y_{k,l}(\mathbf{\Omega}) := Y_{k,l}(\psi, \phi) : \mathbb{R}^2 \to \mathbb{R}, \quad |l| \leq k \leq N, \quad k, l \in \mathbb{N}.$$

The construction and properties of the spherical harmonics are detailed in Appendix A. We recall that the transport equation reads

$$\partial_t \mathcal{I}(t, \mathbf{x}, \mathbf{\Omega}) + \mathbf{\Omega} \cdot \nabla \mathcal{I}(t, \mathbf{x}, \mathbf{\Omega}) = -\Big(\sigma_a(\mathbf{x}) + \sigma_s(\mathbf{x})\Big)\mathcal{I}(t, \mathbf{x}, \mathbf{\Omega}) + \sigma_s(\mathbf{x}) < \mathcal{I} > (t, \mathbf{x}), \tag{4.1}$$

where $\mathcal{I}$ is the radiative intensity average in frequency, t the time variable, $\mathbf{x}$ the space variable, $\mathbf{\Omega}$ the direction and we use the notation

$$< \cdot > (t, \mathbf{x}) := \frac{1}{4\pi} \int_{S^2} d\mathbf{\Omega},$$

where S^2 is the unit sphere in $\mathbb{R}^3$. The absorption and the scattering coefficients are denoted respectively

$$\sigma_a(\mathbf{x}) \geq 0 \text{ and } \sigma_s(\mathbf{x}) \geq 0.$$

We introduce some notations and adopt the presentation from [GH16] but with the spherical harmonics vector arranged as in [Her16]. In the following, we denote m^{3D} the number of unknown, m_e^{3D} the number of even moments and m_o^{3D} the number of odd moments for the three dimensional P_N model. That is

$$m^{3D} := m_e^{3D} + m_o^{3D} = (N+1)^2, \quad m_e^{3D} := \frac{1}{2}N(N+1), \quad m_o^{3D} := \frac{1}{2}(N+1)(N+2).$$

For any integer $0 \leq k \leq N$ we define $\mathbf{y}_k(\mathbf{\Omega})$ the vectorial function whose components are the $2k+1$ real valued spherical harmonics of order k. Moreover we denote $\mathbf{y}_e(\mathbf{\Omega})$ the vectorial function made of the so-called even moments $\big(\mathbf{y}_{2k}(\mathbf{\Omega})\big)_{0\leq 2k\leq N}$ and $\mathbf{y}_o(\mathbf{\Omega})$ the vectorial function made of the so-called odd moments $\big(\mathbf{y}_{2k+1}(\mathbf{\Omega})\big)_{0\leq 2k+1\leq N}$. That is

$$\mathbf{y}_k(\mathbf{\Omega}) := \Big(Y_{k,-k}(\mathbf{\Omega}), Y_{k,-k+1}(\mathbf{\Omega}), ..., Y_{k,k-1}(\mathbf{\Omega}), Y_{k,k}(\mathbf{\Omega})\Big)^T \in \mathbb{R}^{2k+1},$$

$$\mathbf{y}_e(\mathbf{\Omega}) := \Big(\mathbf{y}_0^T(\mathbf{\Omega}), \mathbf{y}_2^T(\mathbf{\Omega}), ..., \mathbf{y}_{N-1}^T(\mathbf{\Omega})\Big)^T \in \mathbb{R}^{m_e^{3D}}, \quad \mathbf{y}_o(\mathbf{\Omega}) := \Big(\mathbf{y}_1^T(\mathbf{\Omega}), \mathbf{y}_3^T(\mathbf{\Omega}), ..., \mathbf{y}_N^T(\mathbf{\Omega})\Big)^T \in \mathbb{R}^{m_o^{3D}},$$

Finally, we define $\mathbf{y}(\mathbf{\Omega})$ the vectorial function made of $\mathbf{y}_e(\mathbf{\Omega})$, $\mathbf{y}_o(\mathbf{\Omega})$ and arranged as follow

$$\mathbf{y}(\mathbf{\Omega}) = \Big(\mathbf{y}_e^T(\mathbf{\Omega}), \mathbf{y}_o^T(\mathbf{\Omega})\Big)^T \in \mathbb{R}^{m^{3D}}.$$

We generalize this decomposition for any vector $\mathbf{v} \in \mathbb{R}^{m^{3D}}$. We set

$$\begin{aligned} \mathbf{v}_k &:= (v_k^{-k}, v_k^{-k+1}, ..., v_k^{k-1}, v_k^k)^T \in \mathbb{R}^{2k+1}, \\ \mathbf{v}_e := (\mathbf{v}_0^T, \mathbf{v}_2^T, ..., \mathbf{v}_{N-1}^T)^T \in \mathbb{R}^{m_e^{3D}}, &\quad \mathbf{v}_o := (\mathbf{v}_1^T, \mathbf{v}_3^T, ..., \mathbf{v}_N^T)^T \mathbb{R}^{m_o^{3D}}, \end{aligned} \tag{4.2}$$

and denote $\mathbf{v}$ as

$$\mathbf{v} = (\mathbf{v}_e^T, \mathbf{v}_o^T)^T \in \mathbb{R}^{m^{3D}}. \tag{4.3}$$

Now we introduce the decomposition of the function $\mathcal{I}(t,\mathbf{x},\mathbf{\Omega})$ on the spherical harmonics basis

$$\mathcal{I}(t,\mathbf{x},\mathbf{\Omega}) = \sum_{k\geq 0}\sum_{|l|\leq k} Y_{k,l}(\mathbf{\Omega})u_k^l(t,x).$$

The spherical harmonic approximation of (4.1) considers the truncated series $\mathcal{I}_N$ defined as

$$\mathcal{I}_N(t,\mathbf{x},\mathbf{\Omega}) := \mathbf{y}^T(\mathbf{\Omega})\mathbf{u}(t,\mathbf{x}) = \sum_{k=0}^{N}\mathbf{y}_k^T(\mathbf{\Omega})\mathbf{u}_k(t,x) = \sum_{k=0}^{N}\sum_{|l|\leq k} Y_{k,l}(\mathbf{\Omega})u_k^l(t,x),$$

where the unknown of the P_N model is $\mathbf{u}\in\mathbb{R}^{m^{3D}}$. With the approximation $\mathcal{I}=\mathcal{I}_N$ the equation (4.1) reads

$$\mathbf{y}^T(\mathbf{\Omega})\partial_t\mathbf{u}(t,\mathbf{x}) + \sum_{i=1}^{3}\Omega_i\mathbf{y}^T(\mathbf{\Omega})\partial_{x_i}\mathbf{u}(t,\mathbf{x}) = \Big(-(\sigma_a+\sigma_s)\mathbf{y}^T(\mathbf{\Omega})\mathbf{u}(t,\mathbf{x}) + \sigma_s <\mathbf{y}^T(\mathbf{\Omega})>\Big)\mathbf{u}(t,\mathbf{x}).$$

Multiplying by $\mathbf{y}(\mathbf{\Omega})$ and integrating over the sphere gives

$$\begin{aligned} &<\mathbf{y}(\mathbf{\Omega})\mathbf{y}^T(\mathbf{\Omega})>\partial_t\mathbf{u}(t,\mathbf{x}) + \sum_{i=1}^{3}<\Omega_i\mathbf{y}(\mathbf{\Omega})\mathbf{y}^T(\mathbf{\Omega})>\partial_{x_i}\mathbf{u}(t,\mathbf{x}) = \\ &\Big(-(\sigma_a+\sigma_s)<\mathbf{y}(\mathbf{\Omega})\mathbf{y}^T(\mathbf{\Omega})> + \sigma_s <\mathbf{y}(\mathbf{\Omega})><\mathbf{y}^T(\mathbf{\Omega})>\Big)\mathbf{u}(t,\mathbf{x}). \end{aligned} \tag{4.4}$$

From the orthogonal properties of the spherical harmonics one has $<\mathbf{y}(\mathbf{\Omega})\mathbf{y}^T(\mathbf{\Omega})>= I_{m^{3D}}$ and $<\mathbf{y}(\mathbf{\Omega})><\mathbf{y}^T(\mathbf{\Omega})>= \mathbf{e}_1\mathbf{e}_1^T$ with $\mathbf{e}_1=(1,0,...,0)^T\in\mathbb{R}^{m^{3D}}$. Therefore one gets the system

$$\partial_t\mathbf{u} + \sum_{i=1}^{3}\mathcal{A}_i\partial_{x_i}\mathbf{u} = -\mathcal{R}\mathbf{u}, \tag{4.5}$$

where

$$\mathbf{u}\in\mathbb{R}^{m^{3D}}, \quad \mathcal{A}_1,\mathcal{A}_2,\mathcal{R}\in\mathbb{R}^{m^{3D}\times m^{3D}}.$$

The matrices $\mathcal{A}_i$ are defined as

$$\mathcal{A}_i = <\Omega_i\mathbf{y}(\mathbf{\Omega})\mathbf{y}^T(\mathbf{\Omega})> \tag{4.6}$$

and can be computed using the recursion relations (A.4) to expand $\Omega_i\mathbf{y}(\mathbf{\Omega})$ in terms of spherical harmonics. As pointed in [Her16] the matrix $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ have the following block structure

$$\mathcal{A}_1 = \begin{pmatrix} 0 & \mathcal{A} \\ \mathcal{A}^T & 0 \end{pmatrix}, \quad \mathcal{A}_2 = \begin{pmatrix} 0 & \mathcal{B} \\ \mathcal{B}^T & 0 \end{pmatrix}, \quad \mathcal{A}_3 = \begin{pmatrix} 0 & \mathcal{C} \\ \mathcal{C}^T & 0 \end{pmatrix}, \tag{4.7}$$

where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}\in\mathbb{R}^{m_e^{3D}\times m_o^{3D}}$ are rectangular matrices. The matrix $\mathcal{R}$ is a diagonal matrix

$$\mathcal{R} = \mathrm{diag}(\sigma_a,\sigma_a+\sigma_s,...,\sigma_a+\sigma_s).$$

In the following we may use the notation $\sigma_t := \sigma_a+\sigma_s$.

4-1.2 Properties

In this section, we derive some properties in three dimensions of the P_N model based on the results given in [GH16]. We use the matrix representations of the rotation operators in the basis of spherical harmonics [BFB97, DX13, PH07]

$$\mathcal{U}(\alpha,\beta,\gamma)\in\mathbb{R}^{m^{3D}\times m^{3D}}, \tag{4.8}$$

where α, β and γ denotes rotation around the axes O_z, O_y and O_z respectively. The matrix $\mathcal{U}(\alpha,\beta,\gamma)$ is a block matrix for the vectors $\mathbf{y}_k(\mathbf{\Omega})$

$$\mathcal{U}(\alpha,\beta,\gamma) = \text{diag}\Big(\Delta_0(\alpha,\beta,\gamma), \Delta_2(\alpha,\beta,\gamma), ...\Delta_{m_e}(\alpha,\beta,\gamma), \Delta_1(\alpha,\beta,\gamma), ..., \Delta_{m_o}(\alpha,\beta,\gamma)\Big).$$

The matrices Δ_k reads [PH07]

$$\Delta_k(\alpha,\beta,\gamma) = \mathcal{W}_k'(\alpha)\mathcal{D}_k(\beta)\mathcal{W}_k'(\gamma) \in \mathbb{R}^{2k+1\times 2k+1}. \tag{4.9}$$

Here $\mathcal{D}_k \in \mathbb{R}^{2k+1\times 2k+1}$ is a d-Wigner matrix and the matrix $\mathcal{W}_k'$ has non-zero elements only on its diagonal and anti-diagonal

$$\mathcal{W}_k'(\alpha) = \begin{pmatrix} \cos k\alpha & & & & & & \sin k\alpha \\ & \ddots & & & & \iddots & \\ & & \cos 2\alpha & & \sin 2\alpha & & \\ & & & \cos\alpha & \sin\alpha & & \\ & 0 & & 1 & & 0 & \\ & & & -\sin\alpha & \cos\alpha & & \\ & & -\sin 2\alpha & & & \cos 2\alpha & \\ & \iddots & & & & \ddots & \\ -\sin k\alpha & & & & & & \cos k\alpha \end{pmatrix} \in \mathbb{R}^{2k+1\times 2k+1}. \tag{4.10}$$

To simplify the matrix $\mathcal{U}$ we may consider a rotation θ in the plan xy only and denote

$$\mathcal{U}_\theta := \mathcal{U}(0,0,\theta) \in \mathbb{R}^{m^{3D}\times m^{3D}}.$$

Using the expression of the block rotations (4.9), the structure of the matrix $\mathcal{U}_\theta$ can be written as

$$\mathcal{U}_\theta = \text{diag}\Big(\mathcal{W}_0'(\theta), \mathcal{W}_2'(\theta), ...\mathcal{W}_{m_e}'(\theta), \mathcal{W}_1'(\theta), ..., \mathcal{W}_{m_o}'(\theta)\Big),$$

where the blocks $\mathcal{W}_k'(\theta)$ are given by (4.10).

The matrix $\mathcal{U}$ represents the orthogonal transformations on $\mathbf{y}(\mathbf{\Omega})$. That is for an orthogonal matrix $Q \in \mathbb{R}^{3\times 3}$ one has

$$\mathbf{y}(Q\mathbf{\Omega}) = \mathcal{U}(\alpha,\beta,\gamma)\mathbf{y}(\mathbf{\Omega}), \tag{4.11}$$

where α, β and γ are the angles of the rotation associated with the matrix Q in $\mathbb{R}^3$.

Example 4.1 (The P_1 model in $3D$)**.** For the three dimensional P_1 model $m^{3D} = 4$. The matrices $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$, $\mathcal{R}$ and $\mathcal{U}_\theta$ are

$$\mathcal{A}_1 = \frac{1}{\sqrt{3}}\begin{pmatrix} 0&0&0&1\\0&0&0&0\\0&0&0&0\\1&0&0&0\end{pmatrix}, \quad \mathcal{A}_2 = \frac{1}{\sqrt{3}}\begin{pmatrix} 0&1&0&0\\1&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix}, \quad \mathcal{A}_3 = \frac{1}{\sqrt{3}}\begin{pmatrix} 0&0&1&0\\0&0&0&0\\1&0&0&0\\0&0&0&0\end{pmatrix},$$

$$\mathcal{R} = \begin{pmatrix} \sigma_a&0&0&0\\0&\sigma_t&0&0\\0&0&\sigma_t&0\\0&0&0&\sigma_t\end{pmatrix}, \quad \mathcal{U}_\theta = \begin{pmatrix} 1&0&0&0\\0&\cos\theta&0&\sin\theta\\0&0&1&0\\0&-\sin\theta&0&\cos\theta\end{pmatrix}.$$

●

Invertibility of $\mathcal{A}\mathcal{A}^T$, $\mathcal{B}\mathcal{B}^T$ and $\mathcal{C}\mathcal{C}^T$. The first result of this section is about the eigenvalues and eigenvectors of the matrices $\mathcal{A}\mathcal{A}^T$, $\mathcal{B}\mathcal{B}^T$ and $\mathcal{C}\mathcal{C}^T$ in terms of the eigenvalues and eigenvectors of the matrices $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$. The eigenvalues of $\mathcal{A}\mathcal{A}^T$ and $\mathcal{B}\mathcal{B}^T$ play an important role in the analysis of the TDG scheme. In particular the invertibility of $\mathcal{A}\mathcal{A}^T$ will be used to study the convergence.

Proposition 4.2 (Invertibility of $\mathcal{A}\mathcal{A}^T$, $\mathcal{B}\mathcal{B}^T$ and $\mathcal{C}\mathcal{C}^T$)**.** *The symmetric matrix $\mathcal{A}\mathcal{A}^T$ is invertible and all its eigenvalues are strictly positive. A similar result holds for the matrices $\mathcal{B}\mathcal{B}^T$ and $\mathcal{C}\mathcal{C}^T$.*

To prove Proposition 4.2 we need the some technical Lemmas.

Lemma 4.3. *Let $Q \in \mathbb{R}^3$ be an orthogonal matrix, assume $\boldsymbol{\nu}, \boldsymbol{\nu}_* \in \mathbb{R}^3$ satisfy $\boldsymbol{\nu}_* = Q\boldsymbol{\nu}$ and define the two matrices*

$$M := < (\boldsymbol{\nu}^T\boldsymbol{\Omega})\mathbf{y}(\boldsymbol{\Omega})\mathbf{y}^T(\boldsymbol{\Omega}) > \in \mathbb{R}^{m^{3D}\times m^{3D}}, \quad M_* := < (\boldsymbol{\nu}_*^T\boldsymbol{\Omega})\mathbf{y}(\boldsymbol{\Omega})\mathbf{y}^T(\boldsymbol{\Omega}) > \in \mathbb{R}^{m^{3D}\times m^{3D}}.$$

Then one has

$$M_* = \mathcal{U}(\alpha,\beta,\gamma) M \mathcal{U}^T(\alpha,\beta,\gamma),$$

where α, β and γ are the angles of the rotation associated with the matrix Q in $\mathbb{R}^3$.

Proof. The proof is taken from [GH16]. With the change of variable $\boldsymbol{\Omega}' = Q\boldsymbol{\Omega}$ and the fact that Q is an orthogonal matrix one gets

$$< (\boldsymbol{\nu}_*^T\boldsymbol{\Omega})\mathbf{y}(\boldsymbol{\Omega})\mathbf{y}^T(\boldsymbol{\Omega}) > = < (\boldsymbol{\nu}_*^T Q\boldsymbol{\Omega})\mathbf{y}(Q\boldsymbol{\Omega})\mathbf{y}^T(Q\boldsymbol{\Omega}) > .$$

With $\boldsymbol{\nu}_* = Q\boldsymbol{\nu}$ one finds

$$< (\boldsymbol{\nu}_*^T\boldsymbol{\Omega})\mathbf{y}(\boldsymbol{\Omega})\mathbf{y}^T(\boldsymbol{\Omega}) > = < (\boldsymbol{\nu}^T\boldsymbol{\Omega})\mathbf{y}(Q\boldsymbol{\Omega})\mathbf{y}^T(Q\boldsymbol{\Omega}) > .$$

Using (4.11) completes the proof. ■

From Lemma 4.3 one immediately deduces the following corollary on the eigenstructure of the matrices $\mathcal{A}_i$.

Corollary 4.4. *The eigenvalues of the matrices $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ are the same and their eigenvectors differ by a unitary transformation.*

Proof. The proof is taken from [GH16]. One uses Lemma 4.3 with $\boldsymbol{\nu}$, $\boldsymbol{\nu}_*$ aligned with one of the three Cartesian axes (that is $\boldsymbol{\nu}^T\boldsymbol{\Omega} = \Omega_i$ and $\boldsymbol{\nu}_*^T\boldsymbol{\Omega} = \Omega_j$, $1 \le i, j \le 3$). One concludes with the definition of the matrices $\mathcal{A}_i$ (4.6). ■

Finally we will use the following lemma which give some structure on the kernel of the matrices $\mathcal{A}_i$.

Lemma 4.5. *Let N be odd, $\mathbf{v} = (\mathbf{v}_e, \mathbf{v}_o)^T \in \mathbb{R}^{m^{3D}}$ where $\mathbf{v}_e$, $\mathbf{v}_o$ are as in decomposition (4.3) and assume $\mathcal{A}_3\mathbf{v} = \mathbf{0}$. Then one has $\mathbf{v}_e = \mathbf{0}$. A similar result holds for the matrices $\mathcal{A}_1$ and $\mathcal{A}_2$.*

Proof. The proof is based on [GH16, Theorem 3]. First, assume $\mathcal{A}_3\mathbf{v} = \mathbf{0}$. Since N is odd and using [GH16, Theorem 3] one has $v_{2k}^l = 0$ for $|l| \le 2k \le N$. From the definition (4.2) of $\mathbf{v}_e$ one gets $\mathbf{v}_e = \mathbf{0}$ and this give the result for $\mathcal{A}_3$.

From Corollary 4.4, the eigenvectors of the matrices $\mathcal{A}_i$ differ by the transformation $\mathcal{U}$ (4.8) which is block diagonal considering the components $\mathbf{v}_k$. Using the definition (4.2) of $\mathbf{v}_e$ one deduces the result for the matrices $\mathcal{A}_1$ and $\mathcal{A}_2$. ■

We can now prove Proposition 4.2.

Proof of Proposition 4.2. We give the proof for the matrix $\mathcal{A}\mathcal{A}^T$, the proof for the matrices $\mathcal{B}\mathcal{B}^T$ and $\mathcal{C}\mathcal{C}^T$ is similar. Consider $\mathbf{v}^i$, $i = 1, ..., m^{3D}$ the eigenvectors of the matrix $\mathcal{A}_1$ associated with the eigenvalues λ_i. We use the decomposition (4.3) to denote $\mathbf{v}^i = (\mathbf{v}_e^i, \mathbf{v}_o^i)^T$ with $\mathbf{v}_e^i \in \mathbb{R}^{m_e^{3D}}$. Because $\mathcal{A}_1$ is symmetric, there exists m_e^{3D} eigenvectors $\mathbf{v}^i$ such that the vectors $\mathbf{v}_e^i$ form a basis of $\mathbb{R}^{m_e^{3D}}$. Up to a reordering, one can assume that the vectors $\mathbf{v}_e^i$, $i = 1, ..., m_e^{3D}$ form a basis of $\mathbb{R}^{m_e^{3D}}$. That is

$$\dim\left(\text{Span}\left\{\mathbf{v}_e^1, ..., \mathbf{v}_e^{m_e^{3D}}\right\}\right) = m_e^{3D}. \tag{4.12}$$

Since $\mathcal{A}_1\mathbf{v}^i = \lambda_i\mathbf{v}^i$ and from the block structure (4.7) of the matrix $\mathcal{A}_1$ one gets

$$\begin{cases} \mathcal{A}\mathbf{v}_o^i = \lambda_i\mathbf{v}_e^i \\ \mathcal{A}^T\mathbf{v}_e^i = \lambda_i\mathbf{v}_o^i. \end{cases} \tag{4.13}$$

Multiplying the second equation by $\mathcal{A}$ and using the first equation gives

$$\mathcal{A}\mathcal{A}^T\mathbf{v}_e^i = \lambda_i^2\mathbf{v}_e^i. \tag{4.14}$$

From (4.12)-(4.14) one deduces that the vectors $\mathbf{v}_e^i$, $i = 1, ..., m_e^{3D}$ are all the eigenvectors of the matrix $AA^T \in \mathbb{R}^{m_e^{3D}\times m_e^{3D}}$ and are associated with the eigenvalues λ_i^2.

Now if $\lambda_i = 0$ it implies from Lemma 4.5 that $\mathbf{v}_e^i = \mathbf{0}$. From (4.12) this is not possible and one deduces $\lambda_i \neq 0$ for $i = 1, ..., m_e^{3D}$. Therefore, the matrix $\mathcal{A}\mathcal{A}^T$ admits no zero eigenvalue and the proof is complete. ∎

Eigenvectors of $\mathcal{A}\mathcal{A}^T$ with a non zero first component. A special attention is devoted to the eigenvectors of $\mathcal{A}\mathcal{A}^T$ with a non zero first component. The following proposition will be useful later in the proof of Proposition 4.23.

Proposition 4.6 (Eigenvectors of $\mathcal{A}\mathcal{A}^T$ with a non zero first component)**.** *The eigenvectors of $\mathcal{A}\mathcal{A}^T$ with a non zero first component are associated with distinct eigenvalues.*

To prove Proposition 4.6 we will need the following lemma which is taken from [GH16]. To avoid confusion between the P_N model and the Legendre polynomials we denote in the following lemma Q_k the Legendre Polynomial of degree k. Moreover we denote δ^{jk} the Kronecker symbol.

Lemma 4.7. *The eigenvalues of $\mathcal{A}_3$ are the roots of the polynomial $\partial_x^{(|j|)}Q_{N+1}$ for $|j| \leq N$. More precisely, if λ is a root of $\partial_x^{(|j|)}Q_{N+1}$, then for any fixed ϕ, the vector $\mathbf{v}$ with components*

$$v_k^l = Y_l^j(\cos^{-1}(\lambda), \phi)\delta^{jk},$$

is an eigenvector of $\mathcal{A}_3$ associated with λ.

Proof. The proof is given in [GH16, Lemma 2]. ∎

The following lemma is also useful.

Lemma 4.8. *Assume λ is an eigenvalue of $\mathcal{A}_1$ associated with the eigenvector $\mathbf{v}_1 = (\mathbf{v}_e, \mathbf{v}_o)^T$. Then $-\lambda$ is an eigenvalue of $\mathcal{A}_1$ associated with the eigenvector $\mathbf{v}_2 = (-\mathbf{v}_e, \mathbf{v}_o)^T$*

Proof. This is a direct consequence of the block structure of the matrix $\mathcal{A}_1$ (4.7). ∎

We can now give the proof of Proposition 4.6.

Proof of Proposition 4.6. We proceed in three steps

1. First, we show that the eigenvectors of $\mathcal{A}_1$ with a non zero first component are associated with distinct eigenvalues. Indeed, from Lemma 4.7, one deduces that the eigenvectors of $\mathcal{A}_3$ associated with a non zero first component (that is $v_0^0 \neq 0$) are roots of the Legendre polynomial Q_{N+1} and therefore distinct. One concludes with Corollary 4.4 that the result holds for the matrix $\mathcal{A}_1$ since the eigenvectors of the matrices $\mathcal{A}_i$ differ by a unitary transformation which is block diagonal for the first component v_0^0.
2. Then, we show that the eigenvectors and eigenvalues of $\mathcal{A}\mathcal{A}^T$ can be deduced from the eigenvectors and eigenvalues of $\mathcal{A}_1$. To do so, we proceed as in the proof of Proposition 4.2. We consider $\mathbf{v}^i = (\mathbf{v}_e^i, \mathbf{v}_o^i)^T$, $i = 1, ..., m_e^{3D}$ which are the eigenvectors of $\mathcal{A}_1$ such that the $\mathbf{v}_e^i$, $i = 1, ..., m_e^{3D}$, form a basis of $\mathbb{R}^{m_e^{3D}}$. Up to a reordering, one can denote $\mathbf{v}^i$, $i = 1, ..., k$, $k \leq m_e^{3D}$, all the eigenvectors with a non zero first component and λ_i the eigenvalues associated. From the equality (4.14), the eigenvalues of $\mathcal{A}\mathcal{A}^T$ associated with an eigenvector with a non zero first component are $\lambda_1^2, ..., \lambda_k^2$.
3. Finally, we show that these eigenvalues are distinct that is $\lambda_l \neq \pm\lambda_j$ for $l \neq j$ and $l, j \leq k$.
   — From the first item, the eigenvalues associated with a non zero first component are distinct that is $\lambda_l \neq \lambda_j$.
   — Now assume $\lambda_l = -\lambda_j$. In particular, λ_l and λ_j are two eigenvalues of $\mathcal{A}_1$. From Lemma 4.8 the vector $\mathbf{w} = (-\mathbf{v}_e^l, \mathbf{v}_o^l)^T$ is an eigenvector of $\mathcal{A}_1$ with a non zero first component associated with the eigenvalue $-\lambda_l$. But since the λ_i are distinct, the eigenvector $\mathbf{v}_j$ is the only eigenvector of $\mathcal{A}_1$ with a non zero first component associated with the eigenvalue $-\lambda_l$. One deduces $\mathbf{v}^j = \mathbf{w} = (-\mathbf{v}_e^l, \mathbf{v}_o^l)^T$. Therefore $\mathbf{v}_e^j = -\mathbf{v}_e^l$ which is impossible because the eigenvectors $\mathbf{v}_e^i$ form a basis of $\mathbb{R}^{m_e^{3D}}$.

   One finally deduces that the eigenvalues $\lambda_1^2, ..., \lambda_k^2$ are all distinct. The proof is complete.

■

Rotational relations in 3D. The following proposition establishes a relation between the matrices $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{U}_\theta$. Later in this chapter, we will use this relation to show the rotational invariance of the solutions to the P_N model in two dimensions.

Proposition 4.9 (Rotational relations in 3D)**.** *The matrices $\mathcal{A}_1$ and $\mathcal{A}_2$ satisfy the relation*

$$\mathcal{A}_1 = \mathcal{U}_\theta(\mathcal{A}_1 \cos\theta - \mathcal{A}_2 \sin\theta)\mathcal{U}_\theta^T, \quad \mathcal{A}_2 = \mathcal{U}_\theta(\mathcal{A}_1 \sin\theta + \mathcal{A}_2 \cos\theta)\mathcal{U}_\theta^T.$$

Proof. The general proof is based on Lemma 4.3. Let Q_θ be the rotation matrix of angle θ in the plan xy

$$Q_\theta = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3},$$

and assume $\boldsymbol{\nu}, \boldsymbol{\nu}_* \in \mathbb{R}^3$ satisfy

$$\boldsymbol{\nu}_* = Q_\theta \boldsymbol{\nu} \in \mathbb{R}^3. \tag{4.15}$$

One can define the two matrices

$$M := < (\boldsymbol{\nu}^T \boldsymbol{\Omega})\mathbf{y}(\boldsymbol{\Omega})\mathbf{y}^T(\boldsymbol{\Omega}) > \in \mathbb{R}^{m^{3D}\times m^{3D}}, \quad M_* := < (\boldsymbol{\nu}_*^T \boldsymbol{\Omega})\mathbf{y}(\boldsymbol{\Omega})\mathbf{y}^T(\boldsymbol{\Omega}) > \in \mathbb{R}^{m^{3D}\times m^{3D}}.$$

From Lemma 4.3 one has

$$M_* = \mathcal{U}_\theta M \mathcal{U}_\theta^T.$$

Using the definition (4.6) of the matrices $\mathcal{A}_i$ and taking $\boldsymbol{\nu}_* = (1,0,0)^T$, $\boldsymbol{\nu} = (\cos\theta, -\sin\theta, 0)^T$ immediately give

$$\mathcal{A}_1 = \mathcal{U}_\theta(\mathcal{A}_1\cos\theta - \mathcal{A}_2\sin\theta)\mathcal{U}_\theta^T.$$

For the second equality consider $\boldsymbol{\nu}_* = (0,1,0)^T$, $\boldsymbol{\nu} = (\sin\theta, \cos\theta, 0)^T$ and one gets

$$\mathcal{A}_2 = \mathcal{U}_\theta(\mathcal{A}_1\sin\theta + \mathcal{A}_2\cos\theta)\mathcal{U}_\theta^T.$$

This completes the proof. ■

4-1.3 Derivation and properties in the two dimensional case

4-1.3.1 Derivation from 3D principles

The goal in this section is to derive the P_N model in two dimensions from the three dimensional case. More precisely, in two dimensions the P_N model can be decoupled in two systems: the unknowns u_k^m such that $k+m$ is odd and the unknowns u_k^m such that $k+m$ is even. Making the assumption that the function $\mathbf{u}$ is an even function of $\cos\phi$, one can remove the unknown such that $k+m$ is odd. This simplifies the matrices $\mathcal{A}_1$ and $\mathcal{A}_2$ and one gets the P_N model in two dimensions.

Definition 4.10. Consider $\mathbf{u} \in \mathbb{R}^{m^{3D}}$, $\mathbf{u} = (\mathbf{u}_0, \mathbf{u}_2, ..., \mathbf{u}_1, \mathbf{u}_3, ...)^T$ as in (4.2) where $\mathbf{u}_l$ refers to the moments $\mathbf{u}_l = (u_l^{-l}, u_l^{-l+1}, ..., u_l^{l-1}, u_l^l)^T$. We introduce a second even/odd decomposition to derive the two dimensional P_N model and define the spaces

$$S_e = \{\mathbf{u} \in \mathbb{R}^{m^{3D}},\ u_k^l = 0,\ \forall\ \text{k+l odd}\}, \quad S_o = \{\mathbf{u} \in \mathbb{R}^{m^{3D}},\ u_k^l = 0,\ \forall\ \text{k+l even}\}.$$

Moreover we set

$$m := \dim S_e.$$

One notices that $\mathbb{R}^{m^{3D}} = S_e \oplus S_o$. An important property in the two dimensional case is that the matrices $\mathcal{A}_1$ and $\mathcal{A}_2$ preserve the two spaces S_e and S_o.

Proposition 4.11. *The space S_e (resp. S_o) is invariant under the application of the matrices $\mathcal{A}_1$ and $\mathcal{A}_2$*

$$\mathcal{A}_1 S_e \subset S_e, \quad \mathcal{A}_2 S_e \subset S_e, \quad \mathcal{A}_1 S_o \subset S_o, \quad \mathcal{A}_2 S_o \subset S_o.$$

Proof. This is a direct consequence of the relations (A.4) given in Appendix A and the definition $\mathcal{A}_i := < \Omega_i \mathbf{y}\mathbf{y}^T >$. ■

We use this property to show that, in 2 dimensions, the P_N model can be decoupled in two systems with independent solutions $\mathbf{u}_e \in S_e$ and $\mathbf{u}_o \in S_o$. We make the assumption that the function $\mathbf{u}$ is an even function of $\cos\phi$. From the definition of the spherical harmonics (A.3), this is equivalent to take $u_k^l = 0$ if $k+l$ odd. Therefore, we are interested in the solution $\mathbf{u} \in S_e$ and we consider a system of dimension m.

Definition 4.12. Let $\mathbf{v}_i \in \mathbb{R}^{m^{3D}}$, $i = 1, ..., m$ be the canonical basis of S_e in $\mathbb{R}^{m^{3D}}$ and $\mathbf{w}_j \in \mathbb{R}^m$, $j = 1, ..., m$ the canonical basis of $\mathbb{R}^m$. We define the matrix $P \in \mathbb{R}^{m \times m^{3D}}$ such that

$$P = \sum_{i=1}^{m} \mathbf{w}_i \mathbf{v}_i^T \in \mathbb{R}^{m \times m^{3D}},$$

and the matrices A_1, A_2, R, U_θ are defined as

$$\begin{aligned} A_1 &= P\mathcal{A}_1 P^T \in \mathbb{R}^{m\times m}, \quad A_2 = P\mathcal{A}_2 P^T \in \mathbb{R}^{m\times m}, \\ R &= P\mathcal{R} P^T \in \mathbb{R}^{m\times m}, \quad U_\theta = P\mathcal{U}_\theta P^T \in \mathbb{R}^{m\times m}. \end{aligned} \tag{4.16}$$

The matrices A_1, A_2, R and U_θ are simply the matrices $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{R}$ and $\mathcal{U}_\theta$ where the columns and rows corresponding to components v_k^l such that $k+l$ is odd have been removed. Therefore the matrices A_1, A_2, R have the same block structure as in (4.7).

Example 4.13 (The P_1 model in 2D)**.** For the P_1 model $m = 3$. The matrix P reads

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and the matrices A_1, A_2, R and U_θ are

$$\begin{aligned} A_1 &= \frac{1}{\sqrt{3}} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad A_2 = \frac{1}{\sqrt{3}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\ R &= \begin{pmatrix} \sigma_a & 0 & 0 \\ 0 & \sigma_t & 0 \\ 0 & 0 & \sigma_t \end{pmatrix}, \qquad U_\theta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}. \end{aligned}$$

●

4-1.3.2 The two dimensional case

We can now give the P_N model in two dimensions. Since N is odd one has

$$m = \frac{1}{2}(N+1)(N+2), \quad m_e = \frac{1}{4}(N+1)^2, \quad m_o = \frac{1}{4}(N+1)(N+3),$$

where m is the total size of the system, m_e is the number of even moments and m_o is the number of odd moments. The two dimensional P_N model reads

$$\Big(A_0\partial_t + A_1\partial_x + A_2\partial_y\Big)\mathbf{u}(t,\mathbf{x}) = -R\mathbf{u}(t,\mathbf{x}), \tag{4.17}$$

where $\mathbf{x} = (x,y)^T$, $\mathbf{u} \in \mathbb{R}^m$, $A_1, A_2, R \in \mathbb{R}^{m\times m}$. Since we have adopted the order given in [Her16] the matrices A_1 and A_2 have the block structure

$$A_0 = \varepsilon I_m \in \mathbb{R}^{m\times m}, \quad A_1 = c\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad A_2 = c\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}, \tag{4.18}$$

where $I_m \in \mathbb{R}^{m\times m}$ is the identity matrix, A, $B \in \mathbb{R}^{m_e\times m_o}$ are rectangular matrices and we have introduced the coefficients

$$c \in \mathbb{R}_+, \quad \varepsilon \in \mathbb{R}_+^*.$$

In particular, when $\varepsilon \to 0$, the P_N model admits a diffusion limit [Her16]. To get the same block structure for the matrix R as in (4.18), we may write

$$R = \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix} \in \mathbb{R}^{m\times m}, \tag{4.19}$$

where R_1 and R_2 are both diagonal matrices

$$R_1 := \mathrm{diag}(\varepsilon\sigma_a, \sigma_t, ..., \sigma_t) \in \mathbb{R}^{m_e \times m_e}, \quad R_2 := \sigma_t I_{m_o} \in \mathbb{R}^{m_o \times m_o}.$$

with I_{m_o} the identity matrix of $\mathbb{R}^{m_o \times m_o}$ and

$$\sigma_t := \sigma_t^\varepsilon := \varepsilon\sigma_a + \frac{\sigma_s}{\varepsilon}, \quad \sigma_a, \sigma_s \in \mathbb{R}_+.$$

The rotation matrix U_θ given in (4.16) reads

$$U_\theta = \begin{pmatrix} V_{2\theta} & 0 \\ 0 & V_{2\theta+1} \end{pmatrix} \in \mathbb{R}^{m \times m}, \tag{4.20}$$

where $V_{2\theta}, V_{2\theta+1}$ denotes respectively the rotation on the even and odd spherical harmonics

$$V_{2\theta} = \begin{pmatrix} W_0 & & & \\ & W_2 & 0 & \\ & & \ddots & \\ & 0 & & W_{N-1} \end{pmatrix} \in \mathbb{R}^{m_e \times m_e}, \quad V_{2\theta+1} = \begin{pmatrix} W_1 & & & \\ & W_3 & 0 & \\ & & \ddots & \\ & 0 & & W_N \end{pmatrix} \in \mathbb{R}^{m_o \times m_o}. \tag{4.21}$$

Each diagonal block W_{2l} is the rotation matrix for the spherical harmonics of order $2l$. Therefore it reads

$$W_{2l} = \begin{pmatrix} \cos 2l\theta & & & & & & \sin 2l\theta \\ & \cos 2(l-1)\theta & & & & \sin 2(l-1)\theta & \\ & & \ddots & & \iddots & & \\ & & & \cos 2\theta & \sin 2\theta & & \\ & 0 & & 1 & & 0 & \\ & & & -\sin 2\theta & \cos 2\theta & & \\ & & \iddots & & \ddots & & \\ & -\sin 2(l-1)\theta & & 0 & & \cos 2(l-1)\theta & \\ -\sin 2l\theta & & & & & & \cos 2l\theta \end{pmatrix} \in \mathbb{R}^{(l+1)\times(l+1)}. \tag{4.22}$$

Similarly, each diagonal block W_{2l+1} is the rotation matrix for the spherical harmonics of order $2l+1$

$$W_{2l+1} = \begin{pmatrix} \cos(2l+1)\theta & & & & & & \sin(2l+1)\theta \\ & \cos(2l-1)\theta & & & & \sin(2l-1)\theta & \\ & & \ddots & & \iddots & & \\ & & & \cos\theta & \sin\theta & & \\ & 0 & & -\sin\theta & \cos\theta & 0 & \\ & & \iddots & & \ddots & & \\ & -\sin(2l-1)\theta & & 0 & & \cos(2l-1)\theta & \\ -\sin(2l+1)\theta & & & & & & \cos(2l+1)\theta \end{pmatrix} \in \mathbb{R}^{(l+1)\times(l+1)}. \tag{4.23}$$

Remark 4.14 (Determination of the matrices M^+ and M^- (2.3) with the rotation matrix)**.** The rotation matrix U_θ can be used to calculate the matrices $M(\mathbf{n})$, $M^+(\mathbf{n})$, $M^-(\mathbf{n})$. We recall

the decomposition $M(\mathbf{n}) = M^+(\mathbf{n}) + M^-(\mathbf{n})$ where $M(\mathbf{n}) = A_1 n_x + A_2 n_y$, $\mathbf{n} = (n_x, n_y)^T$. In the following, we will use the decomposition

$$M^+(\mathbf{n}) = \sum_{\lambda_i>0} \lambda_i \mathbf{r}_i \mathbf{r}_i^T, \quad M^-(\mathbf{n}) = \sum_{\lambda_i<0} \lambda_i \mathbf{r}_i \mathbf{r}_i^T, \quad |M|(\mathbf{n}) := \sum_i |\lambda_i| \mathbf{r}_i \mathbf{r}_i^T, \tag{4.24}$$

where $\mathbf{r}_i$ are the eigenvector of the matrix M associated with the eigenvalue λ_i.

The eigenvalues and eigenvectors of the matrix M are therefore required to determinate the matrices M^+ and M^- but, for large values of N, they can be challenging to calculate. Instead, one can use the eigenvalues and eigenvectors of the matrix A_1.

Indeed, since $\mathbf{n} = (n_x, n_y)$ is the outward normal of a given edge, one can write $\mathbf{n} = (\cos\theta, \sin\theta)$ with $\theta \in [0, 2\pi[$. Therefore, another way to write the matrix M is $M = A_1 \cos\theta + A_2 \sin\theta$. From the Proposition 4.18 one gets

$$M(\mathbf{n}) = U_{-\theta}^T A_1 U_{-\theta} \tag{4.25}$$

where $U_{-\theta}$ is given by the five equalities from (4.20) to (4.23). In particular, one deduces from (4.25) that $\mathbf{r}_i$ is an eigenvector of M associated with the eigenvalue λ_i if and only if $U_\theta^T \mathbf{r}_i$ is an eigenvector of A_1 associated with the eigenvalue λ_i. The eigenvalues of A_1 are roots of the Legendre polynomials (and their derivatives) and their eigenvectors are known up to a rotation with the matrix $\mathcal{U}(\alpha, \beta, \gamma)$ [GH16]. ●

4-1.3.3 Properties

In this section, we derive some properties of the P_N model in two dimensions. Later in this chapter, such properties will be used to construct the basis functions and study the convergence of the scheme. In particular, the Propositions 4.2 and 4.9 are adapted to the two dimensional case.

Technical lemmas. We begin with two technical lemmas.

Lemma 4.15. *One has $PP^T = I_m$ where $I_m \in \mathbb{R}^{m\times m}$ is the identity matrix of $\mathbb{R}^{m\times m}$. Moreover one has $P^T P = I_e$ where $I_e \in \mathbb{R}^{m^{3D}\times m^{3D}}$ is the projection on S_e orthogonal at S_o that is $I_e \mathbf{u}_e = \mathbf{u}_e$, $I_e \mathbf{u}_o = \mathbf{0}$ for all $\mathbf{u}_e \in S_e$, $\mathbf{u}_o \in S_o$.*

Proof. From the Definition 4.12 one has $P^T P = \sum_{i,j} \mathbf{w}_i \mathbf{v}_i^T \mathbf{v}_j \mathbf{w}_j^T$. But since $\mathbf{v}_i$ is the canonical basis of S_e, one finds $\mathbf{v}_i^T \mathbf{v}_j = 0$ if $i \neq j$ and $\mathbf{v}_i^T \mathbf{v}_i = 1$. Therefore $P^T P = \sum_{i=1}^m \mathbf{w}_i \mathbf{w}_i^T$ and because $\mathbf{w}_i$ is the canonical basis of $\mathbb{R}^m$ one gets $P^T P = I_m$.

In the same way, $PP^T = \sum_{i,j} \mathbf{v}_i \mathbf{w}_i^T \mathbf{w}_j \mathbf{v}_j^T = \sum_{i=1}^m \mathbf{v}_i \mathbf{v}_i^T := I_e$. From the definition of $I_e = \sum_{i=1}^m \mathbf{v}_i \mathbf{v}_i^T$, one deduces $I_e \mathbf{u}_e = \mathbf{u}_e$, $I_e \mathbf{u}_o = \mathbf{0}$ for all $\mathbf{u}_e \in S_e$, $\mathbf{u}_o \in S_o$. This completes the proof. ■

Lemma 4.16. *Assume $\mathbf{r} \in S_e$ one has*

$$\mathcal{A}_1 \mathbf{r} = \lambda \mathbf{r} \quad \Leftrightarrow \quad A_1 P \mathbf{r} = \lambda P \mathbf{r}.$$

Proof. A direct consequence of Proposition 4.11 is that the eigenvectors of $\mathcal{A}_1$ and $\mathcal{A}_2$ can be chosen such that they belong to S_e or S_o. Assume $\mathbf{r} \in S_e$ and $\mathcal{A}_1 \mathbf{r} = \lambda \mathbf{r}$. Using $P^T P = I_e$ one has $\mathcal{A}_1 P^T P \mathbf{r} = \lambda \mathbf{r}$ and therefore multiplying by P gives $A_1 P \mathbf{r} = \lambda P \mathbf{r}$. Respectively, if $A_1 \mathbf{r} = \lambda \mathbf{r}$ then by definition $P \mathcal{A}_1 P^T \mathbf{r} = \lambda \mathbf{r}$ and therefore $\mathcal{A}_1 P^T \mathbf{r} = \lambda P^T \mathbf{r}$. ■

Invertibility of AA^T and BB^T. We can now give the equivalence of Proposition 4.2 in two dimensions. The invertibility of the matrices AA^T and BB^T will be particularly useful when studying the convergence of the TDG method.

Proposition 4.17 (Invertibility of AA^T and BB^T)**.** *The symmetric matrices AA^T and BB^T are invertible and all their eigenvalues are strictly positive. Moreover all the eigenvectors and eigenvalues of the matrix AA^T and BB^T can be deduced from the eigenvalues and eigenvectors of the matrices $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively.*

Proof. The proof is the same as for the Proposition 4.2. Indeed, from Lemma 4.16 any eigenvalue of A_1 is also an eigenvalue of the matrix $\mathcal{A}_1$ and if $\mathbf{r}$ is an eigenvector to $\mathcal{A}_1$ then $P\mathbf{r}$ is an eigenvector to A_1. Therefore all eigenvalues and eigenvectors of A_1 can be deduced from the eigenvalues and eigenvectors of $\mathcal{A}_1$.

Moreover, the Lemma 4.5 can be easily derived in two dimensions considering the decomposition $\mathbf{v} = (\mathbf{v}_e, \mathbf{v}_o) \in \mathbb{R}^m$, $\mathbf{v}_e \in \mathbb{R}^{m_e}$, $\mathbf{v}_o \in \mathbb{R}^{m_o}$. Therefore, the two dimensional version of the proof of Proposition 4.2 give the invertibility of AA^T and BB^T. ∎

Rotational invariance. We give the two dimensional equivalence of Proposition 4.9 and then use it to show the rotational invariance of the P_N model.

Proposition 4.18 (Rotational relations in 2D)**.** *One has the relations*

$$A_1 = U_\theta(A_1\cos\theta - A_2\sin\theta)U_\theta^T, \quad A_2 = U_\theta(A_1\sin\theta + A_2\cos\theta)U_\theta^T.$$

Proof. We give the proof for A_1, the proof for A_2 is similar. Let $\mathbf{v} \in \mathbb{R}^m$ and consider $\mathbf{u} = P^T\mathbf{v} \in S_e$. From Proposition 4.9 one has

$$\mathcal{A}_1\mathbf{u} = \mathcal{U}_\theta(\mathcal{A}_1\cos\theta - \mathcal{A}_2\sin\theta)\mathcal{U}_\theta^T\mathbf{u}.$$

Because $P^TP = I_e$ one gets

$$\mathcal{A}_1\mathbf{u} = \mathcal{U}_\theta P^T(P\mathcal{A}_1P^T\cos\theta - P\mathcal{A}_2P^T\sin\theta)P\mathcal{U}_\theta^T\mathbf{u}.$$

Multiplying by P on the left and using $\mathbf{u} = P^T\mathbf{v}$ give

$$P\mathcal{A}_1P^T\mathbf{v} = P\mathcal{U}_\theta P^T(P\mathcal{A}_1P^T\cos\theta - P\mathcal{A}_2P^T\sin\theta)P\mathcal{U}_\theta^TP^T\mathbf{v}.$$

That is

$$A_1\mathbf{v} = U_\theta(A_1\cos\theta - A_2\sin\theta)U_\theta^T\mathbf{v},$$

where we used $U_\theta = P\mathcal{U}_\theta P^T$. ∎

We use Proposition 4.18 to show that the solutions are invariant under rotation.

Proposition 4.19 (Rotational invariance of the two dimensional P_N model)**.** *The 2D system (4.17) is invariant under rotation. More precisely, if $\mathbf{u}(t,x,y)$ is solution to (4.17) then the function $U_\theta\mathbf{u}(t, x\cos\theta + y\sin\theta, -x\sin\theta + y\cos\theta)$, $\theta \in [0, 2\pi)$, is also solution to (4.17).*

Proof. Let $\mathbf{u}(t,x,y)$ satisfy

$$\Big(I_m\partial_t + A_1\partial_x + A_2\partial_y + R\Big)\mathbf{u}(t,x,y) = \mathbf{0}. \tag{4.26}$$

We consider the following rotation

$$x' = x\cos\theta - y\sin\theta,$$
$$y' = x\sin\theta + y\cos\theta.$$

Using the chain rule formula on the system (4.26) one gets

$$\Big(I_m\partial_t + A_1(\cos\theta\partial_{x'} + \sin\theta\partial_{y'}) + A_2(-\sin\theta\partial_{x'} + \cos\theta\partial_{y'}) + R\Big)\mathbf{u}(t, x'\cos\theta + y'\sin\theta, -x'\sin\theta + y'\cos\theta) = \mathbf{0}.$$

Setting $\mathbf{v} = U_\theta\mathbf{u}$ and multiplying the equality by U_θ one has

$$U_\theta\Big(I_m\partial_t + (A_1\cos\theta - A_2\sin\theta)\partial_{x'} + (A_1\sin\theta + A_2\cos\theta)\partial_{y'} + R\Big)$$
$$U_\theta^T\mathbf{v}(t, x'\cos\theta + y'\sin\theta, -x'\sin\theta + y'\cos\theta) = \mathbf{0}.$$

The matrix U_θ is block diagonal for the moments $\mathbf{u}_k = (u_k^{-k}, ..., u_k^{-k})$. For $k = 0$, the first moment of the vector $\mathbf{u}$ is $\mathbf{u}_0 = (u_0^0)$ and therefore the first row and column of U_θ write respectively $(1, 0, ..., 0)$ and $(1, 0, ..., 0)^T$. One deduces $U_\theta\mathbf{e}_1\mathbf{e}_1^TU_\theta^T = \mathbf{e}_1\mathbf{e}_1^T$, where $\mathbf{e}_1 = (1, ..., 0)^T \in \mathbb{R}^m$. Since the matrix R reads $R = \sigma_t I_m - \frac{\sigma_s}{\varepsilon}\mathbf{e}_1\mathbf{e}_1^T$, one has

$$U_\theta R U_\theta^T = R.$$

Using Corollary 4.18 one finally gets

$$\Big(I_m\partial_t + A_1\partial_{x'} + A_2\partial_{y'} + R\Big)\mathbf{v}(t, x'\cos\theta + y'\sin\theta, -x'\sin\theta + y'\cos\theta) = \mathbf{0}.$$

Therefore $\mathbf{v}(t, x'\cos\theta + y'\sin\theta, -x'\sin\theta + y'\cos\theta) = U_\theta\mathbf{u}(t, x'\cos\theta + y'\sin\theta, -x'\sin\theta + y'\cos\theta)$ is solution to (4.17). ■

Eigenvalues and eigenvectors of $(AA^T)^{-1}R_1$. Exponential solutions to the P_N model require to study the eigenvalues and eigenvectors of the matrix $(AA^T)^{-1}R_1 \in \mathbb{R}^{m_e\times m_e}$. In the following, we may take for simplicity $\varepsilon = 1$ but the proofs are the same for $\varepsilon \in \mathbb{R}_+^*$. First, one can study the sign of the eigenvalues of $(AA^T)^{-1}R_1$.

Proposition 4.20 (Eigenvalues of $(AA^T)^{-1}R_1$ in the general case $\sigma_a \geq 0$)**.** *The eigenvalues μ_i of the matrix $(AA^T)^{-1}R_1$ are strictly positive when $\sigma_a > 0$ and non negative when $\sigma_a = 0$.*

Proof. Assume $\sigma_a > 0$ and let $\mathbf{u} \in \mathbb{R}^{m_e}$. Since $\sigma_a > 0$, the matrix $R_1 := \mathrm{diag}(\sigma_a, \sigma_a + \sigma_s, ..., \sigma_a + \sigma_s)$ is invertible and one has

$$\mathbf{u}^T(AA^T)^{-1}R_1\mathbf{u} = \Big(\sqrt{R_1}^{-1}\widetilde{\mathbf{u}}\Big)^T(AA^T)^{-1}\sqrt{R_1}\widetilde{\mathbf{u}} = \widetilde{\mathbf{u}}^T\sqrt{R_1}^{-1}(AA^T)^{-1}\sqrt{R_1}\widetilde{\mathbf{u}},$$

with $\widetilde{\mathbf{u}} = \sqrt{R_1}\mathbf{u}$. The eigenvalues of the matrices $\sqrt{R_1}^{-1}(AA^T)^{-1}\sqrt{R_1}$ and $(AA^T)^{-1}$ are the same and therefore one deduces from Proposition 4.17 that the matrix $\sqrt{R_1}^{-1}(AA^T)^{-1}\sqrt{R_1}$ is positive. That is $\mathbf{u}^T(AA^T)^{-1}R_1\mathbf{u} > 0$ if $\mathbf{u} \neq \mathbf{0}$ and one concludes that the eigenvalues of $(AA^T)^{-1}R_1$ are strictly positive when $\sigma_a > 0$. By continuity that the eigenvalues are non negative when $\sigma_a = 0$. ■

An important property is the degeneracy of one eigenvalue when $\sigma_a \to 0$. In terms of the exponential solutions, this results in the degeneracy of the exponentials associated with this eigenvalue.

Proposition 4.21 (Eigenvalues of $(AA^T)^{-1}R_1$ in the degenerative case $\sigma_a \to 0$)**.** *Assume $\sigma_s > 0$. There is exactly one eigenvalue μ_1 such that $\mu_1 \underset{\sigma_a \to 0}{\to} 0$.*

Proof. To show the result, we consider the degenerate case $\sigma_a = 0$. It is clear from the definition of $R_1 := \text{diag}(\sigma_a, \sigma_a+\sigma_s, ..., \sigma_a+\sigma_s)$ and $\sigma_s > 0$ that $\dim(\ker R_1) = 1$. Since $\dim(\ker(AA^T)^{-1}) = 0$ one finds

$$\dim\Big(\ker(AA^T)^{-1}R_1\Big) = 1.$$

From Proposition 4.20, all the eigenvalues are strictly positive when $\sigma_a > 0$. Therefore there is exactly one eigenvalue which degenerate to 0 when $\sigma_a \to 0$. The proof is complete. ■

The fact that the eigenvectors of the matrix $(AA^T)^{-1}R_1$ form a basis of $\mathbb{R}^{m_e}$ is required when studying the approximation properties of the solutions to the P_N model. Moreover, it allows to count the number of distinct couple of eigenvalue/eigenvector of the matrix $(AA^T)^{-1}R_1$ and give the total number of stationary exponential solutions in 1D (see the proof of Theorem 4.25).

Proposition 4.22 (Eigenvectors of $(AA^T)^{-1}R_1$ when $\sigma_a > 0$)**.** *Assume $\sigma_a > 0$. The eigenvectors of $(AA^T)^{-1}R_1 \in \mathbb{R}^{m_e \times m_e}$ from a basis of $\mathbb{R}^{m_e}$. Therefore, there exists m_e distinct couple of eigenvalue/eigenvector of the matrix $(AA^T)^{-1}R_1$.*

Proof. Let $\mathbf{u}$ be an eigenvector of $(AA^T)^{-1}R_1$ associated with the eigenvalue λ that is $(AA^T)^{-1}R_1\mathbf{u} = \lambda\mathbf{u}$. One has $\sqrt{R_1}(AA^T)^{-1}\sqrt{R_1}\widetilde{\mathbf{u}} = \lambda\widetilde{\mathbf{u}}$, with $\widetilde{\mathbf{u}} = \sqrt{R_1}\mathbf{u}$. Since the matrix $(AA^T)^{-1}$ is symmetric, the matrix $\sqrt{R_1}(AA^T)^{-1}\sqrt{R_1}$ is also symmetric. Because $\sigma_a > 0$, the matrix $\sqrt{R_1}$ is invertible and one concludes that the eigenvectors of $(AA^T)^{-1}R_1 \in \mathbb{R}^{m_e \times m_e}$ from a basis of $\mathbb{R}^{m_e}$. ■

For the same reason (approximation properties of the basis functions), one needs to prove that the eigenvectors of $(AA^T)^{-1}R_1$ from a basis of $\mathbb{R}^{m_e}$, this time when $\sigma_a = 0$.

Proposition 4.23 (Eigenvectors of $(AA^T)^{-1}R_1$ in the degenerate case $\sigma_a = 0$)**.** *Assume $\sigma_a = 0$. The eigenvectors of $(AA^T)^{-1}R_1$ form a basis of $\mathbb{R}^{m_e}$.*

Proof. To prove the proposition a distinction must be made between the eigenvectors with a first component equal to zero and the other eigenvectors. We denote $m_1 \in \mathbb{N}$ the number of eigenvectors of the matrix $(AA^T)^{-1}$ with a first component equal to zero and $m_2 \in \mathbb{N}$ the number of eigenvectors of $(AA^T)^{-1}$ with a non zero first component. Since the matrix $(AA^T)^{-1}$ is symmetric one has

$$m_e = m_1 + m_2.$$

- First we consider the eigenvectors of $(AA^T)^{-1}R_1$ with a first component equal to zero. From the definition of R_1 with $\sigma_a = 0$ one has

  $$R_1 = \text{diag}(0, \sigma_s, ..., \sigma_s).$$

  One deduces that each eigenvectors of $(AA^T)^{-1}$ with a first component equal to zero is also an eigenvector of $(AA^T)^{-1}R_1$. Therefore, the matrix $(AA^T)^{-1}R_1$ admits m_1 linearly independent eigenvectors $\mathbf{w}_i$ with a first component equal to zero

  $$\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{m_1}.$$

  We denote μ_i the eigenvalues associated with the eigenvectors $\mathbf{w}_i$

  $$\mu_1, ..., \mu_p, \quad p \leq m_1.$$

- Now we consider $\mathbf{z}_i$ the eigenvectors of $(AA^T)^{-1}R_1$ with a non zero first component and their associated eigenvalues λ_i. The Remark 4.26 below shows that the λ_i are the roots of the function

$$f(\lambda) = 1 + \sigma_s \sum_{i=1}^{m_2} \frac{(\mathbf{u}_i^T \mathbf{e}_1)^2}{\lambda d_i - \sigma_s},$$

where $\mathbf{e}_1 = (1, 0, ..., 0)^T$, $\mathbf{u}_i$ are the eigenvectors of AA^T (or equally of $(AA^T)^{-1}$) with a first component not equal to zero and $d_i > 0$ are the associated eigenvalues. The derivative of f reads

$$f'(\lambda) = -\sigma_s \sum_{i=1}^{m_2} d_i \frac{(\mathbf{u}_i^T \mathbf{e}_1)^2}{(\lambda d_i - \sigma_s)^2}.$$

Therefore, the function f admits m_2 poles located at $\lambda = \frac{\sigma_s}{d_i}$ and is monotone between these poles. From Proposition 4.6 the eigenvalues of $\mathcal{A}_1$ associated with an eigenvector with a non zero first component are distinct. From Lemma 4.16, one deduces that this is also the case for the matrix A_1. Therefore the d_i are distinct and one finds there are m_2 distinct roots λ_i of f which satisfy

$$0 = \lambda_1 < \frac{\sigma_s}{d_1} < ... < \lambda_{m_2} < \frac{\sigma_s}{d_{m_2}}. \tag{4.27}$$

Since all these eigenvalues are distinct, there exists m_2 linearly independent eigenvectors $\mathbf{z}_i$ with non zero first component

$$\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{m_2}.$$

In summary, the matrix $(AA^T)^{-1}R_1$ admits m_1 linearly independent eigenvectors $\mathbf{w}_1, ..., \mathbf{w}_{m_1}$ associated with eigenvalues μ_i and m_2 linearly independent eigenvectors $\mathbf{z}_1, ..., \mathbf{z}_{m_2}$ associated with distinct eigenvalues λ_i. It is important to notice that the eigenvectors $\mathbf{w}_i$ have their first component equal to zero while the eigenvectors $\mathbf{z}_i$ have a non zero first component.

We use the following notation to denote all the eigenvectors of $(AA^T)^{-1}R_1$

$$\mathbf{v}_1 = \mathbf{w}_1, ..., \mathbf{v}_{m_1} = \mathbf{w}_{m_1}, \mathbf{v}_{m_1+1} = \mathbf{z}_1, ..., \mathbf{v}_{m_1+m_2} = \mathbf{z}_{m_2}. \tag{4.28}$$

We also denote α_i the eigenvalues associated to these eigenvectors

$$\alpha_1 = \mu_1, ..., \alpha_p = \mu_p, \alpha_{p+1} = \lambda_1, ...\alpha_{p+m_2} = \lambda_{m_2}.$$

Note that there are $m_1 + m_2$ eigenvectors $\mathbf{v}_i$ but only $p + m_2 \leq m_1 + m_2$ eigenvalues μ_i since multiple eigenvectors can be associated to the same eigenvalue. Finally, we define $E(\alpha_i)$ the set of eigenvectors in the list (4.28) associated to the eigenvalue α_i

$$E(\alpha_i) = \Big\{ \mathbf{v}_k \text{ in the list (4.28) } / \ (AA^T)^{-1}R_1\mathbf{v}_k = \alpha_i \mathbf{v}_k \Big\}.$$

We want to show that the total number of eigenvalues α_i, counting multiplicities, is $m_1 + m_2$. That is

$$\operatorname{card} E(\alpha_i) = \dim \Big(\operatorname{Span} E(\alpha_i) \Big), \quad 1 \leq i \leq p + m_2. \tag{4.29}$$

Let $i \in \mathbb{N}$, $1 \leq i \leq p+m_2$ and consider the eigenvalue α_i. Up to a renumbering of the vectors $\mathbf{v}_j$, we can denote $\mathbf{v}_1, ..., \mathbf{v}_k$, $k = \operatorname{card} E(\alpha_i) \in \mathbb{N}^*$, the eigenvectors associated with the eigenvalue α_i. Assume

$$\sum_{j=1}^{k} a_j \mathbf{v}_j = \mathbf{0}, \quad a_j \in \mathbb{R}. \tag{4.30}$$

Proving (4.29) is equivalent to prove $a_j = 0$ for all $j = 1, ..., k$. There are two possibilities

1. All the eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_k$ have their first component equal to zero. Then, the vectors $\mathbf{v}_1, ..., \mathbf{v}_k$ are taken from the eigenvectors $\mathbf{w}_i$ of the matrix $(AA^T)^{-1}$. Therefore, using (4.30) and the symmetry of $(AA^T)^{-1}$, one concludes $a_i = 0$ for all $i = 1, ..., k$.
2. All the eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_k$ have not their first component equal to zero. Then, there is exactly one eigenvector with a non zero first component. Indeed, the eigenvectors with a non zero first component are the vectors $\mathbf{z}_i$ and from (4.27) they are associated with distinct eigenvalues λ_i. Therefore, up to a renumbering of the vectors $\mathbf{w}_i$ and $\mathbf{z}_i$ one can write
$$a_1\mathbf{z}_1 + \sum_{j=2}^{k} a_j\mathbf{w}_j = \mathbf{0}.$$
Since $\mathbf{z}_1$ is the only vectors with it first component not equal to zero one has $a_1 = 0$. And because the $\mathbf{w}_i$ are the eigenvectors of the matrix $(AA^T)^{-1}$, one finally deduces $a_2 = ... = a_k = 0$.

Therefore, the total number of eigenvalues α_i, counting multiplicities, is $m_e = m_1 + m_2$. One concludes that the eigenvectors associated form a basis of $\mathbb{R}^{m_e}$ and the proof is complete. ■

Finally let $M \in \mathbb{R}^{m_e \times m_e}$ be the matrix made of the eigenvectors of $(AA^T)^{-1}R_1$ when $\sigma_a = 0$. When studying the approximation properties of the basis functions, we will use the following matrix J.

Definition 4.24. We denote $J \in \mathbb{R}^{m_e-1 \times m_e-1}$ the matrix obtained when removing the row and column of the matrix M associated to the zero eigenvalue.

In particular, Proposition 4.23 implies that the matrix J is invertible.

4-2 Special solutions

To apply the TDG method to a P_N model written as in (4.17), one needs to construct the basis functions *i.e.* to find solutions to the system. It has strong reminiscence to case solutions [Cas60] to the transport equation [Gos13, BA69, BA70]. The Theorem 4.25 shows how to construct exponential solutions when $\sigma_a \geq 0$ which can be used as basis functions. It is interesting to consider such exponential solutions for at least two reasons

(i) They fondamentally differ from the polynomial basis functions used with the standard DG method. Therefore, the TDG method applies with these exponential functions may have different (and new) properties compare to the DG method.

(ii) Due to the well-balanced property, the exponential solutions may lead to very efficient schemes to capture boundary layers.

However one realizes that, even if it is possible to use these exponentials when $\sigma_a = 0$, some of them will degenerate toward constant solutions when $\sigma_a \to 0$. Therefore, one will "loose" (in a sense) some basis functions when $\sigma_a \to 0$. To fix this issue, the Theorems 4.29 and 4.34 show how to obtain polynomial solutions from the degenerative exponentials. In practice, when applying the TDG method, the degenerative exponentials will be replaced by the polynomial solutions in the limit $\sigma_a \to 0$. Additionally, some time dependent solutions are also constructed in Section 4-2.4.

In this section and the next one, some analysis will be based on the simplification of the Taylor expansion for solutions to a given system of equations. Using simplified Taylor expansion has several advantages

- We use it to show the degeneracy of the exponential solutions toward polynomials when $\sigma_a \to 0$ (Theorem 4.34). This a good characteristic from a numerical point of view since one can expect that the scheme recovers the same kind of properties when using these two different types of solutions as basis functions.
- We use it to construct the polynomial solutions with simple recurrence formulas (Theorem 4.34). This is of course very important for practical applications of the TDG method.
- Additionally, we will also use a simplified Taylor expansion when studying the approximation properties of the basis functions in Section 4-3. This is not the only possibility [CD98, IG15a, IGD14] but it has the advantage of giving a natural framework to work with (the study of the matrix S_l^k in Section 4-3.2).

4-2.1 Exponential solutions

The main result of this subsection is the Theorem 4.25 which shows how to construct exponential solutions to the P_N model.

Theorem 4.25. *Let $\sigma_t > 0$ and $\mathbf{w}_1, ..., \mathbf{w}_{m_e} \in \mathbb{R}^{m_e}$ be the eigenvectors of the matrix $(AA^T)^{-1}R_1$ associated with the eigenvalues $\mu_1, ..., \mu_{m_e}$. Let $\boldsymbol{\chi}_i = -\sqrt{\frac{\mu_i}{\sigma_t}}A^T\mathbf{w}_i \in \mathbb{R}^{m_o}$, $\mathbf{z}_i = (\mathbf{w}_i^T, \boldsymbol{\chi}_i^T)^T \in \mathbb{R}^m$ and $\mathbf{d}_k = (\cos\theta_k, \sin\theta_k)^T \in \mathbb{R}^2$. Then the following exponential functions*

$$(\mathbf{v}_i)_k(\mathbf{x}) = U_{\theta_k}\mathbf{z}_i e^{\frac{1}{c}\sqrt{\sigma_t\mu_i}\mathbf{d}_k^T\mathbf{x}}, \quad i = 1, ..., m_e, \tag{4.31}$$

are solutions to the P_N model (4.17). Moreover, the functions (4.31) are the only solutions to the P_N model under the form $\mathbf{z}e^{\lambda(\mathbf{d}_k,\mathbf{x})}$, $\lambda \in \mathbb{R}$, $\mathbf{z} \in \mathbb{R}^m$.

Proof. We start searching for solutions under the form

$$\widetilde{\mathbf{v}}(\mathbf{x}) = \widetilde{\mathbf{z}}e^{\lambda\mathbf{d}_k^T\mathbf{x}} \in \mathbb{R}^m, \tag{4.32}$$

with $\mathbf{d}_k = (\cos\theta_k, \sin\theta_k)^T$, $\lambda \in \mathbb{R}$ and $\widetilde{\mathbf{z}} \in \mathbb{R}^m$. Using Proposition 4.19 the function $\mathbf{v}(x, y) = U_{-\theta_k}\widetilde{\mathbf{v}}(x\cos\theta_k - y\sin\theta_k, x\sin\theta_k + y\cos\theta_k)$ is also a solution to the P_N model. This solution can be written under the form

$$\mathbf{v}(x, y) = \mathbf{z}e^{\lambda x} \in \mathbb{R}^m, \tag{4.33}$$

$\mathbf{z} \in \mathbb{R}^m$. Inserting (4.33) in the P_N model (4.17) gives after removing the exponentials $\lambda A_1\mathbf{z} = -R\mathbf{z}$. Due to the matrix R on the right hand side, this eigenvalue problem is different from the one already encountered in the previous section. We use the decomposition

$$\mathbf{z} = (\mathbf{w}^T, \boldsymbol{\chi}^T)^T \in \mathbb{R}^m, \quad \mathbf{w} \in \mathbb{R}^{m_e}, \quad \boldsymbol{\chi} \in \mathbb{R}^{m_o}.$$

Using the particular form of the matrix A_1 and R in (4.18)-(4.19), one has

$$\begin{cases} \lambda cA\boldsymbol{\chi} = -R_1\mathbf{w}, \\ \lambda cA^T\mathbf{w} = -R_2\boldsymbol{\chi}. \end{cases} \tag{4.34}$$

Multiplying the second equation by λcA and then using the first equation on the right hand side with $R_2^{-1} = \frac{1}{\sigma_t}I_{m_o}$ gives $\lambda^2c^2AA^T\mathbf{w} = \sigma_t R_1\mathbf{w}$. From Proposition 4.17 the matrix AA^T is invertible therefore one can write

$$(AA^T)^{-1}R_1\mathbf{w} = \frac{\lambda^2c^2}{\sigma_t}\mathbf{w}. \tag{4.35}$$

If $\mathbf{w} \in \mathbb{R}^{m_e}$ is an eigenvector of the matrix $(AA^T)^{-1}R_1$ associated with the eigenvalue μ ($\mu \geq 0$ from Proposition 4.20), one can take $\lambda = \pm\frac{\sqrt{\sigma_t\mu}}{c}$. First we consider the case $\lambda = \frac{\sqrt{\sigma_t\mu}}{c}$, the case $\lambda = -\frac{\sqrt{\sigma_t\mu}}{c}$ will be discuss later. The second equation in (4.34) gives

$$\boldsymbol{\chi} = -\sqrt{\frac{\mu}{\sigma_t}}A^T\mathbf{w} \in \mathbb{R}^{m_o}.$$

One concludes that the one dimensional function $\mathbf{v}(\mathbf{x}) = \mathbf{z}e^{\frac{1}{c}\sqrt{\sigma_t\mu}x}$ is solution to the P_N model. Applying a rotation as in Proposition 4.19 gives the solutions (4.31). Moreover, considering $\lambda = -\frac{\sqrt{\sigma_t\mu}}{c}$ is equivalent to take $-\mathbf{d}_k$ in (4.31). We conclude that all the solutions under the form (4.32) are given by (4.31). Finally, from Proposition 4.22 there exists m_e distinct pair $(\mu_i, \mathbf{w}_i)$ solution of the eigen problem associated with the matrix $(AA^T)^{-1}R_1$. This completes the proof. ■

Remark 4.26 (Secular equation)**.** Exponential solutions require the eigenvalues and eigenvectors of the matrix $(AA^T)^{-1}R_1$. In practice, it can be difficult to solve directly the eigenvalue problem (4.35) associated with the matrix $(AA^T)^{-1}R_1$ for large values of N. Here we give an alternative method based on the eigenvalues and eigenvectors of the matrix AA^T. They can be simpler to calculate since one can deduced them from the eigenstructure of the matrices $\mathcal{A}_1$ (Proposition 4.17). For example, the eigenvalues are roots of the Legendre polynomials (and their derivatives) and one way to obtain the eigenvectors of $\mathcal{A}_1$ is to apply a rotation to the eigenvectors of the matrix $\mathcal{A}_3$ [GH16, Lemma 2].

We proceed in two steps. At first, since $R_1 = I_{m_e}\sigma_t - \frac{\sigma_s}{\varepsilon}\mathbf{e}_1\mathbf{e}_1^T$, some eigenvalues and eigenvectors of $(AA^T)^{-1}R_1$ can be deduced from the eigenvalues and eigenvectors of $(AA^T)^{-1}$. More precisely, assume $\mathbf{w}_i$ is an eigenvector of $(AA^T)^{-1}$ associated with the eigenvalue λ_i. Using $R_1 = \mathrm{diag}(\sigma_a, \sigma_t, ..., \sigma_t)$, one deduces that, if the first component of $\mathbf{w}_i$ is null, then $\mathbf{w}_i$ is an eigenvector of $(AA^T)^{-1}R_1$ associated with the eigenvalue $\sigma_t\lambda_i$.

Then, to get the other eigenvalues and eigenvectors, we use a so-called secular equation [And96, JL91]. Assume $\mathbf{w}$ is an eigenvector of the matrix $(AA^T)^{-1}R_1$ associated with the eigenvalue λ and that the first component of $\mathbf{w}$ is not zero. From the equality $(AA^T)^{-1}R_1\mathbf{w} = \lambda\mathbf{w}$ one finds

$$R_1\mathbf{w} = AA^T\lambda\mathbf{w}$$

Diagonalizing the matrix $AA^T = PDP^T$ and using $R_1 := I_{m_e}\sigma_t - \frac{\sigma_s}{\varepsilon}\mathbf{e}_1\mathbf{e}_1^T$, $\mathbf{e}_1 = (1, 0, ..., 0)^T \in \mathbb{R}^{m_e}$, one gets

$$\lambda D\widetilde{\mathbf{w}} = \Big(I_{m_e}\sigma_t - \frac{\sigma_s}{\varepsilon}\mathbf{v}\mathbf{v}^T\Big)\widetilde{\mathbf{w}},$$

where $\mathbf{v} = P^T\mathbf{e}_1$ and $\widetilde{\mathbf{w}} = P^T\mathbf{w}$. One has

$$\Big(\sigma_t I_{m_e} - \lambda D\Big)\widetilde{\mathbf{w}} = \frac{\sigma_s}{\varepsilon}\mathbf{v}\mathbf{v}^T\widetilde{\mathbf{w}},$$

and one gets

$$\widetilde{\mathbf{w}} = \Big(I_{m_e}\sigma_t - \lambda D\Big)^{-1}\frac{\sigma_s}{\varepsilon}\mathbf{v}\mathbf{v}^T\widetilde{\mathbf{w}}. \tag{4.36}$$

Multiplying by $\mathbf{v}^T$ on both sides give

$$\mathbf{v}^T\widetilde{\mathbf{w}} = \mathbf{v}^T\Big(I_{m_e}\sigma_t - \lambda D\Big)^{-1}\frac{\sigma_s}{\varepsilon}\mathbf{v}\mathbf{v}^T\widetilde{\mathbf{w}}.$$

Because $\mathbf{v}^T = \mathbf{e}_1^T P$ and $\widetilde{\mathbf{w}} = P^T\mathbf{w}$ one has $\mathbf{v}^T\widetilde{\mathbf{w}} = \mathbf{e}_1^T\mathbf{w}$ and $\mathbf{v}^T\widetilde{\mathbf{w}} \neq 0$ since we assume that the first component of $\mathbf{w}$ is non zero. One can therefore remove $\mathbf{v}^T\widetilde{\mathbf{w}}$ on both sides of the equality. One gets the **secular equation**

$$1 = \frac{\sigma_s}{\varepsilon}\mathbf{v}^T(I_{m_e}\sigma_t - \lambda D)^{-1}\mathbf{v}.$$

Setting $f(\lambda) = 1 + \frac{\sigma_s}{\varepsilon}\mathbf{v}^T(\lambda D - I_{m_e}\sigma_t)^{-1}\mathbf{v}$, the secular equation reads

$$f(\lambda) = 0.$$

Therefore one needs is equivalent to find the roots of f. Denoting $D = \text{diag}(d_1 > 0, ..., d_{m_e} > 0)$, the function f can be written

$$f(\lambda) = 1 + \frac{\sigma_s}{\varepsilon}\sum_i \frac{v_i^2}{\lambda d_i - \sigma_t},$$

and

$$f'(\lambda) = -\frac{\sigma_s}{\varepsilon}\sum_i d_i \frac{v_i^2}{(\lambda d_i - \sigma_t)^2}.$$

Therefore, the function f is a monotone decreasing function which admits pole located at σ_t/d_i if d_i is an eigenvalue of $(AA^T)^{-1}$ associated with an eigenvector with a non zero first component. See Figure 4.2 for the example of the function f in the case of the P_3 model.

Moreover, from (4.36) and denoting $C = \mathbf{v}^T\widetilde{\mathbf{w}} \neq 0$ one has

$$\widetilde{\mathbf{w}} = C\Big(I_{m_e}\sigma_t - \lambda D\Big)^{-1}\frac{\sigma_s}{\varepsilon}\mathbf{v}.$$

Therefore, once one has the eigenvalue λ, one can deduce the eigenvector $\mathbf{w}$ associated.

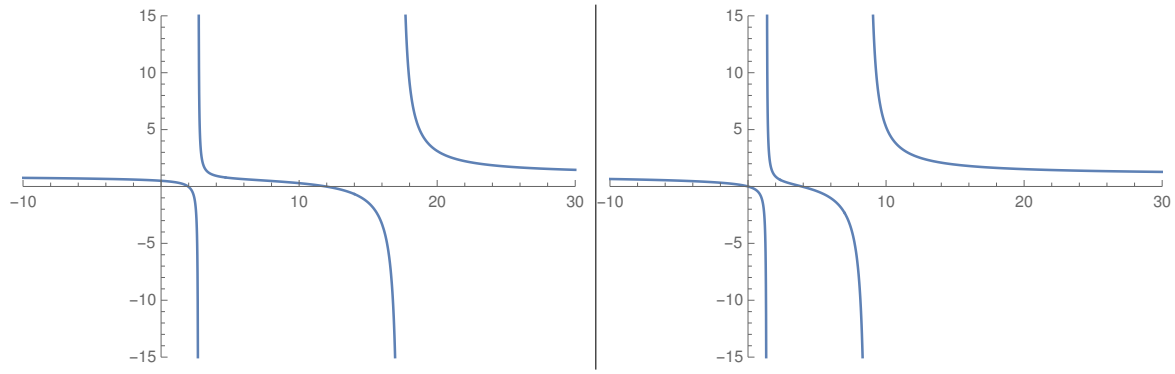

Figure 4.2 – Representation of the function f for the P_3 model. On the left $\sigma_a = \sigma_s = 1$, on the right $\sigma_a = 0$, $\sigma_s = 1$.

●

Example 4.27 (Secular equation for the P_1 model)**.** For the P_1 model, one deduces from the matrices given in Example 4.13 that $AA^T = \frac{1}{3}$. This matrix admits one eigenvalue $d_1 = \frac{1}{3}$ associated with the eigenvector $\mathbf{w}_1 = 1$. Therefore the function f reads

$$f(\lambda) = 1 + \frac{\sigma_s}{\varepsilon(\lambda d_1 - \sigma_t)}.$$

This function admits one roots λ_1. Using $\sigma_t = \varepsilon\sigma_a + \frac{\sigma_s}{\varepsilon}$, one finds $\lambda_1 = 3\varepsilon\sigma_a$. One can check that λ_1 is indeed the eigenvalue of the matrix $(AA^T)^{-1}R_1 = 3\varepsilon\sigma_a$. ●

4-2.2 Polynomial solutions (only when $\sigma_a = 0$) with Birkhoff and Abu-Shumays method's

In the previous section, we have constructed exponential solutions when $\sigma_a \geq 0$. If all these solutions can be used when $\sigma_a > 0$, some of them degenerate toward constant solutions when $\sigma_a \to 0$.

It is therefore required to construct new solutions which will be used in the basis functions to replaced the degenerative exponential when $\sigma_a = 0$. In the following, we construct polynomial solutions to the P_N model in the degenerate case $\sigma_a = 0$ from solutions to the transport equation. However, although this procedure is quiet straightforward, it does not mathematically justify the replacement of degenerative exponentials by polynomials in the basis functions. This issue will be addressed in Section 4-2.3.

In this section, we use the solutions to the transport equation given by Birkhoff and Abu-Shumays [BA69]. Some notations are taken from Section 4-1.1. Since we are interested in the two dimensional P_N model, we search for solutions which do not depend on the variable z. In the following we assume $q(x, y)$ is a harmonic polynomial

$$\Delta q(x, y) = 0.$$

And for $l \in \mathbb{N}$ we define the following polynomial functions

$$f_q^l(x, y, \mathbf{\Omega}) := \sum_{k=0}^{l} \Big(\frac{-1}{\sigma_s}\Big)^k \Big(\mathbf{\Omega} \cdot \nabla\Big)^k q(x, y).$$

Note that for all $l \in \mathbb{N}$ the function f_q^l is a polynomial function of degree $deg(q)$ where we have denoted $deg(q)$ the degree of the polynomial q. From the definition of f_q^l one has $f_q^{deg(q)} = f_q^{deg(q)+l}$ for all $l \in \mathbb{N}$. The functions $f_q^{deg(q)}$ are solutions to the transport equation (4.1), see [BA69].

Proposition 4.28 (Polynomial solutions to the transport equation [BA69])**.** *Assume $\sigma_a = 0$. Then the function $f_q^{deg(q)}(x, y, \mathbf{\Omega})$ is solution to the transport equation.*

To construct solutions to the P_N model we consider the truncated functions f_q^N. Note that f_q^N is only an approximation of $f_q^{deg(q)}$ if $N < deg(q)$ and $f_q^N = f_q^{deg(q)}$ if $N \geq deg(q)$.

Theorem 4.29 (Polynomial solutions to the P_N model)**.** *Assume $\sigma_a = 0$. The function*

$$\mathbf{f}_N(x, y) =< \mathbf{y}(\mathbf{\Omega}) f_q^N(x, y, \mathbf{\Omega}) > \in \mathbb{R}^{m^{3D}},$$

is solution to the P_N model. This function is a polynomial vector with respect to x, y.

The proofs of these two results are based on the following lemmas. For the Proposition 4.28, the proof given in [BA69] is more general and works in any dimensions. We give here a more direct (but less general) proof.

Lemma 4.30. *One has*

$$\Big(\mathbf{\Omega} \cdot \nabla\Big)^k q(x, y) = \Big(\frac{\sin\phi}{2}\Big)^k \Big(e^{-ik\psi}(\partial_x + i\partial_y)^k + e^{ik\psi}(\partial_x - i\partial_y)^k\Big) q(x, y), \quad \text{if } k \geq 1. \tag{4.37}$$

Proof. Indeed since $\partial_z q = 0$ one has

$$\begin{aligned}\Big(\mathbf{\Omega} \cdot \nabla\Big) q(x, y) &= \sin\phi \Big(\cos\psi \partial_x + \sin\psi \partial_y\Big) q(x, y), \\ &= \frac{\sin\phi}{2} \Big(e^{i\psi}(\partial_x - i\partial_y) + e^{-i\psi}(\partial_x + i\partial_y)\Big) q(x, y).\end{aligned}$$

Therefore

$$\Big(\mathbf{\Omega}\cdot\nabla\Big)^k q(x,y) = \Big(\frac{\sin\phi}{2}\Big)^k \sum_{p=0}^{k} C_k^p e^{ip\psi}(\partial_x - i\partial_y)^p e^{-i(k-p)\psi}(\partial_x + i\partial_y)^{k-p} q(x,y),$$
$$= \Big(\frac{\sin\phi}{2}\Big)^k \sum_{p=0}^{k} C_k^p e^{i(2p-k)\psi}(\partial_x - i\partial_y)^p (\partial_x + i\partial_y)^{k-p} q(x,y).$$

But since $q(x,y)$ is harmonic one has

$$(\partial_x - i\partial_y)(\partial_x + i\partial_y)q(x,y) = 0.$$

Therefore, all the terms in the sum vanish except the first and the last. One finally finds the equalities (4.37). ■

Lemma 4.31. *One has*

$$\Big(\mathbf{\Omega}\cdot\nabla\Big) f_q^l(x,y,\mathbf{\Omega}) = \sigma_s\Big(- f_q^l(x,y,\mathbf{\Omega}) + < f_q^l(x,y,\mathbf{\Omega}) > \Big)$$
$$+ \Big(\frac{-1}{\sigma_s}\Big)^l \Big(\frac{\sin\phi}{2}\Big)^{l+1} \Big(e^{-i(l+1)\psi}(\partial_x + i\partial_y)^{l+1} + e^{i(l+1)\psi}(\partial_x - i\partial_y)^{l+1}\Big) q(x,y).$$

Proof. From the definition of the function $f_q^l(x,y,\mathbf{\Omega})$ one has

$$\Big(\mathbf{\Omega}\cdot\nabla\Big) f_q^l(x,y,\mathbf{\Omega}) = \sum_{k=0}^{l} \Big(\frac{-1}{\sigma_s}\Big)^k \Big(\mathbf{\Omega}\cdot\nabla\Big)^{k+1} q(x,y).$$

That is

$$\Big(\mathbf{\Omega}\cdot\nabla\Big) f_q^l(x,y,\mathbf{\Omega}) = \sigma_s q(x,y) - \sigma_s \sum_{k=0}^{l} \Big(\frac{-1}{\sigma_s}\Big)^k \Big(\mathbf{\Omega}\cdot\nabla\Big)^k q(x,y) + \Big(\frac{-1}{\sigma_s}\Big)^l \Big(\mathbf{\Omega}\cdot\nabla\Big)^{l+1} q(x,y).$$

One concludes with Lemma 4.30. ■

We can now prove Proposition 4.28 and Theorem 4.29.

Proof of Proposition 4.28. Since $q(x,y)$ has degree $deg(q)$ one has

$$\Big(e^{-i(deg(q)+1)\psi}(\partial_x + i\partial_y)^{deg(q)+1} + e^{i(deg(q)+1)\psi}(\partial_x - i\partial_y)^{deg(q)+1}\Big) q(x,y) = 0.$$

Therefore using Lemma 4.31 one gets

$$\Big(\mathbf{\Omega}\cdot\nabla\Big) f_q^{deg(q)}(x,y,\mathbf{\Omega}) = \sigma_s\Big(- f_q^{deg(q)}(x,y,\mathbf{\Omega}) + < f_q^{deg(q)}(x,y,\mathbf{\Omega}) > \Big),$$

which is the stationary version of the transport equation (4.1) when $\sigma_a = 0$. ■

Proof of Theorem 4.29. From Lemma 4.30, the definition of f_q^N and the definition of the spherical harmonics (A.3), one concludes that the function $f_q^N(x,y,\mathbf{\Omega})$ can be decomposed on the spherical harmonics of degree less than N. Therefore one can write

$$f_q^N(x,y,\mathbf{\Omega}) = \mathbf{y}^T(\mathbf{\Omega})\mathbf{f}_N(x,y). \tag{4.38}$$

From Lemma 4.31 one has

$$\Big(\mathbf{\Omega}\cdot\nabla\Big)f_q^N(x,y,\mathbf{\Omega}) =\sigma_s\Big(-f_q^N(x,y,\mathbf{\Omega})+<f_q^N(x,y,\mathbf{\Omega})>\Big) + \Big(\frac{-1}{\sigma_s}\Big)^N\Big(\frac{\sin\phi}{2}\Big)^{N+1}\Big(e^{-i(N+1)\psi}(\partial_x+i\partial_y)^{N+1}+e^{i(N+1)\psi}(\partial_x-i\partial_y)^{N+1}\Big)q(x,y).$$

Multiplying by $\mathbf{y}(\mathbf{\Omega})$, integrating on the sphere and using

$$\int_0^{2\pi} e^{ik\psi}d\psi = 0, \quad \text{if } k\neq 0,$$

yields

$$<\mathbf{y}(\mathbf{\Omega})\Big(\mathbf{\Omega}\cdot\nabla\Big)f_q^N(x,y,\mathbf{\Omega})>=\sigma_s\Big(-<\mathbf{y}(\mathbf{\Omega})f_q^N(x,y,\mathbf{\Omega})>+<\mathbf{y}(\mathbf{\Omega})><f_q^N(x,y,\mathbf{\Omega})>\Big).$$

Using the decomposition (4.38) one gets

$$<\mathbf{y}(\mathbf{\Omega})\Big(\mathbf{\Omega}\cdot\nabla\Big)\mathbf{y}^T(\mathbf{\Omega})\mathbf{f}_N(x,y)>=\sigma_s\Big(-<\mathbf{y}(\mathbf{\Omega})\mathbf{y}^T(\mathbf{\Omega})\mathbf{f}_N(x,y)>+<\mathbf{y}(\mathbf{\Omega})><\mathbf{y}^T(\mathbf{\Omega})\mathbf{f}_N(x,y)>\Big).$$

That is

$$\sum_{i=1}^{3}<\Omega_i\mathbf{y}(\mathbf{\Omega})\mathbf{y}^T(\mathbf{\Omega})>\partial_{x_i}\mathbf{f}_N(x,y)=\sigma_s\Big(-<\mathbf{y}(\mathbf{\Omega})\mathbf{y}^T(\mathbf{\Omega})>+<\mathbf{y}(\mathbf{\Omega})><\mathbf{y}^T(\mathbf{\Omega})>\Big)\mathbf{f}_N(x,y),$$

which is the stationary version of the P_N model (4.4) when $\sigma_a = 0$. ■

Remark 4.32. If $N \geq deg(q)$, the solution $f_q^{deg(q)}$ of the transport equation can be completely reconstructed from the solution $\mathbf{f}_N$ of the P_N model. Therefore, the polynomial solutions $f_q^{deg(q)}$ of the transport equation are, in a sense, preserved by the P_N model when $N \geq deg(q)$. ●

4-2.3 Link between exponential and polynomial solutions

In the previous section, we have constructed polynomial solutions to the P_N model to replace the degenerative exponentials when $\sigma_a = 0$. However, even if the procedure used is quiet straightforward, it does not mathematically justify the replacement of the degenerative exponentials with polynomials in the basis functions. The goal of this section, which is independent from the previous one, is to

- Show the degeneracy of exponentials constructed in section 4-2.1 toward polynomial solutions when $\sigma_a \to 0$.
- Show that those polynomial solutions can be constructed using recurrence formulas. One advantage of this procedure is that the formulas are explicit while the solutions given in Theorem 4.29 of the previous section necessitate to integrate spherical harmonics and therefore require adapted formulas.

For the P_1 model, the solutions given in this section coincide with the polynomial solutions given in the previous section. We conjecture that this is also the case for the $P_{N>1}$ model, see Remark 4.50.

The main result is the Theorem 4.34. As a first step we give some definitions. Let $n \in \mathbb{N}$, $\mathbf{x}_0 = (x_0, y_0)^T \in \Omega$ and consider

$$\gamma_k^p(\mathbf{x}) := \begin{cases} \dfrac{C_k^p}{k!}(x-x_0)^p(y-y_0)^{k-p}, & \text{if } 0\leq p\leq k,\\ 0, & \text{otherwise.}\end{cases} \tag{4.39}$$

Given some matrix $M \in \mathbb{R}^{m_1 \times m_2}$, we define the matrix $M_{|_k^j} \in \mathbb{R}^{m_1 \times m_2}$, $1 \leq j \leq k \leq m_1$ which is the restriction of M beetween the rows j and k

$$M = \begin{pmatrix} m_{1,1} & \cdots & m_{1,m_2} \\ m_{2,1} & \cdots & m_{2,m_2} \\ \vdots & & \vdots \\ m_{m_1-1,1} & \cdots & m_{m_1-1,m_2} \\ m_{m_1,1} & \cdots & m_{m_1,m_2} \end{pmatrix}, \quad M_{|_k^j} = \begin{pmatrix} & 0 & \\ m_{j,1} & \cdots & m_{j,m_2} \\ \vdots & & \vdots \\ m_{k,1} & \cdots & m_{k,m_2} \\ & 0 & \end{pmatrix}. \tag{4.40}$$

If $k < j$ we set by convention $M_{|_k^j} = 0$. Now we recursively define some coefficients.

Definition 4.33. Consider an integer $n \geq 0$. The matrices $F_k^p(\mathbf{x}), G_k^p(\mathbf{x}), \Gamma_k^p(\mathbf{x}) \in \mathbb{R}^{m \times m}$ are defined in the range $0 \leq p \leq k \leq n+1$ with two recursions. The first recursion reads:

- by convention set $\Gamma_k^{-1} = \Gamma_k^{k+1} = 0$, $\Gamma_{-1}^p = 0$, $\forall p, k$
- for $k = 0$ to $k = n+1$, do
  - for $p = 0$ to $p = k$, do

$$\Gamma_k^p(\mathbf{x}) := \gamma_k^p(\mathbf{x}) I_m - \Gamma_{k-1}^{p-1} R_{|_m^2}^{-1} A_1 - \Gamma_{k-1}^p R_{|_m^2}^{-1} A_2 \tag{4.41}$$

The second recursion reads:

- by convention set $G_{n+1}^p(\mathbf{x}) = G_{n+2}^p(\mathbf{x}) = 0$, $G_k^{-1}(\mathbf{x}) = G_k^{-2}(\mathbf{x}) = 0$, $\forall p, k$
- for $k = n$ to $k = 0$, do
  - for $p = 0$ to $p = k$, do

$$F_k^p(\mathbf{x}) := \Gamma_k^p(\mathbf{x}) + \lambda^2 G_{k+2}^p(\mathbf{x}), \tag{4.42}$$

$$G_k^p(\mathbf{x}) := F_k^p(\mathbf{x}) - G_k^{p-2}(\mathbf{x}), \tag{4.43}$$

Theorem 4.34 (Polynomial solutions to the P_N model.)**.** *Consider the exponential solutions given in Theorem 4.25 associated with the eigenvalue μ_i which satisfies $\mu_i \underset{\sigma_a \to 0}{\to} 0$. There exits linear combinations of these solutions which degenerate toward polynomial solutions when $\sigma_a \to 0$. These polynomial solutions are given by the first column of the matrices $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$.*

The matrices $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$ can be recursively calculated and we give here the example of the P_1 model.

Example 4.35 (Application to the P_1 model)**.** Consider the two dimensional P_1 model. To recover the standard notations we switch here the axis x and y compare to Example 4.13. One has

$$A_1 = \frac{c}{\sqrt{3}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \frac{c}{\sqrt{3}} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} \varepsilon\sigma_a & 0 & 0 \\ 0 & \sigma_t & 0 \\ 0 & 0 & \sigma_t \end{pmatrix}, \quad R_{|_m^2}^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{\sigma_t} & 0 \\ 0 & 0 & \frac{1}{\sigma_t} \end{pmatrix},$$

with $\sigma_t := \varepsilon\sigma_a + \frac{\sigma_s}{\varepsilon}$. The first matrices G_k^k and G_k^{k-1} reads

$$G_0^0(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad G_1^0(\mathbf{x}) = \begin{pmatrix} y & 0 & 0 \\ 0 & y & 0 \\ -\frac{c}{\sqrt{3}\sigma_t} & 0 & y \end{pmatrix}, \quad G_1^1(\mathbf{x}) = \begin{pmatrix} x & 0 & 0 \\ -\frac{c}{\sqrt{3}\sigma_t} & x & 0 \\ 0 & 0 & x \end{pmatrix},$$

$$G_2^1(\mathbf{x}) = \begin{pmatrix} xy & 0 & 0 \\ -\frac{c}{\sqrt{3}\sigma_t} y & xy & 0 \\ -\frac{c}{\sqrt{3}\sigma_t} x & 0 & xy \end{pmatrix}, \quad G_2^2(\mathbf{x}) = \begin{pmatrix} \frac{1}{2}(x^2-y^2) & 0 & 0 \\ -\frac{c}{\sqrt{3}\sigma_t} x & \frac{1}{2}(x^2-y^2) & 0 \\ \frac{c}{\sqrt{3}\sigma_t} y & 0 & \frac{1}{2}(x^2-y^2) \end{pmatrix},$$

and one can check that their first column is indeed solution to the P_1 model when $\sigma_a = 0$. ●

The proof of Theorem 4.34 is decomposed in two steps. At first, we study in Section 4-2.3.1 the degenerative exponentials of a simplified second order equation and introduce some useful tools. Using this framework, we then prove the Theorem 4.34 in Section 4-2.3.2.

4-2.3.1 A simplified second order equation

In this section, we treat the simpler case of a second order equation. The results obtained in this section will then be generalized to prove Theorem 4.34. More precisely, Propositions 4.37, 4.38, 4.40 and the definition (4.54) of the matrix S_ω will be needed in Section 4-2.3.2.

The goal here is to find linear combinations of functions of the form $e^{\omega(\mathbf{d}_i,\mathbf{x})}$ which degenerate toward polynomial functions when $\omega \to 0$. The degeneracy of exponential solutions to the Helmholtz equation has already been studied in [GHP09]. However, since our goal is to generalize the proof for the two dimensional P_N system, we consider here a different approach. The procedure is based on a simplification of the Taylor expansion for the solutions to the second order equation.

Property of the solutions to a second order equation. Let $u \in H^1(\Omega)$. We consider the following auxiliary second order equation

$$\Delta u(\mathbf{x}) = \omega u(\mathbf{x}), \tag{4.44}$$

with $\omega \in \mathbb{R}$ which may take positive or negative values. At first, our goal is to write a simplified Taylor expansion for regular solutions to this equation. Every function $u(\mathbf{x}) \in C^{n+1}(\Omega)$ can be written under the form of a usual Taylor expansion which comes from [Fle77, Page 94]

$$u(\mathbf{x}) = \sum_{k=0}^{n}\sum_{p=0}^{k} \gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}u(\mathbf{x}_0) + \sum_{p=0}^{n+1} \gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s), \tag{4.45}$$

where $\gamma_k^p(\mathbf{x})$ is given by (4.39), $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1-s)x_0 + sx$, $y_s = (1-s)y_0 + sy$, $s \in [0,1]$. In our analysis, we need intermediate quantities named α_k^p and β_k^p.

Definition 4.36. Consider an integer $n \geq 0$. The functions $\alpha_k^p(\mathbf{x})$ and $\beta_k^p(\mathbf{x})$ are defined in the range $0 \leq p \leq k \leq n$ by a decreasing recursion from $k = n$ to $k = 0$. The recursion reads:

- by convention set $\beta_{n+1}^p(\mathbf{x}) = \beta_{n+2}^p(\mathbf{x}) = 0$, $\beta_k^{-1}(\mathbf{x}) = \beta_k^{-2}(\mathbf{x}) = 0$, $\forall p, k$.
- for $k = n$ to $k = 0$, do
  - for $p = 0$ to $p = k$, do

$$\alpha_k^p(\mathbf{x}) := \gamma_k^p(\mathbf{x}) + \omega\beta_{k+2}^p(\mathbf{x}), \tag{4.46}$$

$$\beta_k^p(\mathbf{x}) := \alpha_k^p(\mathbf{x}) - \beta_k^{p-2}(\mathbf{x}), \tag{4.47}$$

Since $\beta_{n+1}^p(\mathbf{x}) = \beta_{n+2}^p(\mathbf{x}) = 0$, thus $\alpha_{n-1}^p(\mathbf{x}) = \gamma_{n-1}^p(\mathbf{x})$, $\alpha_n^p(\mathbf{x}) = \gamma_n^p(\mathbf{x})$. Also because $\beta_k^{-2} = \beta_k^{-1} = 0$ the equality (4.47) implies

$$\beta_k^0(\mathbf{x}) = \alpha_k^0(\mathbf{x}), \quad \beta_k^1(\mathbf{x}) = \alpha_k^1(\mathbf{x}), \quad 0 \leq k \leq n. \tag{4.48}$$

In the case $\omega \neq 0$, the functions $\alpha_k^p(\mathbf{x})$ and $\beta_k^p(\mathbf{x})$ are polynomials of degree n if both n and k are even or odd and $n-1$ otherwise. If $\omega = 0$, the functions $\alpha_k^p(\mathbf{x})$ and $\beta_k^p(\mathbf{x})$ are polynomials of degree k for $0 \leq k \leq n$. Note that in order to use simple notation, we do not explicitly write the dependence of these functions in n and $\mathbf{x}_0$.

Proposition 4.37 (Simplified Taylor expansion)**.** *Assume $u(\mathbf{x}) \in C^{n+1}(\Omega)$ is solution to (4.44). Then the double sum Taylor expansion in (4.45) can be recast as a simple sum with only zero or first order derivatives with respect to y*

$$\begin{aligned} u(\mathbf{x}) = \beta_0^0(\mathbf{x})u(\mathbf{x}_0) + \sum_{k=1}^{n}\left[\beta_k^k(\mathbf{x})\partial_x^k u(\mathbf{x}_0) + \beta_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y u(\mathbf{x}_0)\right] \\ + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s), \quad \forall \mathbf{x}\in\Omega, \end{aligned} \tag{4.49}$$

where $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1-s)x_0 + sx$ and $y_s = (1-s)y_0 + sy$, $s \in [0,1]$.

By symmetry, a similar result holds with high order derivative with respect to y and only zero and first order derivatives with respect to x. Even if the rigorous proof of Proposition 4.37 is a little technical, the idea behind is actually very simple. Indeed, since u is solution to (4.44) one can use the equality $\partial_y^2 u = (\omega - \partial_x^2)u$ to recursively eliminate the derivatives of y. A graphical illustration of the procedure is provided in Figure 4.3. The proof of Proposition 4.37 is postponed in Appendix B.

We show that in the case $\omega = 0$ the coefficients $\beta_k^k(\mathbf{x})$ and $\beta_k^{k-1}(\mathbf{x})$ are harmonic polynomials. We define the harmonic polynomials as follow

$$q_1(\mathbf{x}) = 1, \quad q_{2k}(\mathbf{x}) = \frac{1}{k!}\Re((x-x_0)+i(y-y_0))^k, \quad q_{2l+1}(\mathbf{x}) = \frac{1}{k!}\Im((x-x_0)+i(y-y_0))^k, \quad \text{for } l \in \mathbb{N}^*. \tag{4.50}$$

Consider the two following two vectorial functions $\boldsymbol{\beta}(\mathbf{x}), \mathbf{q}(\mathbf{x}) \in \mathbb{R}^{2n+1}$ where $\boldsymbol{\beta}(\mathbf{x})$ is the vectorial function made of the coefficients $\beta_k^{k-1}(\mathbf{x})$ and β_k^k and $\mathbf{q}(\mathbf{x})$ is the vectorial function made of the harmonic polynomials $q_i(\mathbf{x})$

$$\mathbf{q}(\mathbf{x}) = \Big(q_1(\mathbf{x}), ..., q_{2n+1}(\mathbf{x})\Big)^T, \quad \boldsymbol{\beta}(\mathbf{x}) = \Big(\beta_0^0(\mathbf{x}), \beta_1^0(\mathbf{x}), \beta_1^1(\mathbf{x}), ..., \beta_n^{n-1}(\mathbf{x}), \beta_n^n(\mathbf{x})\Big)^T. \tag{4.51}$$

Proposition 4.38 (Limit of the coefficients $\beta_k^k(\mathbf{x})$ and $\beta_k^{k-1}(\mathbf{x})$)**.** *Assume $\omega = 0$ and $0 \le k \le n$. The coefficients β_k^k and β_k^{k-1} are harmonic polynomials when $\omega = 0$. More precisely, one has*

$$\boldsymbol{\beta}(\mathbf{x}) \underset{\omega\to 0}{\to} \mathbf{q}(\mathbf{x}). \tag{4.52}$$

Proof. The proof is postponed in Appendix B. ∎

Limit for linear combinations of exponential functions. We consider v_i solutions to the second order equation (4.44) and assume $\Omega \in \mathbb{R}^2$ is a compact set. More precisely, we consider the following exponential functions centered in $\mathbf{x}_0 \in \mathbb{R}^2$

$$v_i(\mathbf{x}) = e^{\sqrt{\omega}(\mathbf{d}_i, \mathbf{x}-\mathbf{x}_0)}, \quad i = 1, ..., 2n+1. \tag{4.53}$$

Since we are interested in the regime $\omega \to 0$, we will assume $\omega \le 1$. We define $S_\omega := S_{v_1, v_2, ..., v_{2n+1}} \in \mathbb{R}^{2n+1\times 2n+1}$ such that

$$S_\omega(\mathbf{x}_0) := S_{v_1,v_2,...,v_{2n+1}}(\mathbf{x}_0) := \begin{pmatrix} v_1(\mathbf{x}_0) & v_2(\mathbf{x}_0) & ... & v_{2n+1}(\mathbf{x}_0) \\ \partial_x v_1(\mathbf{x}_0) & \partial_x v_2(\mathbf{x}_0) & ... & \partial_x v_{2n+1}(\mathbf{x}_0) \\ \partial_y v_1(\mathbf{x}_0) & \partial_y v_2(\mathbf{x}_0) & ... & \partial_y v_{2n+1}(\mathbf{x}_0) \\ \partial_x^2 v_1(\mathbf{x}_0) & \partial_x^2 v_2(\mathbf{x}_0) & ... & \partial_x^2 v_{2n+1}(\mathbf{x}_0) \\ \partial_x\partial_y v_1(\mathbf{x}_0) & \partial_x\partial_y v_2(\mathbf{x}_0) & ... & \partial_x\partial_y v_{2n+1}(\mathbf{x}_0) \\ \vdots & \vdots & ... & \vdots \\ \partial_x^n v_1(\mathbf{x}_0) & \partial_x^n v_2(\mathbf{x}_0) & ... & \partial_x^n v_{2n+1}(\mathbf{x}_0) \\ \partial_x^{n-1}\partial_y v_1(\mathbf{x}_0) & \partial_x^{n-1}\partial_y v_2(\mathbf{x}_0) & ... & \partial_x^{n-1}\partial_y v_{2n+1}(\mathbf{x}_0) \end{pmatrix}. \tag{4.54}$$

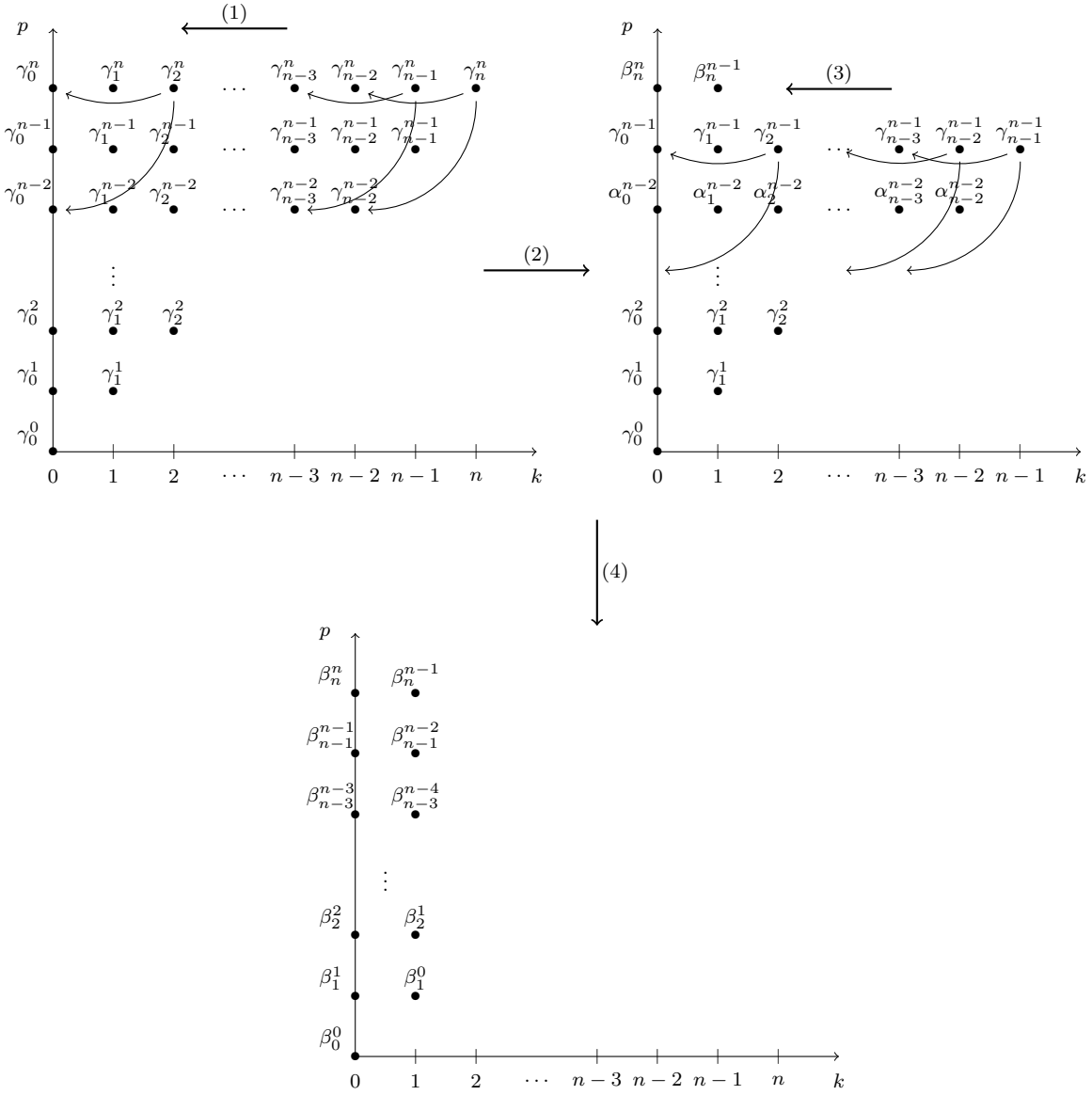

Figure 4.3 – Illustration of the recursive procedure to get the simplified Taylor expansion (4.49).

Proposition 4.39. *Let* $\mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), ..., v_{2n+1}(\mathbf{x}))^T \in \mathbb{R}^{2n+1}$ *with* $v_i(\mathbf{x}) \in C^{2n+1}(\Omega)$ *solution to* (4.44). *The vectorial function* $\mathbf{v}(\mathbf{x})$ *satisfies the following Taylor expansion*

$$\mathbf{v}(\mathbf{x}) = S_\omega^T(\mathbf{x}_0)\boldsymbol{\beta}(\mathbf{x}) + \boldsymbol{\xi}(\mathbf{x}),$$

with $\boldsymbol{\xi}(\mathbf{x}) = (\xi_1(\mathbf{x}), ..., \xi_{2n+1}(\mathbf{x}))^T$, $\xi_i(\mathbf{x}) = \sum_{p=0}^{n+1} \partial_x^p \partial_y^{n+1-p} v_i(\mathbf{x}_s)\gamma_{n+1}^p(\mathbf{x})$.

Proof. The proof follows from Proposition 4.37 and the definition of the matrix $S_\omega(\mathbf{x}_0)$. ■

With the exponentials (4.53) the matrix $S_\omega(\mathbf{x}_0)$ reads

$$S_\omega := S_\omega(\mathbf{x}_0) = \begin{pmatrix} 1 & 1 & ... & 1 \\ \sqrt{\omega}\cos(\theta_1) & \sqrt{\omega}\cos(\theta_2) & ... & \sqrt{\omega}\cos(\theta_{2n+1}) \\ \sqrt{\omega}\sin(\theta_1) & \sqrt{\omega}\sin(\theta_2) & ... & \sqrt{\omega}\sin(\theta_{2n+1}) \\ \vdots & \vdots & ... & \vdots \\ \sqrt{\omega}^n\cos^n(\theta_1) & \sqrt{\omega}^n\cos^n(\theta_2) & ... & \sqrt{\omega}^n\cos^n(\theta_{2n+1}) \\ \sqrt{\omega}^n\sin(\theta_1)\cos^{n-1}(\theta_1) & \sqrt{\omega}^n\sin(\theta_2)\cos^{n-1}(\theta_2) & ... & \sqrt{\omega}^n\sin(\theta_{2n+1})\cos^{n-1}(\theta_{2n+1}) \end{pmatrix}.$$

For simplicity, we remove the dependence in $\mathbf{x}_0$ of the matrix S_ω.

Proposition 4.40. *The matrix $S := S_{\omega=1}$ is invertible.*

Proof. This is a particular case of Proposition 4.58 with $N = 1$. ∎

We now define $s_{k,j} = (S_\omega)_{k,j}$, $a_{k,j} = (S)^{-1}_{k,j}$ and notice

$$(S_\omega S^{-1})_{l,j} = \sum_{k=1}^{2n+1} s_{l,k}a_{k,j} = \sqrt{\omega}^{\lfloor \frac{l}{2} \rfloor}\delta_{lj} = \begin{cases} \sqrt{\omega}^{\lfloor \frac{l}{2} \rfloor} & \text{if } l = j \\ 0 \quad \text{else} \end{cases} \quad \text{for } 1 \leq l, j \leq 2n+1. \tag{4.55}$$

Therefore setting

$$D_\omega := \text{diag}(\omega^{\lfloor \frac{1}{2} \rfloor}, \omega^{\lfloor \frac{2}{2} \rfloor}, ..., \omega^{\lfloor \frac{2n+1}{2} \rfloor}), \tag{4.56}$$

one can write the matrix S_ω^{-T} as

$$S_\omega^{-T} = D_\omega^{-1} S^{-T}. \tag{4.57}$$

Now we have introduced the main objects needed to study degenerative exponentials to the P_N model. To give the general idea we study the degenerative exponentials to the second order equation as a first step.

Definition 4.41. We define $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^{2n+1}$ such that

$$\mathbf{c}(\mathbf{x}) = S_\omega^{-T}\mathbf{v}(\mathbf{x}). \tag{4.58}$$

with $\mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), ..., v_{2n+1}(\mathbf{x})) \in \mathbb{R}^{2n+1}$ and $v_i(\mathbf{x})$ given in (4.53).

The vector $\mathbf{c}(\mathbf{x})$ degenerates toward harmonic polynomials.

Proposition 4.42 (Second order equation: degeneration of exponential solutions toward polynomials)**.** *Consider the functions* (4.58)*. Each component of the vector* $\mathbf{c}(\mathbf{x})$ *is solution to* (4.44)*. Moreover one has*

$$\mathbf{c}(\mathbf{x}) \underset{\omega \to 0}{\to} \mathbf{q}(\mathbf{x}).$$

Proof. The components of the vector $\mathbf{c}$ are linear combination of the functions v_i which are solutions to (4.44) and are therefore also solution to (4.44). Since the functions v_i are solutions to the second order equation (4.44), one can use Proposition 4.39 on the vector $\mathbf{v}$ and writes

$$\mathbf{v}(\mathbf{x}) = S_\omega^T \boldsymbol{\beta}(\mathbf{x}) + \boldsymbol{\xi}(\mathbf{x}).$$

with $\boldsymbol{\xi}(\mathbf{x}) = (\xi_1(\mathbf{x}), ..., \xi_{2n+1}(\mathbf{x}))$ and $\boldsymbol{\beta}_{2n+1}(\mathbf{x}) = (\beta_0^0(\mathbf{x}), \beta_1^0(\mathbf{x}), \beta_1^1(\mathbf{x}), ..., \beta_n^{n-1}(\mathbf{x}), \beta_n^n(\mathbf{x}))$. Since Ω is bounded and from the definition of the functions v_i one has

$$\xi_i(\mathbf{x}) = \sum_{p=0}^{n+1} \partial_x^p \partial_y^{n+1-p} v_i(\mathbf{x}_s)\gamma_{n+1}^p(\mathbf{x}) = \sqrt{\omega}^{n+1}\widetilde{\xi}_i(\mathbf{x}),$$

where $\widetilde{\xi}_i(\mathbf{x})$ is bounded uniformly in $\mathbf{x}$ and ω when $\omega \to 0$. Using $\widetilde{\boldsymbol{\xi}}(\mathbf{x}) = (\widetilde{\xi}_1(\mathbf{x}), ..., \widetilde{\xi}_{2n+1}(\mathbf{x}))^T$ one gets

$$\mathbf{v}(\mathbf{x}) = S_\omega^T \boldsymbol{\beta}(\mathbf{x}) + \sqrt{\omega}^{n+1}\widetilde{\boldsymbol{\xi}}(\mathbf{x}).$$

Therefore, the vector $\mathbf{c}$ can be written under the form

$$\begin{aligned} \mathbf{c}(\mathbf{x}) &= S_\omega^{-T}\left(S_\omega^T \boldsymbol{\beta}(\mathbf{x}) + \sqrt{\omega}^{n+1}\widetilde{\boldsymbol{\xi}}(\mathbf{x})\right), \\ &= \boldsymbol{\beta}(\mathbf{x}) + \sqrt{\omega}^{n+1} D_\omega^{-1} S^{-T}\widetilde{\boldsymbol{\xi}}(\mathbf{x}), \end{aligned}$$

where we have used the equality (4.57) in the second term on the right hand side of the last equality. From the definition (4.56) of D_ω one gets $\sqrt{\omega}^{n+1} D_\omega^{-1} \underset{\omega\to 0}{\to} 0$. Therefore one finally finds

$$\mathbf{c}(\mathbf{x}) \underset{\omega\to 0}{\to} \boldsymbol{\beta}(\mathbf{x}).$$

Using Proposition 4.38 completes the proof. ■

4-2.3.2 Proof of Theorem 4.34

Now we generalize the results given in the previous section to a class of system which includes the P_N model

$$\Big(A_1\partial_x + A_2\partial_y\Big)\mathbf{u}(\mathbf{x}) = -R\mathbf{u}(\mathbf{x}), \tag{4.59}$$

where $\mathbf{u} \in \mathbb{R}^m$, $A_1, A_2, R \in \mathbb{R}^m$ and the matrix $R := R_\omega$ depends on some coefficient ω. The only assumption we make on the system (4.59) is the following.

Assumption 4.43. *There exists $\mathbf{u}_\lambda(\mathbf{x}) \in \mathbb{R}^m$ solution to the system (4.59) which satisfies*

$$\mathbf{u}_\lambda(\mathbf{x}) = \mathbf{z}e^{\lambda(\mathbf{d},\mathbf{x})}, \quad \lambda \underset{\omega\to 0}{\to} 0, \tag{4.60}$$

with

$$\mathbf{d} := \mathbf{d}(\theta) := \Big(\cos\theta, \sin\theta\Big)^T \in \mathbb{R}^2, \quad \mathbf{z} := \mathbf{z}(\theta,\omega) \neq \mathbf{0} \in \mathbb{R}^m, \quad \lambda := \lambda(\omega) \in \mathbb{R}.$$

Moreover, there exists a matrix denoted $R^{-1}_{|_m^2}$ such that

$$R^{-1}_{|_m^2} R\mathbf{z} = \mathbf{z}_{|_m^2}, \quad \forall \mathbf{z} \in \mathbb{R}^m, \quad \forall \omega \in \mathbb{R}^+, \tag{4.61}$$

where we used the notation (4.40) for the vector $\mathbf{z}_{|_m^2}$.

From Proposition 4.21 and the definition of the matrix $R := \mathrm{diag}(\varepsilon\sigma_a, \sigma_t, ..., \sigma_t)$, the P_N model satisfies the Assumption 4.43 when $\sigma_t > 0$ and with $\omega = \sigma_a$. We can now prove the degeneracy of the exponentials solutions (4.60) toward polynomials as we did for the second order equation in the previous section. We start with a technical lemma.

Lemma 4.44. *Rescaling the vector $\mathbf{z} = (z_1, ..., z_m)^T$ in (4.60) if needed, one can take $z_1 = 1$ with all the other components satisfying $z_j \underset{\omega\to 0}{\to} 0$, $2 \leq j \leq m$.*

Proof. From the definition of the function (4.60) one gets $\mathbf{u}_\lambda(\mathbf{x}) \underset{\omega\to 0}{\to} \mathbf{z}$. Therefore from (4.59), the definition of the matrix $R^{-1}_{|_m^2}$ (4.61) and because $\mathbf{u}_\lambda(\mathbf{x})$ is constant in the limit $\omega \to 0$ one has

$$\mathbf{z}_{|_m^2} = R^{-1}_{|_m^2}\Big(A_1\partial_x + A_2\partial_y\Big)\mathbf{u}_\lambda(\mathbf{x}) \underset{\omega\to 0}{\to} \mathbf{0}. \tag{4.62}$$

But since $\mathbf{z} \neq 0$, and considering the correct scaling of the function, there exists at least one component $z_j \neq 0$ such that z_j does not depend on ω and all other components z_i, $i \neq j$ are bounded when $\omega \to 0$. One deduces from (4.62) that it can't be any components between 2 and m therefore $j = 1$ and one can take $z_1 = 1$ considering the correct scaling. ■

In the following we take $z_1 = 1$ and use the matrices Γ_k^p and G_k^p from the Definition 4.33. We make a first simplification on the Taylor expansion of solutions to (4.59).

Proposition 4.45 (A first simplification of the Taylor expansion)**.** *Let* $\mathbf{u}(\mathbf{x}) \in C^{n+2}(\Omega)$ *be a solution to* (4.59)*. The beginning of the Taylor expansion on the vectorial function* $\mathbf{u}(\mathbf{x})$ *can be recast as a Taylor expansion on the vectorial function* $\mathbf{u}_{|_1^1}(\mathbf{x}_0)$

$$\begin{aligned}\mathbf{u}(\mathbf{x}) &= \sum_{k=0}^{n}\sum_{p=0}^{k}\Gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}\mathbf{u}_{|_1^1}(\mathbf{x}_0) + \boldsymbol{\xi}(\mathbf{x}),\\ \boldsymbol{\xi}(\mathbf{x}) &= \sum_{p=0}^{n+1}\Gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}\mathbf{u}(\mathbf{x}_0) + \sum_{p=0}^{n+2}\gamma_{n+2}^p(\mathbf{x})I_m\partial_x^p\partial_y^{n+2-p}\mathbf{u}(\mathbf{x}_s),\end{aligned} \tag{4.63}$$

where $\mathbf{x}_s = (x_s, y_s)^T$*,* $x_s = (1-s)x_0 + sx$ *and* $y_s = (1-s)y_0 + sy$*,* $s \in [0,1]$*.*

The proof which is purely technical is postponed in Appendix B.

Now we make a second simplification and remove some derivatives in the Taylor expansion. This is the same idea we used for the second order equation in Proposition 4.37.

Proposition 4.46 (A second simplification of the Taylor expansion)**.** *Assume* $\mathbf{u}_\lambda(\mathbf{x})$ *is solution to* (4.59) *and can be written under the form* (4.60)*. Then the double sum Taylor expansion in* (4.63) *can be recast as a simple sum with only zero or first order derivatives with respect to* y

$$\mathbf{u}_\lambda(\mathbf{x}) = G_0^0(\mathbf{x})\mathbf{u}_{\lambda|_1^1}(\mathbf{x}_0) + \sum_{k=1}^{n}\left[G_k^k(\mathbf{x})\partial_x^k\mathbf{u}_{\lambda|_1^1}(\mathbf{x}_0) + G_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y\mathbf{u}_{\lambda|_1^1}(\mathbf{x}_0)\right] + \boldsymbol{\xi}_\lambda(\mathbf{x}),$$

where $\boldsymbol{\xi}_\lambda(\mathbf{x})$ *is as in Proposition 4.45 replacing* $\mathbf{u}(\mathbf{x})$ *with* $\mathbf{u}_\lambda(\mathbf{x})$*.*

Proof. Since the solution is under the form (4.60), its first component follows the second order equation (4.44) with $\omega = \lambda^2$. Therefore, each component of the vector $\mathbf{u}_{\lambda|_1^1}(\mathbf{x})$ follows also the equation (4.44) with $\omega = \lambda^2$. Using the Taylor expansion (4.63) one can proceed as in Proposition 4.37 replacing the coefficients $\gamma_k^p(\mathbf{x})$ and $\beta_k^p(\mathbf{x})$ with the matrices $\Gamma_k^p(\mathbf{x})$ and $G_k^p(\mathbf{x})$ respectively. This completes the proof. ∎

We consider now $2n+1$ functions under the form (4.60) centered in $\mathbf{x}_0$

$$\mathbf{v}_i(\mathbf{x}) = \mathbf{z}_i e^{\lambda \mathbf{d}_i^T(\mathbf{x}-\mathbf{x}_0)}, \quad i = 1, ..., 2n+1,$$

with $d_i \neq d_j$ if $i \neq j$ and $\lambda \underset{\omega\to 0}{\to} 0$.

We define the following functions.

Definition 4.47. Let $(G_k^p)_{.,1}(\mathbf{x}) \in \mathbb{R}^m$ be the first column of the matrix $G_k^p(\mathbf{x})$. We denote $B(\mathbf{x}), V(\mathbf{x})$ the matrices

$$B(\mathbf{x}) := \Big((G_0^0)_{.,1}(\mathbf{x}), (G_1^0)_{.,1}(\mathbf{x}), (G_1^1)_{.,1}(\mathbf{x}), ..., (G_n^{n-1})_{.,1}(\mathbf{x}), (G_n^n)_{.,1}(\mathbf{x})\Big)^T \in \mathbb{R}^{2n+1\times m},$$
$$V(\mathbf{x}) := \Big(\mathbf{v}_1(\mathbf{x}), ..., \mathbf{v}_{2n+1}(\mathbf{x})\Big)^T \in \mathbb{R}^{2n+1\times m}.$$

Finally, we define $C(\mathbf{x})$ the matrix

$$C(\mathbf{x}) = S_{\lambda^2}^{-T}V(\mathbf{x}) \in \mathbb{R}^{2n+1\times m}, \tag{4.64}$$

where $S_{\lambda^2} = S_\omega$ is defined in (4.54).

One can write a simplify Taylor expansion of the matrix $V(\mathbf{x})$.

Proposition 4.48. *The matrix $V(\mathbf{x})$ admits the following Taylor expansion*

$$V(\mathbf{x}) = S_{\lambda^2}^T B(\mathbf{x}) + \Xi(\mathbf{x}),$$

with $\Xi(\mathbf{x}) \in \mathbb{R}^{2n+1\times m}$, $\Xi(\mathbf{x}) := (\boldsymbol{\xi}_1(\mathbf{x}), ..., \boldsymbol{\xi}_{2n+1}(\mathbf{x}))^T$.

Proof. This is a consequence of Proposition 4.46. Indeed, each vector $\mathbf{v}_i$ can be written under the form

$$\mathbf{v}_i(\mathbf{x}) = G_0^0(\mathbf{x})\mathbf{v}_{i|_1^1}(\mathbf{x}_0) + \sum_{k=1}^{n}\left[G_k^k(\mathbf{x})\partial_x^k \mathbf{v}_{i|_1^1}(\mathbf{x}_0) + G_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y \mathbf{v}_{i|_1^1}(\mathbf{x}_0)\right] + \boldsymbol{\xi}_i(\mathbf{x}).$$

Using $G_k^p(\mathbf{x})\mathbf{v}_{i|_1^1}(\mathbf{x}) = v_i^1(\mathbf{x})(G_k^p)_{.,1}(\mathbf{x})$ (where $v_i^1(\mathbf{x})$ is the first component of the vector $\mathbf{v}_i$), the definition of the matrix S_ω (4.54) and $\Xi(\mathbf{x}) = (\boldsymbol{\xi}_1(\mathbf{x}), ..., \boldsymbol{\xi}_{2n+1}(\mathbf{x}))^T$ completes the proof. ■

And we can now give the main result of this section.

Proof of Theorem 4.34. From Proposition 4.48 one has

$$V(\mathbf{x}) = S_{\lambda^2}^T B(\mathbf{x}) + \Xi(\mathbf{x}),$$

Since Ω is bounded, from the definition of the functions $\mathbf{v}_i$ and because each component of the vectors $\boldsymbol{\xi}_i$ is at least a derivative of order $n+1$ of $\mathbf{v}_i$, one has $\boldsymbol{\xi}_i(\mathbf{x}) = \lambda^{n+1}\widetilde{\boldsymbol{\xi}}_i(\mathbf{x})$ where $\widetilde{\boldsymbol{\xi}}_i(\mathbf{x})$ is bounded uniformly in $\mathbf{x}$ and λ when $\lambda \to 0$. The equality now reads

$$V(\mathbf{x}) = S_{\lambda^2}^T B(\mathbf{x}) + \lambda^{n+1}\widetilde{\Xi}(\mathbf{x}).$$

Therefore the matrix $C(\mathbf{x})$ can be written under the form

$$\begin{aligned} C(\mathbf{x}) &= S_{\lambda^2}^{-T}\left(S_{\lambda^2}^T B(\mathbf{x}) + \lambda^{n+1}\widetilde{\Xi}(\mathbf{x})\right), \\ &= B(\mathbf{x}) + \lambda^{n+1} D_{\lambda^2}^{-1} S^{-T}\widetilde{\Xi}(\mathbf{x}), \end{aligned}$$

where we have used the equality (4.57) in the second term on the right hand side of the last equality. From the definition of D_{λ^2} (4.56) one gets $\lambda^{n+1}D_{\lambda^2}^{-1} \underset{\omega\to 0}{\to} 0$. Therefore one finally finds

$$C(\mathbf{x}) \underset{\omega\to 0}{\to} B(\mathbf{x}). \tag{4.65}$$

Finally, the rows of the matrix $C(\mathbf{x})$ are linear combinations of the rows of the matrix V which are solutions to (4.59). Therefore, each row of C is also solution to (4.59). Therefore, from (4.65), each row of the matrix $B(\mathbf{x})$ is a limit of a solution and one deduces that the rows of $B(\mathbf{x})$ are also solutions to (4.59) when $\omega = 0$ and the proof is complete. ■

The rows of the matrix $B(\mathbf{x})$ are actually the first column of the matrices $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$. Therefore, to get polynomial solutions to the system (4.59) in the case $\omega = 0$, one just has to study the matrices $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$. One deduces the following corollary.

Corollary 4.49 (Polynomial solutions to the P_N model in the case $\sigma_a = 0$)**.** *The first column of $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$ are polynomial solutions to the P_N model when $\sigma_a = 0$. Moreover their first component are the harmonics polynomials of degree k (4.50) while all their other components have degree strictly inferior to k.*

Proof. Since the P_N model satisfies Assumption 4.43 (with $\omega = \sigma_a$) one deduces that the first column of $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$ are polynomial solutions to the P_N model when $\sigma_a = 0$.

We are interested in the case $\sigma_a = 0$ and therefore we take $\lambda = 0$ in the formulas (4.41)-(4.42)-(4.43). Since the first row of the matrix $R_{|_m^2}^{-1}$ is zero, the first component of the first column of $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$ is construct in the same way as the coefficients $\beta_k^k(\mathbf{x})$ and $\beta_k^{k-1}(\mathbf{x})$ respectively in (4.46)-(4.47) (with $\omega = 0$). From Proposition 4.38, one deduces that the first component of the first column of $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$ is the harmonic polynomial given by (4.50). For the other components of the first column of $G_k^k(\mathbf{x})$ and $G_k^{k-1}(\mathbf{x})$, the contribution of polynomials are only made through the matrices $\Gamma_{k-1}^{p-1}(\mathbf{x})$ and $\Gamma_{k-1}^{p}(\mathbf{x})$ which are at most of degree $k-1$. This completes the proof. ■

Remark 4.50. From Corollary 4.49, the first component of the polynomial solutions to the P_N model derived in this section is a harmonic polynomial. This was also the case for the polynomial solutions obtained with Birkhoff and Abu-Shumays method's in Section 4-2.2. For the P_1 model, it is enough to conclude that these solutions are the same. Indeed from the structure of the P_1 model

$$\begin{cases} \dfrac{1}{\sqrt{3}}\Big(\partial_x u_2(\mathbf{x}) + \partial_y u_3(\mathbf{x})\Big) = -\sigma_a u_1(\mathbf{x}), \\ \dfrac{1}{\sqrt{3}}\partial_x u_1(\mathbf{x}) = -\sigma_t u_2(\mathbf{x}), \\ \dfrac{1}{\sqrt{3}}\partial_y u_1(\mathbf{x}) = -\sigma_t u_3(\mathbf{x}), \end{cases}$$

one deduces that if $u_1 = 0$ then $u_2 = u_3 = 0$.

We conjecture that the polynomial solutions obtained in this section and the previous one are also the same for the $P_{N>1}$ model. ●

4-2.4 Time dependent solutions

In this section, we give some possible ways to get time dependent solutions to the P_N model. These solutions will be used as basis functions for the TDG method in Chapter 5. As we will see, although such basis functions are very effective to reduce the diffusion, they might also deteriorate the condition number of the mass matrix.

- A first possibility is to consider solutions to the P_N model (4.17) which depend only on the time variable. One gets

  $$\varepsilon \partial_t \mathbf{u}(t) = -R\mathbf{u}(t).$$

  Since R is a diagonal matrix, $R = \mathrm{diag}(\varepsilon\sigma_a, \sigma_t, ..., \sigma_t)$, one immediately gets the following solutions

  $$\mathbf{v}_1(t) = \mathbf{e}_1 e^{-\sigma_a t}, \quad \mathbf{v}_2(t) = \mathbf{e}_2 e^{-\frac{\sigma_t}{\varepsilon}t}, \quad ... \quad , \quad \mathbf{v}_m(t) = \mathbf{e}_m e^{-\frac{\sigma_t}{\varepsilon}t}, \tag{4.66}$$

  where the functions $\mathbf{e}_i$ represent the canonical basis of $\mathbb{R}^m$. One can use the solutions $\mathbf{v}_i$ as basis functions. Note however that

  $$\mathbf{v}_1(t) \underset{\sigma_a \to 0}{\to} \mathbf{e}_1.$$

  The function $\mathbf{e}_1$ is a stationary solution and may be already in the approximation space if one uses time dependent and stationary basis functions. Therefore, one will "loose" a basis function when $\sigma_a \to 0$. This can be seen as a defect for this special choice of basis functions.

- A second possibility is to consider one dimensional solution under the form

$$\mathbf{v}(t,x) = \mathbf{q}(t,x)e^{\lambda x},$$

where $\mathbf{q}(t,x) \in \mathbb{R}^m$ is polynomial vector in x ant t. A concrete example is given in Section 3-1.1 of Chapter 3 for the case of the P_1 model. Then one can use the rotational invariance of the P_N model and gets the following solutions

$$\mathbf{v}(t,\mathbf{x}) = U_\theta \mathbf{q}(t, x\cos\theta + y\sin\theta)e^{\lambda(x\cos\theta + y\sin\theta)}. \tag{4.67}$$

Another possibility is to search directly for two dimensional solutions under the form

$$\mathbf{v}(t,\mathbf{x}) = \mathbf{p}(t,\mathbf{x})e^{\lambda(x\cos\theta + y\sin\theta)}, \tag{4.68}$$

where $\mathbf{p}(t,\mathbf{x}) \in \mathbb{R}^m$ is polynomial vector in x, y ant t. Note that the functions obtained with (4.68) may differ from the functions (4.67). A complete example is given in Chapter 5 for the P_1 model.

- A third possibility is to consider time dependent solutions under the form

$$\mathbf{v}(t,\mathbf{x}) = \mathbf{g}(\mathbf{x})e^{\alpha t}, \tag{4.69}$$

with $\alpha \in \mathbb{R}$. One can inject this solution in the P_N model (4.17). One gets after removing the exponentials

$$\Big(A_1\partial_x + A_2\partial_y + (R + \varepsilon\alpha I_m)\Big)\mathbf{g}(\mathbf{x}) = \mathbf{0},$$

where I_m is the identity matrix of $\mathbb{R}^{m\times m}$. The function $\mathbf{g}(\mathbf{x})$ is very similar to the stationary solutions already calculated in Sections 4-2.1 and 4-2.3. The matrix R is just replaced by the matrix $\widetilde{R} := R + \varepsilon\alpha I_m$. Note that the solutions (4.66) are included in the solutions (4.69). For simplicity, we will make the distinction and therefore assume $\mathbf{g}(\mathbf{x})$ is a non constant vectorial function.

For example, if one takes α such that $\sigma_a + \varepsilon\alpha > 0$, then $\mathbf{g}(\mathbf{x})$ is one of the exponential solutions (4.31). In particular, if $\alpha > 0$, the functions (4.69) naturally degenerate toward non trivial time dependent solutions when $\sigma_a \to 0$. This is one advantage of the functions (4.69) compare to the stationary solutions or the other time dependent solutions.

4-3 Convergence of the scheme

In this section, we study the h-convergence of the TDG method applied to the stationary P_N model. To do so, we start from Chapter 2 where the DG formulation of the method has been used to derive some estimations in various norms. To prove the convergence of the scheme, it remains to study the approximation properties of the basis functions. Usually, when considering the standard DG method, the approximation properties of simple monomials (such as 1, x, y...) can be easily studied since they appear in the Taylor expansion of every regular functions. This is not the case anymore when considering the TDG method with other kind of basis functions. Can one approximate stationary solutions to the P_N model at any order using the exponential functions of Theorem 4.25 and the special polynomial functions of Theorem 4.34 as basis functions? The answer is yes as we will see in the rest of this section. More precisely, we proceed in four steps

1. First, we show that is is enough to study the approximation properties of the m_e first components of the basis functions and that these components are solutions to a second order system (Propositions 4.51 and 4.52).
2. After that, we simplify the Taylor expansion of solutions to such second order system (Proposition 4.54).

3. Then, we use this simplified Taylor expansion to show that, studying the approximation properties of stationary solutions is in fact equivalent to study the rank of a particular matrix (Proposition 4.55).
4. Finally, we study the rank of this matrix when considering the exponential and polynomial solutions (Propositions 4.58 for the case $\sigma_a > 0$ and 4.69 for the case $\sigma_a = 0$).

Using this procedure and the estimates of Chapter 2, the Theorem 4.75 finally gives a convergence result in L^2 norm. Note however that this estimate may not be optimal when $N > 1$ (in the sense that the basis functions which give the convergence are not known) as suggested by Remark 4.76. Since it is already complicated enough to study the standard case, we do not consider asymptotic regimes in this section and take

$$\varepsilon = 1, \quad c = 1.$$

Moreover for Θ a generic open set we study the convergence of the scheme using the following norms

$$\|u\|_{W^{n,\infty}(\Theta)} = \sum_{k=0}^{n} \sum_{p=0}^{k} \sup_{\mathbf{x}\in\Theta} |\partial_x^p \partial_y^{k-p} u(\mathbf{x})|, \text{ and } \|\mathbf{u}\|_{W^{n,\infty}(\Theta)} = \sum_{j=1}^{m} \|u_j\|_{W^{n,\infty}(\Theta)}. \tag{4.70}$$

By convention we set $\|\cdot\|_{L^\infty(\Theta)} = \|\cdot\|_{W^{0,\infty}(\Theta)}$. Finally we consider

$$\sigma_t > 0,$$

since when $\sigma_t = \sigma_a + \sigma_s = 0$ the relaxation term vanishes ($R = 0$) which is of less interest for our applications.

4-3.1 A simplified Taylor expansion

In the following Ω is a bounded domain of $\mathbb{R}^2$. First, we explain why it is enough to study the approximation properties of the m_e first components of the basis functions. It comes from the block structure of the P_N model.

Proposition 4.51 ($\mathbf{u}_e$ controls $\mathbf{u}$)**.** *Assume $\sigma_t > 0$ and $\mathbf{u}(\mathbf{x}) = (\mathbf{u}_e^T(\mathbf{x}), \mathbf{u}_o^T(\mathbf{x}))^T \in \mathbb{R}^m$, $\mathbf{u}_e(\mathbf{x}) \in \mathbb{R}^{m_e}$, $\mathbf{u}_o(\mathbf{x}) \in \mathbb{R}^{m_o}$ is a stationary solution to the P_N model (4.17). One has*

$$\|\mathbf{u}\|_{L^\infty(\Omega)} \leq C\Big(\|\mathbf{u}_e\|_{L^\infty(\Omega)} + \|\partial_x \mathbf{u}_e\|_{L^\infty(\Omega)} + \|\partial_y \mathbf{u}_e\|_{L^\infty(\Omega)}\Big). \tag{4.71}$$

Proof. Since $\mathbf{u}(\mathbf{x})$ is a stationary solution to (4.17) one has

$$\Big(A_1\partial_x + A_2\partial_y\Big)\mathbf{u}(\mathbf{x}) = -R\mathbf{u}(\mathbf{x}).$$

Using the block structure (4.18)-(4.19) of the system (4.17) one gets

$$\begin{cases} \Big(A\partial_x + B\partial_y\Big)\mathbf{u}_o(\mathbf{x}) = -R_1\mathbf{u}_e(\mathbf{x}), \\ \Big(A^T\partial_x + B^T\partial_y\Big)\mathbf{u}_e(\mathbf{x}) = -R_2\mathbf{u}_o(\mathbf{x}). \end{cases} \tag{4.72}$$

Since $\sigma_t > 0$ the matrix R_2 is invertible and therefore

$$\|\mathbf{u}_o\|_{L^\infty(\Omega)} \leq C\Big(\|\partial_x \mathbf{u}_e\|_{L^\infty(\Omega)} + \|\partial_y \mathbf{u}_e\|_{L^\infty(\Omega)}\Big).$$

Because $\mathbf{u} = (\mathbf{u}_e, \mathbf{u}_o)^T$, one deduces the inequality (4.71). ∎

The function $\mathbf{u}_e(\mathbf{x})$ is solution to a second order system.

Proposition 4.52 (Second order system)**.** *Under the asserts of Proposition 4.51 and if* $\mathbf{u}$ *is regular enough so that* $\partial_{xy}\mathbf{u}(\mathbf{x}) = \partial_{yx}\mathbf{u}(\mathbf{x})$*, then* $\mathbf{u}_e(\mathbf{x})$ *is solution to the following second order system*

$$\Big(AA^T\partial_{xx} + (AB^T + BA^T)\partial_{xy} + BB^T\partial_{yy}\Big)\mathbf{u}_e(\mathbf{x}) = \sigma_t R_1 \mathbf{u}_e(\mathbf{x}). \tag{4.73}$$

Moreover

$$\partial_{yy}\mathbf{u}_e(\mathbf{x}) = (BB^T)^{-1}\Big(- AA^T\partial_{xx} - (AB^T + BA^T)\partial_{xy} + \sigma_t R_1\Big)\mathbf{u}_e(\mathbf{x}). \tag{4.74}$$

Proof. Since $\sigma_t > 0$, the matrix R_2 is invertible and $R_2^{-1} = \frac{1}{\sigma_t} I_{m_o}$. Therefore, the system (4.73) is obtained from (4.72) after eliminating $\mathbf{u}_o$. The equality (4.74) is given by the invertibility of the matrix BB^T from Proposition 4.17. The proof is complete. ∎

Now we study some properties of solution to the second order system (4.73). We recall that every vectorial function $\mathbf{w}(\mathbf{x}) \in C^{n+1}(\Omega)$ can be written under the form of a usual Taylor expansion which is a generalization of the scalar case [Fle77, Page 94]

$$\mathbf{w}(\mathbf{x}) = \sum_{k=0}^{n}\sum_{p=0}^{k} \partial_x^p \partial_y^{k-p}\mathbf{w}(\mathbf{x}_0)\gamma_k^p(\mathbf{x}) + \sum_{p=0}^{n+1} \partial_x^p \partial_y^{n+1-p}\mathbf{w}(\mathbf{x}_s)\gamma_{n+1}^p(\mathbf{x}), \tag{4.75}$$

where $\gamma_k^p(\mathbf{x}) \in \mathbb{R}$ is given by (4.39) and $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1-s)x_0 + sx$ and $y_s = (1-s)y_0 + sy$, $s \in [0,1]$. There is of course a double sum in the Taylor expansion but, for Trefftz methods, it is possible to reduce the complexity using the fact that $\mathbf{u}_e(\mathbf{x})$ is a solution to the system (4.73). This is classical [CD98, HMP16a, KMPS16] see also [IGD14, IG15a, IG15b] with a different approach to the coefficients reduction procedure. In our analysis, we use a simplification of the Taylor expansion and need the following intermediate quantities.

Definition 4.53. Consider an integer $n \geq 0$. The matrices

$$K_k^p \in \mathbb{R}^{m_e \times m_e}, \quad L_k^p \in \mathbb{R}^{m_e \times m_e},$$

are defined in the range $0 \leq p \leq k \leq n$ by a decreasing recursion from $k = n$ to $k = 0$. The recursion writes:

- by convention set $L_{n+1}^p(\mathbf{x}) = L_{n+2}^p(\mathbf{x}) = 0$, $L_k^{-1}(\mathbf{x}) = L_k^{-2}(\mathbf{x}) = 0$, $\forall p, k$
- for $k = n$ to $k = 0$, do
  - for $p = 0$ to $p = k$, do

$$K_k^p(\mathbf{x}) := \gamma_k^p(\mathbf{x}) + \sigma_t L_{k+2}^p(\mathbf{x})(BB^T)^{-1}R_1, \tag{4.76}$$

  - for $p = 0$ to $p = k-1$, do

$$L_k^p(\mathbf{x}) := K_k^p(\mathbf{x}) - L_k^{p-1}(\mathbf{x})(BB^T)^{-1}(AB^T + BA^T) - L_k^{p-2}(\mathbf{x})(BB^T)^{-1}AA^T, \tag{4.77}$$

    and

$$L_k^k(\mathbf{x}) := K_k^k(\mathbf{x}) - L_k^{k-2}(\mathbf{x})(BB^T)^{-1}AA^T, \tag{4.78}$$

Since $L_{n+1}^p(\mathbf{x}) = L_{n+2}^p(\mathbf{x}) = 0$, thus $K_{n-1}^p(\mathbf{x}) = \gamma_{n-1}^p(\mathbf{x})$, $K_n^p(\mathbf{x}) = \gamma_n^p(\mathbf{x})$. Also because $L_k^{-2} = L_k^{-1} = 0$, the equalities (4.77) and (4.78) imply

$$L_k^0 = K_k^0, \quad L_1^1 = K_1^1, \quad 0 \leq k \leq n. \tag{4.79}$$

To study the approximation properties of the basis functions, we use a simplified Taylor expansion for the solutions $\mathbf{u}_e(\mathbf{x})$ to the second order system (4.73).

Proposition 4.54 (A simplification of the Taylor expansion). *Assume* $\mathbf{u}_e(\mathbf{x}) \in C^{n+1}(\Omega)$ *is solution to* (4.73). *Then, the double sum Taylor expansion in (4.75) can be recast as a simple sum with only zero or first order derivatives with respect to* y

$$\begin{aligned}\mathbf{u}_e(\mathbf{x}) = L_0^0(\mathbf{x})\mathbf{u}_e(\mathbf{x}_0) + \sum_{k=1}^{n}\Big[L_k^k(\mathbf{x})\partial_x^k\mathbf{u}_e(\mathbf{x}_0) + L_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y\mathbf{u}_e(\mathbf{x}_0)\Big] \\ + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}\mathbf{u}_e(\mathbf{x}_s), \quad \forall \mathbf{x}\in\Omega,\end{aligned} \tag{4.80}$$

where $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1-s)x_0 + sx$ *and* $y_s = (1-s)y_0 + sy$, $s \in [0,1]$.

The proof of Proposition 4.54 is actually very similar to the proof of Proposition 4.37 which was given in the context of a second order equation. The idea is the same: use the equality (4.74) to recursively eliminate the derivatives of y. We recommend the reader to understand the Proposition 4.37 first since Proposition 4.54 presents no additional difficulties. The proof is postponed in Appendix B.

4-3.2 Approximation properties of the basis functions

Let $\mathbf{v}_i(\mathbf{x}) \in \mathbb{R}^{m_e}$, $i \in \mathbb{N}$, be solutions to the second order system (4.73). To study their approximation properties, the simplify Taylor expansion (4.80) suggests to study the matrix S_l^k defines as

$$S_l^k(\mathbf{x}_0) := \begin{pmatrix} \mathbf{v}_1(\mathbf{x}_0) & \mathbf{v}_2(\mathbf{x}_0) & \cdots & \mathbf{v}_{lm_e}(\mathbf{x}_0) \\ \partial_x\mathbf{v}_1(\mathbf{x}_0) & \partial_x\mathbf{v}_2(\mathbf{x}_0) & \cdots & \partial_x\mathbf{v}_{lm_e}(\mathbf{x}_0) \\ \partial_y\mathbf{v}_1(\mathbf{x}_0) & \partial_y\mathbf{v}_2(\mathbf{x}_0) & \cdots & \partial_y\mathbf{v}_{lm_e}(\mathbf{x}_0) \\ \partial_{xx}\mathbf{v}_1(\mathbf{x}_0) & \partial_{xx}\mathbf{v}_2(\mathbf{x}_0) & \cdots & \partial_{xx}\mathbf{v}_{lm_e}(\mathbf{x}_0) \\ \partial_x\partial_y\mathbf{v}_1(\mathbf{x}_0) & \partial_x\partial_y\mathbf{v}_2(\mathbf{x}_0) & \cdots & \partial_x\partial_y\mathbf{v}_{lm_e}(\mathbf{x}_0) \\ \vdots & \vdots & \vdots & \vdots \\ \partial_x^k\mathbf{v}_1(\mathbf{x}_0) & \partial_x^k\mathbf{v}_2(\mathbf{x}_0) & \cdots & \partial_x^k\mathbf{v}_{lm_e}(\mathbf{x}_0) \\ \partial_x^{k-1}\partial_y\mathbf{v}_1(\mathbf{x}_0) & \partial_x^{k-1}\partial_y\mathbf{v}_2(\mathbf{x}_0) & \cdots & \partial_x^{k-1}\partial_y\mathbf{v}_{lm_e}(\mathbf{x}_0) \end{pmatrix} \in \mathbb{R}^{(2k+1)m_e\times lm_e}, \tag{4.81}$$

where $\mathbf{x}_0 \in \mathbb{R}^2$, $k, l \in \mathbb{N}$. Using the matrix S_l^k one can study the approximation properties of the functions $\mathbf{v}_i$.

Proposition 4.55 (Approximation properties of the basis functions). *Let* $\mathbf{v}_1(\mathbf{x}), \mathbf{v}_2(\mathbf{x}), ..., \mathbf{v}_{lm_e}(\mathbf{x}) \in W^{k+1,\infty}(\Omega)$ *and* $\mathbf{u}_e(\mathbf{x}) \in W^{k+1,\infty}(\Omega)$ *be solutions to the second order system* (4.73). *Assume*

$$\operatorname{rank} S_l^k(\mathbf{x}_0) \geq (2k+1)m_e. \tag{4.82}$$

Then there exists real numbers $\mathbf{a} = (a_1, a_2, ..., a_{lm_e})^T \in \mathbb{R}^{lm_e}$ *and a constant* $C > 0$ *such that*

$$\Big\|\sum_{i=1}^{lm_e} a_i\mathbf{v}_i - \mathbf{u}_e\Big\|_{L^\infty(\Omega)} \leq Ch^{k+1}\|\mathbf{u}_e\|_{W^{k+1,\infty}(\Omega)}, \qquad h = \operatorname{diam}(\Omega).$$

and

$$\Big\|\nabla\Big(\sum_{i=1}^{lm_e} a_i\mathbf{v}_i - \mathbf{u}_e\Big)\Big\|_{L^\infty(\Omega)} \leq Ch^{k}\|\mathbf{u}_e\|_{W^{k+1,\infty}(\Omega)}, \qquad h = \operatorname{diam}(\Omega).$$

Proof. Let

$$\mathbf{b} = \Big(\mathbf{u}_e^T(\mathbf{x}_0), \partial_x\mathbf{u}_e^T(\mathbf{x}_0), \partial_y\mathbf{u}_e^T(\mathbf{x}_0), ..., \partial_x^k\mathbf{u}_e^T(\mathbf{x}_0), \partial_x^{k-1}\partial_y\mathbf{u}_e^T(\mathbf{x}_0)\Big)^T \in \mathbb{R}^{(2k+1)m_e}. \tag{4.83}$$

Because the solutions $\mathbf{v}_i(\mathbf{x})$, $1 \leq i \leq l$ and $\mathbf{u}_e(\mathbf{x})$ are in $W^{k+1,\infty}(\Omega)$, one can write them under the form (4.80). Consider the solution of the linear system

$$S_l^k(\mathbf{x}_0)\mathbf{a} = \mathbf{b}, \quad \mathbf{a} \in \mathbb{R}^{lm_e}, \tag{4.84}$$

which exists because $\text{rank}(S_l^k(\mathbf{x}_0)) \geq (2k+1)m_e$. The functions $\mathbf{v}_i(\mathbf{x})$ and $\mathbf{u}_e(\mathbf{x})$ both satisfy the expansion (4.80). It implies

$$\sum_{i=1}^{lm_e} a_i \mathbf{v}_i(\mathbf{x}) - \mathbf{u}_e(\mathbf{x}) = \sum_{p=0}^{k+1} \gamma_{k+1}^p(\mathbf{x}) \partial_x^p \partial_y^{k+1-p} \mathbf{v}(\mathbf{x}_s), \quad \mathbf{v}(\mathbf{x}) = \sum_{i=1}^{lm_e} a_i \mathbf{v}_i(\mathbf{x}) - \mathbf{u}_e(\mathbf{x}). \tag{4.85}$$

Since γ_{n+1}^p is a difference between $\mathbf{x}$ and $\mathbf{x}_0$ to the power $k+1$, one immediately gets

$$\left\| \sum_{i=1}^{lm_e} a_i \mathbf{v}_i - \mathbf{u}_e \right\|_{L^\infty(\Omega_j)} \leq C h^{k+1} \left\| \mathbf{v} \right\|_{W^{k+1}(\Omega_j)}.$$

Additionally, the triangular inequality yields $\|\mathbf{v}\|_{W^{k+1,\infty}(\Omega_j)} \leq \sum_{i=1}^{lm_e} |a_i| \, \|\mathbf{v}_i\|_{W^{k+1,\infty}(\Omega_j)} + \|\mathbf{u}_e\|_{W^{k+1,\infty}(\Omega_j)}$ where the coefficients a_i are bounded by $\|\mathbf{u}_e\|_{W^{k+1,\infty}(\Omega_j)}$ as a consequence of (4.84) with the definitions (4.81)-(4.83). Moreover, the basis functions $\mathbf{v}_i(\mathbf{x})$ are bounded by a constant. So $\|\mathbf{v}\|_{W^{k+1,\infty}(\Omega_j)} \leq C \, \|\mathbf{u}_e\|_{W^{n+1,\infty}(\Omega_j)}$ up to the redefinition of the constant and one gets the first inequality. The second inequality follows from (4.85). This completes the proof. ■

If the previous proposition study the approximation properties of the basis functions, it does not say anything about the linear independence of these functions. One can study the linear independence of the basis functions thanks to the matrix $S_{2k+1}^k \in \mathbb{R}^{(2k+1)m_e \times (2k+1)m_e}$.

Proposition 4.56 (Linear independence of the basis functions)**.** *Consider $(2k+1)m_e$ basis functions $\mathbf{v}_i$ solutions to the second order system (4.73) and assume the matrix S_{2k+1}^k associated is invertible. Then the solutions $\mathbf{v}_i$ are linearly independent.*

Proof. Assume

$$\sum_{i=1}^{(2k+1)m_e} a_i \mathbf{v}_i = \mathbf{0},$$

for $a_i \in \mathbb{R}$, $i = 1, ..., (2k+1)m_e$. The vector $\mathbf{0}$ is also a solution to the second order system (4.73). Therefore, one can proceed as in the proof of the previous proposition with $\mathbf{u}_e = \mathbf{0}$. In particular, the equality (4.84) reads

$$S_{2k+1}^k(\mathbf{x}_0)\mathbf{a} = \mathbf{0}.$$

From the invertibility of the matrix S_{2k+1}^k, one finds $\mathbf{a} = \mathbf{0}$ and therefore the basis functions $\mathbf{v}_i$ are linearly independent. ■

In the following, we focus on the criterion (4.82). Note however that the invertibility of the matrix S_{2k+1}^k with $2k+1$ directions is proved in Propositions 4.58 and 4.69 for the particular case $N = 1$. See also the Corollaries 4.68 and 4.71 for the general case.

4-3.2.1 Verification of the criterion (4.82) when $\sigma_a > 0$

In this section we study the approximation properties of the exponential solutions (4.31). We define $Z_i(\mathbf{x}) \in \mathbb{R}^{m_e \times m_e}$ the block matrix made of the m_e first components of the functions (4.31). That is each column of the matrix Z_i reads

$$\Big(Z_i(\mathbf{x})\Big)_{\bullet,j} = V_{2\theta_i}\mathbf{w}_j e^{\lambda_j(\mathbf{d}_i,\mathbf{x})} \in \mathbb{R}^{m_e}, \quad j = 1, ..., m_e, \tag{4.86}$$

where the notation $(Z_i)_{\bullet,j}$ denotes the column j of the matrix Z_i. The other notations come from (4.31): the vectors $\mathbf{w}_j \in \mathbb{R}^{m_e}$ are the eigenvectors of the matrix $(AA^T)^{-1}R_1$, $\lambda_j = \frac{1}{c}\sqrt{\sigma_t \mu_j}$ (μ_j are the eigenvalues of the matrix $(AA^T)^{-1}R_1$) and $V_{2\theta_i} \in \mathbb{R}^{m_e \times m_e}$ is the rotation matrix for the even moments (4.21)-(4.22). In the following we consider a matrix S_k^l made of the blocks Z_i. For simplicity, we consider centered exponentials in $\mathbf{x}_0$ and drop the dependence of S_k^l in $\mathbf{x}_0$.

In the following lemma we consider l blocks Z_i which is equivalent to consider the exponential solutions (4.31) with l directions. Of course we assume

$$\theta_i \neq \theta_j, \qquad \text{for } i \neq j.$$

Lemma 4.57. *Consider the matrix S_k^l obtained with the columns of Z_i (4.86). With l blocks Z_i, the matrix S_k^l (4.81) reads*

$$S_l^k := \begin{pmatrix} H_1 & H_2 & \cdots & H_l \\ \cos\theta_1 H_1 D & \cos\theta_2 H_2 D & \cdots & \cos\theta_l H_l D \\ \sin\theta_1 H_1 D & \sin\theta_2 H_2 D & \cdots & \sin\theta_l H_l D \\ \cos^2\theta_1 H_1 D^2 & \cos^2\theta_2 H_2 D^2 & \cdots & \cos^2\theta_l H_l D^2 \\ \cos\theta_1 \sin\theta_1 H_1 D^2 & \cos\theta_2 \sin\theta_2 H_2 D^2 & \cdots & \cos\theta_l \sin\theta_l H_l D^2 \\ \vdots & \vdots & \cdots & \vdots \\ \cos^k\theta_1 H_1 D^k & \cos^k\theta_2 H_2 D^k & \cdots & \cos^k\theta_l H_l D^k \\ \cos^{k-1}\theta_1 \sin\theta_1 H_1 D^k & \cos^{k-1}\theta_2 \sin\theta_2 H_2 D^k & \cdots & \cos^{k-1}\theta_l \sin\theta_l H_l D^k \end{pmatrix}, \tag{4.87}$$

where

$$H_i = V_{2\theta_i} H \in \mathbb{R}^{m_e \times m_e}, \quad D = \mathrm{diag}(\lambda_1, ..., \lambda_{m_e}) \in \mathbb{R}^{m_e \times m_e}, \tag{4.88}$$

and $H \in \mathbb{R}^{m_e \times m_e}$ is the matrix of the eigenvectors of $(AA^T)^{-1}R_1$.

Proof. This is equivalent to consider the matrix S_k^l (4.81) with

$$\mathbf{v}_{(i-1)m_e+j}(\mathbf{x}) = \Big(Z_i(\mathbf{x})\Big)_{\bullet,j}, \quad i = 1, ..., l, \quad j = 1, ..., m_e.$$

One concludes with the definition of the column $\Big(Z_i(\mathbf{x})\Big)_{\bullet,j}$ in (4.86). ∎

To satisfy Proposition 4.55, one must have $\mathrm{rank}\, S_l^k(\mathbf{x}_0) \geq (2k+1)m_e$ which implies $l \geq 2k+1$. Ideally, one would like to prove $\mathrm{rank}\, S_{2k+1}^k = (2k+1)m_e$, i.e. that the matrix $S_{2k+1}^k \in \mathbb{R}^{(2k+1)m_e \times (2k+1)m_e}$ is invertible. However this may be difficult to show (see Remark 4.67) and we will instead focus on the matrix $S_{2(k+N)-1}^k$.

Proposition 4.58 (Criterion (4.82) when $\sigma_a > 0$)**.** *The matrix (4.87) with $l = 2(k+N)-1$ satisfies* $\mathrm{rank}\, S_{2(k+N)-1}^k = (2k+1)m_e$.

The proof of Proposition 4.58 requires to study roots of polynomials on circles. To do so, we introduce some notations and give some technical results.

Technical material

- In the following items $l, k \in \mathbb{N}$. The space of polynomials of degree k with value in $\mathbb{R}^l$ is

$$P_k^l[x,y] := \Big\{ \mathbf{q} \ / \ \mathbf{q}(x,y) = \sum_{i+j\leq k} \boldsymbol{\eta}_{i,j} x^i y^j, \ \text{ for all } x,y, \ \boldsymbol{\eta}_{i,j} \in \mathbb{R}^l \Big\},$$

  by convention $P_k[x,y] := P_k^1[x,y]$.
- The space of polynomial matrices of degree k with value in $\mathbb{R}^{l\times l}$ is

$$P_k^{l\times l}[x,y] := \Big\{ L \ / \ L = (l_{i,j})_{i,j\leq l} \in P_k^{l\times l}[x,y], \ L(x,y) \in \mathbb{R}^{l\times l} \Big\}.$$

- The space of polynomials of total degree k and affine in y with value in $\mathbb{R}^l$ is

$$\mathcal{P}_k^l[x,y] := \Big\{ \mathbf{g} \ / \ \mathbf{g} \in P_k^l[x,y], \ \partial_y^2 \mathbf{g} = \mathbf{0} \Big\}.$$

  By convention $\mathcal{P}_k[x,y] := \mathcal{P}_k^1[x,y]$.
- The space of polynomial matrices of degree k and affine in y with value in $\mathbb{R}^{l\times l}$ is

$$\mathcal{P}_k^{l\times l}[x,y] := \Big\{ L \ / \ L = (l_{i,j})_{i,j\leq l} \in \mathcal{P}_k^{l\times l}[x,y], \ L(x,y) \in \mathbb{R}^{l\times l} \Big\}.$$

One important lemma about the space $\mathcal{P}_k[x,y]$ is the following.

Lemma 4.59 (Roots on circle in $\mathcal{P}_k[x,y]$)**.** *Assume $\lambda > 0$ and $g \in \mathcal{P}_k[x,y]$ admits $2k+1$ roots on the circle of radius λ. One has*

$$g = 0.$$

Proof. They are several ways to show that a non zero polynomial in $\mathcal{P}_k[x,y]$ has a limited number of roots on a circle. In the following we use the Bézout's theorem. The equation for the circle of radius λ reads

$$x^2 + y^2 - \lambda^2 = 0. \tag{4.89}$$

The circle equation (4.89) is absolutely irreducible (if it wasn't, it would be the union of two lines which is impossible). Therefore, the only way for (4.89) and the equation $g = 0$ to admit a common component is to write g under the form $g(x,y) = (x^2 + y^2 - \lambda^2)\widetilde{g}(x,y)$ which is impossible because $g \in \mathcal{P}_k[x,y]$.

Therefore, the circle equation (4.89) and the equation $g = 0$ with $g \in \mathcal{P}_k[x,y]$ are projective curves of degrees 2 and k respectively with no common components. From the Bézout's theorem (see [CLO08, Page 430] for example), they admit at most $2k$ common roots. This completes the proof. ∎

Now we can introduce the following notations

- The circle of radius $\lambda_i > 0$ is

$$\mathcal{C}_i := \Big\{ (x,y) \ / \ x^2 + y^2 = \lambda_i^2 \Big\}.$$

- For $g, q \in P_k$ we denote the equality of two polynomials in $\mathcal{C}_i$ as

$$g \equiv_{\mathcal{C}_i} q \quad \Leftrightarrow \quad g(x,y) - p(x,y) = 0, \quad \text{for all } (x,y) \in \mathcal{C}_i.$$

The following lemma will be useful when considering polynomials in $\mathcal{P}_k$.

Lemma 4.60. *Assume $g \in P_k$. There exists a unique $q \in \mathcal{P}_k$ such that*

$$g \equiv_{\mathcal{C}_i} q.$$

Proof. First we prove the existence. Indeed, for $g \in P_k$ one can use the equality $y^2 = \lambda_i^2 - x^2$ to remove the term depending on y with a power strictly greater than one. One gets a polynomial q in $\mathcal{P}_k$ which satisfy $g \equiv_{\mathcal{C}_i} q$.

Now we prove the uniqueness. Assume $q_1, q_2 \in \mathcal{P}_k$ and $q_1 \equiv_{\mathcal{C}_i} q_2$. Then the polynomial $q_1 - q_2$ admits an infinite number of roots on the circle $\mathcal{C}_i$. From Lemma 4.59 one deduces $q_1 - q_2 = 0$. The proof is complete. ■

- Assume $g \in P_k[x, y]$. We define the function $h_{\mathcal{C}_i}$ as

$$h_{\mathcal{C}_i} : P_k[x, y] \to \mathcal{P}_k[x, y], \quad h_{\mathcal{C}_i}(g) \equiv_{\mathcal{C}_i} g.$$

  From Lemma 4.60 The function $h_{\mathcal{C}_i}$ is well defined.
  Moreover $h_{\mathcal{C}_i}(\alpha g_1 + \beta g_2) \equiv_{\mathcal{C}_i} \alpha h_{\mathcal{C}_i}(g_1) + \beta h_{\mathcal{C}_i}(g_2)$ with $h_{\mathcal{C}_i}(\alpha g_1 + \beta g_2), \alpha h_{\mathcal{C}_i}(g_1) + \beta h_{\mathcal{C}_i}(g_2) \in \mathcal{P}_k$. One concludes using Lemma 4.60 that the function $h_{\mathcal{C}_i}$ is linear

$$h_{\mathcal{C}_i}(\alpha g_1 + \beta g_2) = \alpha h_{\mathcal{C}_i}(g_1) + \beta h_{\mathcal{C}_i}(g_2), \quad g_1, g_2 \in P_k[x, y],\ \alpha, \beta \in \mathbb{R}.$$

- Assume $\mathbf{g} = (g_1, ..., g_{m_e})^T \in P_k^{m_e}[x, y]$. We define the function $h_{\mathcal{C}}$ (with no index) as

$$h_{\mathcal{C}} : P_k^{m_e}[x, y] \to \mathcal{P}_k^{m_e}[x, y], \quad h_{\mathcal{C}}(\mathbf{g}) = \Big(h_{\mathcal{C}_1}(g_1), ..., h_{\mathcal{C}_{m_e}}(g_{m_e})\Big)^T.$$

  The function $h_{\mathcal{C}}$ is also linear

$$h_{\mathcal{C}}(\alpha \mathbf{g}_1 + \beta \mathbf{g}_2) = \alpha h_{\mathcal{C}}(\mathbf{g}_1) + \beta h_{\mathcal{C}}(\mathbf{g}_2), \quad \mathbf{g}_1, \mathbf{g}_2 \in P_k^{m_e}[x, y],\ \alpha, \beta \in \mathbb{R}.$$

Example 4.61. We give a practical example of the functions $h_{\mathcal{C}_i}$ and $h_{\mathcal{C}}$. Let

$$g_1 = 1 + 2x + y + 3y^2 \in P_4[x, y],$$
$$g_2 = x^2 + y^2 + y^4 \in P_4[x, y].$$

Using the equality $y^2 = \lambda_j^2 - x^2$ for $j = 1, 2$, one eliminates y^2, y^4 and gets

$$h_{\mathcal{C}_1}(g_1) = 1 + 3\lambda_1^2 + 2x - 3x^2 + y \in \mathcal{P}_4[x, y],$$
$$h_{\mathcal{C}_2}(g_2) = \lambda_2^2 + \lambda_2^4 - 2\lambda_2^2 x^2 + x^4 \in \mathcal{P}_4[x, y].$$

Now assume $m_e = 2$ and consider the following polynomial

$$\mathbf{g} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = \begin{pmatrix} 1 + 2x + y + 3y^2 \\ x^2 + y^2 + y^4 \end{pmatrix} \in P_4^2[x, y].$$

One has

$$h_{\mathcal{C}}(\mathbf{g}) = \begin{pmatrix} h_{\mathcal{C}_1}(g_1) \\ h_{\mathcal{C}_2}(g_2) \end{pmatrix} = \begin{pmatrix} 1 + 3\lambda_1^2 + 2x - 3x^2 + y \\ \lambda_2^2 + \lambda_2^4 - 2\lambda_2^2 x^2 + x^4 \end{pmatrix} \in \mathcal{P}_4^2[x, y].$$

●

- Assume $r = \sqrt{x^2 + y^2} \in \mathbb{R}^+$, $\theta \in [0, 2\pi[$. The matrix $\mathcal{L}_N(r, \theta)$ written in Polar coordinate is defined as

$$\mathcal{L}_N(r, \theta) := r^{N-1} H^T V_{2\theta}^T \in \mathbb{R}^{m_e \times m_e}, \tag{4.90}$$

where $H \in \mathbb{R}^{m_e \times m_e}$ is the matrix of the eigenvectors of $(AA^T)^{-1}R_1$.

Example 4.62. For the P_3 model the matrix $V_{2\theta}$ reads

$$V_{2\theta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 2\theta & 0 & -\sin 2\theta \\ 0 & 0 & 1 & 0 \\ 0 & \sin 2\theta & 0 & \cos 2\theta \end{pmatrix},$$

and using $(x, y) = (r\cos\theta, r\sin\theta)$ one has

$$r^2 V_{2\theta} = \begin{pmatrix} x^2+y^2 & 0 & 0 & 0 \\ 0 & x^2-y^2 & 0 & -2xy \\ 0 & 0 & x^2+y^2 & 0 \\ 0 & 2xy & 0 & x^2-y^2 \end{pmatrix} \in P_2^{4\times 4}[x, y].$$

In particular, the matrix $\mathcal{L}_N$ given in (4.90) is a polynomial matrix of degree 2 for the case $N = 3$. This result can be generalize for the P_N model as stated by the following lemma. ●

Lemma 4.63. *The coefficients of the matrix $\mathcal{L}_N$ are polynomials of degree $N-1$, that is*

$$\mathcal{L}_N \in P_{N-1}^{m_e \times m_e}[x, y].$$

Proof. Indeed the unitary matrix $V_{2\theta}$ in (4.21) depends only on even angles $2l\theta$ with $2l \leq N-1$. Therefore, the coefficients of the matrix $\mathcal{L}_{N-1}$ read

$$r^{N-1}\cos 2l\theta = r^{N-1-2l}r^{2l}\cos 2l\theta \quad \text{or} \quad r^{N-1}\sin 2l\theta = r^{N-1-2l}r^{2l}\sin 2l\theta.$$

Since N is odd, $N-1-2l$ is even and therefore

$$r^{N-1-2l} := (x^2+y^2)^{(N-1-2l)/2},$$

is a polynomial of degree $N-1-2l$. Moreover developing $\cos 2l\theta$, $\sin 2l\theta$ in terms of $\cos\theta$, $\sin\theta$ and using $x = r\cos\theta$, $y = r\sin\theta$, one deduces that $r^{2l}\cos 2l\theta$ and $r^{2l}\sin 2l\theta$ are polynomials of degree $2l$. Note that the coefficients of the polynomials r^{N-1-2l}, $r^{2l}\cos 2l\theta$ and $r^{2l}\sin 2l\theta$ do not depend on θ. Therefore, each coefficient of the matrix $\mathcal{L}_N$ is a polynomial of degree $N-1-2l+2l = N-1$ and the coefficients of this polynomial do not depend on θ. ■

The following lemma will be useful.

Lemma 4.64. *Assume $q \in P_j[x, y]$, $\mathbf{g} \in P_k^{m_e}[x, y]$. Then one has*

$$h_{\mathcal{C}}(q\mathbf{g}) = h_{\mathcal{C}}\Big(Qh_{\mathcal{C}}(\mathbf{g})\Big),$$

where $Q = \operatorname{diag}\big(h_{\mathcal{C}_1}(q), ..., h_{\mathcal{C}_{m_e}}(q)\big) \in \mathcal{P}_j^{m_e \times m_e}[x, y]$.

Proof. Consider the vectors $\mathbf{r} := h_{\mathcal{C}}(q\mathbf{g}) \in \mathcal{P}_{k+j}^{m_e}$, $\mathbf{s} := h_{\mathcal{C}}\Big(Qh_{\mathcal{C}}(\mathbf{g})\Big) \in \mathcal{P}_{k+j}^{m_e}$. We denote $\mathbf{r} = (r_1, ..., r_{m_e})^T$, $\mathbf{s} = (s_1, ..., s_{m_e})^T$, $\mathbf{g} = (g_1, ..., g_{m_e})^T$. From the definition of $\mathbf{r}$ and $\mathbf{s}$ one gets $r_i = h_{\mathcal{C}_i}(qg_i)$, $s_i = h_{\mathcal{C}_i}\Big(h_{\mathcal{C}_i}(q)h_{\mathcal{C}_i}(g_i)\Big)$. From the definition of $h_{\mathcal{C}_i}$ one finds $r_i \equiv_{\mathcal{C}_i} s_i$, $1 \leq i \leq m_e$. Since $r_i, s_i \in \mathcal{P}_{k+j}$, one concludes $r_i = s_i$ with Lemma 4.60. Therefore $\mathbf{r} = \mathbf{s}$ and the proof is complete. ■

In the next lemma, we study the particular structure of the polynomial $h_{\mathcal{C}}(\mathcal{L}_N\mathbf{g})$ for $\mathbf{g} \in \mathcal{P}_k^{m_e}[x,y]$. More precisely, we study the decomposition of $h_{\mathcal{C}}(\mathcal{L}_N\mathbf{g})$ in two terms: a term which depends only on $\mathbf{g}(x,y)$ and a term which depends on $x\mathbf{g}(x,y)$.

Lemma 4.65. *Assume $\mathbf{g} \in \mathcal{P}_k^{m_e}[x,y]$. One has the equality*

$$h_{\mathcal{C}}\Big(\mathcal{L}_N(x,y)\mathbf{g}(x,y)\Big) = H^T\Big(D^{N-1}M\mathbf{g}(x,y) + h_{\mathcal{C}}\big(xV(x,y)\mathbf{g}(x,y)\big)\Big), \quad \text{for all } x,y,$$

where $M \in \mathbb{R}^{m_e\times m_e}$ is a diagonal matrix with each diagonal element equal to 1 or -1, D is defined in (4.88) and $V \in P_{N-1}^{m_e\times m_e}[x,y]$ is a polynomial matrix.

Proof. We recall the formulas

$$\cos 2k\theta = \Re(\cos\theta + i\sin\theta)^{2k}, \quad \sin 2k\theta = \Im(\cos\theta + i\sin\theta)^{2k}.$$

Since N is odd one deduces

$$r^{N-1}\cos 2k\theta = (-1)^k y^{N-1} + xw_1(x,y), \quad r^{N-1}\sin 2k\theta = xw_2(x,y),$$

where $w_1, w_2 \in P_{N-1}[x,y]$. From (4.21) the matrix $V_{2\theta}$ has only coefficients of the form $\cos 2k\theta$ on its diagonal and coefficients of the form or $\sin 2k\theta$ on its anti diagonal. Therefore

$$r^{N-1}V_{2\theta} = My^{N-1} + xW(x,y),$$

where $M \in \mathbb{R}^{m_e\times m_e}$ is a diagonal matrix with each diagonal element equal to 1 or -1 and $W \in P_{N-1}^{m_e\times m_e}[x,y]$ is a polynomial matrix. One deduces that the matrix $\mathcal{L}_N$ (4.90) can be written

$$\mathcal{L}_N(x,y) = H^T\Big(My^{N-1} + xW(x,y)\Big), \quad \text{for all } x,y.$$

Therefore, since the function $h_{\mathcal{C}}$ is linear one gets

$$h_{\mathcal{C}}\Big(\mathcal{L}_N(x,y)\mathbf{g}(x,y)\Big) = H^T M h_{\mathcal{C}}\Big(y^{N-1}\mathbf{g}(x,y)\Big) + H^T h_{\mathcal{C}}\Big(xW(x,y)\mathbf{g}(x,y)\Big), \quad \text{for all } x,y. \tag{4.91}$$

We are interested in the terms which depend on x only through the polynomial $\mathbf{g}(x,y)$. Therefore, we focus on the first term $h_{\mathcal{C}}(y^{N-1}\mathbf{g}(x,y))$. Using Lemma 4.64 one has

$$h_{\mathcal{C}}\Big(y^{N-1}\mathbf{g}(x,y)\Big) = h_{\mathcal{C}}\Big(\begin{pmatrix} h_{\mathcal{C}_1}(y^{N-1}) & & 0 \\ & \ddots & \\ 0 & & h_{\mathcal{C}_{m_e}}(y^{N-1}) \end{pmatrix} h_{\mathcal{C}}(\mathbf{g}(x,y))\Big). \tag{4.92}$$

With the equality $y^{N-1} = (\lambda_i^2 - x^2)^{\frac{N-1}{2}}$, $\mathbf{g} \in \mathcal{P}_k^{m_e}$ and the linearity of $h_{\mathcal{C}_i}$ one finds

$$h_{\mathcal{C}_i}(y^{N-1}) = \lambda_i^{N-1} + h_{\mathcal{C}_i}\Big(x\widetilde{w}_i(x,y)\Big), \quad h_{\mathcal{C}}(\mathbf{g}) = \mathbf{g}, \tag{4.93}$$

where $\widetilde{w}_i \in P_{N-1}[x,y]$ are polynomials. Therefore, using (4.93) in (4.92) and the linearity of $h_{\mathcal{C}}$ one gets

$$h_{\mathcal{C}}\Big(y^{N-1}\mathbf{g}(x,y)\Big) = D^{N-1}\mathbf{g}(x,y) + h_{\mathcal{C}}\Big(x\widetilde{W}(x,y)\mathbf{g}(x,y)\Big), \quad \text{for all } x,y. \tag{4.94}$$

where $D = \operatorname{diag}(\lambda_1, ..., \lambda_{m_e})$ and $\widetilde{W} \in P_{N-1}^{m_e\times m_e}[x,y]$ is a polynomial matrix. Using (4.94) in (4.91) to replace the first term yields

$$h_{\mathcal{C}}\Big(\mathcal{L}_N(x,y)\mathbf{g}(x,y)\Big) = H^T\Big(MD^{N-1}\mathbf{g}(x,y) + h_{\mathcal{C}}\big(xV(x,y)\mathbf{g}(x,y)\big)\Big), \quad \text{for all } x,y,$$

where $V = W + \widetilde{W} \in P_{N-1}^{m_e\times m_e}[x,y]$. The proof is complete. ∎

The idea behind Lemma 4.65 is that the polynomial $h_{\mathcal{C}}(\mathcal{L}_N\mathbf{g})$ can be written as a term which depends only on $\mathbf{g}(x,y)$ and a term which depends on $x\mathbf{g}(x,y)$. That is

$$h_{\mathcal{C}}\Big(\mathcal{L}_N(x,y)\mathbf{g}(x,y)\Big) = K\mathbf{g}(x,y) + h_{\mathcal{C}}\Big(xV(x,y)\mathbf{g}(x,y)\Big),$$

with a matrix K invertible. This particular structure combines with the invertibility of the matrix K play an important role in the following Lemma.

Lemma 4.66. *Assume $k,j \in \mathbb{N}$, $\mathbf{g} \in \mathcal{P}_k^{m_e}[x,y]$, $V \in P_j^{m_e\times m_e}[x,y]$ and the matrix $K \in \mathbb{R}^{m_e\times m_e}$ is invertible. Let*

$$\mathbf{q}(x,y) = K\mathbf{g}(x,y) + h_{\mathcal{C}}\Big(xV(x,y)\mathbf{g}(x,y)\Big), \quad \textit{for all } x,y. \tag{4.95}$$

Then one has

$$\mathbf{q} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{g} = \mathbf{0}.$$

Proof. The case $\mathbf{g} = \mathbf{0} \Rightarrow \mathbf{q} = \mathbf{0}$ is straightforward.

We prove the other case by contradiction. If $V = 0$ one immediately deduces the result so we take $V \neq 0$. Assume $\mathbf{q} = \mathbf{0}$ and $\mathbf{g} \neq \mathbf{0}$. We write $\mathbf{g}(x,y)$ as a polynomial in x with coefficients depending on y

$$\mathbf{g}(x,y) = \sum_{i=0}^{k} \boldsymbol{\alpha}_i(y)x^i, \quad \text{for all } x,y. \tag{4.96}$$

We write $h_{\mathcal{C}}\Big(xV(x,y)\mathbf{g}(x,y)\Big)$ in the same way

$$h_{\mathcal{C}}\Big(xV(x,y)\mathbf{g}(x,y)\Big) = \sum_{i=0}^{k} \boldsymbol{\beta}_i(y)x^i, \quad \text{for all } x,y. \tag{4.97}$$

Since $\mathbf{g} \neq \mathbf{0}$ there exists $a,b \in \mathbb{N}$ such that

$$a = \min_{\boldsymbol{\alpha}_i \neq \mathbf{0}} i, \quad b = \min_{\boldsymbol{\beta}_i \neq \mathbf{0}} i.$$

For a given coefficient, the function $h_{\mathcal{C}} : P_k^{m_e}[x,y] \to \mathcal{P}_k^{m_e}[x,y]$ do not decrease the power in x (it can only decrease the power in y). Therefore, using the definition of the coefficients $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ in (4.96) and (4.97) one deduces

$$b > a.$$

Now we consider the coefficients associated only with the power x^a in (4.95). Using $\mathbf{q} = \mathbf{0}$ one gets

$$K\boldsymbol{\alpha}_a = \mathbf{0}.$$

Since the matrix K is invertible this is a contradiction with the assumption $\boldsymbol{\alpha}_a \neq \mathbf{0}$. The proof is complete. ∎

Proof of Proposition 4.58.

Proof of Proposition 4.58. To denote the transpose matrix of S_l^k, we will use the notation $S_l^{k,T} := (S_l^k)^T$. The goal is to show that

$$\dim(\ker S_{2(k+N)-1}^{k,T}) = 0.$$

Because of the rank-nullity theorem this implies $\operatorname{rank} S^{k,T}_{2(k+N)-1} = (2k+1)m_e$ and one will conclude

$$\operatorname{rank} S^{k}_{2(k+N)-1} = (2k+1)m_e.$$

From the definition (4.87) of the matrix S^k_l one deduces that the matrix $S^{k,T}_{2(k+N)-1} \in \mathbb{R}^{lm_e \times (2k+1)m_e}$ reads

$$S^{k,T}_{2(k+N)-1} = \begin{pmatrix} H_1^T & \cos\theta_1 DH_1^T & \cdots & \sin\theta_1 \cos^{k-1}\theta_1 D^k H_1^T \\ H_2^T & \cos\theta_2 DH_2^T & \cdots & \sin\theta_2 \cos^{k-1}\theta_2 D^k H_2^T \\ \vdots & \vdots & \cdots & \vdots \\ H^T_{2(k+N)-1} & \cos\theta_{2(k+N)-1} DH^T_{2(k+N)-1} & \cdots & \sin\theta_{2(k+N)-1}\cos^{k-1}\theta_{2(k+N)-1} D^k H^T_{2(k+N)-1} \end{pmatrix}.$$

Let

$$\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T, ..., \mathbf{u}_{2k+1}^T)^T \in \mathbb{R}^{(2k+1)m_e}, \quad \mathbf{u}_i \in \mathbb{R}^{m_e},$$

and assume

$$\mathbf{u} \in \ker S^{k,T}_{2(k+N)-1}.$$

The equality $S^{k,T}_{2(k+N)-1}\mathbf{u} = \mathbf{0}$ gives

$$\sum_{l=0}^{k} \cos^l\theta_i D^l H_i \mathbf{u}_{2l+1} + \sin\theta_i \sum_{l=0}^{k-1} \cos^l\theta_i D^{l+1} H_i \mathbf{u}_{2(l+1)} = \mathbf{0},$$

for $i = 1, ..., 2(k+N)-1$. Multiplying by D^{N-1} one gets

$$\sum_{l=0}^{k} \cos^l\theta_i D^{l+N-1} H_i \mathbf{u}_{2l+1} + \sum_{l=0}^{k-1} \sin\theta_i \cos^l\theta_i D^{l+N} H_i \mathbf{u}_{2(l+1)} = \mathbf{0}, \tag{4.98}$$

for $i = 1, ..., 2(k+N)-1$. The equalities (4.98) can be interpreted as the equations of the roots of some polynomials. Indeed let

$$\begin{gathered} \mathbf{g} \in \mathcal{P}_k^{m_e}[x,y], \\ \mathbf{g}(x,y) := \sum_{l=0}^{k} x^l \mathbf{u}_{2l+1} + y \sum_{l=0}^{k-1} x^l \mathbf{u}_{2(l+1)}, \quad \text{for all } x, y, \end{gathered} \tag{4.99}$$

where $\mathbf{u}_1, ..., \mathbf{u}_{2k+1} \in \mathbb{R}^{m_e}$ satisfy (4.98). We define the polynomial vector $\mathbf{f}$ as

$$\mathbf{f} := \mathcal{L}_N \mathbf{g} \in P^{m_e}_{k+N-1}[x,y], \tag{4.100}$$

where $\mathcal{L}_N$ is defined in (4.90) and we recall that from Lemma 4.63 one has $\mathcal{L}_N \in P^{m_e \times m_e}_{N-1}[x,y]$. We denote $\mathbf{f} = (f_1, ..., f_{m_e})^T$ and claim that the equations (4.98) give some roots of the components f_i. Consider the point $(x,y) = (\lambda_j \cos\theta_i, \lambda_j \sin\theta_i)$ one has

$$\begin{aligned} \mathbf{f}(x,y) &= \sum_{l=0}^{k} x^l \mathcal{L}_N(x,y)\mathbf{u}_{2j+1} + y \sum_{l=0}^{k-1} x^l \mathcal{L}_N(x,y)\mathbf{u}_{2(j+1)}, \\ \mathbf{f}(\lambda_j \cos\theta_i, \lambda_j \sin\theta_i) &= \sum_{l=0}^{k} \lambda_j^l \cos^l\theta_i \mathcal{L}_N(\lambda_j\cos\theta_i, \lambda_j \sin\theta_i)\mathbf{u}_{2l+1} \\ &\quad + \sum_{l=0}^{k-1} \lambda_j^{l+1} \sin\theta_i \cos^l\theta_i \mathcal{L}_N(\lambda_j\cos\theta_i, \lambda_j\sin\theta_i)\mathbf{u}_{2(l+1)}, \end{aligned}$$

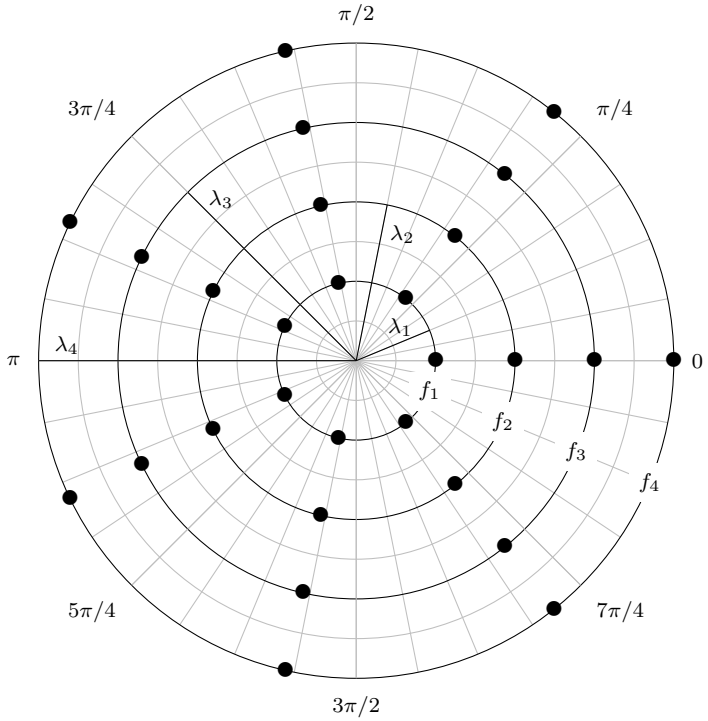

Figure 4.4 – Representation of the roots of the components f_i. Here we consider $N = 3$, $k = 1$ and directions with angles $\frac{2\pi j}{7}$, $j = 0, ..., 6$.

From (4.88) one has $H_i^T = H^T V_{2\theta_i}^T$ and therefore

$$\mathcal{L}_N(\lambda_j \cos\theta_i, \lambda_j \sin\theta_i) = \lambda_j^{N-1} H_i^T. \tag{4.101}$$

Using (4.101) to reformulate the matrix $\mathcal{L}_N(\lambda_j \cos\theta_i, \lambda_j \sin\theta_i)$ one gets

$$\mathbf{f}(\lambda_j \cos\theta_i, \lambda_j \sin\theta_i) = \sum_{l=0}^{k} \cos^l\theta_i \lambda_j^{l+N-1} H_i^T \mathbf{u}_{2l+1} + \sum_{l=0}^{k-1} \sin\theta_i \cos^l\theta_i \lambda_j^{l+N} H_i^T \mathbf{u}_{2l+1}.$$

Using the definition of the matrix D in (4.88), one deduces that $f_j(\lambda_j \cos\theta_i, \lambda_j \sin\theta_i)$ and the component j on the left hand side of (4.98) coincident. Therefore

$$f_j(\lambda_j \cos\theta_i, \lambda_j \sin\theta_i) = 0, \quad 1 \le i \le 2(k+N)-1, \ 1 \le j \le m_e. \tag{4.102}$$

The equality (4.102) can be reformulate as follow: each component f_j of the polynomial vector $\mathbf{f}$ admits $2(k+N)-1$ distinct roots on the circle of radius λ_j see Figure 4.4 for an example in the case $N = 3$. We denote

$$\mathbf{f}_{\mathcal{C}} := h_{\mathcal{C}}(\mathbf{f}) \in \mathcal{P}_{k+N-1}^{m_e}[x, y].$$

From (4.102) each component $f_{\mathcal{C},j} \in \mathcal{P}_{k+N-1}[x, y]$ of $\mathbf{f}_{\mathcal{C}}$ admits $2k+N-1$ roots on the circle of radius λ_j. Using Lemma 4.59 one deduces

$$\mathbf{f}_{\mathcal{C}} = \mathbf{0}.$$

Now we want to prove that it implies $\mathbf{g} = \mathbf{0}$. From Lemma 4.65 one has

$$\mathbf{f}_{\mathcal{C}}(x, y) = H^T \Big(M D^{N-1} \mathbf{g}(x, y) + h_{\mathcal{C}}\big(xV(x, y)\mathbf{g}(x, y)\big) \Big), \quad \text{for all } x, y.$$

Since the matrix H^T is invertible from Proposition 4.22, one deduces

$$M D^{N-1} \mathbf{g}(x, y) + h_{\mathcal{C}}\Big(xV(x, y)\mathbf{g}(x, y)\Big) = \mathbf{0}, \quad \text{for all } x, y.$$

Using Lemma 4.66 with the invertibility of the matrix MD^{N-1} one finally finds

$$\mathbf{g} = \mathbf{0}.$$

Since the coefficients of the polynomial $\mathbf{g}$ are made of the components $\mathbf{u}$, one gets $\mathbf{u} = \mathbf{0}$ and therefore $\dim(\ker S^{k,T}_{2(k+N)-1}) = 0$. This completes the proof. ■

Remark 4.67 (Invertibility of the matrix S^k_{2k+1})**.** The beginning of the proof of Proposition 4.58 is also true when studying the matrix S^k_{2k+1}. However, with the matrix S^k_{2k+1} each component f_i will only have $2k+1$ roots on the circle of radius λ_i. Therefore, one can not used (at least directly) the Bézout's theorem. Some more advanced tools in algebraic geometry may be needed to prove that the matrix S^k_{2k+1} is, or is not, invertible [CLO08]. ●

Given $2k+1$ directions and the solutions (4.31) it is therefore not clear how to prove the invertibility of the matrix S^k_{2k+1} (4.87). However, we can give a weaker result.

Corollary 4.68. *Assume $\sigma_a > 0$ and consider the solutions (4.31) with $2(k+N)-1$ directions for a total of $(2(k+N)-1)m_e$ solutions. Among these $(2(k+N)-1)m_e$ functions there exists $(2k+1)m_e$ functions such that the matrix S^k_{2k+1} (4.81) is invertible.*

Proof. The proof is straightforward. From Proposition 4.58 one has $\operatorname{rank} S^k_{2(k+N)-1} = (2k+1)m_e$ and therefore, one can extract $(2k+1)m_e$ columns from the matrix $S^k_{2(k+N)-1}$ such that the associated matrix S^k_{2k+1} satisfies $\operatorname{rank} S^k_{2k+1} = (2k+1)m_e$. Since $S^k_{2k+1} \in \mathbb{R}^{(2k+1)m_e \times (2k+1)m_e}$, this matrix is invertible. The proof is complete. ■

The main defect of Corollary 4.68 is that we do not know which basis functions give the invertibility of the matrix S^k_{2k+1}.

4-3.2.2 Verification of the criterion (4.82) when $\sigma_a = 0$

In this section, we study the approximation properties of the exponential solutions (4.31) combined with the polynomial solutions constructed in Section 4-2.3 when $\sigma_a = 0$. From Proposition 4.21, there exists a unique λ_j such that $\lambda_j \to 0$ when $\sigma_a \to 0$. For convenience, we take $j = 1$ and therefore one has $\lambda_1 \to 0$ when $\sigma_a \to 0$. Since we show in Section 4-2.3 that the degenerative exponentials tend toward a family of polynomials when $\sigma_a \to 0$, we simply replace the degenerative exponentials with the polynomial solutions. Therefore we now consider the following matrices Z_i

$$\Big(Z_i(\mathbf{x})\Big)_{\bullet,1} = \mathbf{p}_i(\mathbf{x}) \in \mathbb{R}^{m_e}, \quad \Big(Z_i(\mathbf{x})\Big)_{\bullet,j} = V_{2\theta_i}\mathbf{w}_j e^{\lambda_j(\mathbf{d}_i,\mathbf{x})} \in \mathbb{R}^{m_e},\ j = 2, ..., m_e, \tag{4.103}$$

with $\lambda_j \neq 0$ and $\mathbf{p}_i(\mathbf{x}) \in \mathbb{R}^{m_e}$ represents the m_e first components of the polynomials given by the Theorem 4.34 of Section 4-2.3. The matrix S^k_l of the non degenerative exponential solutions with l directions and the first l polynomial solutions is

$$S^k_l(\mathbf{x}_0) := \begin{pmatrix} \mathbf{p}_1(\mathbf{x}_0) & H_1 & \cdots & \mathbf{p}_l(\mathbf{x}_0) & H_l \\ \partial_x\mathbf{p}_1(\mathbf{x}_0) & \cos\theta_1 H_1 D & \cdots & \partial_x\mathbf{p}_l(\mathbf{x}_0) & \cos\theta_l H_l D \\ \partial_y\mathbf{p}_1(\mathbf{x}_0) & \sin\theta_1 H_1 D & \cdots & \partial_y\mathbf{p}_l(\mathbf{x}_0) & \sin\theta_l H_l D \\ \partial^2_x\mathbf{p}_1(\mathbf{x}_0) & \cos^2\theta_1 H_1 D^2 & \cdots & \partial^2_x\mathbf{p}_l(\mathbf{x}_0) & \cos^2\theta_l H_l D^2 \\ \partial_x\partial_y\mathbf{p}_1(\mathbf{x}_0) & \cos\theta_1\sin\theta_1 H_1 D^2 & \cdots & \partial_x\partial_y\mathbf{p}_l(\mathbf{x}_0) & \cos\theta_l\sin\theta_l H_l D^2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \partial^k_x\mathbf{p}_1(\mathbf{x}_0) & \cos^k\theta_1 H_1 D^k & \cdots & \partial^k_x\mathbf{p}_l(\mathbf{x}_0) & \cos^k\theta_l H_l D^{k+N} \\ \partial^{k-1}_x\partial_y\mathbf{p}_1(\mathbf{x}_0) & \cos^{k-1}\theta_1\sin\theta_1 H_1 D^k & \cdots & \partial^{k-1}_x\partial_y\mathbf{p}_l(\mathbf{x}_0) & \cos^{k-1}\theta_l\sin\theta_l H_l D^k \end{pmatrix}. \tag{4.104}$$

Since we remove the degenerative exponentials (that is the eigenvalue which degenerate to zero) in the definitions of the matrices D and H_i, one has

$$D = \mathrm{diag}(\lambda_2, ..., \lambda_{m_e}) \in \mathbb{R}^{(m_e-1)\times(m_e-1)},$$

and

$$H = \begin{pmatrix} \mathbf{h}^T \\ J \end{pmatrix} \in \mathbb{R}^{m_e\times(m_e-1)}. \tag{4.105}$$

Where the matrix $J \in \mathbb{R}^{m_e-1\times m_e-1}$ is the same matrix as in Definition 4.24. Therefore, the vector $\mathbf{h} \in \mathbb{R}^{m_e-1}$ is simply the vector of the first component of the eigenvectors of the matrix $(AA^T)^{-1}R_1$ associated with a non zero eigenvalue. We recall that Proposition 4.23 implies that the matrix J is invertible. The matrices H_i and $\mathcal{L}_N$ are defined in the same way as in (4.88) and (4.90) respectively. Therefore

$$H_i \in \mathbb{R}^{m_e\times(m_e-1)}, \quad \mathcal{L}_N \in \mathbb{P}_{N-1}^{m_e-1\times m_e}[x, y].$$

The results given in the previous section holds when $\sigma_a = 0$.

Proposition 4.69 (Criterion (4.82) when $\sigma_a = 0$)**.** *The matrix* (4.104) *with* $l = 2(k + N) - 1$ *satisfies* $\mathrm{rank}\, S^k_{2(k+N)-1} = (2k + 1)m_e$.

We will use the following lemma.

Lemma 4.70. *We denote* $q_i(\mathbf{x})$ *the harmonic polynomials* (4.50)*. For* $n \geq 1$ *one has*

$$\partial_x^n q_{2n}(\mathbf{x}) = 1, \quad \partial_x^{n+1+l} q_{2n}(\mathbf{x}) = 0, \quad \partial_x^{n-1+l}\partial_y q_{2n}(\mathbf{x}) = 0, \quad \text{for all } l \in \mathbb{N},$$

and

$$\partial_x^{n-1}\partial_y q_{2n+1}(\mathbf{x}) = 1, \quad \partial_x^{n+l} q_{2n+1}(\mathbf{x}) = 0, \quad \partial_x^{n+l}\partial_y q_{2n+1}(\mathbf{x}) = 0, \quad \text{for all } l \in \mathbb{N}.$$

Proof. One has $(x + iy)^n = \sum_{p=0}^n C_n^p (i)^{n-p} x^p y^{n-p}$, thus

$$\partial_x^n \Re(x + iy)^n = n!, \quad \partial_x^{n+1+l}\Re(x + iy)^n = 0, \quad \partial_x^{n-1+l}\partial_y \Re(x + iy)^n = 0, \quad \text{for all } l \in \mathbb{N},$$

and

$$\partial_x^{n-1}\partial_y \Im(x + iy)^n = C_n^1 (n - 1)!, \quad \partial_x^{n+l}\Im(x + iy)^n = 0, \quad \partial_x^{n+l}\partial_y \Im(x + iy)^n = 0, \quad \text{for all } l \in \mathbb{N}.$$

One concludes with the definition of the harmonic polynomials (4.50). ∎

We can now prove Proposition 4.69.

Proof of Proposition 4.69. We start by proceeding as in the proof of Proposition 4.58. The matrix $S^{k,T}_{2(k+N)-1} := (S^k_{2(k+N)-1})^T$ reads

$$S^{k,T}_{2(k+N)-1}(\mathbf{x}_0) = \begin{pmatrix} \mathbf{p}_1^T(\mathbf{x}_0) & \partial_x \mathbf{p}_1^T(\mathbf{x}_0) & \cdots & \partial_x^{k-1}\partial_y \mathbf{p}_1^T(\mathbf{x}_0) \\ H_1^T & \cos\theta_1 D H_1^T & \cdots & \sin\theta_1 \cos^{k-1}\theta_1 D^k H_1^T \\ \mathbf{p}_2^T(\mathbf{x}_0) & \partial_x \mathbf{p}_2^T(\mathbf{x}_0) & \cdots & \partial_x^{k-1}\partial_y \mathbf{p}_2^T(\mathbf{x}_0) \\ H_2^T & \cos\theta_2 D H_2^T & \cdots & \sin\theta_2 \cos^{k-1}\theta_2 D^k H_2^T \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{p}^T_{2(k+N)-1}(\mathbf{x}_0) & \partial_x \mathbf{p}^T_{2(k+N)-1}(\mathbf{x}_0) & \cdots & \partial_x^{k-1}\partial_y \mathbf{p}^T_{2(k+N)-1}(\mathbf{x}_0) \\ H^T_{2(k+N)-1} & \cos\theta_{2(k+N)-1} D H^T_{2(k+N)-1} & \cdots & \sin\theta_{2(k+N)-1} \cos^{k-1}\theta_{2(k+N)-1} D^k H^T_{2(k+N)-1} \end{pmatrix}.$$

We assume

$$\mathbf{u} \in \ker S^{k,T}_{2(k+N)-1},$$

and use the following notations

$$\begin{aligned} \mathbf{u} &= (\mathbf{u}_1^T, \mathbf{u}_2^T, ..., \mathbf{u}_{2k+1}^T)^T \in \mathbb{R}^{(2k+1)m_e}, \\ \mathbf{u}_i &= (v_i, \mathbf{w}_i)^T \in \mathbb{R}^{m_e}, \quad v_i \in \mathbb{R}, \quad \mathbf{w}_i \in \mathbb{R}^{m_e-1}. \end{aligned} \tag{4.106}$$

We define the matrix $\mathcal{S}_1 \in \mathbb{R}^{(m_e-1)(2(k+N)-1)\times(2k+1)m_e}$ as

$$\mathcal{S}_1 = \begin{pmatrix} H_1^T & \cos\theta_1 D H_1^T & \cdots & \sin\theta_1 \cos^{k-1}\theta_1 D^k H_1^T \\ H_2^T & \cos\theta_2 D H_2^T & \cdots & \sin\theta_2 \cos^{k-1}\theta_2 D^k H_2^T \\ \vdots & \vdots & \cdots & \vdots \\ H_{2(k+N)-1}^T & \cos\theta_{2(k+N)-1} D H_{2(k+N)-1}^T & \cdots & \sin\theta_{2(k+N)-1} \cos^{k-1}\theta_{2(k+N)-1} D^k H_{2(k+N)-1}^T \end{pmatrix},$$

and the matrix $\mathcal{S}_2 \in \mathbb{R}^{2(k+N)-1\times(2k+1)m_e}$ as

$$\mathcal{S}_2 = \begin{pmatrix} \mathbf{p}_1^T(\mathbf{x}_0) & \partial_x \mathbf{p}_1^T(\mathbf{x}_0) & \cdots & \partial_x^{k-1}\partial_y \mathbf{p}_1^T(\mathbf{x}_0) \\ \mathbf{p}_2^T(\mathbf{x}_0) & \partial_x \mathbf{p}_2^T(\mathbf{x}_0) & \cdots & \partial_x^{k-1}\partial_y \mathbf{p}_2^T(\mathbf{x}_0) \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{p}_{2(k+N)-1}^T(\mathbf{x}_0) & \partial_x \mathbf{p}_{2(k+N)-1}^T(\mathbf{x}_0) & \cdots & \partial_x^{k-1}\partial_y \mathbf{p}_{2(k+N)-1}^T(\mathbf{x}_0) \end{pmatrix}.$$

Note that

$$S^{k,T}_{2(k+N)-1}\mathbf{u} = \mathbf{0} \quad \Rightarrow \quad \mathcal{S}_1\mathbf{u} = \mathbf{0} \quad \text{and} \quad \mathcal{S}_2\mathbf{u} = \mathbf{0}.$$

As a first step, we study the equality $\mathcal{S}_1\mathbf{u} = \mathbf{0}$ and proceed as in the proof of Proposition 4.58. We define the polynomial vector $\mathbf{g}$ as

$$\begin{aligned} \mathbf{g} &\in \mathcal{P}_k^{m_e}[x,y], \\ \mathbf{g}(x,y) &:= \sum_{l=0}^{k} x^l \mathbf{u}_{2l+1} + y\sum_{l=0}^{k-1} x^l \mathbf{u}_{2(l+1)}, \quad \text{for all } x, y, \end{aligned}$$

and the polynomial vector $\mathbf{f}$ as

$$\mathbf{f} := \mathcal{L}_N \mathbf{g} \in P_{k+N-1}^{m_e-1}[x,y].$$

As in the proof of Proposition 4.58, one can show that each component f_j of the polynomial vector $\mathbf{f}$ admits $2(k+N)-1$ distinct roots on the circle of radius λ_{j+1}. Therefore denoting

$$\mathbf{f}_{\mathcal{C}} := h_{\mathcal{C}}(\mathbf{f}) \in \mathcal{P}_{k+N-1}^{m_e-1}[x,y].$$

One concludes using Lemma 4.59 that

$$\mathbf{f}_{\mathcal{C}} = \mathbf{0}.$$

And with the same reasoning as in the proof of Proposition 4.58 one can write

$$\mathbf{f}_{\mathcal{C}}(x,y) = H^T\Big(MD^{N-1}\mathbf{g}(x,y) + h_{\mathcal{C}}\big(xV(x,y)\mathbf{g}(x,y)\big)\Big) = \mathbf{0}, \quad \text{for all } x, y. \tag{4.107}$$

This is where the similarities with the previous proof end. Indeed since $H^T \in \mathbb{R}^{m_e-1\times m_e}$ one can not invert the matrix H^T as before.

Now, we use the equality $\mathcal{S}_2\mathbf{u} = \mathbf{0}$. We recall that the harmonic polynomial q_i is the first component of the polynomial vector $\mathbf{p}_i$ and from Corollary 4.49 the other components have a degree strictly less than q_i. Therefore using Lemma 4.70 one can write

$$\partial_x^n \mathbf{p}_{2n}(\mathbf{x}) = (1,0,...,0)^T, \quad \partial_x^{n+1+l}\mathbf{p}_{2n}(\mathbf{x}) = (0,...,0)^T, \quad \partial_x^{n-1+l}\partial_y \mathbf{p}_{2n}(\mathbf{x}) = (0,...,0)^T,$$

for all $l \in \mathbb{N}$ and

$$\partial_x^{n-1}\partial_y \mathbf{p}_{2n+1}(\mathbf{x}) = (1,0,...,0)^T, \quad \partial_x^{n+l}\mathbf{p}_{2n+1}(\mathbf{x}) = (0,...,0)^T, \quad \partial_x^{n+l}\partial_y \mathbf{p}_{2n+1}(\mathbf{x}) = (0,...,0)^T,$$

for all $l \in \mathbb{N}$. This implies that the matrix $\mathcal{S}_2$ can be written

$$\mathcal{S}_2 = \begin{pmatrix} (1,0,...,0)^T & & 0 \\ & (1,0,...,0)^T & & \ddots \\ & * & & & (1,0,...,0)^T \end{pmatrix}, \tag{4.108}$$

where $(1,0,...,0)^T \in \mathbb{R}^{m_e}$.

To show $\mathbf{u} = \mathbf{0}$ we proceed by recurrence. First we show $\mathbf{u}_1 = \mathbf{0}$. We recall $\mathbf{u}_1 := (v_1, \mathbf{w}_1)^T$ and proceed in two steps

- Since $\mathcal{S}_2\mathbf{u} = \mathbf{0}$ one immediately deduces from the particular structure (4.108) of the matrix $\mathcal{S}_2$ that
$$v_1 = 0.$$
- Now we consider the coefficients which do not depend on x in (4.107). One gets
$$H^T M D^{N-1}\Big(\mathbf{u}_1 + y\mathbf{u}_2\Big) = \mathbf{0}, \quad \text{for all } x, y.$$
In particular, one has
$$H^T M D^{N-1}\mathbf{u}_1 = \mathbf{0}.$$

Since $v_1 = 0$ one finds using the particular structure (4.105) of the matrix H

$$J^T M D^{N-1}\mathbf{w}_1 = \mathbf{0}.$$

Since the matrix J is invertible, the matrix $J^T D^2 M$ is also invertible and one deduces $\mathbf{w}_1 = \mathbf{0}$. One finds $\mathbf{u}_1 = \mathbf{0}$.

Now assume there exists $j \geq 2$ such that

$$\mathbf{u}_i = \mathbf{0}, \quad \text{for all } i < j.$$

Our goal is to show that $\mathbf{u}_j = \mathbf{0}$. We proceed in two steps

- Since $\mathbf{u}_i = \mathbf{0}$ for all $i < j$ and using the particular structure (4.108) of the matrix $\mathcal{S}_2$, the equality $\mathcal{S}_2\mathbf{u} = 0$ yields
$$v_j = 0.$$
- We denote $j = 2n+1$ if j is odd and $j = 2(n+1)$ if j is even. Since $\mathbf{u}_i = \mathbf{0}$ for all $i < j$, all the coefficients of $\mathbf{g}(x,y)$ associated with a power x^k with $k < n$ are equals to zero. Therefore using the same idea as in the proof of Lemma 4.66, the equality (4.107) implies
$$H^T M D^{N-1}\mathbf{u}_j x^n = \mathbf{0}, \quad \text{for all } x, y, \quad \text{if } j = 2n+1, \tag{4.109}$$
or
$$H^T M D^{N-1}\mathbf{u}_j y x^n = \mathbf{0}, \quad \text{for all } x, y, \quad \text{if } j = 2(n+1), \tag{4.110}$$

We can now conclude. Using (4.109)-(4.110) and the decomposition (4.105) of the matrix H with $v_j = 0$ gives

$$J^T M D^{N-1}\mathbf{w}_j = \mathbf{0}.$$

Since the matrix J is invertible, the matrix $J^T M D^{N-1}$ is also invertible and one deduces $\mathbf{w}_j = \mathbf{0}$. Finally one gets $\mathbf{u}_j = \mathbf{0}$. Repeating recursively this process from $j = 2$ to $j = 2(k+N)-1$ yields $\mathbf{u} = \mathbf{0}$. One concludes as in the proof of Proposition 4.58. The proof is complete. ■

We can give a result similar to Corollary 4.68.

Corollary 4.71. *Consider the solutions (4.31) with $2(k+N)-1$ directions where the degenerative exponentials have been replaced by the polynomial solutions from Theorem 4.34 for a total of $(2(k+N)-1)m_e$ solutions. Among these $(2(k+N)-1)m_e$ functions there exists $(2k+1)m_e$ functions such that the matrix S^k_{2k+1} (4.81) is invertible.*

Proof. The proof is the same as in Corollary 4.68. ■

Again, the main defect of this corollary is that we do not know which basis functions give the invertibility of the matrix S^k_{2k+1}.

4-3.3 High order convergence for the stationary case

The main results of this section are the Theorem 4.75 which study the convergence of the TDG method applied to the P_N model when $\sigma_a > 0$ and the Theorem 4.80 for the P_1 model when $\sigma_a = 0$.

We consider a series of mesh $\mathcal{T}_h^n$, $n \in \mathbb{N}$. For a polygonal cell $\Omega_j^n \in \mathcal{T}_h^n$, we define h_j^n the size of its larger edge and ρ_j^n the radius of the larger inner circle include in Ω_j. We assume that the sequence of meshes is refined, that is

$$h^n := \max_j h_j^n \underset{n\to\infty}{\to} 0,$$

and the mesh is quasi uniform, that is there exists a constant $C \in \mathbb{R}^+$ such that

$$\max_{j,n} \frac{h_j^n}{\rho_j^n} \leq C. \tag{4.111}$$

To keep the notations simple we remove the index n in the following. We also assume in the rest of this section that the coefficients σ_a and σ_s are bounded: there exists $C \in \mathbb{R}^+$ such that

$$\sigma_a \leq C, \quad \sigma_s \leq C,$$

and we recall that

$$\varepsilon = 1, \quad c = 1.$$

For convenience $k \in \mathbb{N}$ is fixed. The following proposition generalizes on the variable $\mathbf{u}$ the estimates given in Proposition 4.55 (which were given for the variable $\mathbf{u}_e$) with a loss of one degree of convergence.

Proposition 4.72. *Let $k \in \mathbb{N}$, $\Omega_j \in \mathcal{T}_h$, $\mathbf{x}_0 \in \Omega_j$ and $\mathbf{u} = (\mathbf{u}_e^T, \mathbf{u}_o^T)^T \in W^{k+1,\infty}(\Omega_j)$ be a solution to the stationary P_N model. Consider the basis functions (4.31) with $2(k+N)-1$ directions for a total of $[2(k+N)-1]m_e$ functions (if $\sigma_a = 0$ the degenerative exponentials are replaced by the polynomial solutions from Theorem 4.34). Among these $[2(k+N)-1]m_e$ functions there exists $(2k+1)m_e$ solutions denoted $\mathbf{v}_1, ..., \mathbf{v}_{(2k+1)m_e} \in W^{k+1,\infty}(\Omega_j)$ and $\mathbf{a} = (a_1, ..., a_{(2k+1)m_e})^T \in \mathbb{R}^{(2k+1)m_e}$ such that*

$$\Big\| \sum_{i=1}^{(2k+1)m_e} a_i \mathbf{v}_i - \mathbf{u} \Big\|_{L^\infty(\Omega_j)} \leq C h^k \|\mathbf{u}\|_{W^{k+1,\infty}(\Omega_j)},$$

and

$$\Big\| \nabla \Big(\sum_{i=1}^{\nu_N m_e} a_i \mathbf{v}_i - \mathbf{u} \Big) \Big\|_{L^\infty(\Omega_j)} \leq C h^{k-1} \|\mathbf{u}\|_{W^{k+1,\infty}(\Omega_j)},$$

where C is a constant which does not depend on Ω_j.

Proof. We denote

$$\mathbf{z} = \sum_{i=1}^{(2k+1)m_e} a_i \mathbf{v}_i - \mathbf{u},$$

and use the decomposition $\mathbf{z} = (\mathbf{z}_e^T, \mathbf{z}_o^T)^T$. By definition $\mathbf{z}$ is solution to the P_N model. Corollary 4.68 if $\sigma_a > 0$, Corollary 4.71 if $\sigma_a = 0$ combined with Proposition 4.55 give an error estimate in h^{k+1} for $\mathbf{z}_e$. Using Proposition 4.51 one can control $\mathbf{z}$ with $\mathbf{z}_e$, $\partial_x \mathbf{z}_e$ and $\partial_y \mathbf{z}_e$. One deduces the first inequality and the second inequality immediately follows. The proof is complete. ■

In the following we consider the functions $\mathbf{v}_1, ..., \mathbf{v}_{(2k+1)m_e}$ from Proposition 4.72 as basis functions in all the cells and denote

$$V_h := \text{Span}\left\{\mathbf{v}_1, ..., \mathbf{v}_{(2k+1)m_e}\right\},$$

Note that the Proposition 4.56 combined with Corollaries 4.68-4.71 give the linear independence of the functions $\mathbf{v}_i$. We can now give an approximation result in terms of the $\|\cdot\|_{DG^*}$ norm.

Proposition 4.73. *Under the assumptions of Proposition 4.72, there exists $\mathbf{v}_h \in V_h$ such that*

$$\|\mathbf{u} - \mathbf{v}_h\|_{DG^*} \leq C h^{k-1/2} \|\mathbf{u}\|_{W^{k+1,\infty}(\Omega)},$$

with $h = \max_{\Omega_j \in \mathcal{T}_h} h_j$, $h_j = \text{diam}(\Omega_j)$ and C a constant independent of h.

Proof. From Proposition 4.72 one deduces that there exist $\mathbf{v}_h \in V_h$ such that $\forall \Omega_j$

$$\|\mathbf{u} - \mathbf{v}_h\|^2_{L^2(\Omega_j)} \leq C h_j^{2k+2} \|\mathbf{u}\|^2_{W^{k+1,\infty}(\Omega_j)},$$
$$|(\mathbf{u} - \mathbf{v}_h)|^2_{1,\Omega_j} \leq C h_j^{2k} \|\mathbf{u}\|^2_{W^{k+1,\infty}(\Omega_j)},$$

therefore

$$\|\mathbf{u} - \mathbf{v}_h\|_{L^2(\Omega_j)} \left(\frac{1}{h_j} \|\mathbf{u} - \mathbf{v}_h\|_{L^2(\Omega_j)} + |(\mathbf{u} - \mathbf{v}_h)|_{1,\Omega_j} \right) \leq C h_j^{2k+1} \|\mathbf{u}\|^2_{W^{k+1,\infty}(\Omega_j)}, \quad \forall \Omega_j.$$

Summing over all Ω_j and using that for a regular mesh of size h, the total number of elements is bounded by C/h^2 one has

$$\sum_j \|\mathbf{u} - \mathbf{v}_h\|_{L^2(\Omega_j)} \left(\frac{1}{h_j} \|\mathbf{u} - \mathbf{v}_h\|_{L^2(\Omega_j)} + |(\mathbf{u} - \mathbf{v}_h)|_{1,\Omega_j} \right) \leq C h^{2k-1} \|\mathbf{u}\|^2_{W^{k+1,\infty}(\Omega)}.$$

One concludes using Proposition 2.16. ■

Combining the previous proposition with the results of Section 2-3 one can now give an estimation of the error in DG norm.

Proposition 4.74. *Consider the TDG method (2.15) under the assumptions of Proposition 4.72. One has*

$$\|\mathbf{u} - \mathbf{u}_h\|_{DG} \leq C h^{k-1/2} \|\mathbf{u}\|_{W^{k+1,\infty}(\Omega)},$$

with $h = \max_{\Omega_j \in \mathcal{T}_h} h_j$, $h_j = \text{diam}(\Omega_j)$, where $\mathbf{u}_h$ stands for the solution to the TDG method.

Proof. We use Proposition 4.73 and conclude with the quasi-optimality result from Proposition 2.12. ■

One can now easily study the convergence in quadratic norm.

4-3.3.1 The P_N model when $\sigma_a > 0$

Theorem 4.75 (Convergence of the TDG method for the P_N model)**.** *Assume $\sigma_a > 0$, the hypothesis of Proposition 4.74 are satisfied and consider $2(k+N)-1$ directions for a total of $(2(k+N)-1)m_e$ functions. Among these $(2(k+N)-1)m_e$ functions there exists $(2k+1)m_e$ basis functions such that*

$$\|\mathbf{u}-\mathbf{u}_h\|_{L^2(\Omega)} \leq Ch^{k-1/2}\|\mathbf{u}\|_{W^{k+1,\infty}(\Omega)}, \tag{4.112}$$

with $h = \max_{\Omega_j \in \mathcal{T}_h} h_j$, $h_j = \operatorname{diam}(\Omega_j)$ and where $\mathbf{u}_h$ stands for the solution to the TDG method.

Proof. Since $\sigma_a > 0$, the matrix R is positive definite and one can give an L^2 lower bound of the DG norm with Proposition 2.14. One concludes with Proposition 4.74. ■

Remark 4.76. The Theorem 4.75 shows a remarkable property of the TDG method: the number of additional basis functions to gain one order of convergence from k to $k+1$ does not depend on k. This is not the case for the standard DG method where the number of additional basis functions increases with k.

For the P_1 model in particular, the Theorem 4.75 gives a convergence result with $2k+1$ directions.

Corollary 4.77. *The TDG method applied to the stationary P_1 model with $2k+1$ directions satisfies the estimation (4.112) of Theorem 4.75.*

Ideally, one would like the same convergence estimate using $2k+1$ directions for the general P_N model. Since there are m_e solutions per directions, such convergence result would use $(2k+1)m_e$ functions.

Although Theorem 4.75 gives a convergence result with $(2k+1)m_e$ basis functions, the main issue is that such basis functions may not be known when $N > 1$. Indeed for $N > 1$, the Theorem 4.75 only assures that the basis functions which give the convergence of the TDG method can be taken from $2(k+N)-1$ directions. We conjecture that the estimate (4.112) holds for $N > 1$ when considering $2k+1$ directions. This is equivalent to prove that the matrix S^k_{2k+1} is invertible, see Remark 4.67. In the numerical tests, we will use $2k+1$ directions.

It is interesting to compare the convergence estimate given by Theorem 4.75 with the standard convergence estimate obtained with the DG method. We compare the number of basis functions needed to achieve a given fractional order for the TDG method (denoted p_{TDG}) and for the general DG method (denoted p_{DG}).

For the P_1 model, $m_e = 1$ and one has (see Table 4.1)

$$p_{\text{TDG}} = 2(\text{order}+1), \quad p_{\text{DG}} = \frac{3}{2}(\text{order}+\frac{1}{2})(\text{order}+\frac{3}{2}).$$

In particular the number of basis functions is the same to get order $1/2$ and one always gets $p_{\text{TDG}} \leq p_{\text{DG}}$.

order	1/2	3/2	5/2	7/2	9/2
p_{TDG}	3	5	7	9	11
p_{DG}	3	9	18	30	45

Table 4.1 – P_1 model. Comparison of the number of basis functions needed to achieve a given order for the TDG method (denoted p_{TDG}) and the DG method (denoted p_{DG}).

For the P_3 model, $m_e = 4$ and one has (see Table 4.2)

$$p_{\text{TDG}} = 8(\text{order} + 1), \quad p_{\text{DG}} = 5(\text{order} + \frac{1}{2})(\text{order} + \frac{3}{2}).$$

Except for the order 1/2, one always get $p_{\text{TDG}} \leq p_{\text{DG}}$.

order	1/2	3/2	5/2	7/2	9/2
p_{TDG}	12	20	28	36	44
p_{DG}	10	30	60	100	150

Table 4.2 – P_3 model. Comparison of the number of basis functions needed to achieve a given order for the TDG method (denoted p_{TDG}) and the DG method (denoted p_{DG}).

●

4-3.3.2 The P_1 model when $\sigma_a = 0$

For the P_1 model, one can derive a convergence estimate in L^2 norm of the TDG scheme for the dominant scattering regime ($\sigma_s > 0$, $\sigma_a = 0$) with a loss of convergence of a half degree compare to Theorem 4.75. The main difficulty when studying the convergence of the scheme with $\sigma_a = 0$ is that the matrix R is not strictly positive anymore which results in a loss of control on the first variable u_1. For the TDG scheme applied to the P_1 model, one can recover some control on u_1 using the particular structure of the system and the fact that the TDG method uses solutions to the equation as basis functions. We recall that, with $\varepsilon = c = 1$, the stationary two dimensional P_1 model reads

$$\begin{cases} \frac{1}{\sqrt{3}}\Big(\partial_x u_2(\mathbf{x}) + \partial_y u_3(\mathbf{x})\Big) = -\sigma_a u_1(\mathbf{x}), \\ \frac{1}{\sqrt{3}}\partial_x u_1(\mathbf{x}) = -\sigma_t u_2(\mathbf{x}), \\ \frac{1}{\sqrt{3}}\partial_y u_1(\mathbf{x}) = -\sigma_t u_3(\mathbf{x}), \end{cases} \tag{4.113}$$

where $\mathbf{u} = (u_1, u_2, u_3)^T \in \mathbb{R}^3$ is the unknown and we switch the axis x and y to recover the usual notations. For a solution $\mathbf{u} = (u_1, u_2, u_3)^T$ to the P_1 model one deduces from the structure of (4.113) the following inequalities

$$|\partial_x u_1(\mathbf{x})| \leq C|u_2(\mathbf{x})|, \quad |\partial_y u_1(\mathbf{x})| \leq C|u_3(\mathbf{x})|, \quad C = \sqrt{3}\sigma_t. \tag{4.114}$$

These inequalities can be used to control u_1 with u_2 and u_3. Additionally, we need the generalization of the Poincaré inequality to discontinuous functions.

Lemma 4.78. *Assume $w \in H^1(\mathcal{T}_h)$. One has*

$$\|w\|^2_{L^2(\Omega)} \leq C\Big(\|\partial_x w\|^2_{L^2(\Omega)} + \|\partial_y w\|^2_{L^2(\Omega)} + \frac{1}{h}\sum_k \sum_{j<k} \|[\![w]\!]\|^2_{L^2(\Sigma_{kj})} + \sum_k \|w\|^2_{L^2(\Sigma_{kk})}\Big),$$

with $h = \max_{\Omega_k \in \mathcal{T}_h} h_k$, $h_k = diam(\Omega_k)$, where $[\![w]\!]$ denotes the jump of the function across a face and where C is a constant independent of h.

Proof. We use the mesh quasi uniformity (4.111) and the proof given in [Bre03] (see also [Arn82] for a weaker result). ■

The following lemma give a control of the L^2 norm in term of the DG norm.

Lemma 4.79. *Assume* $\mathbf{w} = (w_1, w_2, w_3)^T \in V(\mathcal{T}_h)$ *and* $\sigma_a + \sigma_s > 0$. *One has*

$$\|\mathbf{w}\|_{L^2(\Omega)} \leq \frac{C}{\sqrt{h}} \|\mathbf{w}\|_{DG},$$

with $h = \max_{\Omega_k \in \mathcal{T}_h} h_k$, $h_k = diam(\Omega_k)$ *and where the constant* C *is independent of* h.

Proof. Using the definition of the DG norm (2.24) with $\sigma_a + \sigma_s > 0$ one gets

$$\|w_2\|^2_{L^2(\Omega)} \leq C\|\mathbf{w}\|^2_{DG}, \quad \|w_3\|^2_{L^2(\Omega)} \leq C\|\mathbf{w}\|^2_{DG}. \tag{4.115}$$

It remains to show $\|w_1\|_{L^2(\Omega)} \leq \frac{C}{\sqrt{h}}\|\mathbf{w}\|_{DG}$. For the P_1 model the matrix $|M|$ reads

$$|M| = \begin{pmatrix} 1 & 0 & 0 \\ 0 & n_x^2 & n_x n_y \\ 0 & n_x n_y & n_y^2 \end{pmatrix}. \tag{4.116}$$

Since $\mathbf{w} \in V(\mathcal{T}_h)$ and $\sigma_a + \sigma_s > 0$, the L^2 generalization of the inequality (4.114) yields $\|\partial_x w_1\|^2_{L^2(\Omega)} \leq C\|w_2\|^2_{L^2(\Omega)}$ and $\|\partial_y w_1\|^2_{L^2(\Omega)} = C\|w_3\|^2_{L^2(\Omega)}$, $C \neq 0$. Therefore, from the inequality (4.115), the definition (4.116) of the matrix $|M|$ and the definition of the DG norm (2.24) one deduces

$$\|\partial_x w_1\|^2_{L^2(\Omega)} + \|\partial_y w_1\|^2_{L^2(\Omega)} + \sum_k \sum_{j<k} \|[\![w_1]\!]\|^2_{L^2(\Sigma_{kj})} + \sum_k \|w_1\|^2_{L^2(\Sigma_{kk})} \leq C\|\mathbf{w}\|^2_{DG}.$$

One concludes using $V(\mathcal{T}_h) \subset H^1(\mathcal{T}_h)$ and Lemma 4.78. ■

We can now give a convergence result in L^2 norm when $\sigma_a = 0$.

Theorem 4.80 (Convergence in the general regime: $\sigma_a + \sigma_s > 0$)**.** *Assume* $\sigma_a + \sigma_s > 0$. *Consider the stationary two dimensional* P_1 *model with the assumptions of Proposition 4.74 and* $2k+1$ *basis functions. One has the h-convergence estimate*

$$\|\mathbf{u} - \mathbf{u}_h\|_{L^2(\Omega)} \leq Ch^{k-1}\|\mathbf{u}\|_{W^{k+1,\infty}(\Omega)},$$

where $\mathbf{u}$ *stands for the exact solution and* $\mathbf{u}_h$ *for the approximate solution calculated by the TDG method.*

Proof. The case $\sigma_a > 0$ is already treated in Theorem 4.75. To treat the remaining case $\sigma_a = 0$ one can combine Lemma 4.79 and Propositions 4.74. The guaranteed order of convergence is the worst case, that is $k-1$. This completes the proof. ■

Remark 4.81 (Case $\varepsilon \to 0^+$)**.** It would be of course desirable to get uniform estimate in the case $\varepsilon \to 0^+$. The Theorem 4.80 in particular could be very helpful since the cases $\varepsilon \to 0^+$ and $\sigma_a \to 0$ are closely related. However dependence in ε arises through the basis functions $\mathbf{v}_i$ and the solution $\mathbf{u}$ and this dependence must therefore be carefully studied when using the results of the previous sections. Whereas it is possible to easily study this limit regime for the basis functions $\mathbf{v}_i$, it is much harder for the solution $\mathbf{u}$ mostly because boundary layers may occur depending on the boundary values. We note that initial boundary layers can also arise for time dependent problems. These theoretical issues are left for future research. ●

Chapter 5

Application to the P_1 and P_3 models in 2D

Contents

In this chapter, the application of the TDG method to the two dimensional P_1 and P_3 models is detailed. In particular, the results of Chapter 4 are used to give explicitly the basis functions which can then be used for numerical applications. Additionally, two dimensional numerical results are presented to illustrate some properties including the convergence of the method, its ability to capture boundary layers and the asymptotic behavior of the scheme in the diffusive regime.

5-1 General form of the P_N model

From the first section of Chapter 4, we recall that the P_N model can be written under the general form (4.17)-(4.18), that is

$$\Big(\varepsilon I_m \partial_t + A_1 \partial_x + A_2 \partial_y\Big)\mathbf{u}(t,\mathbf{x}) = -R\mathbf{u}(t,\mathbf{x}), \tag{5.1}$$

with $\mathbf{u} \in \mathbb{R}^m$. The matrices A_1 and A_2 have the following block structure [Her16]

$$A_1 = c\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad A_2 = c\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}, \tag{5.2}$$

where A, $B \in \mathbb{R}^{m_e \times m_o}$ are rectangular matrices and R is a diagonal matrix which might be written under the form

$$R = \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix} \in \mathbb{R}^{m \times m}, \tag{5.3}$$

where $R_1 \in \mathbb{R}^{m_e \times m_e}$, $R_2 \in \mathbb{R}^{m_o \times m_o}$ are diagonal matrices. Moreover we have introduced the parameters

$$c \in \mathbb{R}^+, \quad \varepsilon \in \mathbb{R}_*^+.$$

The parameter ε is used to study the diffusive regime of the P_N model since its first variable admits a diffusion limit when $\varepsilon \to 0$. The parameter c will be considered as a scaling constant.

5-2 The P_1 model

In this section, we derive stationary and time dependent solutions to the P_1 model. For the two dimensional P_1 model, one has $m = 3$, $m_e = 1$, $m_o = 2$ and the matrices read

$$A = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & \frac{1}{\sqrt{3}} \end{pmatrix}, \quad R_1 = \varepsilon \sigma_a, \quad R_2 = \begin{pmatrix} \sigma_t & 0 \\ 0 & \sigma_t \end{pmatrix}, \tag{5.4}$$

where

$$\sigma_t := \sigma_t^\varepsilon := \varepsilon \sigma_a + \frac{\sigma_s}{\varepsilon}.$$

Remark 5.1. To recover the usual notations used when studying the P_1 model, the axis x and y have been switched compare to the P_1 model given in Chapter 4. The inversion of the axis doesn't change anything except that one now has to consider the transpose of the rotation matrix U_θ. Therefore, the new rotation matrix reads

$$\widetilde{U}_\theta = U_\theta^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}. \tag{5.5}$$

●

5-2.1 Special stationary solutions

We calculate stationary solutions to the P_1 model derived in Chapter 4. We start with the exponential solutions when $\sigma_a > 0$.

Proposition 5.2 (Solutions to the P_1 model when $\sigma_a > 0$)**.** *Take* $\mathbf{d}_k = (\cos\theta_k, \sin\theta_k)^T \in \mathbb{R}^2$. *The following functions are solution to the* P_1 *model*

$$\mathbf{v}_k = \begin{pmatrix} \sqrt{\sigma_t} \\ -\sqrt{\varepsilon\sigma_a}\mathbf{d}_k \end{pmatrix} e^{\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\mathbf{d}_k^T \mathbf{x}}, \tag{5.6}$$

where $\sigma_t = \varepsilon\sigma_a + \frac{\sigma_s}{\varepsilon}$.

Proof. We use the Theorem 4.25. From the definition of the matrices A and R_1 (5.4) associated with the P_1 model, one has $(AA^T)^{-1}R_1 = 3\varepsilon\sigma_a$. This matrix has one eigenvalue $\mu_1 = 3\varepsilon\sigma_a$ associated with the eigenvector $\mathbf{w}_1 = 1$. Taking the notations from Theorem 4.25, one has $\mathbf{z}_1 = (1, -\sqrt{\frac{\varepsilon\sigma_a}{\sigma_t}}, 0)^T$. Using the definition (5.5) of the rotation matrix U_θ and multiplying the solution by $\sqrt{\sigma_t}$ give the functions (5.6). The proof is complete. ■

Now, we give the polynomial solutions when $\sigma_a = 0$.

Proposition 5.3 (Polynomial solutions to the P_1 model when $\sigma_a = 0$)**.** *Assume $\sigma_a = 0$. We denote $q_k(\mathbf{x})$, $k \in \mathbb{N}$, the scaled harmonic polynomials in two dimensions*

$$q_1 = 1, \quad q_{2l} = \frac{1}{l!}\Re((x-x_0)+i(y-y_0))^l, \quad q_{2l+1} = \frac{1}{k!}\Im((x-x_0)+i(y-y_0))^k, \quad \text{for } l \in \mathbb{N}^*. \tag{5.7}$$

The following functions are solutions to the P_1 model

$$\mathbf{v}_k = \begin{pmatrix} \frac{\sigma_s}{\varepsilon} q_k \\ -\frac{c}{\sqrt{3}} \nabla q_k \end{pmatrix}. \tag{5.8}$$

Proof. Let $\mathbf{v}_k = (v_1, v_2, v_3)^T$. The solutions to the P_1 model when $\sigma_a = 0$ are given by Theorem 4.34. More precisely, from Corollary 4.49, one gets that the first component v_1 is equal to the harmonic polynomial q_k. Using the definition (5.4) of the matrices A, B and R associated with the P_1 model one gets

$$\frac{c}{\sqrt{3}}\partial_x v_1 = -\sigma_t v_2, \quad \frac{c}{\sqrt{3}}\partial_y v_1 = -\sigma_t v_3.$$

Using $\sigma_t = \frac{\sigma_s}{\varepsilon}$ completes the proof. ■

5-2.2 Time dependent solutions

We derive here some special time dependent solutions to the P_1 model. For other examples of time dependent solutions which can be easily constructed or deduced from the stationary solutions (5.6)-(5.7), see Section 4-2.4 of Chapter 4. In this section, the solutions that we consider are product of time dependent polynomials and stationary exponentials.

Proposition 5.4 (Time dependent solutions when $\sigma_a > 0$)**.** *The following functions are solutions to the two dimensional P_1 model*

$$\mathbf{w}_{1,k}(t,\mathbf{x}) = \begin{pmatrix} -2c\varepsilon\sqrt{\sigma_a\sigma_t}\cos\theta_k - \sqrt{3\varepsilon}\sigma_t(\varepsilon\sigma_a+\sigma_t)x - 2c\sqrt{\sigma_a\sigma_t}\sigma_t\cos\theta_k t \\ c\sqrt{\varepsilon}(\varepsilon\sigma_a+\sigma_t) + \varepsilon\sqrt{3\sigma_a\sigma_t}(\varepsilon\sigma_a+\sigma_t)\cos\theta_k x + 2c\sqrt{\varepsilon}\sigma_a\sigma_t\cos^2\theta_k t \\ \varepsilon\sqrt{3\sigma_a\sigma_t}(\varepsilon\sigma_a+\sigma_t)\sin\theta_k x + 2c\sqrt{\varepsilon}\sigma_a\sigma_t\cos\theta_k\sin\theta_k t \end{pmatrix} e^{\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\mathbf{d}_k^T\mathbf{x}},$$

$$\mathbf{w}_{2,k}(t,\mathbf{x}) = \begin{pmatrix} -2c\varepsilon\sqrt{\sigma_a\sigma_t}\sin\theta_k - \sqrt{3\varepsilon}\sigma_t(\varepsilon\sigma_a+\sigma_t)y - 2c\sqrt{\sigma_a\sigma_t}\sigma_t\sin\theta_k t \\ \varepsilon\sqrt{3\sigma_a\sigma_t}(\varepsilon\sigma_a+\sigma_t)\cos\theta_k y + 2c\sqrt{\varepsilon}\sigma_a\sigma_t\cos\theta_k\sin\theta_k t \\ c\sqrt{\varepsilon}(\varepsilon\sigma_a+\sigma_t) + \varepsilon\sqrt{3\sigma_a\sigma_t}(\varepsilon\sigma_a+\sigma_t)\sin\theta_k y + 2c\sqrt{\varepsilon}\sigma_a\sigma_t\sin^2\theta_k t \end{pmatrix} e^{\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\mathbf{d}_k^T\mathbf{x}}, \tag{5.9}$$

with $\mathbf{d}_k = (\cos\theta_k, \sin\theta_k)^T$.

Proof. We start searching for solutions under the form

$$\mathbf{u}(t,\mathbf{x}) = \mathbf{q}(t,\mathbf{x})e^{\lambda\mathbf{d}_k^T\mathbf{x}}, \tag{5.10}$$

where $\mathbf{q}(t,\mathbf{x})$ can be written

$$\mathbf{q}(t,\mathbf{x}) = \mathbf{q}_0 + \mathbf{q}_1 x + \mathbf{q}_2 y + \mathbf{q}_3 t. \tag{5.11}$$

Using (5.10) in (5.1) and dropping the exponential terms, one has

$$\Big(\varepsilon\partial_t + A_1\partial_x + A_2\partial_y + (A_1\lambda\cos\theta_k + A_2\lambda\sin\theta_k + R)\Big)\mathbf{q}(t,\mathbf{x}) = \mathbf{0}.$$

Extending $\mathbf{q}$ one finds

$$\Big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\Big)\Big(\mathbf{q}_0 + \mathbf{q}_1 x + \mathbf{q}_2 y + \mathbf{q}_3 t\Big) + A_1\mathbf{q}_1 + A_2\mathbf{q}_2 + \varepsilon\mathbf{q}_3 = \mathbf{0}.$$

This equality holds for all x, y and t, thus one gets the following system

$$\begin{cases} \Big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\Big)\mathbf{q}_3 = \mathbf{0}, \\ \Big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\Big)\mathbf{q}_2 = \mathbf{0}, \\ \Big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\Big)\mathbf{q}_1 = \mathbf{0}, \\ \Big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\Big)\mathbf{q}_0 = -A_1\mathbf{q}_1 - A_2\mathbf{q}_2 - \varepsilon\mathbf{q}_3. \end{cases} \tag{5.12}$$

Therefore, the solutions to (5.1) under the form (5.10) with $\mathbf{q}$ given by (5.11) satisfy the system (5.12). One notices from the rotational relations of the P_N model (see for example Remark 4.14) that $U^T_{-\theta_k} A_1 U_{-\theta_k} = A_1\cos\theta_k + A_2\sin\theta_k$. Using $U_{-\theta_k} R U^T_{-\theta_k} = R$ and $U^T_{\theta_k} = U_{-\theta_k}$, one finds

$$\lambda\Big(A_1\cos\theta_k + A_2\sin\theta_k\Big) + R = U^T_{-\theta_k}\Big(\lambda A_1 + U_{-\theta_k} R U^T_{-\theta_k}\Big)U_{-\theta_k} = U_{\theta_k}\Big(\lambda A_1 + R\Big)U^T_{\theta_k}. \tag{5.13}$$

And one gets

$$\ker\Big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\Big) = U_{\theta_k}\big(\ker\lambda A_1 + R\big). \tag{5.14}$$

Therefore, a necessary condition for the system (5.12) to admits a non zero solution is $\det\big(\lambda A_1 + R\big) = 0$. This has already been studied in the one dimensional case (see the proof of Proposition 3.2) and one finds

$$\lambda = \pm\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}.$$

In the following, we take $\lambda = \frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}$ and study $\ker\big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\big)$. From (5.14) this is equivalent to study the kernel of $\lambda A_1 + R$ and then apply the rotation U_{θ_k}. The study of the kernel of $A_1\lambda + R$ has already been done in the one dimensional case and one gets

$$\ker(A_1\lambda + R) = \mathrm{Span}\Big((-\sqrt{\sigma_t}, \sqrt{\varepsilon\sigma_a}, 0)^T\Big).$$

Setting $\mathbf{w} \in \ker\lambda A_1 + R$ one finds

$$\mathrm{Span}\Big(U_{\theta_k}\mathbf{w}\Big) = \mathrm{Span}\Big((-\sqrt{\sigma_t}, \sqrt{\varepsilon\sigma_a}\cos\theta_k, \sqrt{\varepsilon\sigma_a}\sin\theta_k)^T\Big) = \ker\Big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\Big).$$

Using the relations (5.12) one gets

$$\mathbf{q}_1 = \alpha U_{\theta_k}\mathbf{w}, \quad \mathbf{q}_2 = \beta U_{\theta_k}\mathbf{w}, \quad \mathbf{q}_3 = \gamma U_{\theta_k}\mathbf{w}, \quad \alpha,\beta,\gamma \in \mathbb{R}.$$

From the last equality of (5.12), one sees that $-A_1\mathbf{q}_1 - A_2\mathbf{q}_2 - \varepsilon\mathbf{q}_3 \in \mathrm{Im}\,\big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\big)$. It implies

$$-A_1\mathbf{q}_1 - A_2\mathbf{q}_2 - \varepsilon\mathbf{q}_3 \in \ker\Big((\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R)^T\Big)^\perp$$

Since the matrices A_1, A_2 and R are symmetric, $\ker\big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\big)^T = \ker\big(\lambda(A_1\cos\theta_k + A_2\sin\theta_k) + R\big) = \mathrm{Span}(U_{\theta_k}\mathbf{w})$. A necessary condition is then $\big(U_{\theta_k}\mathbf{w}\big)^T\big(-A_1\mathbf{q}_1 - A_2\mathbf{q}_2 - \varepsilon\mathbf{q}_3\big) = 0$ which is equivalent to

$$2c\sqrt{\sigma_a\sigma_t}\Big(\alpha\cos\theta_k + \beta\sin\theta_k\Big) = \sqrt{3\varepsilon}(\varepsilon\sigma_a + \sigma_t)\gamma.$$

In the following, we consider two choices

- $\alpha = \sqrt{3\sigma_t}(\sigma_a + \sigma_t)$, $\beta = 0$, $\gamma = 2\sqrt{\sigma_a}\sigma_t \cos\theta_k$,
- $\alpha = 0$, $\beta = \sqrt{3\sigma_t}(\sigma_a + \sigma_t)$, $\gamma = 2\sqrt{\sigma_a}\sigma_t \sin\theta_k$.

With the first choice one gets from the fourth equation of (5.12)

$$\alpha = \sqrt{3\varepsilon\sigma_t}(\varepsilon\sigma_a + \sigma_t), \quad \beta = 0, \quad \gamma = 2c\sqrt{\sigma_a}\sigma_t \cos\theta_k,$$
$$\mathbf{q}_0 = \Big(-2c\varepsilon\sqrt{\sigma_a\sigma_t}\cos\theta_k, c\sqrt{\varepsilon}(\varepsilon\sigma_a + \sigma_t), 0\Big)^T + \delta U_{\theta_k}\mathbf{w}^T, \quad \delta \in \mathbb{R}.$$

and with the second choice one gets

$$\alpha = 0, \quad \beta = \sqrt{3\varepsilon\sigma_t}(\varepsilon\sigma_a + \sigma_t), \quad \gamma = 2c\sqrt{\sigma_a}\sigma_t \sin\theta_k,$$
$$\mathbf{q}_0 = \Big(-2c\varepsilon\sqrt{\sigma_a\sigma_t}\sin\theta_k, 0, c\sqrt{\varepsilon}(\varepsilon\sigma_a + \sigma_t)\Big)^T + \delta U_{\theta_k}\mathbf{w}^T, \quad \delta \in \mathbb{R}.$$

Setting $\delta = 0$ and using $\mathbf{u}(t,\mathbf{x}) = (\mathbf{q}_0 + \mathbf{q}_1 x + \mathbf{q}_2 y + \mathbf{q}_3 t)e^{\lambda \mathbf{d}_k^T \mathbf{x}}$, one finds the solutions $\mathbf{w}_{1,k}(t,\mathbf{x})$ for the first case and $\mathbf{w}_{2,k}(t,\mathbf{x})$ for the second case. This completes the proof. ■

Remark 5.5 (Two dimensional time dependent solutions using the rotational invariance)**.** It is also possible to derive two dimensional time dependent solutions from the one dimensional solutions given in Proposition 3.2. We recall that the matrices A_1 in one and two dimensions read

$$A_1^{1D} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad A_1^{2D} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where A_1^{1D} and A_1^{2D} denote respectively the matrices A_1 in one and two dimensions. Therefore, to get a solution to the two dimensional P_1 model, one can take a solution to the one dimensional P_1 model for the first two components and a third component which is zero. Using the time dependent solution given in the Proposition 3.2 of Chapter 3, one deduces that a solution to the two dimensional P_1 model is

$$\mathbf{v}(t,\mathbf{x}) = \begin{pmatrix} -\frac{c}{\varepsilon}(\varepsilon\sigma_a - \sigma_t) - \sqrt{\frac{3\sigma_a\sigma_t}{\varepsilon}}(\varepsilon\sigma_a + \sigma_t)x - 2\frac{c}{\varepsilon}\sigma_a\sigma_t t \\ \sqrt{3}\sigma_a(\varepsilon\sigma_a + \sigma_t)x + 2c\sigma_a\sqrt{\frac{\sigma_a\sigma_t}{\varepsilon}}t \\ 0 \end{pmatrix} e^{\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}x}.$$

One can now use Proposition 4.19 and apply a rotation to this solution with the rotation matrix (5.5). This gives the following solutions

$$\mathbf{w}_{3,k}(t,\mathbf{x}) = \begin{pmatrix} \frac{c}{\varepsilon}(\sigma_t - \varepsilon\sigma_a) - \sqrt{\frac{3\sigma_a\sigma_t}{\varepsilon}}(\varepsilon\sigma_a + \sigma_t)(\cos\theta_k x + \sin\theta_k y) - 2\frac{c}{\varepsilon}\sigma_a\sigma_t t \\ \sqrt{3}\sigma_a(\varepsilon\sigma_a + \sigma_t)\cos\theta_k(\cos\theta_k x + \sin\theta_k y) + 2c\sqrt{\frac{\sigma_a\sigma_t}{\varepsilon}}\sigma_a\cos\theta_k t \\ \sqrt{3}\sigma_a(\varepsilon\sigma_a + \sigma_t)\sin\theta_k(\cos\theta_k x + \sin\theta_k y) + 2c\sqrt{\frac{\sigma_a\sigma_t}{\varepsilon}}\sigma_a\sin\theta_k t \end{pmatrix} e^{\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\mathbf{d}_k^T\mathbf{x}}, \tag{5.15}$$

with $\mathbf{d}_k = (\cos\theta_k, \sin\theta_k)^T$. However, the solutions (5.15) can be directly deduced from the solutions (5.6) and (5.9). Indeed one notices

$$\mathbf{w}_{3,k}(t,\mathbf{x}) = \frac{\sqrt{\sigma_a}}{\varepsilon\sqrt{\sigma_t}}\Big(\cos\theta_k \mathbf{w}_{1,k}(t,\mathbf{x}) + \sin\theta_k \mathbf{w}_{2,k}(t,\mathbf{x})\Big) + c\frac{\sigma_t + \varepsilon\sigma_a}{\varepsilon\sqrt{\sigma_t}}\mathbf{v}_k(\mathbf{x}).$$

Therefore it is enough to consider only the solutions (5.6)-(5.9). ●

From the solutions (5.9), one can derive time-dependent polynomial solutions when $\sigma_a = 0$.

Proposition 5.6 (Time dependent polynomial solutions when $\sigma_a = 0$)**.** *The following functions are solutions to the two dimensional P_1 model when $\sigma_a = 0$*

$$\mathbf{v}_1(\mathbf{x}) = \begin{pmatrix} -\frac{2}{\sqrt{3}}\sqrt{\varepsilon}c^2\partial_x - \sqrt{3\varepsilon}\sigma_t^2 x - \frac{2}{\sqrt{3\varepsilon}}c^2\sigma_t t\partial_x \\ \sqrt{\varepsilon}c\sigma_t + \sqrt{\varepsilon}c\sigma_t x\partial_x + \frac{2}{3\sqrt{\varepsilon}}c^3 t\partial_x^2 \\ \sqrt{\varepsilon}c\sigma_t x\partial_y + \frac{2}{3\sqrt{\varepsilon}}c^3 t\partial_{xy} \end{pmatrix} q_k(\mathbf{x}),$$
$$\mathbf{v}_2(\mathbf{x}) = \begin{pmatrix} -\frac{2}{\sqrt{3}}\sqrt{\varepsilon}c^2\partial_y - \sqrt{3\varepsilon}\sigma_t^2 y - \frac{2}{\sqrt{3\varepsilon}}c^2\sigma_t t\partial_y \\ \sqrt{\varepsilon}c\sigma_t y\partial_x + \frac{2}{3\sqrt{\varepsilon}}c^3 t\partial_{xy} \\ \sqrt{\varepsilon}c\sigma_t + \sqrt{\varepsilon}c\sigma_t y\partial_y + \frac{2}{3\sqrt{\varepsilon}}c^3 t\partial_y^2 \end{pmatrix} q_k(\mathbf{x}), \tag{5.16}$$

where $q_k(\mathbf{x})$ is a harmonic polynomial.

Proof. Consider $2l+1$ functions under the form

$$f_k(\mathbf{x}) = e^{\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\mathbf{d}_k^T\mathbf{x}}, \quad i = 1, ..., 2l+1.$$

One notices

$$\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\cos\theta_k f_k(\mathbf{x}) = \partial_x f_k(\mathbf{x}), \quad \frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\sin\theta_k f_k(\mathbf{x}) = \partial_y f_k(\mathbf{x}) \tag{5.17}$$

From Chapter 4 Section 4-2.3 (see also [GHP09]), there exists $a_{i,j} \in \mathbb{R}$, $1 \le i, j \le 2l+1$ such that

$$\sum_{k=1}^{2l+1} a_{k,j} f_k(\mathbf{x}) \underset{\sigma_a \to 0}{\to} q_j(\mathbf{x}), \quad 1 \le j \le 2l+1.$$

We would like to use the same linear combinations and pass to the limit in (5.17). Of course, it requires to prove

$$\lim_{\sigma_a\to 0}\Big(\sum_{k=1}^{2k+1} a_{k,j}\partial_x f_k(\mathbf{x})\Big) = \partial_x\Big(\lim_{\sigma_a\to 0}\sum_{k=1}^{2k+1} a_{k,j} f_k(\mathbf{x})\Big). \tag{5.18}$$

For the simplicity and the brevity of the proof, we use the relations (5.18) without proving them and check *a posteriori* that the functions obtained are solutions to the P_1 model. We consider the solutions (5.9) and use the equalities (5.17) to replace $\cos\theta_k$ and $\sin\theta_k$ by ∂_x and ∂_y. One gets

$$\mathbf{w}_{1,k}(t,\mathbf{x}) = \begin{pmatrix} -\frac{2}{\sqrt{3}}\sqrt{\varepsilon}c^2\partial_x - \sqrt{3\varepsilon}\sigma_t(\varepsilon\sigma_a+\sigma_t)x - \frac{2}{\sqrt{3\varepsilon}}c^2\sigma_t t\partial_x \\ \sqrt{\varepsilon}c(\varepsilon\sigma_a+\sigma_t) + \sqrt{\varepsilon}c(\varepsilon\sigma_a+\sigma_t)x\partial_x + \frac{2}{3\sqrt{\varepsilon}}c^3 t\partial_x^2 \\ \sqrt{\varepsilon}c(\varepsilon\sigma_a+\sigma_t)x\partial_y + \frac{2}{3\sqrt{\varepsilon}}c^3 t\partial_{xy} \end{pmatrix} e^{\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\mathbf{d}_k^T\mathbf{x}},$$

$$\mathbf{w}_{2,k}(t,\mathbf{x}) = \begin{pmatrix} -\frac{2}{\sqrt{3}}\sqrt{\varepsilon}c^2\partial_y - \sqrt{3\varepsilon}\sigma_t(\varepsilon\sigma_a+\sigma_t)y - \frac{2}{\sqrt{3\varepsilon}}c^2\sigma_t t\partial_y \\ \sqrt{\varepsilon}c(\varepsilon\sigma_a+\sigma_t)y\partial_x + \frac{2}{3\sqrt{\varepsilon}}c^3 t\partial_{xy} \\ \sqrt{\varepsilon}c(\varepsilon\sigma_a+\sigma_t) + \sqrt{\varepsilon}c(\varepsilon\sigma_a+\sigma_t)y\partial_y + \frac{2}{3\sqrt{\varepsilon}}c^3 t\partial_y^2 \end{pmatrix} e^{\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\mathbf{d}_k^T\mathbf{x}}.$$

Assuming the relations (5.18) are true, one finds the functions (5.16) in the limit $\sigma_a \to 0$. One can check that these functions are solutions to the P_1 model when $\sigma_a = 0$. The proof is complete. ■

5-3 The P_3 model

In this section, we derive stationary solutions to the P_3 model. For the P_3 model one has $m = 10$, $m_e = 4$, $m_o = 6$ and the matrices read

$$A = \begin{pmatrix} 0 & \frac{1}{\sqrt{3}} & 0 & 0 & 0 & 0 \\ \frac{1}{\sqrt{5}} & 0 & \sqrt{\frac{3}{14}} & -\frac{1}{\sqrt{70}} & 0 & 0 \\ 0 & -\frac{1}{\sqrt{15}} & 0 & 0 & \sqrt{\frac{6}{35}} & 0 \\ 0 & \frac{1}{\sqrt{5}} & 0 & 0 & -\frac{1}{\sqrt{70}} & \sqrt{\frac{3}{14}} \end{pmatrix}, \quad B = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{5}} & 0 & 0 & -\frac{1}{\sqrt{70}} & -\sqrt{\frac{3}{14}} \\ -\frac{1}{\sqrt{15}} & 0 & 0 & \sqrt{\frac{6}{35}} & 0 & 0 \\ -\frac{1}{\sqrt{5}} & 0 & \sqrt{\frac{3}{14}} & \frac{1}{\sqrt{70}} & 0 & 0 \end{pmatrix},$$

$$R_1 = \begin{pmatrix} \varepsilon\sigma_a & 0 & 0 & 0 \\ 0 & \sigma_t & 0 & 0 \\ 0 & 0 & \sigma_t & 0 \\ 0 & 0 & 0 & \sigma_t \end{pmatrix}, \quad R_2 = \sigma_t I_{m_o}, \tag{5.19}$$

where I_{m_o} is the identity matrix of $\mathbb{R}^{m_o \times m_o}$.

We calculate the stationary solutions derived in Chapter 4 for the particular case of the P_3 model. We start with the exponential solutions when $\sigma_a > 0$.

Proposition 5.7 (Stationary solutions when $\sigma_a > 0$)**.** *Take* $\mathbf{d}_k = (\cos\theta_k, \sin\theta_k)^T \in \mathbb{R}^2$. *The following functions are solutions to the* P_3 *model*

$$\mathbf{v}_1(\mathbf{x}) = \begin{pmatrix} 0 \\ -\sqrt{30}\cos 2\theta_k \\ 0 \\ \sqrt{30}\sin 2\theta_k \\ \sqrt{14}\cos\theta_k \\ -\sqrt{14}\sin\theta_k \\ \sqrt{15}\cos 3\theta_k \\ -\cos\theta_k \\ \sin\theta_k \\ -\sqrt{15}\sin 3\theta_k \end{pmatrix} e^{\frac{1}{c}\sqrt{\frac{7}{3}}\sigma_t \mathbf{d}_k^T \mathbf{x}}, \quad \mathbf{v}_2(\mathbf{x}) = \begin{pmatrix} 0 \\ \sqrt{2}\sin 2\theta_k \\ \sqrt{6} \\ \sqrt{2}\cos 2\theta_k \\ 0 \\ 0 \\ -\sqrt{3}\sin 3\theta_k \\ -\sqrt{5}\sin\theta_k \\ -\sqrt{5}\cos\theta_k \\ -\sqrt{3}\cos 3\theta_k \end{pmatrix} e^{\frac{1}{c}\sqrt{7}\sigma_t \mathbf{d}_k^T \mathbf{x}},$$

$$\mathbf{v}_3(\mathbf{x}) = \begin{pmatrix} \frac{\sqrt{\sigma_t}}{14\sqrt{15}}\rho^+ \\ \varepsilon\sqrt{\sigma_t}\sigma_a \sin 2\theta_k \\ -\frac{\varepsilon\sqrt{\sigma_t}\sigma_a}{\sqrt{3}} \\ \varepsilon\sqrt{\sigma_t}\sigma_a \cos 2\theta_k \\ -\frac{1}{630\sqrt{2}}\upsilon^-\tau^+ \sin\theta_k \\ -\frac{1}{630\sqrt{2}}\upsilon^-\tau^+ \cos\theta_k \\ -\frac{\varepsilon}{2\sqrt{21}}\sigma_a\upsilon^- \sin 3\theta_k \\ \frac{\varepsilon}{2\sqrt{35}}\sigma_a\upsilon^- \sin\theta_k \\ \frac{\varepsilon}{2\sqrt{35}}\sigma_a\upsilon^- \cos\theta_k \\ -\frac{\varepsilon}{2\sqrt{21}}\sigma_a\upsilon^- \cos 3\theta_k \end{pmatrix} e^{\frac{1}{c}\upsilon^-\sqrt{\frac{\sigma_t}{18}}\mathbf{d}_k^T \mathbf{x}}, \quad \mathbf{v}_4(\mathbf{x}) = \begin{pmatrix} \frac{\sqrt{\sigma_t}}{14\sqrt{15}}\rho^- \\ \varepsilon\sqrt{\sigma_t}\sigma_a \sin 2\theta_k \\ -\frac{\varepsilon\sqrt{\sigma_t}\sigma_a}{\sqrt{3}} \\ \sqrt{\sigma_t}\sigma_a \cos 2\theta_k \\ -\frac{\sqrt{\varepsilon}}{630\sqrt{2}}\upsilon^+\tau^- \sin\theta_k \\ -\frac{\sqrt{\varepsilon}}{630\sqrt{2}}\upsilon^+\tau^- \cos\theta_k \\ -\frac{\sqrt{\varepsilon}}{2\sqrt{21}}\sigma_a\upsilon^+ \sin 3\theta_k \\ \frac{\sqrt{\varepsilon}}{2\sqrt{35}}\sigma_a\upsilon^+ \sin\theta_k \\ \frac{\sqrt{\varepsilon}}{2\sqrt{35}}\sigma_a\upsilon^+ \cos\theta_k \\ -\frac{\sqrt{\varepsilon}}{2\sqrt{21}}\sigma_a\upsilon^+ \cos 3\theta_k \end{pmatrix} e^{\frac{1}{c}\upsilon^+\sqrt{\frac{\sigma_t}{18}}\mathbf{d}_k^T \mathbf{x}}, \tag{5.20}$$

with $\sigma_t = \varepsilon\sigma_a + \frac{\sigma_s}{\varepsilon}$ *and where we use the following notations* $\kappa = \sqrt{605\varepsilon^2\sigma_a^2 + 14\varepsilon\sigma_a\sigma_t + 245\sigma_t^2}$, $\upsilon^\pm = \sqrt{55\varepsilon\sigma_a + 35\sigma_t \pm \sqrt{5}\kappa}$, $\tau^\pm = \sqrt{5}\varepsilon\sigma_a + 35\sqrt{5}\sigma_t \pm 5\kappa$, $\rho^\pm = (\upsilon^\pm)^2 - 110\varepsilon\sigma_a$.

Proof. With the definitions (5.19) of the matrices A and R_1, one finds that the matrix $(AA^T)^{-1}R_1$ admits the following eigenvalues

$$\mu_1 = \frac{7}{3}\sigma_t, \quad \mu_2 = 7\sigma_t, \quad \mu_3 = \frac{(\upsilon^-)^2}{18}, \quad \mu_4 = \frac{(\upsilon^+)^2}{18},$$

and the following eigenvectors

$$\mathbf{w}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 0 \\ 0 \\ \sqrt{3} \\ 1 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} \frac{-11\sqrt{5}\varepsilon\sigma_a + 7\sqrt{5}\sigma_t + \kappa}{14\sqrt{3}\varepsilon\sigma_a} \\ 0 \\ -\frac{1}{\sqrt{3}} \\ 1 \end{pmatrix}, \quad \mathbf{w}_4 = \begin{pmatrix} \frac{-11\sqrt{5}\varepsilon\sigma_a - 7\sqrt{5}\sigma_t + \kappa}{14\sqrt{3}\varepsilon\sigma_a} \\ 0 \\ -\frac{1}{\sqrt{3}} \\ 1 \end{pmatrix}.$$

Then, it is just the application of the Theorem 4.34 using the Definition 4.20 of the rotation matrix U_{θ_k}. After easy simplifications, and considering the correct scaling, one finds the functions (5.20). The proof is complete. ■

Now we give some polynomial solutions when $\sigma_a = 0$.

Proposition 5.8 (Polynomial solutions when $\sigma_a = 0$)**.** *The polynomial solutions to the P_3 model when $\sigma_a = 0$ are given by the Theorem 4.34 and can be recursively calculate with the formulas (4.41)-(4.42)-(4.43). For example the first five polynomial solutions read*

$$\mathbf{v}_1(\mathbf{x}) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2(\mathbf{x}) = \begin{pmatrix} \sigma_t x \\ 0 \\ 0 \\ 0 \\ 0 \\ -\frac{c}{\sqrt{3}} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_3(\mathbf{x}) = \begin{pmatrix} \sigma_t y \\ 0 \\ 0 \\ 0 \\ -\frac{c}{\sqrt{3}} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\mathbf{v}_4(\mathbf{x}) = \begin{pmatrix} \sigma_t^2 xy \\ \frac{2c^2}{\sqrt{15}} \\ 0 \\ 0 \\ -\frac{\sigma_t c}{\sqrt{3}} x \\ -\frac{\sigma_t c}{\sqrt{3}} y \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_5(\mathbf{x}) = \begin{pmatrix} \frac{1}{2}\sigma_t^2 (x^2 - y^2) \\ 0 \\ 0 \\ \frac{2c^2}{\sqrt{15}} \\ \frac{\sigma_t c}{\sqrt{3}} y \\ -\frac{\sigma_t c}{\sqrt{3}} x \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{5.21}$$

Proof. We apply the Theorem 4.34 and use the recurrence formulas (4.41)-(4.42)-(4.43). Rescaling the functions if needed give the solutions (5.21). The proof is complete. ■

Remark 5.9. The solutions (5.21) which are calculated with the recurrence formulas from Theorem 4.34 are the same than the solutions obtained using Birkhoff and Abu-Shumays method in Section 4-2.2 with $q = 1, x, y, xy$ and $x^2 - y^2$. ●

Remark 5.10 (Time dependent solutions)**.** As we will see in the numerical tests, time dependent basis functions may deteriorate the condition number of the mass matrix when they are used together with stationary basis functions. In particular it seems difficult to perform realistic two dimensional numerical tests for the P_1 model with time dependent basis functions under the form

$$\mathbf{v}(t,\mathbf{x}) = \mathbf{q}(t,\mathbf{x})e^{\lambda(x\cos\theta+y\sin\theta)}. \tag{5.22}$$

Therefore, we do not try to construct the solutions (5.22) for the P_3 model since such basis functions probably require a good preconditioner to be used in our tests.

Note however that we will be able to use the exponential in time solutions constructed in Section 4-2.4 of Chapter 4. ●

5-4 Numerical results

The goal of this section is to validate on numerical examples some properties such has the convergence, the ability to capture boundary layers and the asymptotic-preserving (AP) property of the scheme. The tests will be perform in two dimensions for stationary and time dependent problems. Moreover, meshes made of random quads are using. A random quad mesh is made of $N \times N$ quads, $N \in \mathbb{N}^*$, where the vertices are randomly moved around their initial position by a factor of at most 33%.

In the following, we may identify the number and the type of the stationary basis functions used in the TDG scheme by their directions. To remove all ambiguity, we make the following comments

— For the stationary P_1 model, the functions (5.6) admit one solution per direction. Therefore, when we say we consider the P_1 model with n directions, it means that the TDG method is applied with n basis functions. On the contrary, for the stationary P_3 model the functions (5.20) admit 4 solutions per direction. Therefore, when we say we consider the P_3 model with n directions, it means that the TDG method is applied with $4n$ basis functions.

— When $\sigma_a = 0$ the polynomial solutions (5.8)-(5.21) do not strictly speaking dependent on a direction. For simplicity, we may still speak about direction to describe the number of basis functions used in our scheme. For the P_1 model, n directions will simply mean the first n polynomial solutions (5.8). For the P_3 model, n directions will mean $3n$ exponential basis functions and the first n polynomial solutions (5.21).

More precisely, we consider the following possible choices. With 3 basis functions per cell, we consider the following equi-distributed directions

$$\mathbf{d}_1 = (1,0)^T, \quad \mathbf{d}_2 = (\cos\frac{2\pi}{3}, \sin\frac{2\pi}{3})^T, \quad \mathbf{d}_3 = (\cos\frac{4\pi}{3}, \sin\frac{4\pi}{3})^T. \tag{5.23}$$

With 4 basis functions per cell, we consider the following equi-distributed directions

$$\mathbf{d}_1 = (1,0)^T, \quad \mathbf{d}_2 = (0,1)^T, \quad \mathbf{d}_3 = (-1,0)^T, \quad \mathbf{d}_4 = (0,-1)^T. \tag{5.24}$$

With 5 basis functions per cell, we consider the following equi-distributed directions

$$\begin{gathered}\mathbf{d}_1 = (1,0)^T, \quad \mathbf{d}_2 = (\cos\frac{2\pi}{5}, \sin\frac{2\pi}{5})^T, \quad \mathbf{d}_3 = (\cos\frac{4\pi}{5}, \sin\frac{4\pi}{5})^T, \\ \mathbf{d}_4 = (\cos\frac{6\pi}{5}, \sin\frac{6\pi}{5})^T, \quad \mathbf{d}_5 = (\cos\frac{8\pi}{5}, \sin\frac{8\pi}{5})^T.\end{gathered} \tag{5.25}$$

Remark 5.11 (Normalized exponentials)**.** For some numerical tests, the basis functions of the TDG method can be written

$$\mathbf{z}e^{\lambda \mathbf{d}^T \mathbf{x}}.$$

In particular, if $\lambda >> 1$ the basis functions are stiff exponentials. To keep the calculation of the integrals bounded, the exponentials are normalized in each cell. That is, we consider basis functions under the form

$$\mathbf{z}e^{\lambda \mathbf{d}^T (\mathbf{x}-\mathbf{x}_0)},$$

where $\mathbf{x}_0$ is the node of the cell where the function $e^{\lambda \mathbf{d}^T \mathbf{x}}$ takes its maximum value. ●

5-4.1 Convergence with absorption

Consider the stationary P_1 model in two dimensions. Let $\mathbf{x} = (x, y)^T, \Omega = [0, 1]^2$, $\varepsilon = 1, c = \sqrt{3}, \sigma_a = 1, \sigma_s = 1$. The exact solution we consider here is

$$\mathbf{u}_{ex}(\mathbf{x}) = \left(\cos(y)e^{\sqrt{3}x}, -(\sqrt{3}/2)\cos(y)e^{\sqrt{3}x}, 0.5\sin(y)e^{\sqrt{3}x} \right)^T.$$

We assume $M^-\mathbf{u} = M^-\mathbf{u}_{ex}$ is imposed on the boundary and consider $n \in \mathbb{N}$ basis functions (5.6) define as

$$\mathbf{e}_k(\mathbf{x}) = (\sqrt{2}, \mathbf{d}_k)e^{\sqrt{2}(\mathbf{d}_k, \mathbf{x})}, \quad k = 1, ..., n,$$

with $\mathbf{d}_k = \left(\cos(\theta_k), \sin(\theta_k) \right)^T, \theta_k = 2(k-1)\pi/n$.

Results obtained with $3, 5$ and 7 basis functions are displayed on the left of Figure 5.1. As stated in Theorem 4.75 for the particular case $N = 1$, one only needs two additional basis functions to increase the order by a factor 1. Note however that the orders obtained here are slightly better than those predicted in Theorem 4.75: with 3, 5 and 7 basis functions, one gets respectively order 0.8, 1.5 and 2.5.

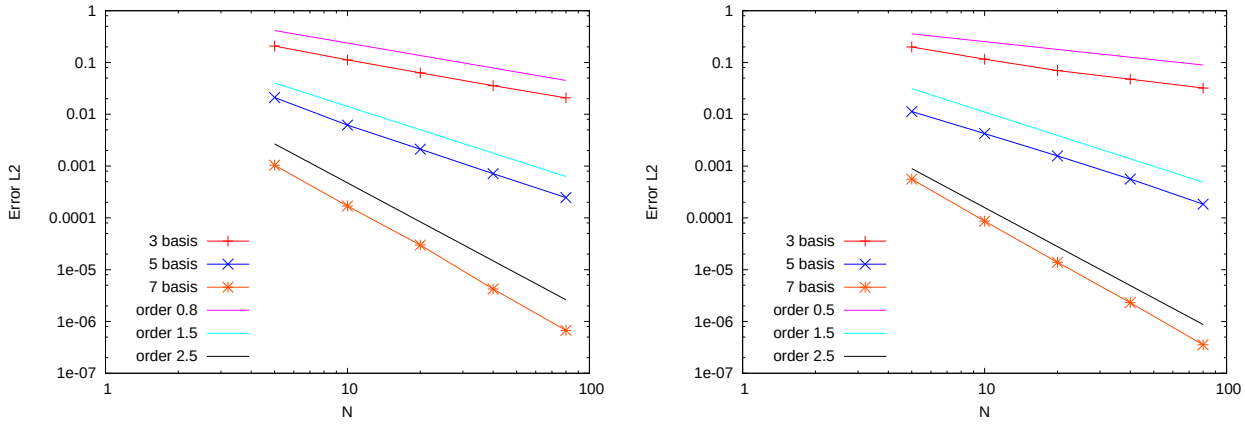

Figure 5.1 – **P_1 model**. Case $\sigma_a = 1$ on the left and $\sigma_a = 0$ on the right. L^2 error in logarithmic scale of the TDG method for the stationary two dimensional P_1 model. Random meshes.

5-4.2 Convergence without absorption

Consider the stationary P_1 model in two dimensions with the same parameters as before but without absorption: $\mathbf{x} = (x, y)^T, \Omega = [0, 1]^2$, $\varepsilon = 1, c = \sqrt{3}, \sigma_a = 0, \sigma_s = 1$. The exact solution is

$$\mathbf{u}_{ex}(\mathbf{x}) = \Big(\cos(y)e^x, -\cos(y)e^x, \sin(y)e^x \Big)^T.$$

Again, $M^-\mathbf{u} = M^-\mathbf{u}_{ex}$ is imposed on the boundary. We consider the polynomial basis functions (5.8).

Results obtained with $3, 5$ and 7 basis are displayed on the right of Figure 5.1. The orders are very close to those obtained in the case $\sigma_a > 0$ (left of the Figure 5.1) and therefore better, by a factor $1/2$, than those predict by Theorem 4.80. With 3, 5 and 7 basis functions, one respectively gets order 0.5, 1.5 and 2.5.

5-4.3 A first asymptotic study when $\varepsilon << 1$

We study here the asymptotic behavior of the TDG method when $\varepsilon \to 0$. More precisely, we consider the test case from [BDFL16] for the time dependent P_1 model. Let $\mathbf{x} = (x, y)^T, \Omega_S = [0, 1]^2$, $T = 0.036$, $\sigma_a = 0, \sigma_s = 1, c = 1$, and consider the solution

$$p_0 = f + \frac{\varepsilon^2}{\sigma_s}\partial_t f, \quad \mathbf{v}_0 = -\frac{\varepsilon}{\sigma_s}\nabla f,$$

with

$$f(t, \mathbf{x}) = \alpha(t)\cos(2\pi x)\cos(2\pi y),$$

and where $\alpha(t)$ is defined as

$$\alpha(t) = \frac{\lambda_2}{\lambda_2 - \lambda_1}e^{\lambda_1 t} - \frac{\lambda_1}{\lambda_2 - \lambda_1}e^{\lambda_2 t},$$

$$\lambda_1 = -\frac{\sigma_s\Big(\sqrt{1 - \frac{\varepsilon^2}{\sigma_s^2}32\pi^2} + 1\Big)}{2\varepsilon^2}, \quad \lambda_2 = -\frac{\sigma_s\Big(\sqrt{1 - \frac{\varepsilon^2}{\sigma_s^2}32\pi^2} - 1\Big)}{2\varepsilon^2}.$$

One can check that $(p_0, \mathbf{v}_0)^T$ is indeed a solution to the P_1 model when $\sigma_a = 0$, see [BDFL16] for details. An exact relation is enforced between ε and the space step $h = \frac{1}{N}$. The relation between ε and h reads

$$\varepsilon = 0.01(40h)^\tau, \qquad \text{for } \tau \in \Big\{0, \frac{1}{4}, \frac{1}{2}, 1, 2\Big\}.$$

The error between the exact solution and the numerical solution is computed numerically in function of h for the different values of τ. The result is displayed in Figure 5.2 when using the TDG method with the first 3 stationary polynomial basis functions (5.8) and $dt = 0.36h^2$. One observes the convergence of the solution even for small values of ε.

5-4.4 A second asymptotic study when $\varepsilon << 1$

We study a second numerical example when $\varepsilon << 1$. We consider the spatial domain $\Omega_S = [0, 1] \times [0, 1]$ and the final time $T = 0.01$. We take $\sigma_a = 0$, $\sigma_s = 1/3$ and $\varepsilon = 10^{-3}$. In this regime, the first variable of the P_N model follows a diffusion equation [Her16, Theorem 1]. Therefore, we compare our numerical solution with the two dimensional fundamental solution of the heat equation centered in $(0.5, 0.5)^T$

$$p(t, \mathbf{x}) = \frac{1}{4\pi(t + 10^{-4})}e^{-\frac{(x-0.5)^2 + (y-0.5)^2}{4(t+10^{-4})}}. \tag{5.26}$$

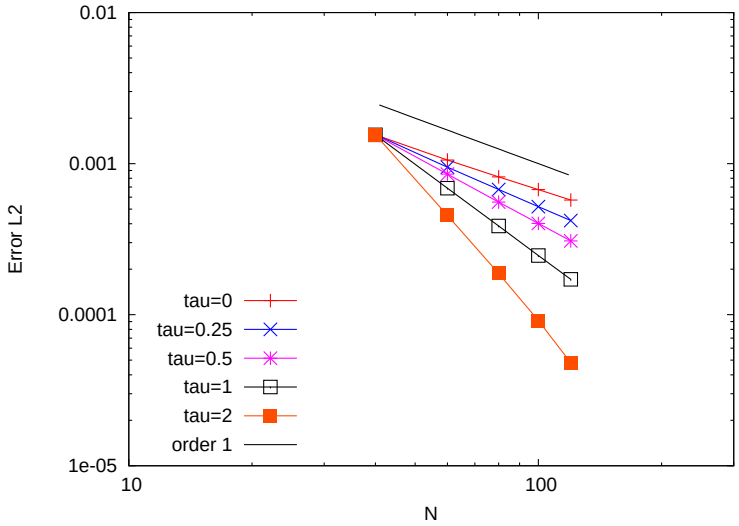

Figure 5.2 – **P_1 model**. Study of the L^2 error for the test case 5-4.3 at the final time in logarithmic scale. TDG method with 3 basis functions and $\varepsilon = 0.01(40h)^\tau$.

On the boundary of the domain we impose $M^-\mathbf{u}$ with

$$\mathbf{u}(t,\mathbf{x}) = \Big(p(t,\mathbf{x}),0,...,0\Big)^T.$$

The P_1 model.

For the P_1 model, we compare the results obtained with the DG and TDG method on a 80×80 mesh with $dt = T/80$. More precisely, we consider the two following cases

- The DG method with constant basis functions only (= finite volume) for a total of 3 basis functions per cell.
- The DG method with affine basis functions (that is $1, x, y$) for a total of 9 basis functions per cell.
- The TDG method with the first three polynomial basis functions (5.8) for a total of 3 basis functions per cell.

The limit solution (5.26), calculated on a 80×80 mesh, and the first variable of the numerical solution is represented in Figure 5.3. Figure 5.3 illustrates that the DG method with only constant basis function is too diffusive. On the contrary, one recovers a good approximation for the TDG method with the same number of basis functions. This illustrates the AP property of the TDG scheme on the P_1 model. To recover a good accuracy, another possibility is to increase the number of basis functions of the DG method and consider a total of 9 basis functions. In such case, the diffusion limit is indeed recovered but at the cost of considering three time more basis functions than the TDG scheme.

An other interesting question is whether the special choice of basis functions for the TDG method has an effect on the condition number of the mass matrix. In Figure 5.4, an estimation of the condition number with different values of ε is given for the two following cases

- The DG method with affine basis functions (that is $1, x, y$) for a total of 9 basis functions per cell.
- The TDG method with the first three polynomial basis functions (5.8) if $\sigma_a = 0$ or the 3 directions (5.25) if $\sigma_a > 0$ for a total of three basis functions per cell in both cases.

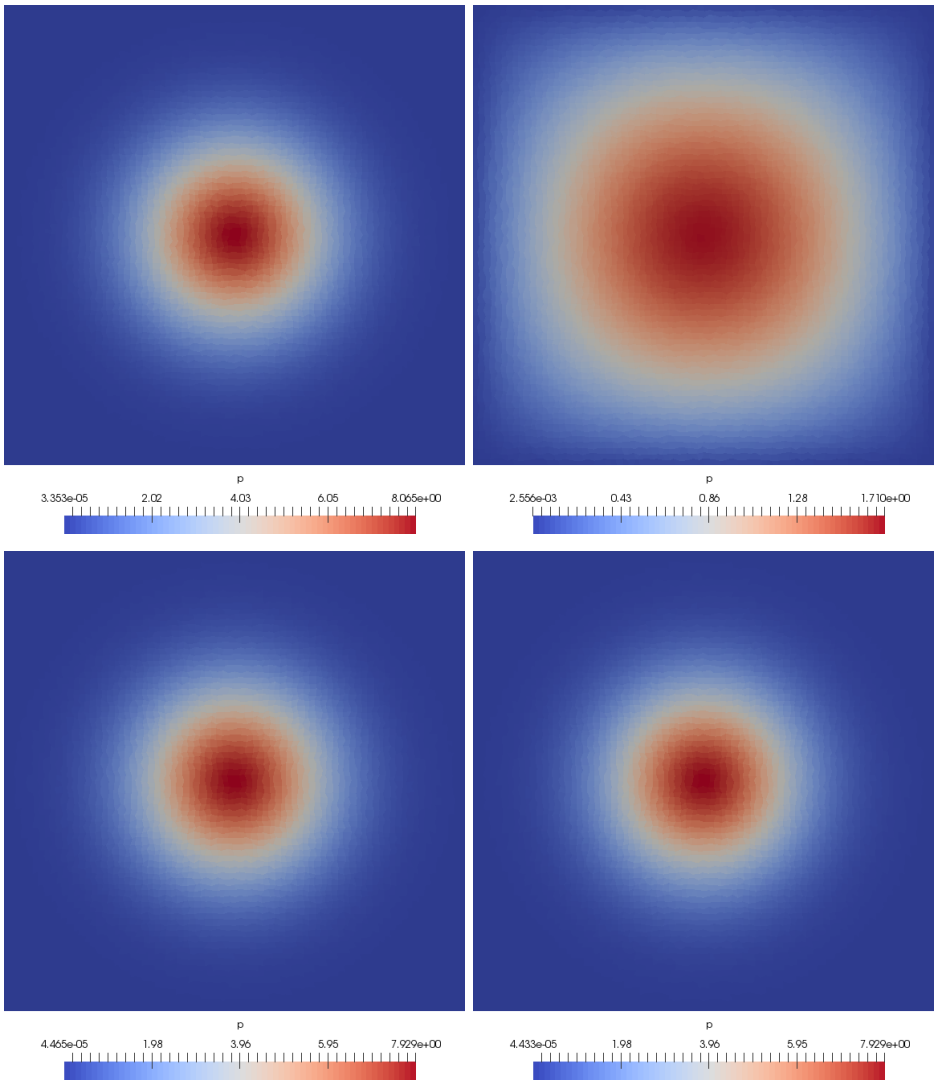

Figure 5.3 – **P_1 model**. Representation of the first variable when $\varepsilon = 10^{-3}$ for the test case 5-4.4. Top left: limit solution. Top right: DG scheme with 3 basis functions per cell. Bottom left: DG scheme with 9 basis functions per cell. Bottom right: TDG scheme with only 3 basis functions per cell. Good behavior of the numerical solution illustrates the AP property.

The condition number is calculated on a 10×10 mesh using the singular values of the matrix. On the left of Figure 5.4, the value $\sigma_a = 0$ is taken and therefore only polynomial solutions are used in the basis functions. As one might have expected, the condition number of the TDG method is not greater than the condition number of the DG method in this case. On the right of Figure 5.4, the same test case but with $\sigma_a = 1$ is considered. This time one sees that the value of the condition number is greater for the TDG method compare to the DG method. This is probably because, when $\sigma_a = 1$, the following exponentials are used in the basis functions of the TDG method (with a rescaling by $\sqrt{\varepsilon}$ compare to the solutions given in (5.6))

$$\mathbf{v}_k(\mathbf{x}) = \begin{pmatrix} \sqrt{\varepsilon^2\sigma_a + \sigma_s} \\ -\varepsilon\sqrt{\sigma_a}\cos\theta_k \\ -\varepsilon\sqrt{\sigma_a}\sin\theta_k \end{pmatrix} e^{\sqrt{3\varepsilon\sigma_a(\varepsilon\sigma_a + \frac{\sigma_s}{\varepsilon})}\mathbf{d}_k^T\mathbf{x}} \underset{\varepsilon\to 0}{\to} \begin{pmatrix} \sqrt{\sigma_s} \\ 0 \\ 0 \end{pmatrix} e^{\sqrt{3\sigma_a\sigma_s}\mathbf{d}_k^T\mathbf{x}}. \tag{5.27}$$

Note that, since the P_1 model is a very simple approximation of the transport equation, no boundary layers exist for this model when $\varepsilon \to 0$. Consequently, the exponentials (5.27) are

not stiff when $\varepsilon \to 0$. However, when $\varepsilon \to 0$, the vector in front of the exponentials in (5.27) tends toward the same limit for all the solutions $\mathbf{v}_k$. This may explain why such basis functions deteriorate the condition number in such limit case.

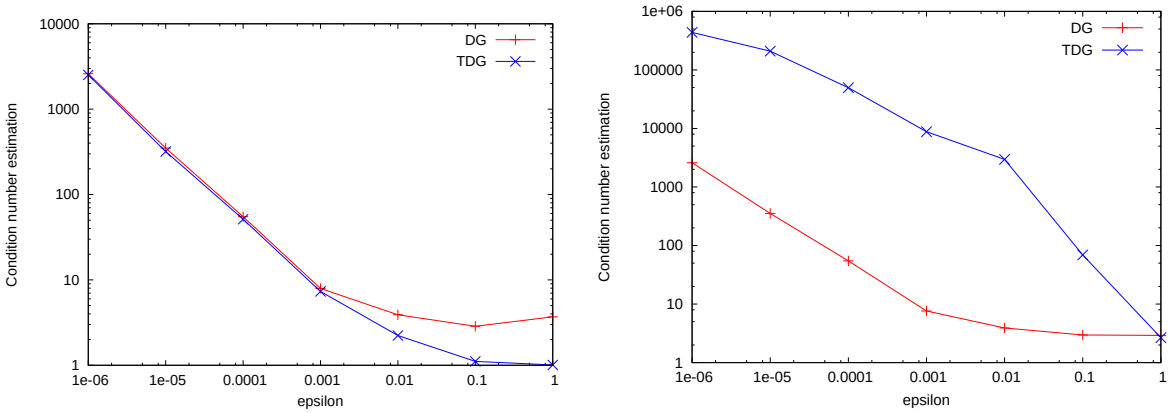

Figure 5.4 – **P_1 model**. Comparison of the condition number between the TDG and the DG method. On the left $\sigma_a = 0$ (polynomial basis functions used in the TDG method) and $\sigma_a = 1$ on the right (exponential basis functions used in the TDG method).

The P_3 model.

For the P_3 model we also compare the results obtained with the DG and TDG method on a 80×80 mesh with $dt = T/80$. More precisely, we consider the two following cases

- The DG method with constant basis functions only (= finite volume) for a total of 10 basis functions per cell.
- The DG method with affine basis function (that is $1, x, y$) for a total of 30 basis functions per cell.
- The TDG method with the basis functions (5.20)-(5.21) depending on the 3 directions (5.23) for a total of 12 basis functions per cell.

Note that, unlike the P_1 model, the TDG method applied to the P_3 model uses exponential *and* polynomial basis functions. The limit solution, which is the same as before, and the first variable of the numerical solution are represented in Figure 5.3. As for the P_1 model, Figure 5.5 illustrates that the DG method with only constant basis function is too diffusive. On the contrary, one recovers a good approximation with the TDG method. This illustrates the AP property of the TDG scheme on the P_3 model. As for the P_1 model, the DG scheme with affine basis functions recovers the correct diffusion limit but with the disadvantage of using approximately three time more basis functions than the TDG scheme.

5-4.5 Boundary layers

In this test, a two dimensional test with discontinuous coefficients is studied. The domain is $\Omega = [0,1]^2$ and we define Ω_1 (resp. Ω_2) as $\Omega_1 = [0.35, 0.65]^2$ (resp. $\Omega_2 = \Omega \setminus \Omega_1$). We take $\epsilon = 1$, $c = 1$ and

$$\sigma_a = 2 \times \mathbf{1}_{\Omega_1}(\mathbf{x}), \quad \sigma_s = 2 \times \mathbf{1}_{\Omega_2}(\mathbf{x}) + 10^5 \times \mathbf{1}_{\Omega_1}(\mathbf{x}).$$

The absorption coefficient has compact support in Ω_1 while the scattering coefficient is discontinuous and takes a high value in Ω_1. Even if we consider random meshes, the interface between Ω_1 and Ω_2 is a straight line. The geometry and parameters of this test are represented in Figure 5.6.

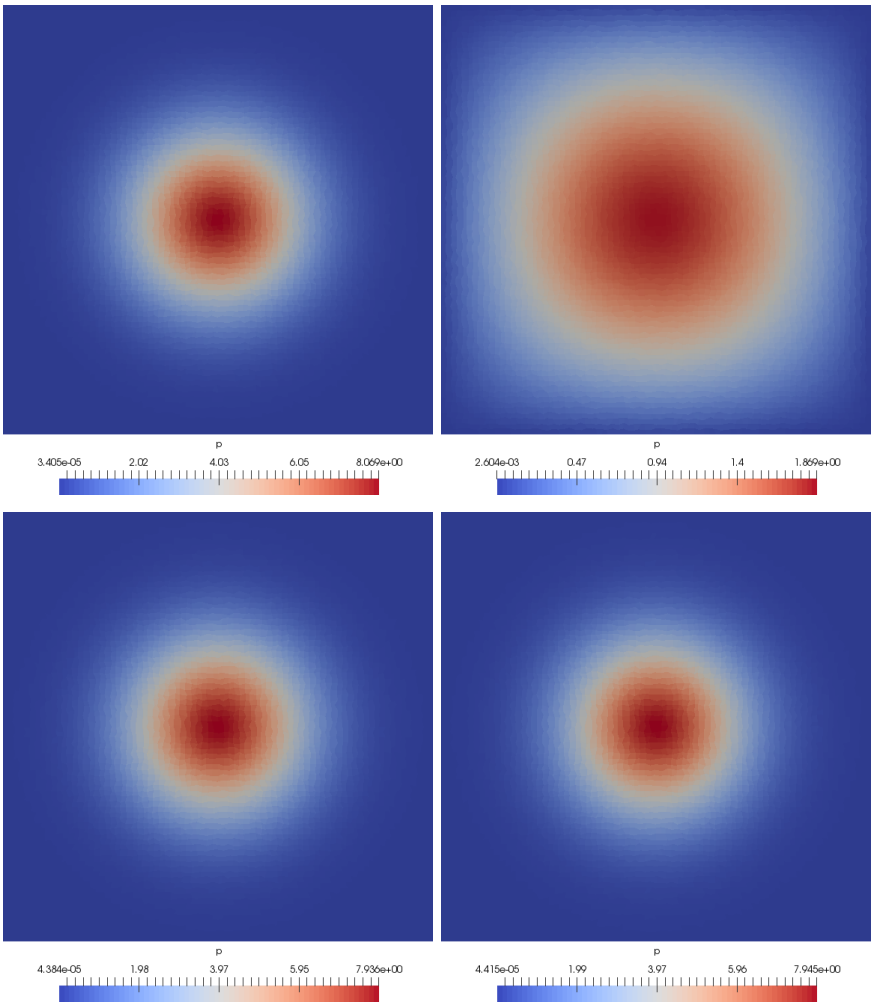

Figure 5.5 – **P_3 model**. Representation of the first variable when $\varepsilon = 10^{-3}$ for the test case 5-4.4. Top left: limit solution. Top right: DG scheme with 10 basis functions per cell. Bottom left: DG scheme with 30 basis functions per cell. Bottom right: TDG scheme with only 12 basis functions per cell. Good behavior of the numerical solution illustrates the AP property.

5-4.5.1 Trefftz discontinuous Galerkin method

The P_1 model.

For the TDG method, one must choose the directions of the basis functions in Ω_1 since $\sigma_a > 0$. As we will see, the choice of directions at the interface plays an important role to correctly capture the boundary layers. In particular, it seems essential to locally get the one dimensional direction perpendicular to the interface associated with the boundary layer. Therefore, we make the special choice of directions (5.24) at the interface in Ω_1. Such directions are well adapted if one considers the one dimensional problem at the interface. A graphical illustration of the adaptive directions at the interface is provided on the right of Figure 5.6.

To show why it can be challenging for standard schemes to capture boundary layers, we compare the TDG method with the standard DG method on a coarse 20×20 mesh. More precisely, we consider the following cases

- The DG method with constant basis functions only (= finite volume) for a total of 3 basis functions per cell.

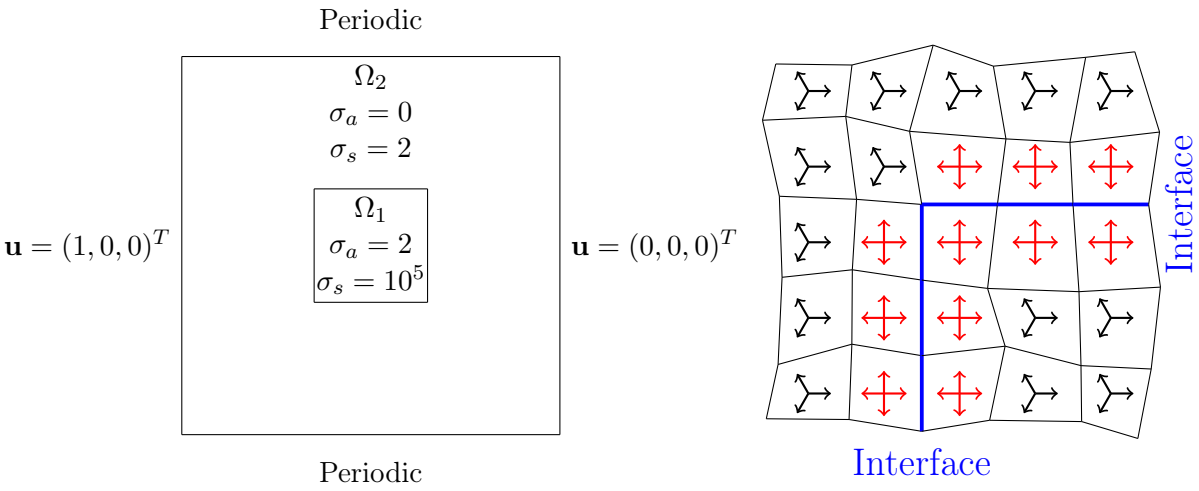

Figure 5.6 – On the left: Domain and boundary condition for the two dimensional boundary layers test. On the right: representation of adaptive directions at the interface. In this example: the 3 equi-distributed directions (5.23) in each cell except at the interface where the directions are locally adapted into (5.24).

- The DG method with affine basis functions (that is 1, x, y) for a total of 9 basis functions per cell.
- The TDG method with the exponential and polynomial basis functions (5.6)-(5.8) depending on the 3 directions (5.23), for a total of 3 basis functions per cell, and on the 4 directions (5.24) at the interface.
- The TDG method with the the exponential and polynomial basis functions (5.6)-(5.8) depending on the 5 directions (5.25), for a total of 5 basis functions per cell, and on the 4 directions (5.24) at the interface.

The reference solution represented in Figure 5.7 is calculated on a 200×200 mesh with the TDG method using 5 basis functions per cell except at the interface where the four adaptive directions (5.24) are used.

In Figure 5.7, we represent the first variable. One observes that the boundary layer is not correctly captured by the DG scheme. The approximation given by the TDG scheme seems more accurate.

In Figure 5.8, we take a one dimensional cut at $y = 0.5$ to compare more precisely the numerical results. The graphic on the left shows that, with less basis functions, the TDG method gives a better approximation than the DG method. Our interpretation is that it is because the boundary layer is correctly captured by TDG but poorly captured by DG. This will be confirmed by the enrichment approach of Section 5-4.5.2.

The graphic on the right of Figure 5.8 illustrates why it is very important to use the directions (5.24) at the interface to obtain a satisfactory discretization of the boundary layer on a coarse meshes. We consider the TDG method with 5 basis functions per cell and compare two cases

- In the first one, the directions are (5.25) in all cells of Ω_1.
- In the second one, the directions are (5.25) except at the interface where the directions (5.24) are used.

The graphic shows that the TDG method gives a non correct approximation with only the directions (5.25). However, if one locally adapts the directions at the interface, the TDG method recovers a very good accuracy. Once again, our interpretation is that it is because the boundary layer is correctly captured with these parameters.

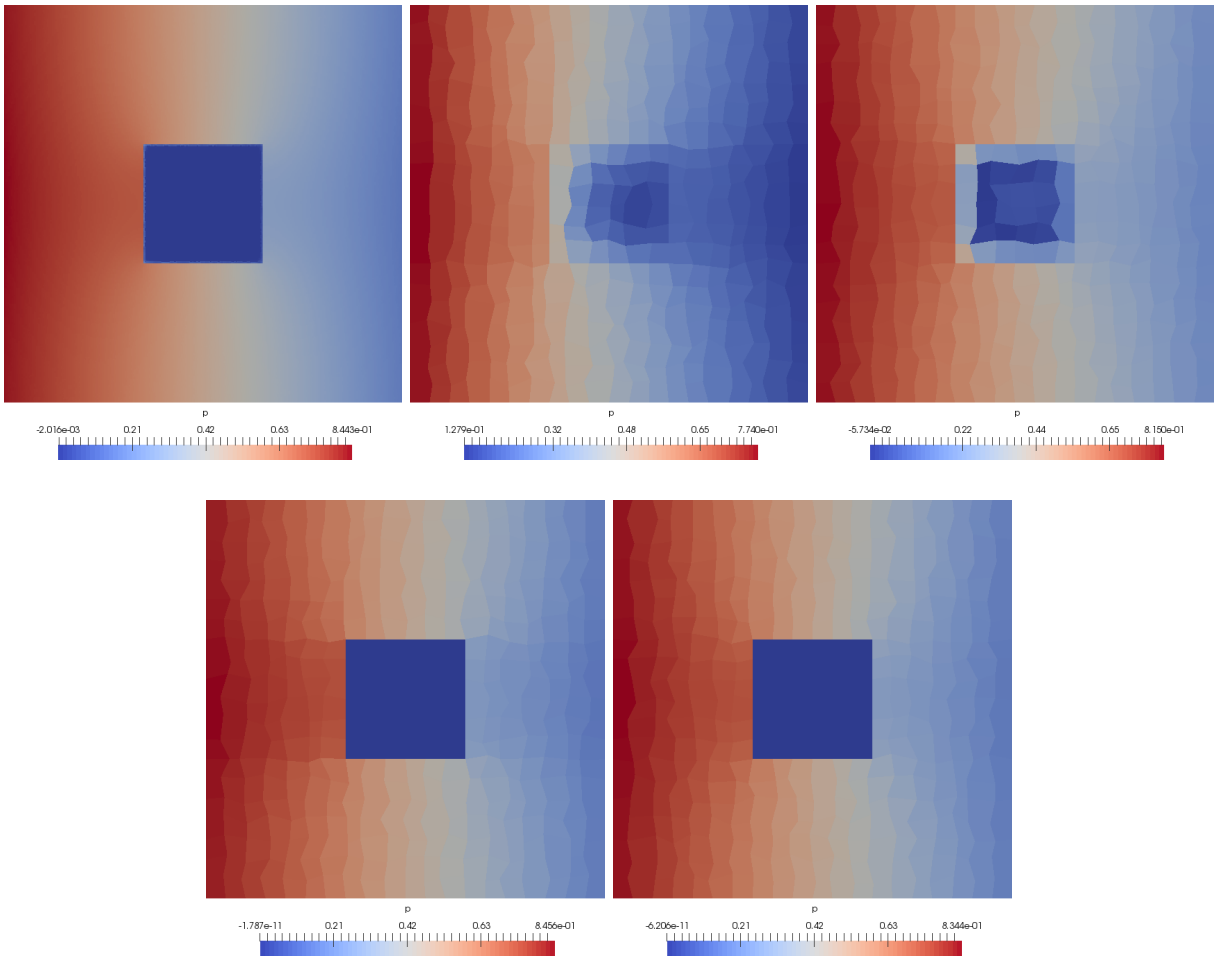

Figure 5.7 – **P_1 model**. Representation of the first variable for the test case 5-4.5. Top left: reference solution. Top center: DG scheme with 3 basis functions per cell. Top right: DG scheme with 9 basis functions per cell. Bottom left: TDG scheme with 3 basis functions per cell. Bottom right: TDG scheme with 5 basis functions per cell. For the TDG scheme, the directions at the interface in Ω_1 are locally adapted into the 4 directions (5.24).

As we have seen in Figure 5.4, one possible drawback of the Trefftz method is the deterioration of the condition number. This is particularly true here since stiff exponentials are used in the basis functions. The Figure 5.9 compares the condition number obtained with the Trefftz method with and without preconditioning where the preconditioner considered here is a simple one diagonal on the left and on the right. The Figure 5.9 shows that the condition number is significantly improved by using this simple preconditioner. Therefore, studying efficient preconditioner in the case of the Trefftz method can be an interesting perspective for future research.

The P_3 model.

For this particular numerical test, there is no visible difference between the solutions to the P_1 and P_3 models. However, since the basis functions differ from the P_1 to the P_3 models, it is still interesting to perform the boundary layer test on the P_3 model.

The reference solution is calculated on a 200×200 random mesh with the 3 directions (5.23) and adaptive directions (5.24) at the interface. We do not calculate the reference solution with 5 basis functions per cell, as we did for the P_1 model, due to some conditioning issue. We compare the following cases on a coarse 20×20 mesh

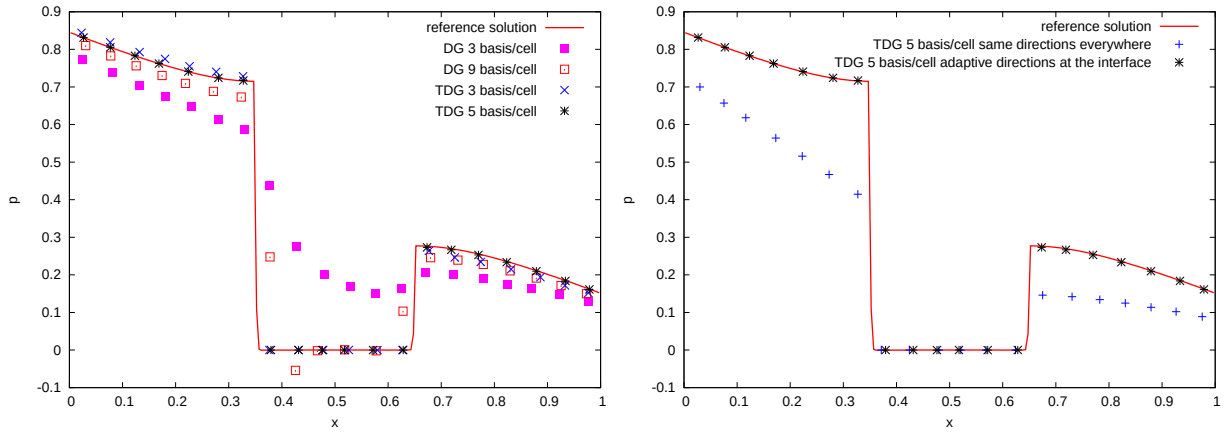

Figure 5.8 – **P_1 model**. One dimensional representation of the variable p at $y = 0.5$ for the test case 5-4.5. Left: comparison between the DG method with 3 basis/cell, the DG method with 9 basis/cell, the TDG method with 3 basis/cell and the TDG method with 5 basis/cell. In both cases for the TDG method, the directions at the interface in Ω_1 are locally adapted into the 4 directions (5.24). Right: comparison between the TDG method with directions (5.25) only and the TDG method where the directions at the interface in Ω_1 are locally adapted into the 4 directions (5.24).

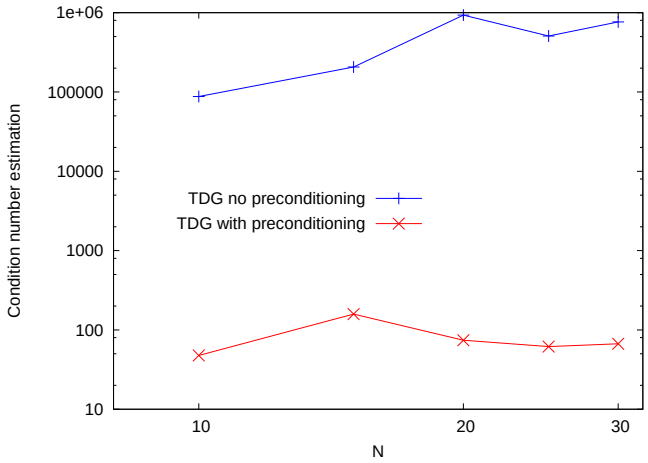

Figure 5.9 – **P_1 model**. Comparison of the condition number between the TDG method with no preconditioner and the TDG method with one simple preconditioner diagonal on the left and on the right.

- The DG method with constant basis functions only (= finite volume) for a total of 10 basis functions per cell.
- The DG method with affine basis functions (that is 1, x, y) for a total of 30 basis functions per cell.
- The TDG method with the basis functions (5.20)-(5.21) depending on the 3 directions (5.23), for a total of 12 basis functions per cell, and on the 4 directions (5.24) at the interface.

- The TDG method with the basis functions (5.20)-(5.21) depending on the 5 directions (5.25), for a total of 20 basis functions per cell, and the 4 directions (5.24) at the interface.

The results given in Figure 5.10 are very similar to the P_1 case. One notices a better approximation of the solution for the TDG method with less degrees of freedom compared to the standard DG scheme.

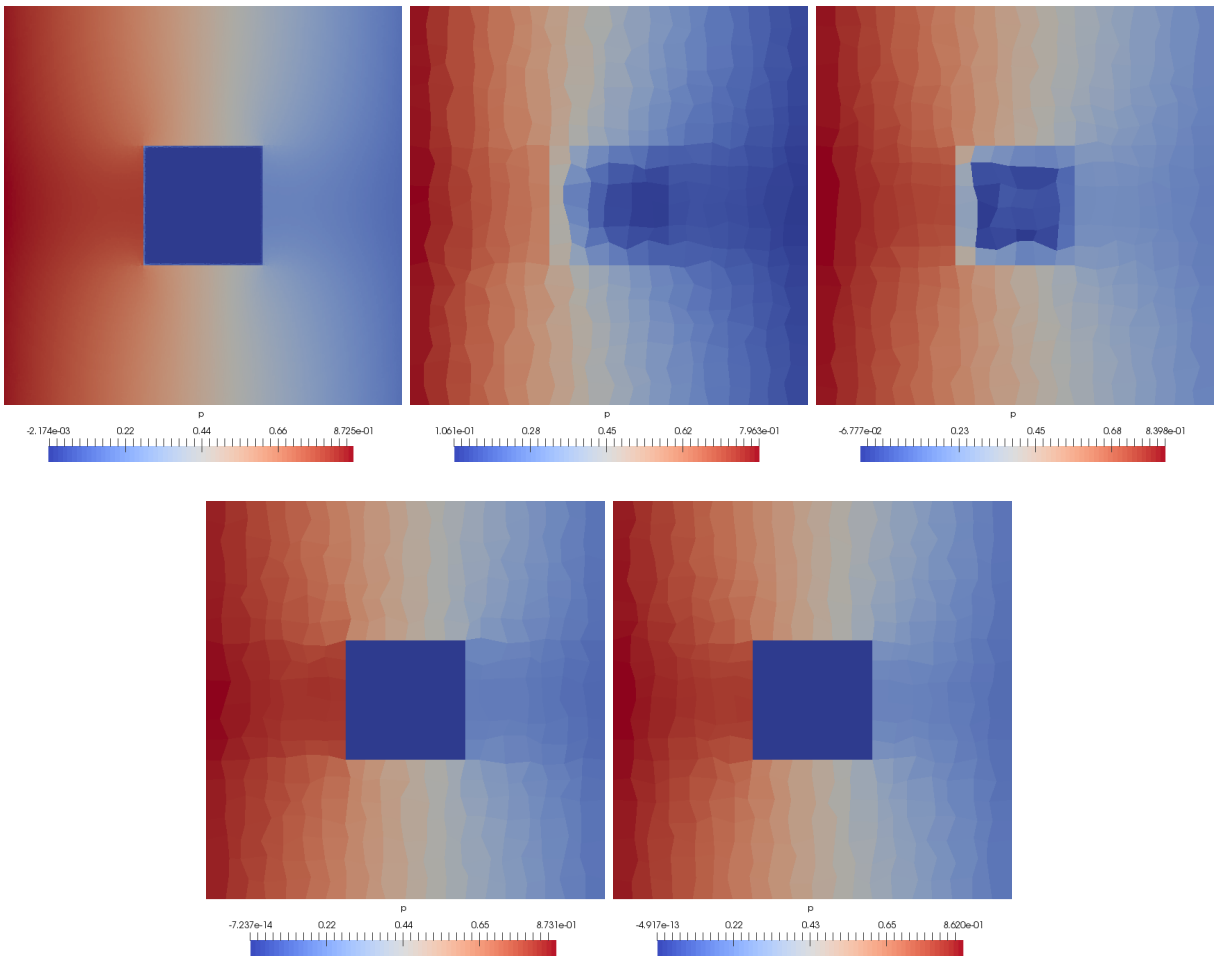

Figure 5.10 – **P_3 model**. Representation of the first variable for the test case 5-4.5. Top left: reference solution. Top center: DG scheme with 10 basis functions per cell. Top right: DG scheme with 30 basis functions per cell. Bottom left: TDG scheme with 12 basis functions per cell. Bottom right: TDG scheme with 20 basis functions per cell. For the TDG scheme, the directions at the interface are locally adapted into the 4 directions (5.24).

5-4.5.2 Enriched discontinuous Galerkin method

Numerical tests with boundary layers are well adapted to consider enrichment strategy. We consider the enriched discontinuous Galerkin method which consists to start from a standard DG basis and add locally (*i.e.* in the boundary layer) some exponential solutions. In this example, we apply the enrichment strategy to the stationary two dimensional P_1 model.

In the previous examples, the directions were adapted without assuming any *a priori* physical knowledge of the solution. Indeed, the directions (5.24) were chosen such that they could capture increasing or decreasing boundary layers at the interface. But it is also possible to use the physical

knowledge of the user and consider only one or two directions in the boundary layer. Here for example, one can assume that the local variation of the boundary layer is known to reduce the number of basis functions added. More precisely

- For the left interface ($x = 0.35$, $0.35 \leq y \leq 0.65$), the boundary layer is a decreasing function with respect to x so we add the direction $\mathbf{d} = (-1, 0)^T$.
- For the right interface ($x = 0.65$, $0.35 \leq y \leq 0.65$), the boundary layer is an increasing function with respect to x so we add the direction $\mathbf{d} = (1, 0)^T$.
- For the bottom interface ($y = 0.35$, $0.35 \leq x \leq 0.65$), the boundary layer is a decreasing function with respect to y so we add the direction $\mathbf{d} = (0, -1)^T$.
- For the top interface ($y = 0.65$, $0.35 \leq x \leq 0.65$), the boundary layer is an increasing function with respect to y so we add the direction $\mathbf{d} = (0, 1)^T$.

Note that we add at most one basis function in the cells except at the corners of Ω_1 where we add two basis functions. For a graphical illustration of the procedure, see Figure 5.11.

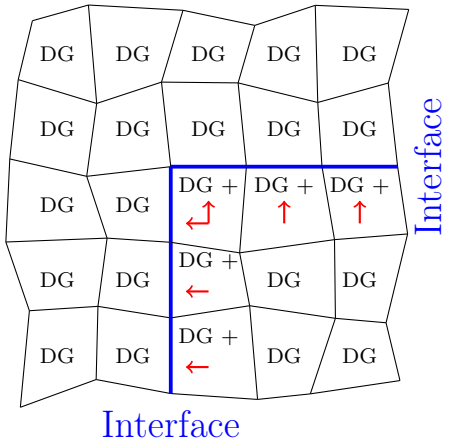

Figure 5.11 – Representation of the enrichment strategy. In this example, basis functions corresponding to the discontinuous Galerkin method are used in all the cells. In the boundary layer one or two exponential solutions (5.6) are locally added. The arrows represent the directions of these solutions.

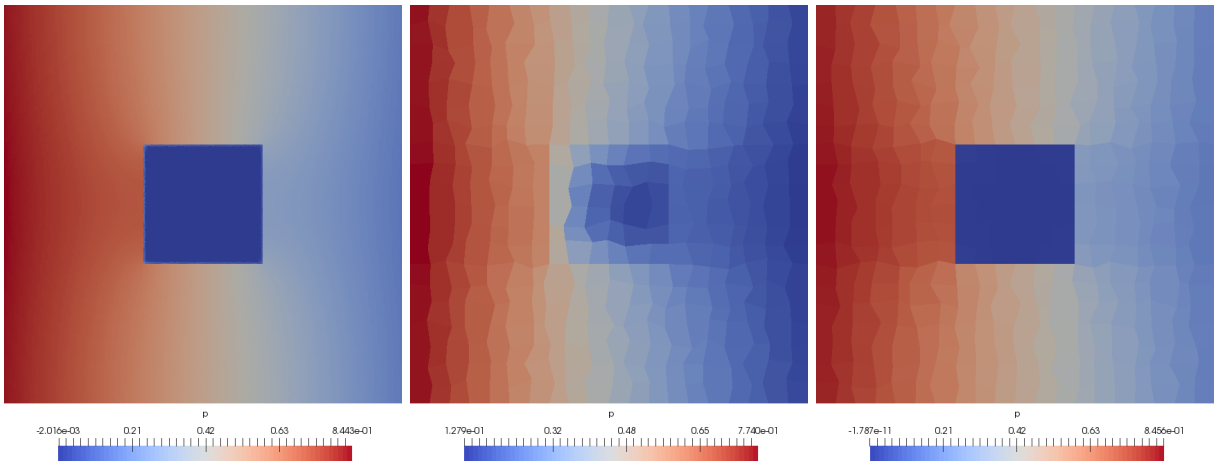

Figure 5.12 – **P_1 model.** On the left: reference solution. Center: DG method with 3 basis functions per cell. On the right: DG method with 3 basis functions per cell where some exponential solutions are locally added in the boundary layer.

In Figure 5.12, we compare the two following cases on a coarse 20×20 mesh

- DG scheme with constant basis functions only for a total of three basis functions per cell.
- Same DG scheme (constant basis functions) except at the interface where we locally add one or two exponential solutions as describe above.

The reference solution is the same we used before for the P_1 model. One sees that the approximation is much better for the enriched method.

5-4.6 A lattice problem

5-4.6.1 Comparison between the TDG and DG method

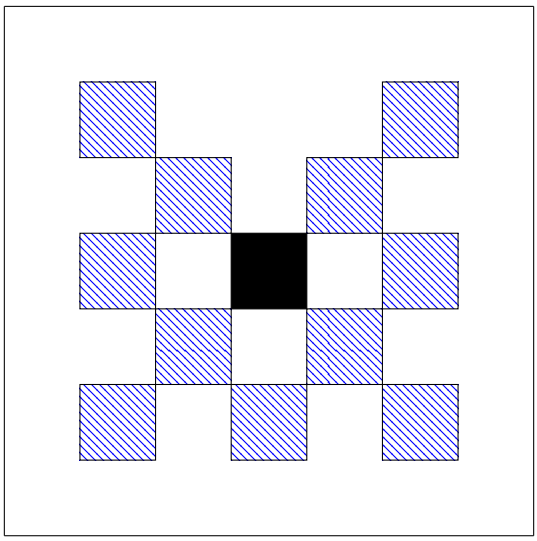

Figure 5.13 – Domain for the lattice problem 5-4.6.

We consider a lattice problem [BDF15, Bru02, Her16, SFL11]. The spatial domain $\Omega_S = [0,7] \times [0,7]$ is represented in Figure 5.13 and we take $T = 3.2$. The white area is a purely scattering region while the striped and black areas are purely absorbing regions. Additionally, the black region contain a source of particles. More precisely, let Ω_c be the union of the eleven striped squares and the black square in Figure 5.13, then one has

$$\begin{cases} \sigma_a(\mathbf{x}) = 10, & \sigma_s(\mathbf{x}) = 0, \quad \text{if } \mathbf{x} \in \Omega_c, \\ \sigma_a(\mathbf{x}) = 0, & \sigma_s(\mathbf{x}) = 1, \quad \text{else.} \end{cases}$$

Note that for some authors $\sigma_a = 0$, $\sigma_s = 1$, in the central region [Bru02, Her16] while other authors take $\sigma_a = 10$, $\sigma_s = 0$ [BDF15, SFL11]. These two choices give similar numerical results and we consider here the second option. We recall that Friedrichs systems with a source term read

$$\Big(\partial_t + A_1\partial_x + A_2\partial_y\Big)\mathbf{u}(t,\mathbf{x}) = -R\mathbf{u}(t,\mathbf{x}) + \mathbf{f}(\mathbf{x}). \tag{5.28}$$

In this example, the source $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$ is contained in the black region

$$\begin{cases} \mathbf{f}(\mathbf{x}) = \sigma_a(\mathbf{x}) \times \mathbf{e}_1, & \text{if } \mathbf{x} \in [3,4]^2, \\ \mathbf{f}(\mathbf{x}) = \mathbf{0}, & \text{else,} \end{cases}$$

where $\mathbf{e}_1 = (1,0,...,0)^T \in \mathbb{R}^m$. For the basis functions which depend on the source of particles, we use the methodology described in Section 2-2.3 of Chapter 2. That is, we add the basis

function $\mathbf{v}_f = R^{-1}\mathbf{f}$ in the central region

$$\begin{cases} \mathbf{v}_f(\mathbf{x}) = \mathbf{e}_1, & \text{if } \mathbf{x} \in [3,4]^2, \\ \mathbf{v}_f(\mathbf{x}) = \mathbf{0}, & \text{else.} \end{cases}$$

We consider vacuum boundaries, that is we impose $\mathbf{u} = \mathbf{0}$ at the boundaries of the domain.

Finally, we consider the possibility of using the time exponential solutions (4.69) in the TDG scheme.

The P_1 model.

The numerical results obtained for the P_1 model are displayed in Figure 5.14. The reference solution is computed with the DG method with affine basis functions for a total of 9 basis functions per cell on a 280×280 random mesh with $dt = 0.01$. We compare the DG and TDG methods on a 140×140 mesh with $dt = 0.02$. We consider the following cases

- The DG method with constant basis functions only for a total of 3 basis functions per cell.
- The DG method with affine basis functions (that is 1, x, y) for a total of 9 basis functions per cell.
- The TDG method with the basis functions (5.6)-(5.8) depending on the 5 directions (5.25), for a total of 5 basis functions per cell (plus one in the black region).
- The TDG method with the basis functions (5.6)-(5.8) depending on the 5 directions (5.25) and the time dependent solutions (4.69), for a total of 8 basis functions per cell (plus one in the black region).

Figure 5.14 shows that the DG method with only constant basis functions is too diffusive. However, if one increases the number of basis functions and considers affine basis functions, the DG method recovers a very good accuracy. From Figure 5.14, one also notices that the TDG method with 5 directions and only stationary basis functions seems too diffusive. Adding the time dependent basis functions (4.69) to the TDG method allow to recover a good accuracy similar to the affine DG method.

The P_3 model.

The comments are very similar for the P_3 model. Figure 5.15 represents the numerical results obtained for the P_3 model. The reference solution is computed with the DG method with affine basis functions for a total of 30 basis functions per cell on a 280×280 random mesh with $dt = 0.01$. We compare the DG and TDG methods on a 140×140 mesh with $dt = 0.02$. More precisely, we consider the following cases

- The DG method with constant basis functions only for a total of 10 basis functions per cell.
- The DG method with affine basis functions (that is 1, x, y) for a total of 30 basis functions per cell.
- The TDG method with the basis functions (5.20)-(5.21) depending on the 3 directions (5.23), for a total of 12 basis functions per cell (plus one in the black region).
- The TDG method with the basis functions (5.20)-(5.21) depending on the 3 directions (5.23) and the time dependent solutions (4.69), for a total of 22 basis functions per cell (plus one in the black region).

As for the P_1 model, Figure 5.15 illustrates that the DG method recovers a good accuracy when using affine basis functions. For the TDG method, considering only 3 stationary basis functions seems too diffusive. Nevertheless, if one adds the time dependent basis functions (4.69), the TDG method recovers a good accuracy similar to the affine DG method.

In particular, a benefit of the TDG method compared to the standard DG method is that it uses less basis functions to recover a good approximation of the numerical solution. However,

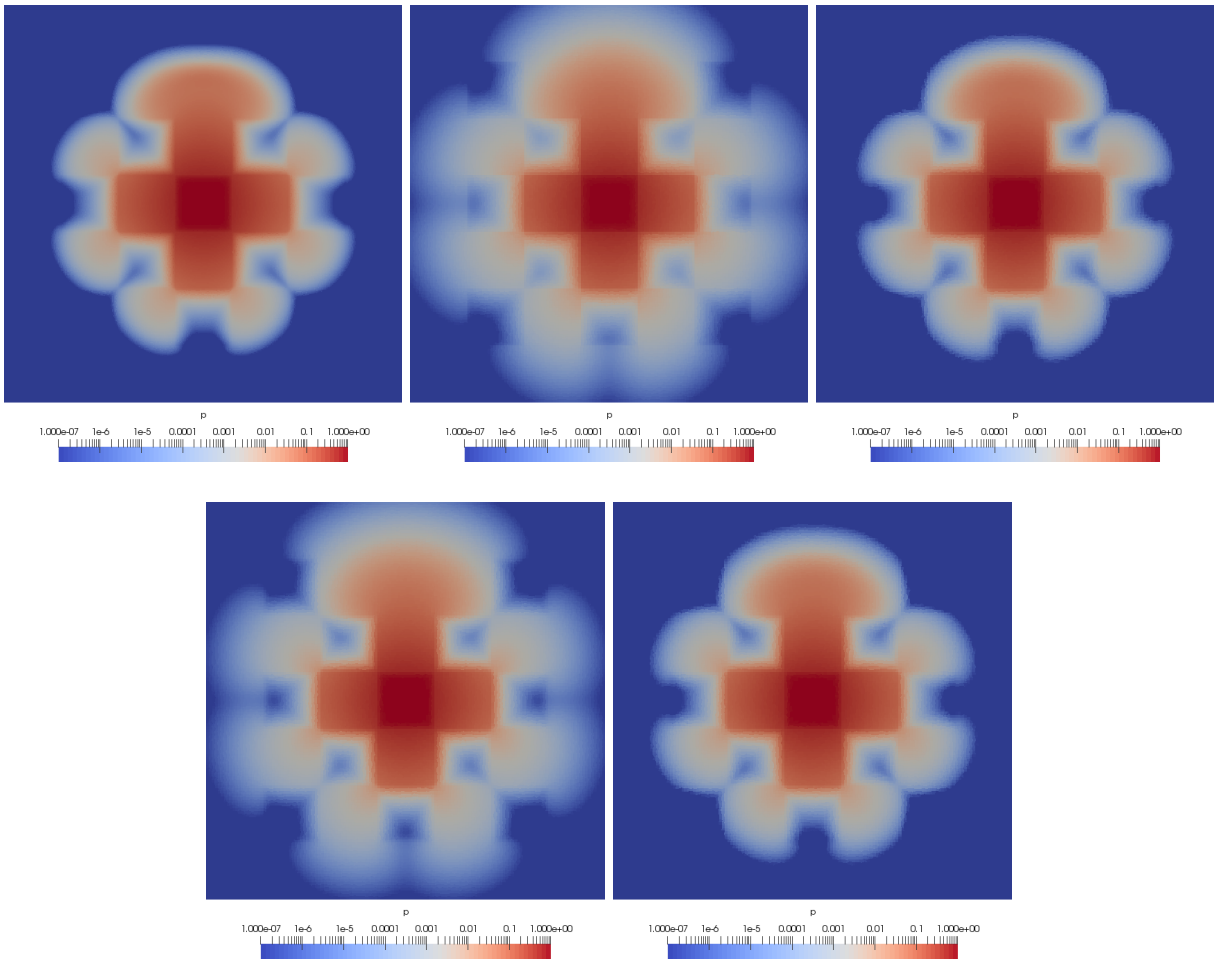

Figure 5.14 – **P_1 model**. Representation of the first variable for the test case 5-4.6. Top left: reference solution. Top center: DG scheme with 3 basis functions per cell. Top right: DG scheme with 9 basis functions per cell. Bottom left: TDG scheme with about 5 stationary basis functions per cell. Bottom right: TDG scheme with about 8 basis functions per cell (stationary and time dependent). Logarithmic scale.

as we will see in the next section, the TDG method may suffer from conditioning issue when considering stationary and time dependent basis functions on fine meshes.

Finally note that, both for the P_1 and P_3 model, the numerical results are similar to those obtained in [Bru02, BDF15].

5-4.6.2 The TDG method with other time dependent basis functions

In this section, we study the TDG method applied to the P_1 model with others time dependent basis functions. We consider the stationary basis functions (5.20)-(5.21) and the time dependent basis functions (5.9)-(5.16), (4.66), (4.69). For the basis functions (4.69) we make the arbitrary choice $\alpha = \sigma_t$ which gives

$$\mathbf{v}(t,\mathbf{x}) = \begin{pmatrix} \sqrt{\sigma_t(1+\varepsilon)} \\ -\sqrt{\varepsilon(\sigma_a+\sigma_t)}\mathbf{d} \end{pmatrix} e^{\frac{1}{c}\sqrt{3\varepsilon(\sigma_a+\sigma_t)\sigma_t(1+\varepsilon)}\mathbf{d}^T\mathbf{x}+\sigma_t t}, \tag{5.29}$$

with $\mathbf{d} = (\cos\theta, \sin\theta)^T \in \mathbb{R}^2$. More precisely, we consider the following cases

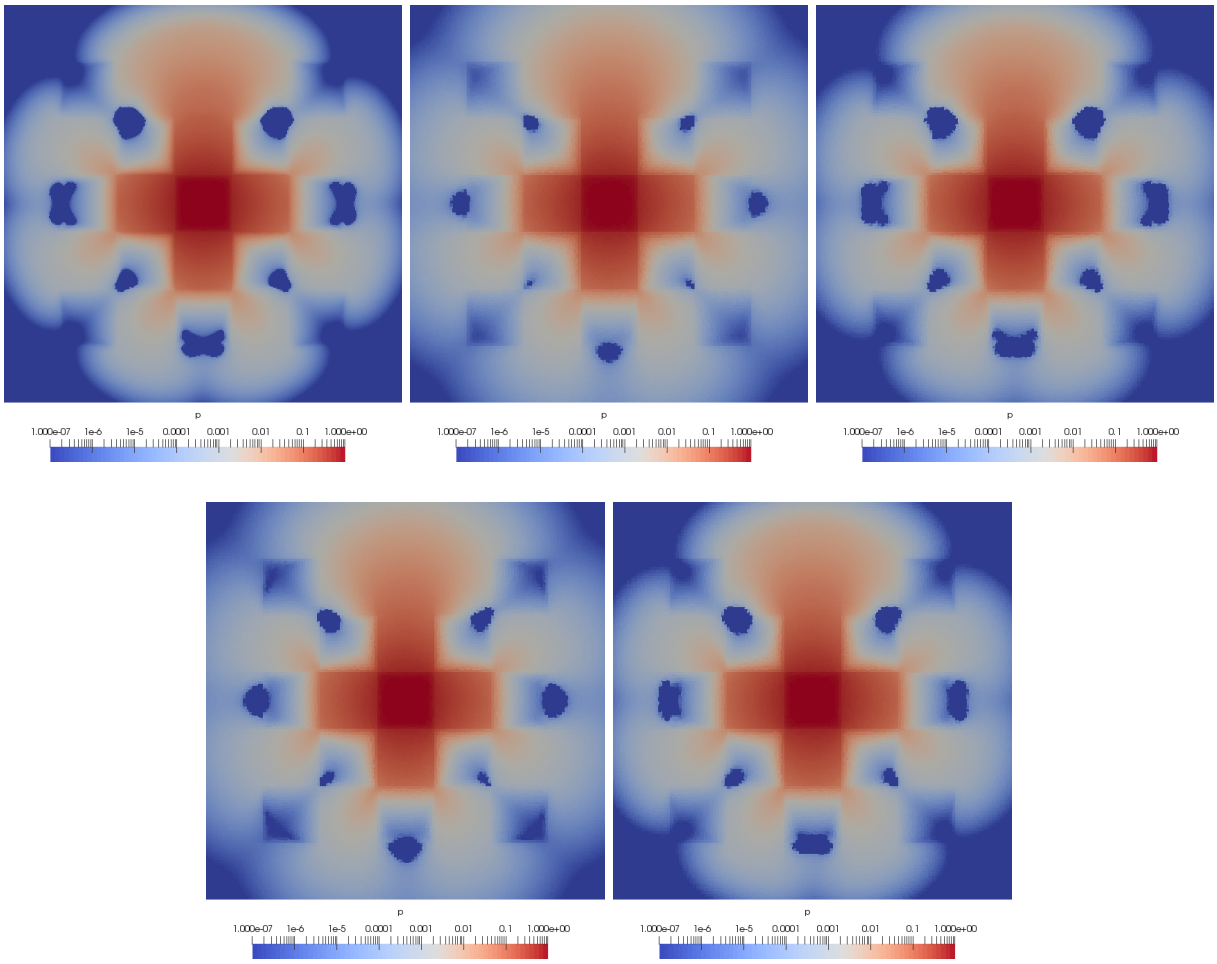

Figure 5.15 – ***P*$_3$ model**. Representation of the first variable for the test case 5-4.6. Top left: reference solution. Top center: DG scheme with 10 basis functions per cell. Top right: DG scheme with 30 basis functions per cell. Bottom left: TDG scheme with about 12 stationary basis functions per cell. Bottom right: TDG scheme with about 22 basis functions per cell (stationary and time dependent). Logarithmic scale.

- **Case 1.** The stationary basis functions (5.6)-(5.8) only with the 3 directions (5.23) for a total of about 3 basis functions per cell.
- **Case 2.** The stationary basis functions (5.6)-(5.8) with the 3 directions (5.23) and the time dependent solutions (4.66) for a total of about 6 basis functions per cell.
- **Case 3.** The stationary basis functions (5.6)-(5.8) and the time dependent solutions (5.9)-(5.16) with the 3 directions (5.23) for a total of about 9 basis functions per cell.
- **Case 4.** The stationary basis functions (5.6)-(5.8) and the time-dependent solutions (5.29) with the 3 directions (5.23) for a total of about 6 basis functions per cell.
- **Case 5.** The stationary basis functions (5.6)-(5.8) and the time-dependent solutions (5.29) with the 4 directions (5.24) for a total of about 8 basis functions per cell.

Remark 5.12 (Case 3: polynomial solutions when $\sigma_a = 0$)**.** When considering the Case 3, one has 9 exponential basis functions when $\sigma_a > 0$. However, when $\sigma_a = 0$ the basis functions became polynomials. It is not clear how to choose those polynomials since both the stationary and time dependent exponentials may degenerate to the same solutions (at least if we follow the

procedure given in the proof of Proposition 5.6). We make the following arbitrary choice and consider a total of 7 polynomial basis functions

$$\mathbf{v}_1(\mathbf{x}) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2(\mathbf{x}) = \begin{pmatrix} \frac{\sigma_s}{\varepsilon} x \\ -\frac{c}{\sqrt{3}} \\ 0 \end{pmatrix}, \quad \mathbf{v}_3(\mathbf{x}) = \begin{pmatrix} \frac{\sigma_s}{\varepsilon} y \\ 0 \\ -\frac{c}{\sqrt{3}} \end{pmatrix},$$

$$\mathbf{v}_4(t,\mathbf{x}) = \begin{pmatrix} -\frac{2}{\sqrt{3}}\sqrt{\varepsilon}c^2 - \sqrt{3\varepsilon}\sigma_t^2 x^2 - \frac{2}{\sqrt{3\varepsilon}}c^2\sigma_t t \\ 2\sqrt{\varepsilon}c\sigma_t x \\ 0 \end{pmatrix}, \quad \mathbf{v}_5(t,\mathbf{x}) = \begin{pmatrix} -\frac{2}{\sqrt{3}}\sqrt{\varepsilon}c^2 - \sqrt{3\varepsilon}\sigma_t^2 y^2 - \frac{2}{\sqrt{3\varepsilon}}c^2\sigma_t t \\ 0 \\ 2\sqrt{\varepsilon}c\sigma_t x \end{pmatrix},$$

$$\mathbf{v}_6(t,\mathbf{x}) = \begin{pmatrix} -\frac{2}{\sqrt{3}}\sqrt{\varepsilon}c^2 y - \sqrt{3\epsilon}\sigma_t^2 x^2 y - \frac{2}{\sqrt{3\varepsilon}}c^2\sigma_t t y \\ 2\sqrt{\varepsilon}c\sigma_t xy \\ \sqrt{\varepsilon}c\sigma_t x^2 + \frac{2}{3\sqrt{\varepsilon}}c^3 t \end{pmatrix}, \quad \mathbf{v}_7(t,\mathbf{x}) = \begin{pmatrix} -\frac{2}{\sqrt{3}}\sqrt{\varepsilon}c^2 x - \sqrt{3\epsilon}\sigma_t^2 x y^2 - \frac{2}{\sqrt{3\varepsilon}}c^2\sigma_t t x \\ \sqrt{\varepsilon}c\sigma_t y^2 + \frac{2}{3\sqrt{\varepsilon}}c^3 t \\ 2\sqrt{\varepsilon}c\sigma_t xy \end{pmatrix}.$$

Here the functions $\mathbf{v}_1(\mathbf{x}), \mathbf{v}_2(\mathbf{x})$ and $\mathbf{v}_3(\mathbf{x})$ can be seen as the limit of the three stationary basis functions (5.6). The time dependent polynomials are taken from (5.16) to assure that all the components have a dependence in time in at least one basis functions. ●

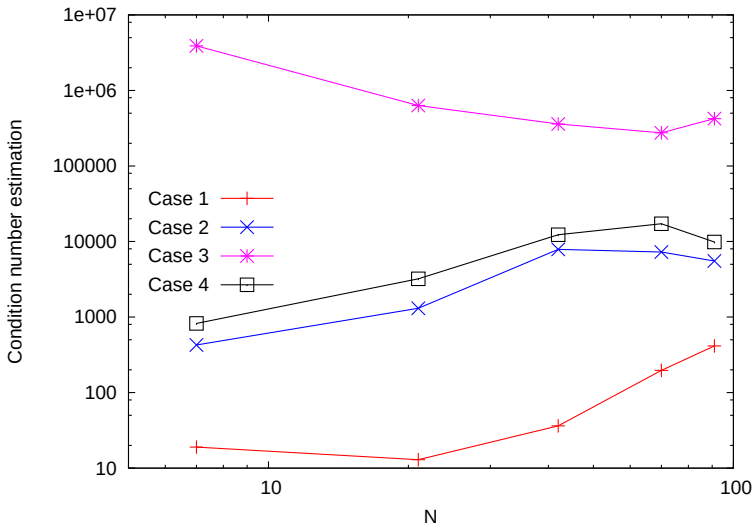

Figure 5.16 – Estimation of the condition number for the cases 1 to 4. Logarithmic scale.

In Figure 5.16, we compare an estimation of the condition number for the cases 1 to 4 on random meshes. The estimation is provided using the AztecOO package of the Trilinos library [HBH+03]. Figure 5.16 illustrates that adding time dependent basis functions deteriorate the conditioning of the mass matrix. One notices that the temporal exponentials (4.66) (Case 2) are the time dependent functions which give the better (or the least bad) result in term of the condition number.

In Figure 5.17, we compare the cases 1 to 5 on a 70×70 mesh. To prevent the condition number from growing too fast, we consider a mesh which is *not* random. One sees that all the time dependent basis functions reduce the diffusion. Compared to Case 2, one notices that the diffusion is lower for cases 3 to 5 but some oscillations appear. For the basis functions (5.29) (Cases 4 and 5), the choice of directions seems important. Indeed, with only the 3 directions (5.23) (Case 4), the numerical solution is highly asymmetric. Considering the 4 directions (5.24)

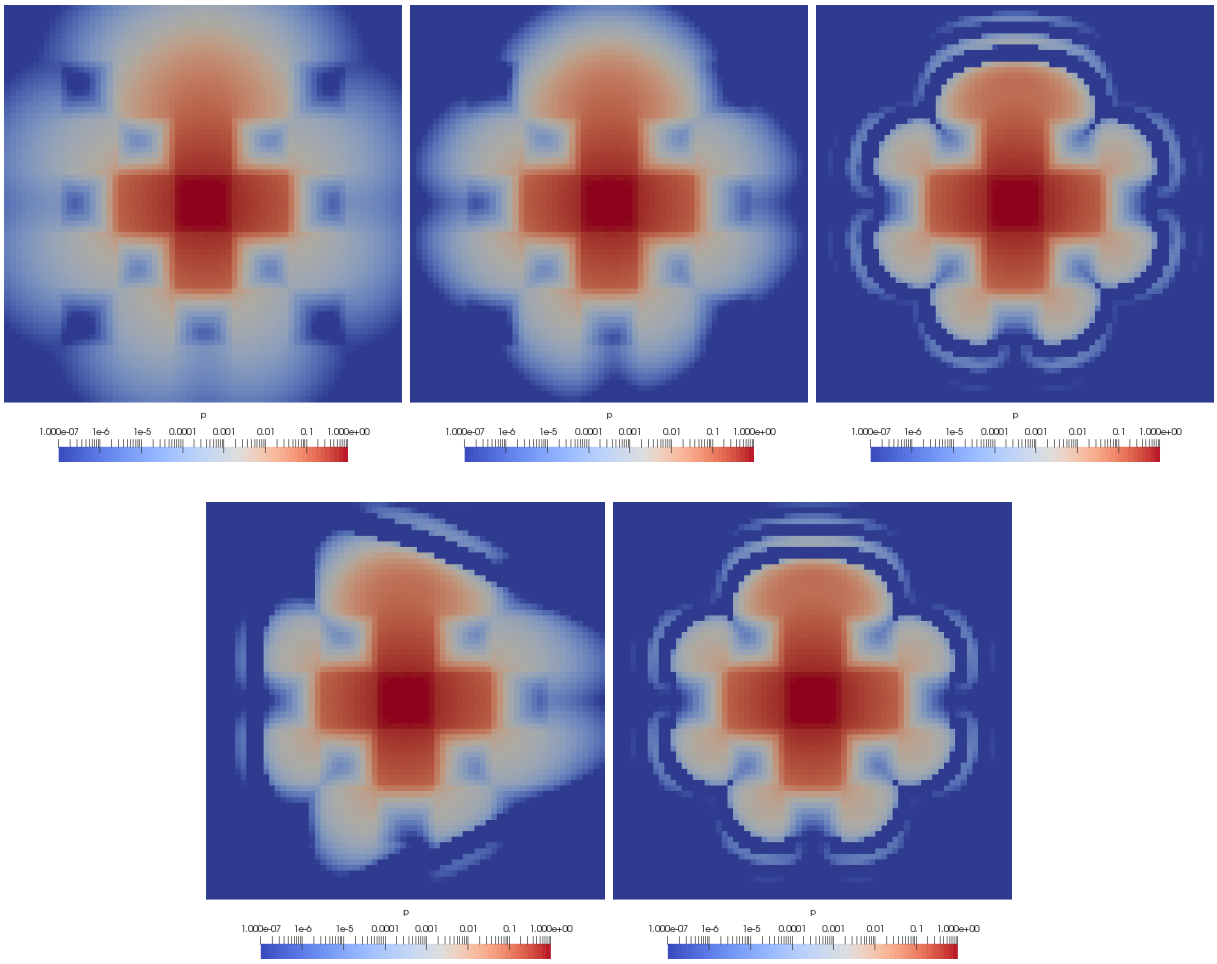

Figure 5.17 – **P_1 model**. Representation of the first variable for the test case 5-4.6. Cases 1 to 5. The cases are numbered from from left to right and top to bottom (top left: Case 1, top center: Case 2...). Logarithmic scale.

(Case 5), fix this issue. Note that Case 3 also considers the 3 directions (5.23) without getting the asymmetric result of Case 4.

Chapter 6

An asymptotic preserving multidimensional ALE method for a system of two compressible flows coupled with friction

This chapter is taken from a published article [PLM18].

S. Del Pino[1], E. Labourasse[1], G. Morel[1,2]

[1] CEA, DAM, DIF, F-91297 Arpajon, France

[2] Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France

Abstract. We present a multidimensional asymptotic preserving scheme for the approximation of a mixture of compressible flows. Fluids are modelled by two Euler systems of equations coupled with a friction term. The asymptotic preserving property is mandatory for this kind of model, to derive a scheme that behaves well in all regimes (*i.e.* whatever the friction parameter value is). The method we propose is defined in ALE coordinates, using a Lagrange plus remap approach. This imposes a multidimensional definition and analysis of the scheme.

Keywords: Compressible gas dynamics, multi-fluid, finite volumes, unstructured meshes, asymptotic preserving, arbitrary-Lagrangian-Eulerian (ALE)

6-1 Introduction

A multifluid model is a model for a fluid mixture for which each fluid is described by its own full set of variables (for instance density, velocity and energy). The model is generally closed in a way that defines interactions between the constituents, depending on the involved physics. These models are widely used in different communities. One very popular model of this kind is the Baer-Nunziato model [BN86] for deflagration-to-detonation transition of reactive flows. Many numerical methods to approximate this model have been designed, we refer to a few of them [SA99, CGHS02, CHSN13, AD14, ACCG14]. Scannapieco and Cheng [SC02] also derive similar kind of model for turbulent flows and apply it to describe a mixing zone driven by Rayleigh-Taylor or Richtmyer-Meshkov instabilities [CS05]. Such kind of model is also used in plasma physics to account for plasmas collision or Non-Local-Thermodynamic-Equilibrium (NLTE) Ion-Electron interactions [DMP98, Sen14]. Although all the analysis done in this paper can be applied to any of the former models, we are in particular interested in the latter applica-

tion. In this context, multifluid models are a good approximation, in particular to account for the collision of two ion populations, each of them being at LTE. However, to our knowledge, these models are never used for plasma collisions. The reason for this is stated by R. Sentis in [Sen14]: "*The [...] system may be quite difficult to solve in two- or three-dimensional geometry, especially in the case when the friction coefficient [...] is large [...].*" Consequently, a simplified model is in general prefered, in which the velocity gap between the two fluids is modelled by a diffusion process on the concentrations. Unfortunately, it implies empirical closures and exhibits bad behaviour at high temperatures (when the coupling between the ion populations is weak).

In the following, we explain why the classical schemes for the multifluid system fail to capture the strong coupling limit. It is in fact inherent to this kind of model and relies to the asymptotic preserving (AP) property [Gos13, GT02, Jin10, JL91] in the high friction regime or infinite friction regime. In the former regime, the fluids interpenetration follows a diffusion law. In the latter one, the mixture evolves as a single fluid, see (6.4)–(6.5). If no attention is paid to these regimes, the scheme will fail to capture it at a reasonnable calculation cost. Some authors [CDW99, CDV07, Ena07] propose an asymptotic discretization for the system (6.1) in 1D in the Eulerian frame — multidimensional calculations being achieved by means of directional splitting —, but no asymptotic preserving scheme has been yet published for 2D unstructured meshes for this model. A similar ALE formalism is used to treat multifluid interaction in [CS12]. Authors use the Compatible Hydro scheme [CBS98] and do not analyze the asymptotic preserving property since they mainly focus on the physics of the coupling.

In this paper, we propose a multidimensional scheme to approximate solutions of this kind of model, written in (6.1), which captures accurately the asymptotic regime. We want our scheme to be able to deal with Arbitrary-Lagrange-Euler (ALE) frame and unstructured meshes in order to properly handle highly deformed calculation domains. Even for simpler models, only few unstructured asymptotic preserving schemes have been developed (refer for instance to Berthon and Turpault [BT11] and Franck *et al.* [BDF12, Fra12]). The scheme we propose in Section 6-4 has connections with [Fra14, FM16], where an Euler with friction system is studied in the limit of high friction for long time, providing a different kind of scaling. So, the proposed scheme is not a direct extension of [Fra14] to the bi-fluid case. The scheme presented in this work is split into two steps. In the first step we solve two Euler systems of equations coupled by friction. Since each fluid has its own velocity, the Lagrangian mesh of each fluid will evolve separately during this step. Then, in the second step, the conservative variables vector of each of the fluids will be projected onto a common mesh (not necessarily identical to the initial mesh).

In the Section 6-2 of this paper, we recall the properties of the model we consider, that are conservation, hyperbolicity, and asymptotic limit model. In Section 6-3, we recall the basis of the solver (Glace [CDDL09] or Eucclhyd [MABO07]) used to compute the Lagrangian step. The Section 6-4 describes the Lagrangian step of the proposed scheme. It is demonstrated that the scheme preserves the properties of conservation, stability and consistency with respect to the continuous model for all regimes (independently of the value of the friction parameter). Then in Section 6-5, our ALE strategy is described. Finally, Section 6-6 is devoted to numerical experiments on several problems (Sod shock tube, triple point and Rayleigh-Taylor). Some comparisons with a non-AP scheme are provided.

6-2 A two fluids model with friction

Let us consider a mixture of two fluids f_1 and f_2. In the following, we will denote by "multi-fluid model", a model for which each fluid $\alpha \in \{f_1, f_2\}$ is represented by its own set of variables: $(\rho^\alpha, \mathbf{u}^\alpha, E^\alpha)$. Conversely, we will refer as "mono-fluid model", a model describing a mixture where mean quantities are considered $(\rho, \mathbf{u}, E)$, each fluid position being precised by an additional

equation on the concentration (e.g. $\chi := \frac{\rho^\alpha}{\rho^\alpha+\rho^\beta}$).

In this part, we present a simplified version of Scannapieco-Cheng's model where the interaction between the two constituents reduces to a friction term. In semi-Lagrangian coordinates, for each fluid $\alpha \in \{f_1, f_2\}$ (β denoting the other fluid), the model reads

$$\begin{aligned} \rho^\alpha D_t^\alpha \tau^\alpha &= \nabla \cdot \mathbf{u}^\alpha, \\ \rho^\alpha D_t^\alpha \mathbf{u}^\alpha &= -\nabla p^\alpha - \nu\rho\delta\mathbf{u}^\alpha, \\ \rho^\alpha D_t^\alpha E^\alpha &= -\nabla \cdot (p^\alpha \mathbf{u}^\alpha) - \nu\rho\delta\mathbf{u}^\alpha \cdot \overline{\mathbf{u}}, \end{aligned} \tag{6.1}$$

where $\rho^\alpha, \mathbf{u}^\alpha$ and E^α respectively denote the mass density, the velocity and the total energy density of fluid α. Also, $\tau^\alpha = \frac{1}{\rho^\alpha}$ denotes the specific volume. The pressure p^α satisfies the equation of state $p^\alpha := p^\alpha(\rho^\alpha, e^\alpha)$, where e^α, the internal energy density, is defined by $e^\alpha := E^\alpha - \frac{1}{2}\|\mathbf{u}^\alpha\|^2$. The total density ρ and the mean velocity $\overline{\mathbf{u}}$ are defined as $\rho := \rho^\alpha + \rho^\beta$ and $\rho\overline{\mathbf{u}} := \rho^\alpha\mathbf{u}^\alpha + \rho^\beta\mathbf{u}^\beta$. The term $\delta\mathbf{u}^\alpha$ is the velocity difference, the $\delta(\cdot)^\alpha$ operator being defined by $\delta\phi^\alpha = -\delta\phi^\beta = \phi^\alpha - \phi^\beta$. Finally, ν is the friction parameter. Also, remark that the Lagrangian derivative $D_t^\alpha := \partial_t + \mathbf{u}^\alpha \cdot \nabla$, is obviously not the same for each fluid.

The entropy η^α defined by Gibbs formula $T^\alpha d\eta^\alpha = de^\alpha + p^\alpha d\tau^\alpha$ satisfies the following entropy inequality

$$T^\alpha D_t^\alpha \eta^\alpha \geq \nu \frac{\tau^\alpha}{\tau^\beta} \delta\mathbf{u}^\alpha \cdot \delta\mathbf{u}^\alpha \geq 0. \tag{6.2}$$

Prior to establishing a numerical scheme that discretizes this set of six equations, we recall some properties of the model itself.

Property 1 (Conservation)**.** *The model (6.1) is conservative in volume and mass for each fluid. Also, it is conservative in the sum of momenta and in the sum of the total energies of the two fluids.*

Proof. Conservation of mass and volume is obvious since the first equation of (6.1) is the continuity equation written for each fluid.

Conservation of momenta sum and total energies sum require more cautiousness, since Lagrangian derivative are not the same for each fluid. To establish them one rewrites (6.1) in an Eulerian framework.

Developing Lagrangian derivatives $D_t^\alpha \phi = \partial_t\phi + \mathbf{u}^\alpha \cdot \nabla\phi$ and using the identity $\partial_t(\rho^\alpha\tau^\alpha) = 0$ elementary calculations can rewrite (6.1) as

$$\begin{aligned} &\partial_t\rho^\alpha + \nabla \cdot (\rho^\alpha \mathbf{u}^\alpha) = 0, \\ &\partial_t(\rho^\alpha\mathbf{u}^\alpha) + \nabla \cdot (\rho^\alpha\mathbf{u}^\alpha \otimes \mathbf{u}^\alpha) + \nabla p^\alpha + \nu\rho\delta\mathbf{u}^\alpha = \mathbf{0}, \\ &\partial_t(\rho^\alpha E^\alpha) + \nabla \cdot (\rho^\alpha E^\alpha \mathbf{u}^\alpha) + \nabla \cdot (p^\alpha\mathbf{u}^\alpha) + \nu\rho\delta\mathbf{u}^\alpha \cdot \overline{\mathbf{u}} = 0. \end{aligned} \tag{6.3}$$

Summing the two later equations over α gives a system of the conservative form $\partial_t\mathbf{U} + \nabla \cdot F(\mathbf{U}) = \mathbf{0}$, where

$$\mathbf{U} = \begin{pmatrix} \rho^\alpha\mathbf{u}^\alpha + \rho^\beta\mathbf{u}^\beta \\ \rho^\alpha E^\alpha + \rho^\beta E^\beta \end{pmatrix},$$

and

$$F(\mathbf{U}) = \begin{pmatrix} \rho^\alpha\mathbf{u}^\alpha \otimes \mathbf{u}^\alpha + \rho^\beta\mathbf{u}^\beta \otimes \mathbf{u}^\beta + \left(p^\alpha + p^\beta\right) I \\ \rho^\alpha E^\alpha\mathbf{u}^\alpha + \rho^\beta E^\beta\mathbf{u}^\beta + p^\alpha\mathbf{u}^\alpha + p^\beta\mathbf{u}^\beta \end{pmatrix},$$

where I is the identity matrix of $\mathbb{R}^{2\times 2}$. ∎

Property 2 (Hyperbolicity)**.** *The model (6.1) is hyperbolic.*

Proof. Since (6.1) is made of two Euler systems only coupled with source terms, it is hyperbolic. ■

Asymptotic model. *When $\nu \to +\infty$, (6.1) behaves as the following five equations model*

$$\rho D_t \mathbf{u} = -\nabla \left(p^\alpha + p^\beta \right), \tag{6.4}$$

while, for each fluid $\alpha \in \{f_1, f_2\}$, β denoting the other one, one has

$$\begin{aligned} \rho^\alpha D_t \tau^\alpha &= \nabla \cdot \mathbf{u}, \\ \rho^\alpha D_t E^\alpha &= -\frac{\rho^\alpha}{\rho} \mathbf{u} \cdot \nabla \left(p^\alpha + p^\beta \right) - p^\alpha \nabla \cdot \mathbf{u}, \end{aligned} \tag{6.5}$$

where $\mathbf{u}$ is the same velocity for both fluids, and thus the Lagrangian derivative is also the same.

Formal derivation (established in [Ena07]). Let $\epsilon = \nu^{-1}$ so that (6.1) rewrites

$$\begin{aligned} \rho^\alpha D_t^\alpha \tau^\alpha &= \nabla \cdot \mathbf{u}^\alpha, \\ \rho^\alpha D_t^\alpha \mathbf{u}^\alpha &= -\nabla p^\alpha - \frac{1}{\epsilon} \rho \delta \mathbf{u}^\alpha, \\ \rho^\alpha D_t^\alpha E^\alpha &= -\nabla \cdot (p^\alpha \mathbf{u}^\alpha) - \frac{1}{\epsilon} \rho \delta \mathbf{u}^\alpha \cdot \overline{\mathbf{u}}. \end{aligned} \tag{6.6}$$

We will now study its limit while $\epsilon \to 0^+$ focusing first on the momentum equations since the friction term's goal is to impose that $\delta \mathbf{u}^0 \overset{\epsilon \to 0}{\longrightarrow} \mathbf{0}$.

Developing the Lagrangian derivatives and dividing each momentum equation by $\rho^\alpha > 0$, one has

$$\partial_t \mathbf{u}^\alpha + (\nabla \mathbf{u}^\alpha) \mathbf{u}^\alpha = -\frac{\nabla p^\alpha}{\rho^\alpha} - \frac{1}{\epsilon} \frac{\rho}{\rho^\alpha} \delta \mathbf{u}^\alpha.$$

Since fluid β satisfies the same equation and recalling that $\delta \phi^\alpha = -\delta \phi^\beta = \phi^\alpha - \phi^\beta$, one gets

$$\partial_t \left(\delta \mathbf{u}^\alpha \right) + \delta \left((\nabla \mathbf{u}) \mathbf{u} \right)^\alpha = -\delta \left(\frac{\nabla p}{\rho} \right)^\alpha - \frac{1}{\epsilon} \lambda \delta \mathbf{u}^\alpha, \quad \text{where } \lambda = \frac{\rho^2}{\rho^\alpha \rho^\beta}.$$

We now perform an Hilbert expansion for all variables in the equation, that is $\phi = \phi^0 + \epsilon \phi^1 + O(\epsilon^2)$. One has

$$\partial_t (\delta \mathbf{u}^{\alpha,0}) + \delta \left((\nabla \mathbf{u}) \mathbf{u} \right)^{\alpha,0} = -\delta \left(\frac{\nabla p}{\rho} \right)^{\alpha,0} - \lambda^0 \left(\frac{1}{\epsilon} \delta \mathbf{u}^{\alpha,0} + \delta \mathbf{u}^{\alpha,1} \right) - \lambda^1 \delta \mathbf{u}^{\alpha,0} + O(\epsilon). \tag{6.7}$$

Multiplying this equation by ϵ one has $\lambda^0 \delta \mathbf{u}^{\alpha,0} = O(\epsilon)$, which gives $\delta \mathbf{u}^{\alpha,0} = \mathbf{0}$ when $\epsilon \to 0$ since $\lambda > 0$.

So, when $\epsilon \to 0$, formula (6.7) recasts

$$\delta \mathbf{u}^{\alpha,1} = -\frac{1}{\lambda^0} \delta \left(\frac{\nabla p}{\rho} \right)^{\alpha,0}. \tag{6.8}$$

Now, we perform an Hilbert expansion for the whole system (6.6), neglecting the non negative powers of ϵ. Choosing $\alpha \in \{f_1, f_2\}$, β being the other one, it reads

$$\begin{aligned} \rho^{\alpha,0} D_t^\alpha \tau^{\alpha,0} =& \nabla \cdot \mathbf{u}^{\alpha,0}, \\ \rho^{\alpha,0} D_t^\alpha \mathbf{u}^{\alpha,0} =& -\nabla p^{\alpha,0} - \rho^0 \left(\frac{1}{\epsilon} \delta \mathbf{u}^{\alpha,0} + \delta \mathbf{u}^{\alpha,1} \right) - \rho^1 \delta \mathbf{u}^{\alpha,0}, \\ \rho^{\alpha,0} D_t^\alpha E^{\alpha,0} =& -\nabla \cdot \left(p^{\alpha,0} \mathbf{u}^{\alpha,0} \right) - \rho^0 \left(\frac{1}{\epsilon} \delta \mathbf{u}^{\alpha,0} \cdot \overline{\mathbf{u}}^{\alpha,0} + \delta \mathbf{u}^{\alpha,1} \cdot \overline{\mathbf{u}}^{\alpha,0} + \delta \mathbf{u}^{\alpha,0} \cdot \overline{\mathbf{u}}^{\alpha,1} \right) \\ & - \rho^1 \delta \mathbf{u}^{\alpha,0} \cdot \overline{\mathbf{u}}^{\alpha,0}. \end{aligned}$$

Since we just established $\delta\mathbf{u}^{\alpha,0} = \mathbf{0}$, one has $\mathbf{u}^0 = \overline{\mathbf{u}}^0 = \mathbf{u}^{\alpha,0} = \mathbf{u}^{\beta,0}$. Also, since $D_t^\alpha \phi = \partial_t \phi + \mathbf{u}^{\alpha,0} \cdot \nabla \phi + \mathcal{O}(\epsilon)$, Lagrangian derivatives are the same when $\epsilon \to 0$, so that using (6.8) the system simplifies to

$$\begin{aligned}
\rho^{\alpha,0} D_t \tau^{\alpha,0} =& \nabla \cdot \mathbf{u}^0, \\
\rho^{\alpha,0} D_t \mathbf{u}^0 =& -\nabla p^{\alpha,0} + \rho^0 \frac{1}{\lambda^0} \delta \left(\frac{\nabla p}{\rho} \right)^{\alpha,0}, \\
\rho^{\alpha,0} D_t E^{\alpha,0} =& -\nabla \cdot \left(p^{\alpha,0} \mathbf{u}^0 \right) + \rho^0 \frac{1}{\lambda^0} \delta \left(\frac{\nabla p}{\rho} \right)^{\alpha,0} \cdot \mathbf{u}^0.
\end{aligned}$$

Recalling $\lambda = \frac{\rho^2}{\rho^\alpha \rho^\beta}$ and developing $\delta \left(\frac{\nabla p}{\rho} \right)^{\alpha,0}$, momentum equation satisfies

$$\begin{aligned}
\rho^{\alpha,0} D_t \mathbf{u}^0 =& -\nabla p^{\alpha,0} + \frac{\rho^{\alpha,0} \rho^{\beta,0}}{\rho^0} \left(\frac{\nabla p^{\alpha,0}}{\rho^{\alpha,0}} - \frac{\nabla p^{\beta,0}}{\rho^{\beta,0}} \right), \\
=& -\frac{\rho^{\alpha,0}}{\rho^0} \nabla \left(p^{\alpha,0} + p^{\beta,0} \right).
\end{aligned}$$

Proceeding the same way with total energy equation, one gets

$$\begin{aligned}
\rho^{\alpha,0} D_t E^{\alpha,0} =& -\nabla \cdot \left(p^{\alpha,0} \mathbf{u}^0 \right) + \frac{\rho^{\alpha,0} \rho^{\beta,0}}{\rho^0} \left(\frac{\nabla p^{\alpha,0}}{\rho^{\alpha,0}} - \frac{\nabla p^{\beta,0}}{\rho^{\beta,0}} \right) \cdot \mathbf{u}^0, \\
=& -\frac{\rho^{\alpha,0}}{\rho^0} \left(\nabla p^{\alpha,0} + \nabla p^{\beta,0} \right) \cdot \mathbf{u}^0 - p^{\alpha,0} \nabla \cdot \mathbf{u}^0,
\end{aligned}$$

■

Remark 1. *Defining* $E := \frac{\rho^\alpha E^\alpha + \rho^\beta E^\beta}{\rho}$ *and* $\tau := \rho^{-1}$*, it is easy to check that if* $(\rho^\alpha, \rho^\beta, \mathbf{u}, E^\alpha, E^\beta)$ *is a solution of the asymptotic model (6.4)–(6.5), one has*

$$\begin{aligned}
\rho D_t \tau =& \nabla \cdot \mathbf{u}, \\
\rho D_t \mathbf{u} =& -\nabla \left(p^\alpha + p^\beta \right), \\
\rho D_t E =& -\nabla \cdot \left(\left(p^\alpha + p^\beta \right) \mathbf{u} \right).
\end{aligned}$$

One recognizes Euler equations for the mixture. The mixing pressure follows Dalton's law as one could have expected since we consider here non-reactive gases.

However, notice that unless each fluid follows a barotropic equation of state ($p^\alpha = p^\alpha(\rho^\alpha)$)*, equation (6.5) must be solved to determine* e^α*.*

6-2.0.0.1 Next-order Hilbert expansion and effect on the concentrations The next order of the Hilbert expansion is interesting to enlighten some peculiar behaviour of the solution in the case of large but finite friction coefficient: $1 << \nu < +\infty$. To this end, let us consider the first equation of the system (6.3). It equivalently recasts into

$$\partial_t \rho^\alpha + \nabla \cdot (\rho^\alpha (\overline{\mathbf{u}} + (1 - \chi) \delta \mathbf{u}^\alpha) = 0, \tag{6.9}$$

where $\chi := \frac{\rho^\alpha}{\rho}$ denotes the mass concentration of fluid α.

Expanding this equation to the first-order gives same result as before, since $\overline{\mathbf{u}}^0 = \mathbf{u}^{\alpha,0}$ and $\delta \mathbf{u}^{\alpha,0} = \mathbf{0}$.

Peforming now a second-order Hilbert expansion of this equation and keeping the first two terms, we infer

$$\partial_t \rho^{\alpha,0,1} + \nabla \cdot (\rho^\alpha \overline{\mathbf{u}})^{0,1} = -\epsilon \nabla \cdot \left(\rho^{\alpha,0,1} \left(1 - \chi^{0,1}\right) \delta \mathbf{u}^{\alpha,1} \right) + O(\epsilon^2), \tag{6.10}$$

where $\phi^{0,1} := \phi^0 + \epsilon \phi^1$.

Injecting the value of $\delta \mathbf{u}^{\alpha,1}$ given by equation (6.8) into this expression, and using $\lambda = \frac{1}{\chi(1-\chi)}$, we obtain:

$$\partial_t \rho^{\alpha,0,1} + \nabla \cdot (\rho^\alpha \overline{\mathbf{u}})^{0,1} = \epsilon \nabla \cdot \left(\rho^{\alpha,0,1} \chi^{0,1} \left(1 - \chi^{0,1}\right)^2 \delta \left(\frac{\nabla p}{\rho} \right)^{\alpha,0} \right) + O(\epsilon^2). \tag{6.11}$$

Cancelling the indices, and recasting this equation into a semi-Lagragian form, we obtain the following law for ρ^α

$$\rho^\alpha D_t^{\overline{\mathbf{u}}} \tau^\alpha = \nabla \cdot \overline{\mathbf{u}} + \frac{\epsilon}{\rho^\alpha} \nabla \cdot \left(\rho^\alpha \chi (1-\chi)^2 \delta \left(\frac{\nabla p}{\rho} \right)^\alpha \right) + O(\epsilon^2), \tag{6.12}$$

where $D_t^{\overline{\mathbf{u}}}$ stands for the Lagrangian derivative at the velocity $\overline{\mathbf{u}}$. Note that, in this form, the equation on ρ^α exhibits a diffusive behaviour (in particular if we consider barotrope equation of state $p = p(\rho)$). From the previous expression, using $\rho D_t^{\overline{\mathbf{u}}}(\tau) = \nabla \cdot \mathbf{u}$ and expanding $\delta \left(\frac{\nabla p}{\rho} \right)^\alpha$, we obtain an equation on the concentration c

$$D_t^{\overline{\mathbf{u}}} \chi = \frac{\epsilon}{\rho} \nabla \cdot \left(\chi (1-\chi) \left((1-\chi) \nabla p^\alpha - \chi \nabla p^\beta \right) \right) + O(\epsilon^2). \tag{6.13}$$

This equation accounts for the diffusive regime of the concentration in the limit $\nu >> 1$. To convince the reader, let us take the same simple equation of state for both fluid: $p = K\rho$, with $K \in \mathbb{R}^{+*}$. We obtain then the following form for (6.13)

$$D_t^{\overline{\mathbf{u}}} \chi = \frac{\epsilon}{\rho} \nabla \cdot (\chi (1-\chi) K \nabla \chi) + O(\epsilon^2), \tag{6.14}$$

which together with the equations on ρ, $\rho\overline{\mathbf{u}}$, and $\rho E = \rho^\alpha E^\alpha + \rho^\beta E^\beta$, verified by construction by our model, gives the basic form of well-known simplified models (refer for instance to the model $\mathcal{E}2\mathcal{M}$, Page 210 of [Sen14]). This analysis justifies, that we require our scheme to be able to reproduce this behaviour. An easy way to check that in Lagrangian schemes is to calculate the relative evolution of the specific volumes in the fluids α and β, since it must satisfy the following equation for $\nu >> 1$

$$\rho^\alpha D_t^\alpha \tau^\alpha - \rho^\beta D_t^\beta \tau^\beta = -\epsilon \nabla \cdot \left((1-\chi) \nabla p^\alpha - \chi \nabla p^\beta \right). \tag{6.15}$$

In the analysis of the discrete version of the scheme, we verify that the scheme is consistent with the discrete version of

$$\delta \overline{\mathbf{u}} = -\frac{\epsilon}{\rho} \left((1-\chi) \nabla p^\alpha - \chi \nabla p^\beta \right). \tag{6.16}$$

6-3 Cell-centered schemes

We recall briefly the multidimensional finite volume schemes [Maz07, DM05, MABO07], since it is the basis of this work. For convenience, we use the notations defined in [CDDL09]. In the following, for all cell j, and for any quantity ϕ, one defines its mean value $\phi_j := \frac{1}{V_j} \int_j \phi$, where $V_j := \int_j 1$ is the cell volume. Also, let us denote the cell's mass as $m_j := \int_j \rho = \rho_j V_j$, which is constant in time in semi-Lagrangian coordinates ($d_t m_j = 0$).

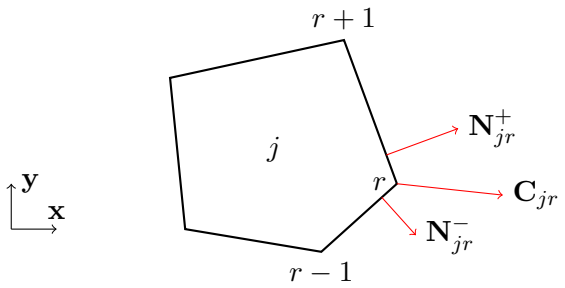

Figure 6.1 – Illustration of $\mathbf{C}_{jr}$ and $\mathbf{N}_{jr}^{i}$ vectors at vertex r for a polygonal cell j.

We consider first-order schemes, so that one has the following relations

$$\frac{d}{dt}\int_j 1 = m_j d_t \tau_j, \qquad \frac{d}{dt}\int_j \rho_j = 0,$$
$$\frac{d}{dt}\int_j \rho_j \mathbf{u}_j = m_j d_t \mathbf{u}_j, \qquad \frac{d}{dt}\int_j \rho_j E_j = m_j d_t E_j.$$

Let $\mathcal{J}_r$ denote the set of cells connected to node r and let $\mathcal{R}_j$ the set of nodes of cell j. Also, let us introduce $\mathbf{C}_{jr} := \nabla_{\mathbf{x}_r} V_j$, the gradient of the volume of the polygonal cell j, according to the position of one of its vertices r. In dimension d, one has the relation $V_j = \frac{1}{d}\sum_{r\in\mathcal{R}_j} \mathbf{C}_{jr}\cdot\mathbf{x}_r$. So in cartesian coordinates, since the volume of a cell is independent of its position, one has

$$\forall j \in \mathcal{J}, \quad \sum_{r\in\mathcal{R}_j} \mathbf{C}_{jr} = \mathbf{0}. \tag{6.17}$$

For more properties of $\mathbf{C}_{jr}$ vectors, one may refer to [CDDL09]. The cell-centered schemes we consider in this paper have the following structure: for any cell j of the mesh one has

$$\begin{aligned} m_j d_t \tau_j &= \sum_{r\in\mathcal{R}_j} \mathbf{C}_{jr}\cdot\mathbf{u}_r, \\ d_t m_j &= 0, \\ m_j d_t \mathbf{u}_j &= -\sum_{r\in\mathcal{R}_j} \mathbf{F}_{jr}, \\ m_j d_t E_j &= -\sum_{r\in\mathcal{R}_j} \mathbf{F}_{jr}\cdot\mathbf{u}_r, \end{aligned} \tag{6.18}$$

where the fluxes $\mathbf{u}_r$ and $\mathbf{F}_{jr}$ are defined for any node r

$$\forall j \in \mathcal{J}_r, \quad \mathbf{F}_{jr} = \mathbf{C}_{jr} p_j - A_{jr}(\mathbf{u}_r - \mathbf{u}_j), \tag{6.19}$$
$$\text{and} \quad \sum_{j\in\mathcal{J}_r} \mathbf{F}_{jr} = \mathbf{0}. \tag{6.20}$$

On one hand, relation (6.19) is the matrix form of the acoustic Riemann solver (see for instance [Klu08, Mai11]), while on the other hand (6.20) imposes conservation.

In the following to simplify notations, we omit sets $\mathcal{R}_j$ and $\mathcal{J}_r$ when there is no confusion.

— If $A_{jr} := \rho_j c_j \frac{\mathbf{C}_{jr}\otimes\mathbf{C}_{jr}}{\|\mathbf{C}_{jr}\|}$, then (6.18)–(6.20) defines the Glace scheme [DM05, CDDL09].

— Let $\mathbf{N}_{jr}^+ = -\frac{1}{2}(\mathbf{x}_{r+1}-\mathbf{x}_r)^\perp$ and $\mathbf{N}_{jr}^- = -\frac{1}{2}(\mathbf{x}_r-\mathbf{x}_{r-1})^\perp$. If $A_{jr} := \rho_j c_j \left(\frac{\mathbf{N}_{jr}^+ \otimes \mathbf{N}_{jr}^+}{\|\mathbf{N}_{jr}^+\|} + \frac{\mathbf{N}_{jr}^- \otimes \mathbf{N}_{jr}^-}{\|\mathbf{N}_{jr}^-\|} \right)$, the scheme (6.18)–(6.20) is Eucclhyd [MABO07, Mai11]. One has $\mathbf{N}_{jr}^+ + \mathbf{N}_{jr}^- = \mathbf{C}_{jr}$, see Figure 6.1.

These schemes are conservative in volume, mass, momentum and total energy. One easily shows that they are entropy stable. These results can be found in [DM05, MABO07, CDDL09, Mai11], for instance. Also, a consistency result has been established in [Des10b]. Both schemes are very close to each other, Glace scheme is considered more precise and Eucclhyd more stable.

6-4 Asymptotic Preserving scheme in semi-Lagrangian coordinates

Let us first introduce the following notations. We set $\rho_r^\alpha := \frac{1}{\#\mathcal{J}_r} \sum_{j\in\mathcal{J}_r} \rho_j^\alpha$ and $\rho_r := \rho_r^\alpha + \rho_r^\beta$. Also, we set $\overline{\mathbf{u}}_r := \frac{\rho_r^\alpha \mathbf{u}_r^\alpha + \rho_r^\beta \mathbf{u}_r^\beta}{\rho_r^\alpha + \rho_r^\beta}$ and $\overline{\mathbf{u}}_{jr} := \frac{\rho_r^\alpha \mathbf{u}_j^\alpha + \rho_r^\beta \mathbf{u}_j^\beta}{\rho_r^\alpha + \rho_r^\beta}$. B_{jr} are symmetric and positive definite matrices that satisfy $\sum_{r\in\mathcal{R}_j} B_{jr} = V_j I$. Matrices A_{jr}^α are the standard "hydro-matrices" as defined in Section 6-3.

Remark 2. *One can choose $B_{jr} := V_{jr} I$, where V_{jr} is the volume of the subcell associated to vertex r of cell j. Another obvious choice could be for instance $B_{jr} := \frac{1}{\#\mathcal{R}_j} V_j I$.*

Remark 3. *Following [BDF12], one could also choose $B_{jr} := \mathbf{C}_{jr} \otimes (\mathbf{x}_r - \mathbf{x}_j)$. However if in that case one could hope to push that analysis further in terms of diffusion limit, since $B_{jr} := \mathbf{C}_{jr} \otimes (\mathbf{x}_r - \mathbf{x}_j)$ are not positive, one could not show anymore the entropy stability of the scheme. We have run the tests of Section 6-6 with this choice without noticing large differences.*

Observe that simple calculations allow to write

$$\rho_r \overline{\mathbf{u}}_r = \rho_r \mathbf{u}_r^\alpha - \rho_r^\beta \delta\mathbf{u}_r^\alpha \quad \text{and} \quad \rho_r \overline{\mathbf{u}}_{jr} = \rho_r \mathbf{u}_j^\alpha - \rho_r^\beta \delta\mathbf{u}_j^\alpha. \tag{6.21}$$

6-4.1 Reference scheme

Let us first introduce the following scheme that will be used as a reference scheme to illustrate the advantages of our AP scheme — described by (6.24)–(6.26).

For each fluid $\alpha \in \{f_1, f_2\}$, one writes

$$\begin{aligned} d_t m_j^\alpha &= 0, \\ m_j^\alpha d_t \tau_j^\alpha &= \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r^\alpha, \\ m_j^\alpha d_t \mathbf{u}_j^\alpha &= -\sum_r \mathbf{F}_{jr}^\alpha - \sum_r \nu \rho_r B_{jr} \delta\mathbf{u}_j^\alpha, \\ m_j^\alpha d_t E_j^\alpha &= -\sum_r \mathbf{F}_{jr}^\alpha \cdot \mathbf{u}_r^\alpha - \sum_r \nu \rho_r \overline{\mathbf{u}}_{jr}^T B_{jr} \delta\mathbf{u}_j^\alpha, \end{aligned} \tag{6.22}$$

where $\overline{\mathbf{u}}_{jr}$ and ρ_r are defined as in Section 6-4.2 and the fluxes are given by

$$\begin{aligned} \mathbf{F}_{jr}^\alpha &= \mathbf{C}_{jr} p_j^\alpha - A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) \\ \sum_j A_{jr}^\alpha \mathbf{u}_r^\alpha &= \sum_j A_{jr}^\alpha \mathbf{u}_j^\alpha + \sum_j \mathbf{C}_{jr} p_j^\alpha. \end{aligned} \tag{6.23}$$

It can be showed that this scheme is entropic, conservative in volume and mass for each fluid and in the sum of momenta and total energies. Also, the scheme is weakly consistent with (6.1). However, this scheme does not *a priori* preserve the asymptotic, see 6.D for some details.

6-4.2 Continuous in time semi-discrete scheme

We shall now present a multidimensional finite volume scheme written in semi-Lagrangian coordinates that preserves the asymptotic.

This scheme will be the Lagrangian step of our ALE method. In this step, each fluid is associated to its own mesh. If the meshes may evolve differently, we assume that they coincide at the begining of the Lagrangian step. The rezoning/remapping procedure that is detailed in Section 6-5 is used to ensure that the meshes will coincide for the next Lagrangian step.

We first focus on the semi-discrete continuous in time scheme. Most of the properties of the scheme are proved using this simpler formulation without any lost of generality. In Paragraph 6-4.3, we describe the fully discrete scheme. It is analysed in the remaining of this section.

Let $\alpha \in \{f_1, f_2\}$ denote one of the two fluids and β the other one, we define the scheme

$$\begin{aligned}
m_j^\alpha d_t \tau_j^\alpha &= \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r^\alpha, \\
d_t m_j^\alpha &= 0, \\
m_j^\alpha d_t \mathbf{u}_j^\alpha &= -\sum_r \mathbf{F}_{jr}^\alpha - \sum_r \nu \rho_r B_{jr} \delta \mathbf{u}_j^\alpha, \\
m_j^\alpha d_t E_j^\alpha &= -\sum_r \mathbf{F}_{jr}^\alpha \cdot \mathbf{u}_r^\alpha - \sum_r \nu \rho_r \overline{\mathbf{u}}_r^T B_{jr} \delta \mathbf{u}_r^\alpha + \sum_r \nu \rho_r \overline{\mathbf{u}}_{jr}^T B_{jr} (\delta \mathbf{u}_r^\alpha - \delta \mathbf{u}_j^\alpha),
\end{aligned} \tag{6.24}$$

where the fluxes are given by

$$\mathbf{F}_{jr}^\alpha = \mathbf{C}_{jr} p_j^\alpha - A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) - \nu \rho_r B_{jr} \delta \mathbf{u}_r^\alpha, \quad \text{and} \tag{6.25}$$

$$\sum_j \mathbf{F}_{jr}^\alpha = \mathbf{0}. \tag{6.26}$$

Injecting (6.25) in (6.24), and using (6.21), one gets the alternative form

$$\begin{aligned}
m_j^\alpha d_t \tau_j^\alpha &= \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r^\alpha, \\
d_t m_j^\alpha &= 0, \\
m_j^\alpha d_t \mathbf{u}_j^\alpha &= \sum_r A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) + \nu \sum_r \rho_r B_{jr} \left(\delta \mathbf{u}_r^\alpha - \delta \mathbf{u}_j^\alpha \right), \\
m_j^\alpha d_t E_j^\alpha &= -\sum_r \mathbf{C}_{jr} p_j^\alpha \cdot \mathbf{u}_r^\alpha + \sum_r \mathbf{u}_r^{\alpha T} A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) + \nu \sum_r \rho_r^\beta \ \delta \mathbf{u}_r^{\alpha T} B_{jr} \delta \mathbf{u}_r^\alpha \\
&\quad - \nu \sum_r \rho_r^\beta \ \delta \mathbf{u}_j^{\alpha T} B_{jr} (\delta \mathbf{u}_r^\alpha - \delta \mathbf{u}_j^\alpha) + \nu \sum_r \rho_r \ \mathbf{u}_j^{\alpha T} B_{jr} (\delta \mathbf{u}_r^\alpha - \delta \mathbf{u}_j^\alpha).
\end{aligned} \tag{6.27}$$

This form enlightens the fact that knowing the fluxes $(\mathbf{u}_r^\alpha, \mathbf{u}_r^\beta)$ at any vertex r is enough to define the scheme. We shall now show that these nodal velocities are well defined.

Injecting (6.25) in (6.26) allows to calculate $(\mathbf{u}_r^\alpha, \mathbf{u}_r^\beta)$. Obviously, as soon as $\nu \neq 0$, both nodal velocities are coupled at vertex r. Omitting boundary conditions for the sake of simplicity, each

vertex of the mesh $(\mathbf{u}_r^\alpha, \mathbf{u}_r^\beta)$ is the unique solution of the following linear system:

$$\underbrace{\sum_j \begin{pmatrix} A_{jr}^\alpha + \nu\rho_r B_{jr} & -\nu\rho_r B_{jr} \\ -\nu\rho_r B_{jr} & A_{jr}^\beta + \nu\rho_r B_{jr} \end{pmatrix}}_{\mathbb{A}_r^\nu :=} \begin{pmatrix} \mathbf{u}_r^\alpha \\ \mathbf{u}_r^\beta \end{pmatrix} = \underbrace{\sum_j \begin{pmatrix} A_{jr}^\alpha \mathbf{u}_j^\alpha + \mathbf{C}_{jr} p_j^\alpha \\ A_{jr}^\beta \mathbf{u}_j^\beta + \mathbf{C}_{jr} p_j^\beta \end{pmatrix}}_{\mathbf{b}_r :=}.$$

Proof. Since matrices A_{jr}^α and B_{jr} are symmetric, $\mathbb{A}_r^\nu$ is also symmetric. To prove that $(\mathbf{u}_r^\alpha, \mathbf{u}_r^\beta)$ is unique, it remains to show that it is positive definite. Elementary calculations give, $\forall (\mathbf{v}^\alpha, \mathbf{v}^\beta) \in \mathbb{R}^2 \times \mathbb{R}^2$,

$$\begin{aligned}(\mathbf{v}^\alpha, \mathbf{v}^\beta)^T \mathbb{A}_r^\nu (\mathbf{v}^\alpha, \mathbf{v}^\beta) =& \mathbf{v}^{\alpha T} \left(\sum_j A_{jr}^\alpha \right) \mathbf{v}^\alpha + \mathbf{v}^{\beta T} \left(\sum_j A_{jr}^\beta \right) \mathbf{v}^\beta \\ &+ (\mathbf{v}^\alpha - \mathbf{v}^\beta)^T \left(\sum_j \nu\rho_r B_{jr} \right) (\mathbf{v}^\alpha - \mathbf{v}^\beta),\end{aligned}$$

which is strictly positive if $(\mathbf{v}^\alpha, \mathbf{v}^\beta) \neq (\mathbf{0}, \mathbf{0})$ since matrices $\sum_j A_{jr}^\alpha$ and $\sum_j \nu\rho_r B_{jr}$ are positive definite. ■

The scheme being well-defined, we now establish its properties.

6-4.2.1 Nodal velocities *a priori* estimates

Here, we establish estimates for the nodal velocities with regard to the frictionless case. These are actually some *instantaneous* stability results with regard to the mono-fluid schemes [CDDL09, MABO07], *i.e.* velocity fluxes are controled by the frictionless ones.

Property 3 (*A priori* estimates)**.** *For each fluid $\alpha \in \{f_1, f_2\}$, let $\mathbf{u}_r^{\alpha,\nu}$ denote the nodal velocities at vertex r. Let $A_r^\alpha := \sum_j A_{jr}^\alpha$ and $B_r := \sum_j B_{jr}$. Let β denote the other fluid, then one has the following relations, $\forall \nu \geq 0$*

$$\mathbf{u}_r^{\alpha,\nu T} A_r^\alpha \mathbf{u}_r^{\alpha,\nu} + \mathbf{u}_r^{\beta,\nu T} A_r^\beta \mathbf{u}_r^{\beta,\nu} \leq \mathbf{u}_r^{\alpha,0 T} A_r^\alpha \mathbf{u}_r^{\alpha,0} + \mathbf{u}_r^{\beta,0 T} A_r^\beta \mathbf{u}_r^{\beta,0}, \tag{6.28}$$

$$\left(\mathbf{u}_r^{\alpha,\nu} - \mathbf{u}_r^{\beta,\nu}\right)^T B_r \left(\mathbf{u}_r^{\alpha,\nu} - \mathbf{u}_r^{\beta,\nu}\right) \leq \frac{1}{2\nu\rho_r} \left(\mathbf{u}_r^{\alpha,0 T} A_r^\alpha \mathbf{u}_r^{\alpha,0} + \mathbf{u}_r^{\beta,0 T} A_r^\beta \mathbf{u}_r^{\beta,0}\right), \tag{6.29}$$

$$\textit{and} \quad \left(\mathbf{u}_r^{\alpha,\nu} - \mathbf{u}_r^{\beta,\nu}\right)^T B_r \left(\mathbf{u}_r^{\alpha,\nu} - \mathbf{u}_r^{\beta,\nu}\right) \leq \left(\mathbf{u}_r^{\alpha,0} - \mathbf{u}_r^{\beta,0}\right)^T B_r \left(\mathbf{u}_r^{\alpha,0} - \mathbf{u}_r^{\beta,0}\right). \tag{6.30}$$

Proof. See 6-2.1 Page 149. ■

Let us comment on these estimates. The estimate (6.28) is a stability result. It shows that the nodal velocity $\|(\mathbf{u}_r^{\alpha,\nu}, \mathbf{u}_r^{\beta,\nu})\|_{\mathbb{A}_r^0}$ is bounded by $\|(\mathbf{u}_r^{\alpha,0}, \mathbf{u}_r^{\beta,0})\|_{\mathbb{A}_r^0}$ independently of ν. It shows that friction nodal velocities are stable with regard to the classic frictionless case for a given state.

The second estimate (6.29) shows that the nodal velocity difference $\|\delta\mathbf{u}_r^{\alpha,\nu}\|_{B_r}$ is at most $\mathcal{O}(\nu^{-1/2})$ according to $\|(\mathbf{u}_r^{\alpha,0}, \mathbf{u}_r^{\beta,0})\|_{\mathbb{A}_r^0}$. In 6.C, we show numerically that one can obtain $\mathcal{O}(\nu^{-1})$ and that (6.29) may not be optimal.

The last inequality (6.30) states that the nodal velocity difference is bounded by the frictionless case independently of ν in the $\|\cdot\|_{B_r}$ norm, which is purely geometric.

6-4.2.2 Conservativity

Property 4 (Conservation). *The scheme defined by (6.24)–(6.26) ensures conservation of mass and volume for each fluid α or β. It also ensures that the sum of the fluids' momenta and total energies are conserved.*

Proof. See 6-2.2 Page 151. ■

6-4.2.3 Stability

Before announcing this result, we recall that the fully discrete scheme's stability is presented bellow (see Paragraph 6-4.3).

Property 5 (Entropy). *The first-order continuous in time scheme defined by (6.24)–(6.26) satisfies the following entropy inequality $\forall \alpha \in \{f_1, f_2\}$*

$$m_j^\alpha T_j^\alpha d_t \eta_j^\alpha \geq \frac{1}{2} \sum_r \nu \rho_r^\beta \ \delta \mathbf{u}_r^{\alpha T} B_{jr} \delta \mathbf{u}_r^\beta + \frac{1}{2} \sum_r \nu \rho_r^\beta \ \delta \mathbf{u}_j^{\alpha T} B_{jr} \delta \mathbf{u}_j^\alpha \geq 0.$$

This inequality is consistent with (6.2).

Proof. See 6-2.3 Page 151. ■

6-4.2.4 Asymptotic preserving

We now establish the main result of this paper. It consists in stating that when the friction parameter ν tends to infinity, the scheme (6.24)–(6.26) behaves asymptotically as a scheme that is consistent with the asymptotic model (6.4)–(6.5).

To this end, we first compute the asymptotic scheme by means of Hilbert expansions, then we show its consistency with the asymptotic model. This later result relies strongly on B. Després's work [Des10b].

Asymptotic scheme. *If $\forall \alpha \in \{f_1, f_2\}$, $\forall j$, $(\rho_j^\alpha, \mathbf{u}_j^\alpha, E_j^\alpha)$ are constant cell data, then the scheme (6.24)–(6.26), behaves asymptotically as*

$$(m_j^\alpha + m_j^\beta) d_t \mathbf{u}_j = -\sum_r \mathbf{F}_{jr}^\alpha - \sum_r \mathbf{F}_{jr}^\beta, \tag{6.31}$$

$$d_t V_j = m_j^\alpha d_t \tau_j^\alpha = \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r, \tag{6.32}$$

$$d_t m_j^\alpha = 0,$$

$$m_j^\alpha d_t E_j^\alpha = -\sum_r \mathbf{C}_{jr} p_j^\alpha \cdot \mathbf{u}_r + \sum_r \mathbf{u}_r^T A_{jr}^\alpha (\mathbf{u}_r - \mathbf{u}_j) - \frac{\rho_j^\alpha \rho_j^\beta}{\rho_j} \sum_r \mathbf{u}_j^T \delta \left(\frac{A_{jr}}{\rho_j} \right)^\alpha (\mathbf{u}_r - \mathbf{u}_j), \tag{6.33}$$

where $\mathbf{u}_j = \mathbf{u}_j^\alpha = \mathbf{u}_j^\beta$, and where nodal velocities $\mathbf{u}_r = \mathbf{u}_r^\alpha = \mathbf{u}_r^\beta$ satisfy

$$\mathbf{F}_{jr}^\alpha + \mathbf{F}_{jr}^\beta = \mathbf{C}_{jr} \left(p_j^\alpha + p_j^\beta \right) - \left(A_{jr}^\alpha + A_{jr}^\beta \right) (\mathbf{u}_r - \mathbf{u}_j),$$

$$\text{and} \quad \sum_j \mathbf{F}_{jr}^\alpha = \mathbf{0}. \tag{6.34}$$

Formal derivation. See 6-2.4 Page 152. ■

In order to establish that the scheme is asymptotic preserving, it remains to show that the limit scheme (6.31)–(6.34) is consistent with the asymptotic model (6.4)–(6.5).

Before establishing this result, we recall the fundamental result by B. Després [Des10b], that we adapt to the present context.

Property 6 (B. Després)**.** *Let* $m_j := m_j^\alpha + m_j^\beta$, $\rho_j := \rho_j^\alpha + \rho_j^\beta$, $\tau_j = \rho_j^{-1}$ *and* $E_j := \frac{\rho_j^\alpha E_j^\alpha + \rho_j^\beta E_j^\beta}{\rho_j}$. *Then, the monofluid (Glace or Eucclhyd) scheme defined for a mixture*

$$
\begin{aligned}
& d_t m_j = 0, \\
& m_j d_t \tau_j = \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r, \\
& m_j d_t \mathbf{u}_j = -\sum_r \mathbf{F}_{jr}, \\
& m_j d_t E_j = -\sum_r \mathbf{F}_{jr} \cdot \mathbf{u}_r, \\
\textit{where} \quad & \mathbf{F}_{jr} = \mathbf{C}_{jr}(p_j^\alpha + p_j^\beta) - (A_{jr}^\alpha + A_{jr}^\beta)(\mathbf{u}_r - \mathbf{u}_j), \\
\textit{and} \quad & \sum_j (A_{jr}^\alpha + A_{jr}^\beta)\mathbf{u}_r = \sum_j (A_{jr}^\alpha + A_{jr}^\beta)\mathbf{u}_j + \sum_j \mathbf{C}_{jr}(p_j^\alpha + p_j^\beta),
\end{aligned}
$$

is weakly consistent with the following system of equations

$$
\begin{aligned}
& \rho D_t \tau = \nabla \cdot \mathbf{u}, \\
& \rho D_t \mathbf{u} = -\nabla(p^\alpha + p^\beta), \\
& \rho D_t E = -\nabla \cdot (p^\alpha + p^\beta)\mathbf{u}.
\end{aligned}
$$

Proof. The proof can be found in [Des10b]. ■

Remark 4. *In order to establish the following Property 7, we kept intentionally* A_{jr}^α *and* p_j^α *for both fluids in the fluxes expressions. Actually, to retrieve the result in [Des10b], one has to define simply the mixture pressure* $p_j := p_j^\alpha + p_j^\beta$ *and* $A_{jr} := A_{jr}^\alpha + A_{jr}^\beta$ *which is actually the monofluid* A_{jr} *matrix defined by the mixture sound speed:* $\rho_j c_j := \rho_j^\alpha c_j^\alpha + \rho_j^\beta c_j^\beta$.

Property 7. *The limit scheme (6.31)–(6.34) is weakly consistent with the asymptotic model (6.4)–(6.5).*

Proof. See 6-2.5 Page 154. ■

We now study the diffusive regime. According to equation (6.B.11) page 153, one gets the following identity for $\delta \mathbf{u}_j^{\alpha,1} - \delta \mathbf{u}_r^{\alpha,1}$

$$
\sum_r \delta \left(\frac{A_{jr}^0}{\rho_j} \right)^\alpha (\mathbf{u}_r^0 - \mathbf{u}_j^0) = \frac{\rho_j}{\rho_j^\alpha \rho_j^\beta} \sum_r \rho_r^0 B_{jr} (\delta \mathbf{u}_j^{\alpha,1} - \delta \mathbf{u}_r^{\alpha,1}). \tag{6.35}
$$

As explained previously, it has been proven in [Des10b], that the left hand side of (6.35) fulfills the following weak consistence relation

$$
\frac{1}{V_j} \sum_r \delta \left(\frac{A_{jr}^0}{\rho_j} \right)^\alpha (\mathbf{u}_r^0 - \mathbf{u}_j^0) \approx \delta \left(\frac{\nabla p}{\rho} \right)^\alpha 1_{\Omega_j(\mathbf{x})}. \tag{6.36}
$$

On the other hand, $\sum_r B_{jr}\phi_r$ acts as an averaging operator over the cell j, and

$$\frac{\rho_j}{\rho_j^\alpha \rho_j^\beta}\sum_r \rho_r^0 B_{jr}(\delta\mathbf{u}_j^{\alpha,1} - \delta\mathbf{u}_r^{\alpha,1}) \approx V_j \frac{\rho_j}{\rho_j^\alpha \rho_j^\beta}\overline{\rho_r}_j^0 \left(\delta\mathbf{u}_j^{\alpha,1} - \overline{\delta\mathbf{u}_r^{\alpha,1}}_j\right). \tag{6.37}$$

For smooth enough solution, equations (6.36) and (6.37) suggest that

$$\begin{aligned}\delta\mathbf{u}_j^{\alpha,1} - \delta\mathbf{u}_r^{\alpha,1} &\approx \tfrac{\rho^\alpha\rho^\beta}{\rho^2}\delta\left(\tfrac{\nabla p}{\rho}\right)^\alpha 1_{\Omega_j(\mathbf{x})},\\ &\approx \tfrac{1}{\rho}\left((1-\chi)\nabla p^\alpha - \chi\nabla p^\beta\right).\end{aligned} \tag{6.38}$$

Comparing this last equation with (6.16) and since

$$\begin{aligned}\sum_r \mathbf{C}_{jr}\cdot\left(\delta\mathbf{u}_j^{\alpha,1} - \delta\mathbf{u}_r^{\alpha,1}\right) &= -\sum_r \mathbf{C}_{jr}\cdot\delta\mathbf{u}_r^{\alpha,1},\\ &= -\varepsilon\sum_r \mathbf{C}_{jr}\cdot\delta\mathbf{u}_r^{\alpha},\end{aligned}$$

we find an expression weakly consistent with (6.15) for the evolution of the difference of the specific volumes of the two fluids.

6-4.3 Discrete scheme

We now describe the fully discrete scheme. One defines the following scheme for each fluid $\alpha \in \{f_1, f_2\}$, β denoting the other one,

$$m_j^\alpha \frac{\tau_j^{\alpha n+1} - \tau_j^{\alpha n}}{\Delta t} = \sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n}, \tag{6.39}$$

$$m_j^\alpha \frac{\mathbf{u}_j^{\alpha n+1} - \mathbf{u}_j^{\alpha n}}{\Delta t} = -\sum_r \mathbf{F}_{jr}^{\alpha,n} - \sum_r \nu\rho_r^n B_{jr}^n \delta\mathbf{u}_j^{\alpha n+1}, \tag{6.40}$$

$$m_j^\alpha \frac{E_j^{\alpha n+1} - E_j^{\alpha n}}{\Delta t} = -\sum_r \mathbf{F}_{jr}^{\alpha,n}\cdot\mathbf{u}_r^{\alpha n} - \sum_r \nu\rho_r^n\, \overline{\mathbf{u}}_r^{n T} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} + \sum_r \nu\rho_r^n\, \overline{\mathbf{u}}_{jr}^{n+1 T} B_{jr}^n \left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right), \tag{6.41}$$

where the fluxes are computed explicitly as

$$\mathbf{F}_{jr}^{\alpha,n} = \mathbf{C}_{jr}^n p_j^{\alpha n} - A_{jr}^{\alpha,n}(\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) - \nu\rho_r^n B_{jr}^n \delta\mathbf{u}_r^{\alpha n}, \tag{6.42}$$

and

$$\sum_j A_{jr}^{\alpha,n}\mathbf{u}_r^{\alpha n} + \sum_j \nu\rho_r^n B_{jr}^n \delta\mathbf{u}_r^{\alpha n} = \sum_j A_{jr}^{\alpha,n}\mathbf{u}_j^{\alpha n} + \sum_j \mathbf{C}_{jr}^n p_j^{\alpha n}. \tag{6.43}$$

To complete the scheme definition, observe that we introduced the following mean velocities $\overline{\mathbf{u}}_{jr}^{n+1} := \frac{\rho_r^{\alpha n}\mathbf{u}_j^{\alpha n+1} + \rho_r^{\beta n}\mathbf{u}_j^{\beta n+1}}{\rho_r^{\alpha n} + \rho_r^{\beta n}}$ and $\overline{\mathbf{u}}_r^n := \frac{\rho_r^{\alpha n}\mathbf{u}_r^{\alpha n} + \rho_r^{\beta n}\mathbf{u}_r^{\beta n}}{\rho_r^{\alpha n} + \rho_r^{\beta n}}$, which rewrite

$$\rho_r^n\overline{\mathbf{u}}_{jr}^{n+1} = \rho_r^n\mathbf{u}_j^{\alpha n+1} - \rho_r^{\beta n}\delta\mathbf{u}_j^{\alpha n+1} \quad \text{and} \quad \rho_r^n\overline{\mathbf{u}}_r^n = \rho_r^n\mathbf{u}_r^{\alpha n} - \rho_r^{\beta n}\delta\mathbf{u}_r^{\alpha n}. \tag{6.44}$$

Similarly to the semi-discrete case, for convenience, we substitute the flux expression into momentum and total energy balance equations and use (6.44)

$$m_j^\alpha \frac{\mathbf{u}_j^{\alpha n+1} - \mathbf{u}_j^{\alpha n}}{\Delta t} = \sum_r A_{jr}^{\alpha,n}(\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) + \nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}), \tag{6.45}$$

$$m_j^\alpha \frac{E_j^{\alpha n+1} - E_j^{\alpha n}}{\Delta t} = -\sum_r \mathbf{C}_{jr}^n p_j^{\alpha n} \cdot \mathbf{u}_r^{\alpha n} + \sum_r \mathbf{u}_r^{\alpha n T} A_{jr}^{\alpha,n} (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) + \nu \sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha n T} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} + \nu \sum_r \rho_r^n\ \mathbf{u}_j^{\alpha n+1 T} B_{jr}^n \left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right) - \nu \sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_j^{\alpha n+1 T} B_{jr}^n \left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right). \quad (6.46)$$

One should have noticed that cell velocities $\mathbf{u}_j^\alpha$ are solved implicitly. Since $\delta\mathbf{u}_j^{\alpha n+1}$ are used to compute total energy variation, one has to compute it first. The local linear system associated with $\mathbf{u}_j^{\alpha n+1}$ velocities is given by equations (6.40)

$$\begin{pmatrix} m_j^\alpha + \Delta t \sum_r \nu \rho_r^n B_{jr}^n & -\Delta t \sum_r \nu \rho_r^n B_{jr}^n \\ -\Delta t \sum_r \nu \rho_r^n B_{jr}^n & m_j^\beta + \Delta t \sum_r \nu \rho_r^n B_{jr}^n \end{pmatrix} \begin{pmatrix} \mathbf{u}_j^{\alpha n+1} \\ \mathbf{u}_j^{\beta n+1} \end{pmatrix} = \begin{pmatrix} m_j^\alpha \mathbf{u}_j^{\alpha n} - \Delta t \sum_r \mathbf{F}_{jr}^{\alpha,n} \\ m_j^\beta \mathbf{u}_j^{\beta n} - \Delta t \sum_r \mathbf{F}_{jr}^{\beta,n} \end{pmatrix}.$$

It is easy to check that this linear system is symetric and positive definite if B_{jr}^n matrices are symetric and positive.

6-4.4 Stability of the discrete scheme

In this section we establish that the scheme is stable for arbitrary equations of state: there exists $\Delta t > 0$ such that for each fluid $\alpha \in \{f_1, f_2\}$, $\tau_j^{\alpha n+1} > 0$, $e_j^{\alpha n+1} > e(T = 0)$ and $\eta_j^{\alpha n+1} \geq \eta_j^{\alpha n}$. For the sake of simplicity, and without loss of generality, we will consider in the following the case $e_j^{\alpha n+1} > 0$.

Actually, we will provide explicit timesteps for the positivity of density and internal energy, but we will only show that the increasing physical entropy timestep will be greater that the one of the mono-fluid case for given velocity fluxes, for which we established Property 3. The main reason is that there only exists existence results for entropy stability for cell-centered semi-Lagrangian schemes (even in 1D), see [Des01, Gal03].

6-4.4.1 Positivity of density

Since $p = p(\rho, e)$ one has to ensure that density cannot be made negative.

Property 8 (Positivity of density)**.** *Assuming that* $\forall \alpha \in \{f_1, f_2\}$, $\forall j \in \mathcal{M}$, $\rho_j^{\alpha n} > 0$. *Denoting* $\mathcal{C}^{\alpha n}$ *the set of compressive cells for each fluid* α, $\mathcal{C}^{\alpha n} := \left\{ j \in \mathcal{M} / \sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} < 0 \right\}$, *there exists* $\Delta t^\rho > 0$ *such that,*

$$\forall \alpha \in \{f_1, f_2\},\ \forall j \in \mathcal{C}^{\alpha n}, \quad \Delta t^\rho < \frac{V_j^{\alpha n}}{-\sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n}}.$$

Then, the scheme (6.39)–(6.43) defined by $\Delta t \in]0, \Delta t^\rho]$ *ensures that*

$$\forall \alpha \in \{f_1, f_2\},\ \forall j \in \mathcal{M}, \quad \rho_j^{\alpha n+1} > 0.$$

Observe that, as expected, only compressive cells ($j \in \mathcal{C}^{\alpha n}$) can lead to negative densities, so in the case of non-compressive flows, Δt^ρ may be arbitrarly large. Also, in the case of trianglular meshes, this constrain implies that no cell will tangle during the timestep.

Proof. Obviously, this is equivalent to show that $\tau_j^{\alpha n+1} = \frac{1}{\rho_j^{\alpha n+1}} > 0$. According to (6.39), one has

$$\tau_j^{\alpha n+1} = \tau_j^{\alpha n} + \frac{\Delta t}{m_j^\alpha} \sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n}.$$

So, one has the following alternative:

— if $j \notin \mathcal{C}^{\alpha n}$ that is $\sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} \geq 0$, then $\forall \Delta t > 0$ one has $\tau_j^{\alpha n+1} > 0$,

— else if $j \in \mathcal{C}^{\alpha n}$, one has $\sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} < 0$, then $\forall \Delta t < \tau_j^{\alpha n} \frac{m_j^\alpha}{-\sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n}}$, one has $\tau_j^{\alpha n+1} > 0$. Since $\frac{m_j^\alpha}{-\sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n}} > 0$, the existence of such a $\Delta t > 0$ is obvious.

■

6-4.4.2 Positivity of internal energy

First, as a primary result, we give internal energy variation for fluid $\alpha \in \{f_1, f_2\}$, β denoting the other one.

Lemma 1. *After one time step of scheme (6.39)–(6.43), internal energy is updated as*

$$\begin{aligned} e_j^{\alpha n+1} = e_j^{\alpha n} &+ \frac{\Delta t}{m_j^\alpha}\left[\sum_r (\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n})^T A_{jr}^{\alpha\, n}(\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n}) - \sum_r p_j^{\alpha n} \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n}\right] \\ &+ \nu \frac{\Delta t}{m_j^\alpha}\left[\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} + \sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_j^{\alpha n+1^T} B_{jr}^n (\delta\mathbf{u}_j^{\alpha n+1} - \delta\mathbf{u}_r^{\alpha n})\right] \\ &- \frac{\Delta t^2}{2m_j^{\alpha 2}}\left(\sum_r A_{jr}^{\alpha\, n}(\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n})\right)^2 + \frac{\Delta t^2}{2m_j^{\alpha 2}}\left(\nu \sum_r \rho_r^n B_{jr}^n (\delta\mathbf{u}_j^{\alpha n+1} - \delta\mathbf{u}_r^{\alpha n})\right)^2. \end{aligned} \tag{6.47}$$

Proof. See 6-2.6 Page 155. ■

Actually, (6.47) can be rewritten as

$$\begin{aligned} e_j^{\alpha n+1} = e_{h_j}^{\alpha n+1} &+ \nu \frac{\Delta t}{m_j^\alpha}\left[\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} + \sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_j^{\alpha n+1^T} B_{jr}^n (\delta\mathbf{u}_j^{\alpha n+1} - \delta\mathbf{u}_r^{\alpha n})\right] \\ &+ \frac{\Delta t^2}{2m_j^{\alpha 2}}\left(\nu \sum_r \rho_r^n B_{jr}^n (\delta\mathbf{u}_j^{\alpha n+1} - \delta\mathbf{u}_r^{\alpha n})\right)^2, \end{aligned} \tag{6.48}$$

where $e_{h_j}^{\alpha n+1}$ denotes the obtained internal energy without friction: *i.e.* substituting nodal velocities $\mathbf{u}_r^{\alpha n}$ into the classic mono-fluid scheme. The remaining terms can be viewed as the heating due to the friction.

Corollary 1 (Page 151) allows to give a lower bound to $e_j^{\alpha n+1}$

$$\begin{aligned} e_j^{\alpha n+1} \geq e_{h_j}^{\alpha n+1} &+ \nu \frac{\Delta t}{m_j^\alpha}\left[\frac{1}{2}\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} + \frac{1}{2}\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_j^{\alpha n+1^T} B_{jr}^n \delta\mathbf{u}_j^{\alpha n+1}\right] \\ &+ \frac{\Delta t^2}{2m_j^{\alpha 2}}\left(\nu \sum_r \rho_r^n B_{jr}^n (\delta\mathbf{u}_j^{\alpha n+1} - \delta\mathbf{u}_r^{\alpha n})\right)^2, \end{aligned} \tag{6.49}$$

which implies $e_j^{\alpha n+1} \geq e_{h_j}^{\alpha n+1}$, since friction terms are positive.

Property 9 (Positivity of internal energy)**.** *Assuming that* $\forall \alpha \in \{f_1, f_2\}$, $\forall j \in \mathcal{M}$, $e_j^{\alpha n} > 0$, *there exists* $\Delta t^e > 0$ *such that the scheme (6.39)–(6.43) ensures that*

$$\forall \Delta t \in]0, \Delta t^e[, \forall \alpha \in \{f_1, f_2\},\ \forall j \in \mathcal{M}, \quad e_j^{\alpha n+1} > 0.$$

Proof. The proof is obvious since $e_j^{\alpha n+1} \geq e_{h_j}^{\alpha n+1}$ and since $e_{h_j}^{\alpha n+1}(\Delta t)$ is a polynomial of degree 2 satisfying $e_{h_j}^{\alpha n+1}(0) = e_j^{\alpha n} > 0$. Δt^e is nothing but the smallest root of these polynomials for each cell of each fluid. ■

6-4.4.3 Entropy stability for general equations of state

In the previous paragraph, we provided explicitly a choice of $\Delta t > 0$ that ensures positivity of internal energy and density for the proposed scheme, but this is not sufficient for stability. In this section, we give an existence result of a strictly positive timestep Δt that ensures production of physical entropy for arbitrary physical equations of state.

Property 10 (Entropy)**.** *Let $U := \left(\tau, \mathbf{u}^T, E\right)^T$ and let η the entropy. There exists $\Delta t^\eta > 0$, such that $\forall \alpha, \beta \in \{f_1, f_2\}$, $\alpha \neq \beta$, if the pressure law $p^\alpha : (\rho, e) \to p^\alpha(\rho, e)$ is a differentiable function, then the scheme (6.39)–(6.43) defined by $\Delta t = \Delta t^\eta$ ensures that,*

1. *the scheme is entropy stable:*
$$\forall j \in \mathcal{M}, \quad \eta\left(U_j^{\alpha n+1}\right) \geq \eta\left(U_j^{\alpha n}\right),$$
2. *and $\forall j \in \mathcal{M}$, one has the following alternative. If $\forall r \in \mathcal{R}_j$, $\mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} = \mathbf{C}_{jr}^n \cdot \mathbf{u}_j^{\alpha n}$ and $\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1} = \mathbf{0}$, then*
$$T_j^{\alpha n} m_j^\alpha \frac{\eta(U_j^{\alpha n+1}) - \eta(U_j^{\alpha n})}{\Delta t} \geq \nu \sum_r \rho_r^\beta \ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} + \mathcal{O}(\Delta t),$$
else
$$T_j^{\alpha n} m_j^\alpha \frac{\eta(U_j^{\alpha n+1}) - \eta(U_j^{\alpha n})}{\Delta t} \geq \nu \sum_r \rho_r^\beta \ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n}.$$

Proof. See 6-2.7 Page 156. ■

Remark 5. *Let us comment point 2 of Property 10. Actually, this is a consistency result with regard to (6.2). In the first case (if $\forall r \in \mathcal{R}_j$, $\mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} = \mathbf{C}_{jr}^n \cdot \mathbf{u}_j^{\alpha n}$ and $\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1} = \mathbf{0}$), the scheme gives following values $\rho_j^{\alpha n+1} = \rho_j^{\alpha n}$, $\mathbf{u}_j^{\alpha n+1} = \mathbf{u}_j^{\alpha n}$ and $e_j^{\alpha n+1} = e_j^{\alpha n} + \frac{\Delta t}{m_j^\alpha} \nu \sum_r \rho_r^{\beta n} \ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n}$. In this case, the scheme acts simply as a first-order ODE solver. Since then $d\eta = de$ and since η is strictly convex, a time integration error is to be expected.*

To sum up, we proved that the proposed scheme is stable, meaning that there exists $0 < \Delta t \leq \min(\Delta t^\rho, \Delta t^e, \Delta t^\eta)$ such that the scheme is entropy stable and preserves positivity of density and internal energy. Moreover, it is consistent with (6.2).

6-4.4.4 A lower bound to $\Delta t^{\alpha,\nu}$

As stated before, to prove that the scheme is asymptotic preserving, it remains to show that $\lim_{\nu \to +\infty} \Delta t^{\alpha,\nu} \neq 0$. Even if we will not provide here an explicit value, we will give a lower bound independent of ν.

Property 11. *$\forall j \in \mathcal{M}$, let $\left(\tau_j^n, \mathbf{u}_j^{nT}, E_j^n\right)^T$ denotes the initial state of fluid $\alpha \in \{f_1, f_2\}$. Let $\{\mathbf{u}_r\}_{r \in \mathcal{R}_j}$, be an arbitrary set of nodal velocities (or velocity fluxes). Then, if $\forall \nu \geq 0$, $\left(\tau_j^{\nu,n+1}, e_j^{\nu,n+1}\right)$ denotes the thermodynamic state obtained by scheme (6.39)–(6.41), one has*
$$\eta\left(\tau_j^{\nu,n+1}, e_j^{\nu,n+1}\right) \geq \eta\left(\tau_j^{0,n+1}, e_j^{0,n+1}\right),$$

where $\eta := \eta(\tau, e)$ *is the physical entropy expressed according to the independent variables* τ *and* e.

Proof. Gibbs formula reads $\nabla_{\tau,e}\eta = \frac{1}{T}\binom{p}{1}$, where $T := T(\tau, e)$ is a positive function. So, for any τ, $\eta(\tau, \cdot)$ is an increasing function.

Since (6.39) is independent of ν and according to (6.49), one has

$$\forall \{\mathbf{u}_r\}_{r\in\mathcal{R}_j},\ \forall \nu \geq 0,\ \forall \Delta t \qquad \tau_j^{\nu,n+1} = \tau_j^{0,n+1} \quad \text{and} \quad e_j^{\nu,n+1} \geq e_j^{0,n+1},$$

so

$$\forall \{\mathbf{u}_r\}_{r\in\mathcal{R}_j},\ \forall \nu \geq 0,\ \forall \Delta t \qquad \eta\left(e_j^{\nu,n+1}, \tau_j^{\nu,n+1}\right) \geq \eta\left(e_j^{0,n+1}, \tau_j^{0,n+1}\right).$$

■

Property 11 establishes that, for an arbitrary set of nodal velocities $\{\mathbf{u}_r\}_{r\in\mathcal{R}_j}$, the maximum timestep required for the scheme to be stable is greater for any $\nu \geq 0$ than for $\nu = 0$. We recall that for the asymptotic scheme (6.31)–(6.34), the nodal velocity is solution of

$$\left(\sum_j A_{jr}^{\alpha} + A_{jr}^{\beta}\right)\mathbf{u}_r = \sum_j \left(A_{jr}^{\alpha} + A_{jr}^{\beta}\right)\mathbf{u}_j + \sum_j \mathbf{C}_{jr}\left(p_j^{\alpha} + p_j^{\beta}\right).$$

One recognizes the solution of the nodal solver in the monofluid case with a Dalton mixture law. So, the timestep $\lim_{\nu\to+\infty} \Delta t^{\alpha,\nu}$ is lower bounded independently of ν.

6-4.4.5 On the importance of the implicit velocities in (6.39)–(6.41)

Using the notations defined in Section 6-4.3, let us consider the fully explicit scheme that consists in replacing momentum and total energy updates in (6.39)–(6.43) by their explicit counterparts

$$m_j^{\alpha}\frac{\mathbf{u}_j^{\alpha n+1} - \mathbf{u}_j^{\alpha n}}{\Delta t} = -\sum_r \mathbf{F}_{jr}^{\alpha,n} - \sum_r \nu\rho_r^n B_{jr}^n \delta\mathbf{u}_j^{\alpha n},$$

$$m_j^{\alpha}\frac{E_j^{\alpha n+1} - E_j^{\alpha n}}{\Delta t} = -\sum_r \mathbf{F}_{jr}^{\alpha,n}\cdot\mathbf{u}_r^{\alpha n} - \sum_r \nu\rho_r^n\ \overline{\mathbf{u}}_r^{n\,T} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} + \sum_r \nu\rho_r^n\ \overline{\mathbf{u}}_{jr}^{n\ T} B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n}\right).$$

Using this scheme, one easily checks that internal energy variation reads

$$\begin{aligned} e_j^{\alpha n+1} = e_j^{\alpha n} &+ \frac{\Delta t}{m_j^{\alpha}}\left[\sum_r (\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n})^T A_{jr}^{\alpha\ n}(\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n}) - \sum_r p_j^{\alpha n}\mathbf{C}_{jr}^n\cdot\mathbf{u}_r^{\alpha n}\right] \\ &+ \nu\frac{\Delta t}{m_j^{\alpha}}\left[\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n\delta\mathbf{u}_r^{\alpha n} + \sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_j^{\alpha nT} B_{jr}^n(\delta\mathbf{u}_j^{\alpha n} - \delta\mathbf{u}_r^{\alpha n})\right] \\ &- \frac{\Delta t^2}{2m_j^{\alpha 2}}\left(\sum_r A_{jr}^{\alpha\ n}(\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n}) + \nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_j^{\alpha n} - \delta\mathbf{u}_r^{\alpha n})\right)^2. \end{aligned}$$

That is

$$\begin{aligned} e_j^{\alpha n+1} = e_{h_j}^{\alpha n+1} &+ \nu\frac{\Delta t}{m_j^{\alpha}}\left[\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n\delta\mathbf{u}_r^{\alpha n} + \sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_j^{\alpha nT} B_{jr}^n(\delta\mathbf{u}_j^{\alpha n} - \delta\mathbf{u}_r^{\alpha n})\right] \\ &- \frac{\Delta t^2}{m_j^{\alpha 2}}\left(\nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_j^{\alpha n} - \delta\mathbf{u}_r^{\alpha n})\right)\cdot\left(\sum_r A_{jr}^{\alpha\ n}(\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n})\right) \\ &- \frac{\Delta t^2}{2m_j^{\alpha 2}}\left(\nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_j^{\alpha n} - \delta\mathbf{u}_r^{\alpha n})\right)^2, \end{aligned}$$

where $e_{h_j}^{\alpha n+1}$ still denotes the obtained internal energy without friction. The later term being a negative factor of ν^2, in the explicit case, $\forall \Delta t > 0$ for large values of ν, one can have $e_j^{\alpha n+1} < e_{h_j}^{\alpha n+1}$. So even if a similar result to Property 10 can be established (existence of an entropy stable timestep), one cannot prove an equivalent of Property 11. If cell velocities are explicit, one eventually gets $\lim_{\nu \to +\infty} \Delta t^e = \lim_{\nu \to +\infty} \Delta t^\eta = 0$ for a given set of nodal velocities $\{\mathbf{u}_r\}_{r \in \mathcal{R}_j}$.

6-5 ALE scheme

The semi-Lagrangian scheme presented in this paper is defined assuming that both fluids meshes are identical at the begining of the timestep. One understands easily that this is of huge help in the construction of an asymptotic preserving scheme. One could imagine a purely Lagrangian approach, but even dealing with a non-AP approach seems very difficult since one would have to consider meshes intersections and complex geometrical calculations.

Thus, the algorithm we propose in this paper consists in ensuring that for each timestep both fluids meshes coincide. To do so an ALE formulation is mandatory.

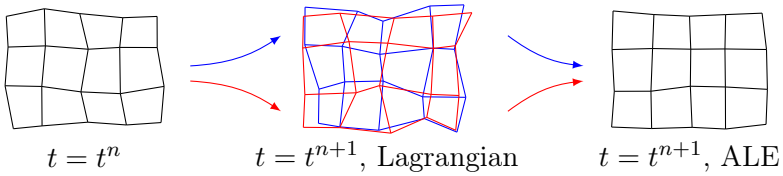

Figure 6.2 – **Left**: at time $t = t^n$, both fluids share the same mesh. **Middle**: at the end of the Lagrangian phase, one gets two different meshes, one for each fluid. **Right**: meshes are displaced so that they coincide. Solution is remapped and a new timestep can be performed.

Figure 6.2 depicts the general ALE case. Our ALE method is a Lagrange-rezoning-advection procedure which ensures that the solution is defined at time t^{n+1} on a unique mesh.

- At time t^n solutions are discretized on the meshes $\mathcal{M}_\alpha^n = \mathcal{M}_\beta^n$
- In a first step (Lagrangian phase), each mesh evolves in a different way $\tilde{\mathcal{M}}_\alpha^{n+1} \neq \tilde{\mathcal{M}}_\beta^{n+1}$. Each mesh being defined by $\tilde{\mathbf{x}}_r^{\alpha,n+1} = \mathbf{x}_r^n + \Delta t \mathbf{u}_r^{\alpha,n}$.
- Then the meshes are smoothed in a way to obtain new meshes such that $\mathcal{M}_\alpha^{n+1} \equiv \mathcal{M}_\beta^{n+1}$. For each fluid α, it allows to define an arbitrary velocity $\mathbf{v}_r^{\alpha,n+1}$ such that $\mathbf{x}_r^{n+1} = \tilde{\mathbf{x}}_r^{\alpha,n+1} + \Delta t \mathbf{v}_r^{\alpha,n+1}$.
- Finally, for both fluids, the numerical solution is computed on the common mesh by remapping the conservative variables $(\rho^\alpha, \rho^\alpha \mathbf{u}^\alpha, \rho^\alpha E^\alpha)^T$ at velocity $-\mathbf{v}_r^{\alpha,n+1}$, with a second-order accurate scheme. One can then compute another timestep.

In the test problems we have experienced three ALE strategies. **First strategy** consists in remapping both fluids on the inital grid for each time step n. Consequently $\forall \alpha \in \{1,2\}$, $\mathbf{v}_r^{\alpha,n+1} = \frac{\mathbf{x}_r^n - \tilde{\mathbf{x}}_r^{\alpha,n+1}}{\Delta t}$. **Second strategy** consists in considering that one fluid is Lagrangian (for instance fluid 1) and to remap the second fluid on the first fluid grid at each time step. In this case, $\mathbf{v}_r^{1,n+1} = \mathbf{0}$ and $\mathbf{v}_r^{2,n+1} = \frac{\tilde{\mathbf{x}}_r^{1,n+1} - \tilde{\mathbf{x}}_r^{2,n+1}}{\Delta t}$. **Third strategy** consists in performing an iteration of barycentric smoother to one of the mesh (for instance $\tilde{\mathcal{M}}_1^{n+1}$) at the end of each Lagrangian step, then consider this new mesh as the initial common mesh for the following step, and finally deducing the advection velocities for both fluids. The algorithm involved in the projection step is classical and aims at solving the equation $\partial_t \varphi = 0$, $\forall \varphi$ on the whole domain Ω, from step n to $n+1$. The point of view we choose in this work is called "sweeping" in the literature [Ben92]. It

consists in considering this step as a transport problem from the domain Ω^n to the domain Ω^{n+1}. The algorithm we use for this step is consistent with the previous equation, conservative for the variables ρ^α, ρ^β, $\rho^\alpha \mathbf{u}^\alpha$, $\rho^\beta \mathbf{u}^\beta$, $\rho^\alpha E^\alpha$ and $\rho^\beta E^\beta$, and preserves the local maximum principle.In consequence, the properties of the Lagrangian step remain valid for the global algorithm. Even if the problem is formulated as an advection step, we insist on the fact that the aim is only to project the solutions from one grid to another one, and no physics is involved. However, in practice, the remapping could artificially increase the gap between the velocities of both fluids. This is why we perform most of the test problems with the three strategies to evaluate this effect. We found this impact negligible, and then we consider that the Asymptotic Preserving behaviour is not affected by the remapping step, for these ALE strategies.

6-6 Numerical tests

6-6.1 Reference scheme

Let us recall that the reference scheme (6.22)–(6.23) is entropic, conservative in volume and mass for each fluid and in the sum of momenta and total energies. Also, the scheme is weakly consistent with (6.1). One can moreover show that its associated discrete in time scheme, where only $\mathbf{u}_j$ terms are implicit, is stable in the same way as scheme (6.39)–(6.43).

However, this scheme does not *a priori* preserve the asymptotic. For these reasons this scheme is a very good candidate for the comparisons we perform in this section.

6-6.2 Test conditions

In all the following tests, we choose $A^\alpha_{jr} = \rho^\alpha_j c^\alpha_j \sum_i \frac{N^i_{jr} \otimes N^i_{jr}}{\|N^i_{jr}\|}$ (Eucclhyd scheme) and $B_{jr} = V_{jr} I_2$, with $V_{jr} = \frac{1}{\#\mathcal{R}_j} V_j$. Also, for each test one chooses $\gamma^\alpha = \gamma^\beta = 1.4$.

Results are compared with the non-AP scheme (6.22)–(6.23). Also for the 2D tests, we compare our results ($\nu >> 1$) to the mono-fluid case, where mass fraction $\frac{\rho^\alpha}{\rho^\alpha + \rho^\beta}$ is treated as a passive scalar.

As it is often the case for multi-velocity models [SA99], the scheme is only defined in regions where both fluids are present. Thus in regions where a fluid should be absent, one keeps a neglectable amount of it. In the tests, we use the ratio $\varepsilon = 10^{-3}$ to define the negligible amount of fluid at initial time. Lower values such as 10^{-6} can lead to instabilities of the scheme since the thermodynamic initial state is very challenging.

6-6.3 Sod shock tube

This test is taken from [Ena07]. The computational domain we consider is $\Omega :=]0,1[\times]0,0.1[$. Initial data is given as $U := (\rho, \mathbf{u}, p)^T$, so that one defines $U^L := (1, \mathbf{0}, 1)^T$, $U^R := (0.125, \mathbf{0}, 0.1)^T$ and $U^\varepsilon := (\varepsilon, \mathbf{0}, \varepsilon)^T$. For both fluids initial states are then

$$U^\alpha = \mathbf{1}_{]0,0.5[}(U^L - U^\varepsilon) + \mathbf{1}_{]0.5,1[} U^\varepsilon \quad \text{and} \quad U^\beta = \mathbf{1}_{]0,0.5[} U^\varepsilon + \mathbf{1}_{]0.5,1[}(U^R - U^\varepsilon),$$

where $\mathbf{1}_O$ denotes the characteritic function of the set O and where we take $\varepsilon = 10^{-3}$.

On Figure 6.3, we compare the solution at time $t = 0.14$ obtained by the proposed scheme (6.39)–(6.43) to the reference scheme (6.22)–(6.23) in the case $\nu = 1000$. One plots the density sum: $\rho^\alpha + \rho^\beta$. The grid is $200x3$ cells and the solution is compared to a reference solution obtained

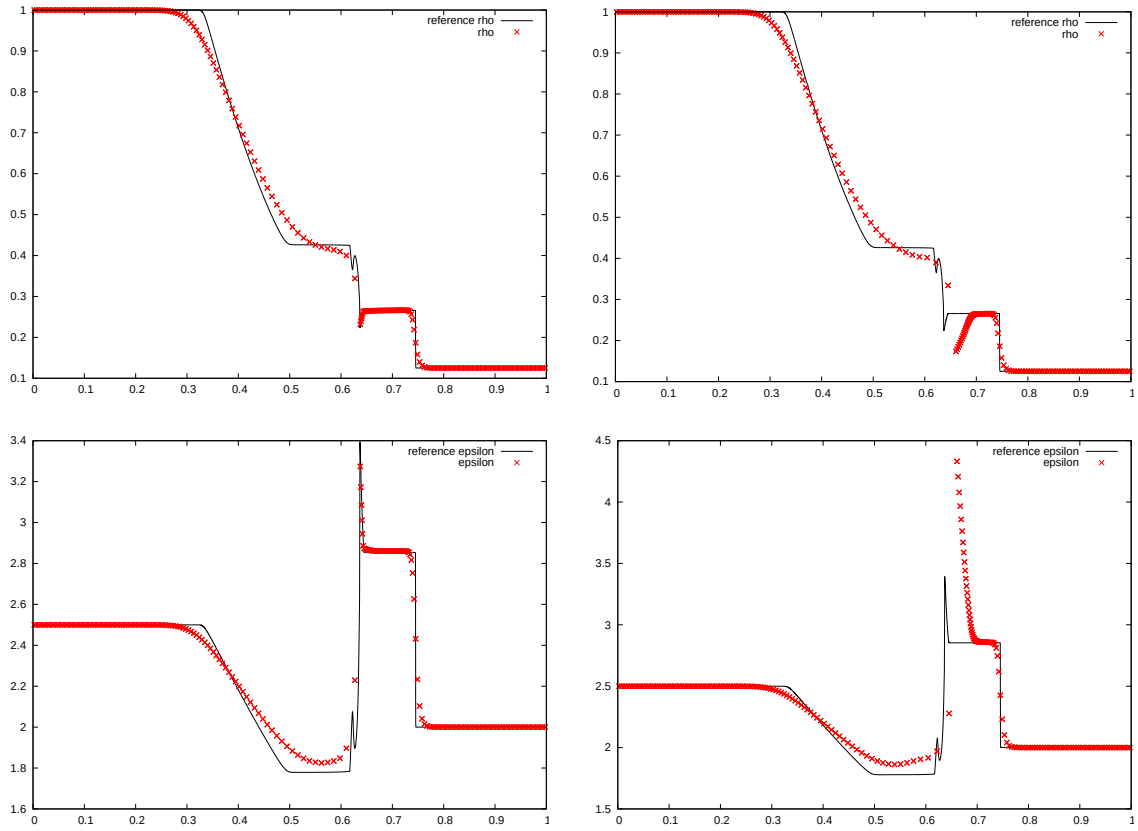

Figure 6.3 – $\nu = 1000$. **Top:** density $\rho^\alpha + \rho^\beta$ profile. **Bottom:** internal energy $\frac{\rho^\alpha \epsilon^\alpha + \rho^\beta \epsilon^\beta}{\rho^\alpha + \rho^\beta}$. AP-scheme (left) gives a much better solution than the non-AP scheme (right).

using a $10^4 \times 3$ grid. The simulation is Lagrangian: the left fluid imposes the mesh to the right one.

The same test is performed for a friction parameter $\nu = 10^6$. The density sum is presented on Figure 6.4 at time $t = 0.14$.

One retrieves the results presented in [Ena07], even if the scheme does not degenerate in 1D to the scheme proposed in [Ena07].

6-6.4 Triple-point problem

The triple-point problem is a standard benchmark [Lou05]. It is a multidimensional Riemann problem whose data are close to the Sod shock tube. The self-similarity of the problem yields an infinitely rolling vortex, the quantity of the details generated by the secondary Kelvin-Helmoltz instabilities depends only on the numerical dissipation of the scheme. Figure 6.5 depictes the initial geometry and the initial three states.

Let us define $\rho^L = 1$, $\rho^l = 0.125$, $p^L = 1$ and $p^l = 0.1$. Also, $\Omega_1 =]0,1[\times]0,3[$, $\Omega_2 =]1,7[\times]0,1.5[$ and $\Omega_3 =]1,7[\times]1.5,3[$. This allows to define the initial states of both fluids:

$$U^\alpha = \mathbf{1}_{\Omega_2} \begin{pmatrix} \rho^L - \varepsilon \\ \mathbf{0} \\ p^l - \varepsilon \end{pmatrix} + \mathbf{1}_{\Omega_1 \cup \Omega_3} \begin{pmatrix} \varepsilon \\ \mathbf{0} \\ \varepsilon \end{pmatrix} \quad \text{and} \quad U^\beta = \mathbf{1}_{\Omega_1} \begin{pmatrix} \rho^L - \varepsilon \\ \mathbf{0} \\ p^L - \varepsilon \end{pmatrix} + \mathbf{1}_{\Omega_3} \begin{pmatrix} \rho^l - \varepsilon \\ \mathbf{0} \\ p^l - \varepsilon \end{pmatrix} + \mathbf{1}_{\Omega_2} \begin{pmatrix} \varepsilon \\ \mathbf{0} \\ \varepsilon \end{pmatrix}.$$

Symmetry boundary conditions are set at each straight boundary of the computational domain.

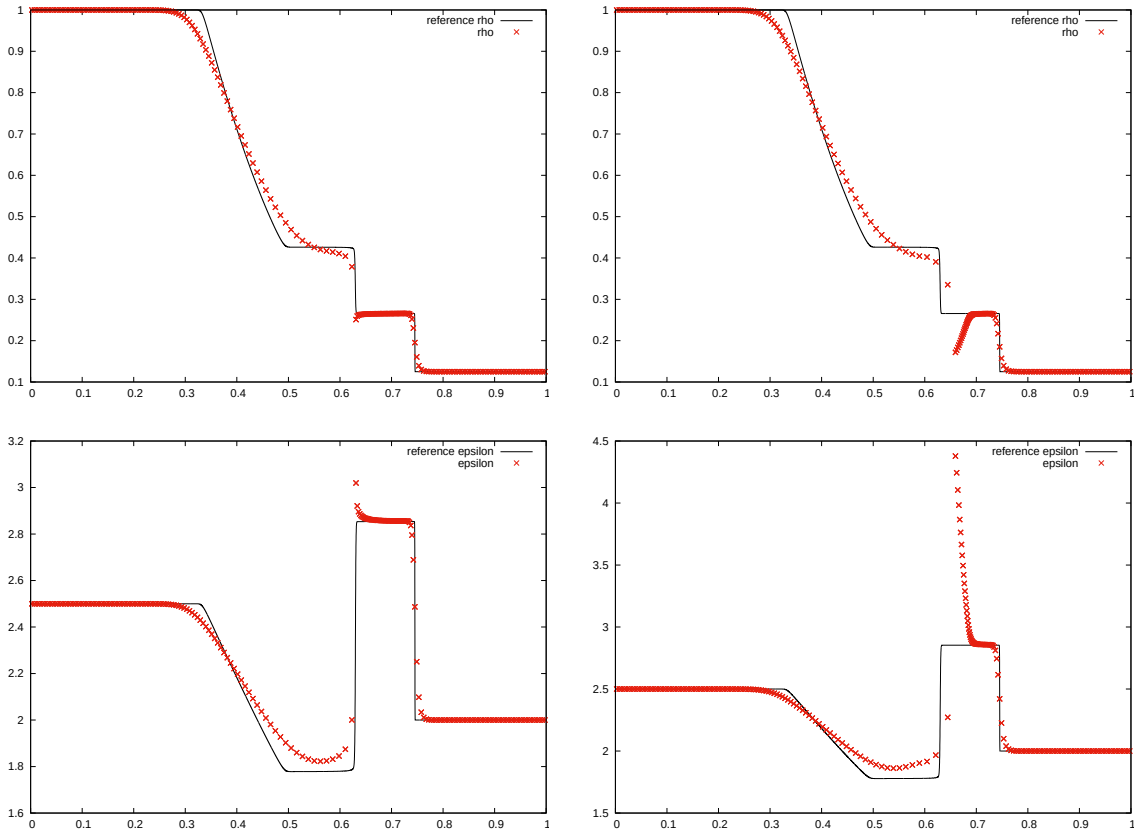

Figure 6.4 – $\nu = 10^6$. **Top:** density $\rho^\alpha + \rho^\beta$ profile. **Bottom:** internal energy $\frac{\rho^\alpha \epsilon^\alpha + \rho^\beta \epsilon^\beta}{\rho^\alpha + \rho^\beta}$. AP-scheme (left) gives a much better solution than the non-AP scheme (right). The expected solution is close to the classical mono-fluid case.

The ALE strategy we use for this test consists in a barycentric smoother for the grid of the fluid α and then to impose $\mathbf{x}_r^\beta = \mathbf{x}_r^\alpha$.

We run the test on a 91×40 grid. Choosing the friction parameter $\nu = 10^6$, we compare the obtained result to the solution of the mono-fluid solver and to the non-AP scheme, see Figure 6.6. For the comparison, we plot the mass fraction in each case: $\frac{\rho^\alpha}{\rho^\alpha + \rho^\beta}$. One notices the nice agreement of the solution for the proposed scheme with regard to the mono-fluid case, even for this small amount of cells, whereas the non-AP scheme is not even able to compute the large structures of the flow at this grid resolution.

Then we study the effect of the friction parameter. Figure 6.7 presents the obtained solutions, on a finer 210×90 grid, for $\nu \in \{10, 100, 10^6\}$.

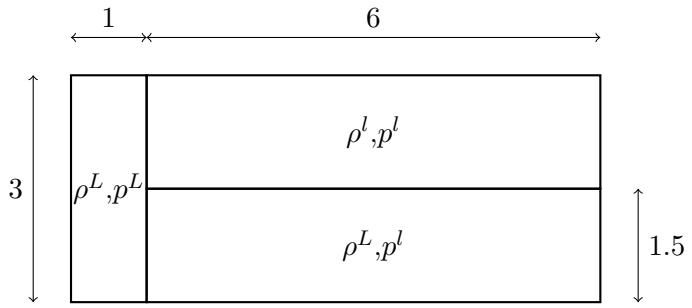

Figure 6.5 – Geometry, pressures and densities for the triple-point problem at time $t = 0$.

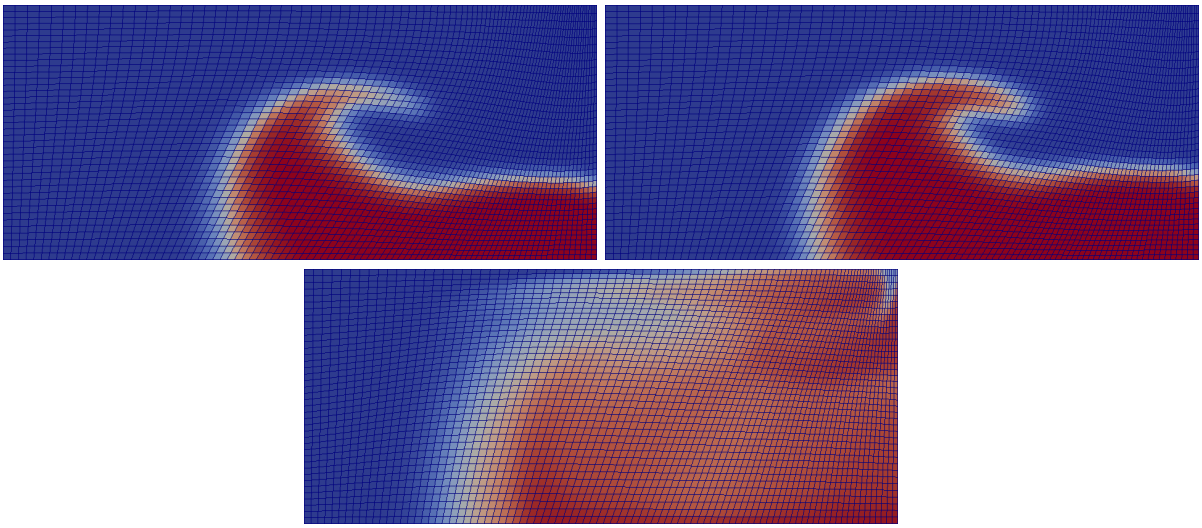

Figure 6.6 – 91×40 mesh. Mass fraction of fluid α at time $t = 5$. **Left**: mono-fluid solution. **Right**: bi-fluid solution with $\nu = 10^6$. **Bottom:** bi-fluid solution with non-AP scheme with $\nu = 10^6$.

6-6.5 A Rayleigh Taylor instability

For this test, we modify the scheme in order to incorporate the gravity treatment. Obviously, we use a well-balanced approach [CL95] to take this term into account. For this modified scheme, the properties we established for (6.39)–(6.44) remain true. We did not take the gravity term into account in Section 6-4 to avoid a more complex presentation, since there is no additional difficulties to overcome.

The interface perturbation is defined by the function $f(y) = 0.05\cos(8\pi y)$ and centered at $x = 0.35$ in the computational domain $\Omega =]0, 0.7[\times]0, 0.25[$. Thus, two regions are defined: $\Omega_\alpha = \{(x, y) \in \Omega / x < 0.35 + f(y)\}$ and $\Omega_\beta = \Omega \setminus \Omega_\alpha$.

Initially, velocities are set to $\mathbf{0}$ in Ω, and densities are defined as

$$\rho^\alpha = \mathbf{1}_{\Omega_\alpha}(0.8 - \varepsilon) + \mathbf{1}_{\mathbf{\Omega}_\beta}\varepsilon, \quad \text{and} \quad \rho^\beta = \mathbf{1}_{\Omega_\alpha}\varepsilon + \mathbf{1}_{\mathbf{\Omega}_\beta}(0.25 - \varepsilon).$$

Choosing the gravity acceleration as $\mathbf{g} = 9.8\ \mathbf{e}_x$, we define the pressure in the whole domain at a quasi-equilibrium state (omitting the y dependancy), that is

$$p(x) = \int_0^x \left(\rho^\alpha + \rho^\beta\right) \mathbf{g} \cdot \mathbf{e}_x.$$

Again, symmetry boundary conditions are imposed all over $\partial\Omega$. We represent the mass fraction of fluid α that is $\frac{\rho^\alpha}{\rho^\alpha + \rho^\beta}$. We use the same ALE strategy as in the previous test: a barycentric remapping is performed on the mesh of fluid α and we set $\mathcal{M}_\beta^{n+1} = \mathcal{M}_\alpha^{n+1}$ to allow the calculation of timestep $n + 1$.

At first, we validate the approach by comparing the obtained result to the mono-fluid scheme. The results are presented on Figure 6.9, one observes again a very good agreement even on a 112×40 coarse grid. As expected, the non-AP scheme clearly shows lack of convergence.

Finally, we study the influence of the friction parameter ν for successive values of 100, 1000 and 10^6. A slightly finer grid (224×80) is used for it.

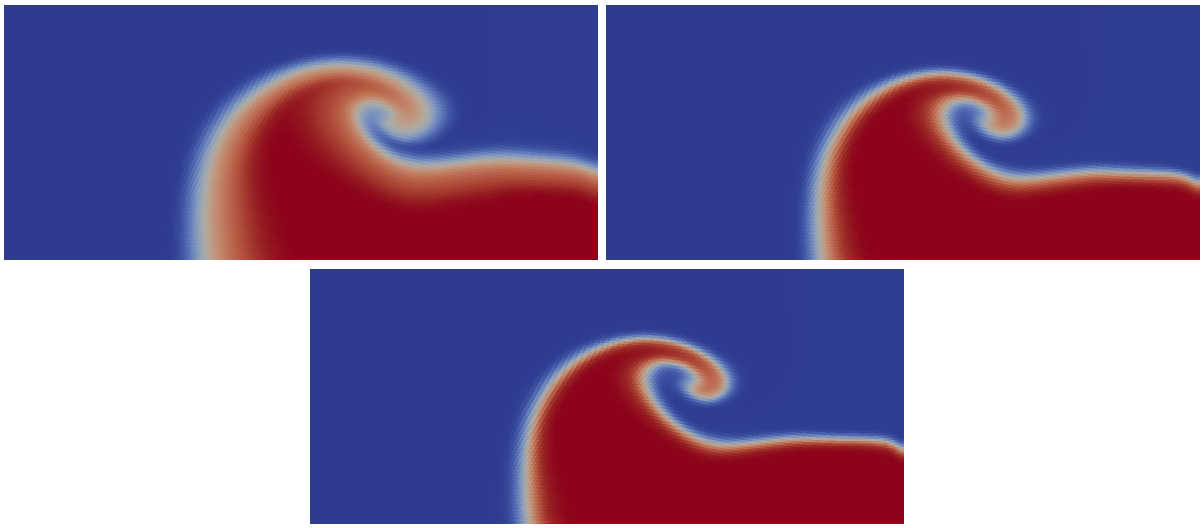

Figure 6.7 – 210×90 mesh. Time $t = 5$. Mass fraction of fluid α. Effect of the friction parameter ν. **Left**: $\nu = 10$. **Right**: $\nu = 100$. **Bottom**: $\nu = 10^6$.

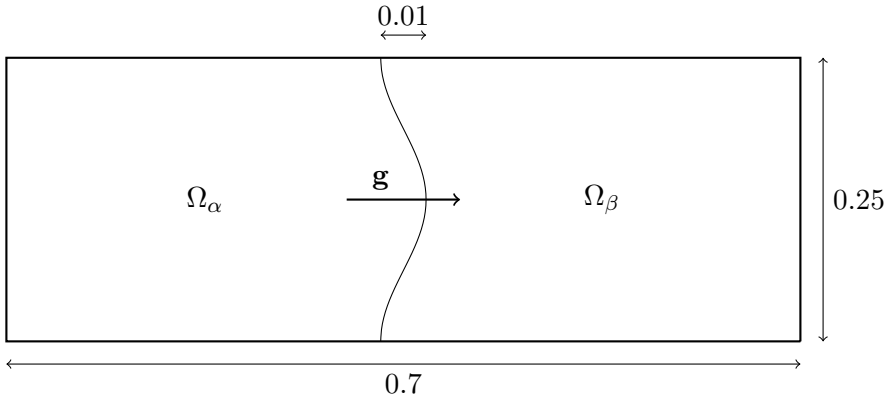

Figure 6.8 – Rayleigh-Taylor test initial geometry. Fluid α being heavier than fluid β, instability will grow.

6-7 Conclusion

In this paper, we presented a multi-dimensionnal asymptotic preserving scheme to solve a bi-fluid model defined as a set of two Euler systems coupled with a friction term. The originality of the approach is that the scheme is ALE: the only constrain being that meshes must coincide at the begining of each timestep.

The scheme is conservative and weakly-consistent by construction. Moreover, we showed that it is at least as stable as the underlying hydro-scheme in the sense that the timestep required to increase entropy does not tend to zero when friction increases. We showed consistency of the limit scheme ($\nu \to +\infty$) to the limit model. So, we proved that the scheme is asymptotic preserving. On the way we proved some stability results with regard to the fluxes $\mathbf{u}_r$, which give some bounds independently of ν (Property 3), and complete the numerical analysis of the scheme.

The numerical results show that the scheme behaves as expected and appears to be a good candidate to study interpenetration mixing [SC02], which is the goal of this work. Actually, all the results [1] can be established with a varying positive friction ν. In the paper we kept ν constant

1. If friction parameter depends on the cell data ($\nu = \nu_j$), Property 3 takes a slightly different form.

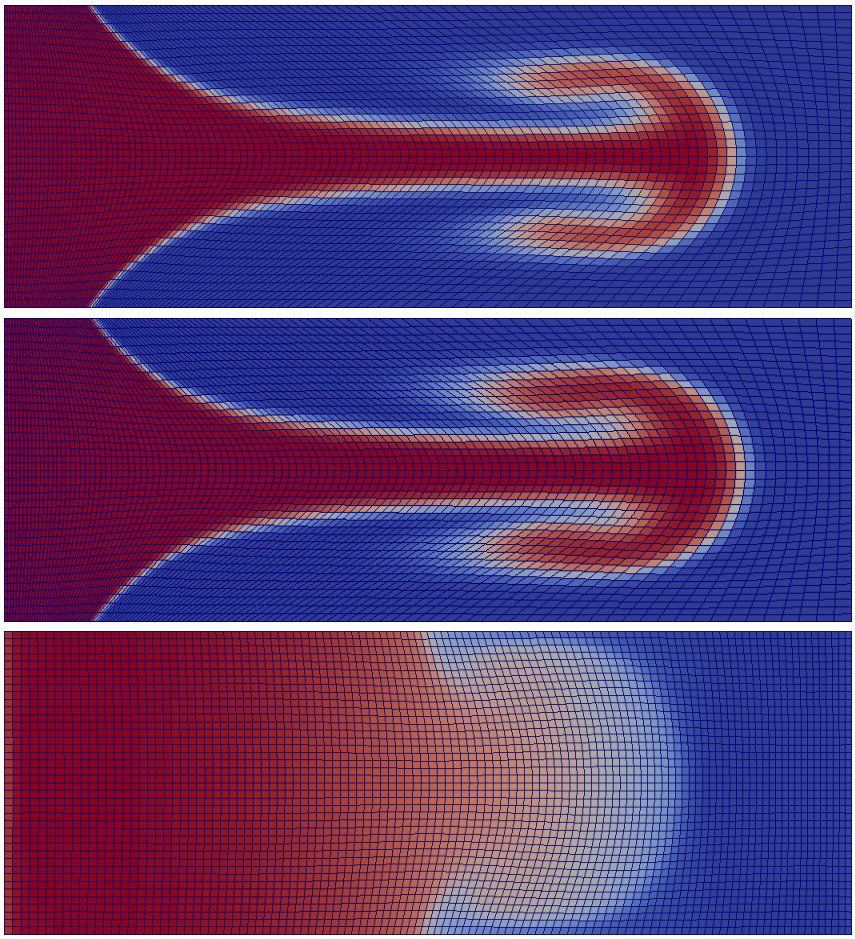

Figure 6.9 – 112×40 mesh. Mass fraction of fluid α. Time $t = 0.7s$. **Top**: mono-fluid solution. **Middle**: bi-fluid solution with $\nu = 10^6$. **Bottom**: bi-fluid solution with non-AP scheme with $\nu = 10^6$.

for the sake of simplicity. The numerical analysis and tests are performed in 2D, however the analysis in 3D is completely unchanged.

On the numerical point of view, a second-order accurate version of the scheme would be of interest. However, this is not an easy task for two main reasons. First, on the theoretical point of view, establishing properly the asymptotic preserving property would be challenging. Second, using a Runge-Kutta-like approach to get second-order accuracy in time would probably impose to incorporate the remeshing into the time integration or to consider a one-step approach.

Another extension is to introduce more physics in the model. The friction coupling is a very simple approach, one could use more appropriate closures based on the presented work. For instance see [SC02] in which this kind of model is used to accounts for eddy diffusivity, or [BDDM11] where Lorentz forces are taken into account in a ion-electron mixture.

Acknowledgement

Authors are grateful to C. Buet, C. Enaux, H. Jourdren and F. Lagoutière for their valuable comments and remarks about this work.

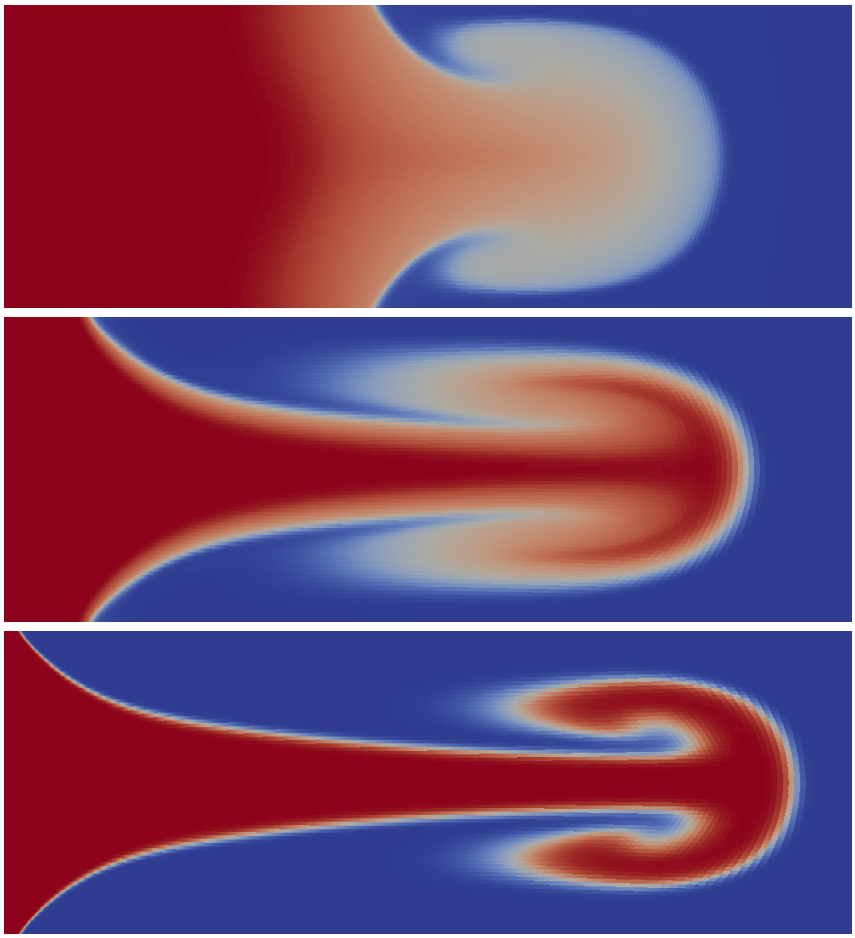

Figure 6.10 – 224×80 mesh. Time $t = 0.7$. Mass fraction of fluid α. Influence of the friction parameter. **Top**: $\nu = 100$. **Middle**: $\nu = 1000$. **Bottom**: $\nu = 10^6$.

Appendices

Appendix 6.A Asymptotic Preserving scheme in dimension one

We give in this section the one dimensional version of the scheme proposed in Section 6-4. The derivation is similar in many aspects with the work of Enaux [Ena07], and we invite the reader to refer to this work for more details.

For the 1D-version of the scheme, we propose some modifications of the notations, in order to mimic the usual 1D framework of Riemann solvers. Instead of noting r the nodes of the cells, we use $j-1/2$ and $j+1/2$, to design respectively the left-sided and right-sided nodes of the cell j. We also use the standard notations $p^\star$ and $u^\star$ for the Riemann pressure and velocity. With these notations, in 1D, the 2D vectors $\mathbf{C}_{jr}$ reduce to -1 in $j-1/2$ and 1 in $j+1/2$. The 2×2 A^α_{jr} and B_{jr} matrices reduce respectively to the scalars $\rho^\alpha_j c^\alpha_j$ and $\Delta x_j/2$. In this framework the scheme (6.39)-(6.41) becomes

$$m^\alpha_j \frac{\tau^{\alpha n+1}_j - \tau^{\alpha n}_j}{\Delta t} = u^{\alpha,\star}_{j+1/2} - u^{\alpha,\star}_{j-1/2}, \tag{6.A.1}$$

$$m^\alpha_j \frac{u^{\alpha n+1}_j - u^{\alpha n}_j}{\Delta t} = -\left(p^{\alpha,\star}_{j+1/2} - p^{\alpha,\star}_{j-1/2}\right) - \frac{\nu}{2}(\rho^n_{j-1/2} + \rho^n_{j+1/2})\Delta x^n_j \delta u^{\alpha n+1}_j, \tag{6.A.2}$$

$$\begin{aligned} m^\alpha_j \frac{E^{\alpha n+1}_j - E^{\alpha n}_j}{\Delta t} =& -\left(p^{\alpha,\star}_{j+1/2}u^{\alpha,\star}_{j+1/2} - p^{\alpha,\star}_{j-1/2}u^{\alpha,\star}_{j-1/2}\right) \\ & - \frac{\nu}{2}\left(\rho^n_{j+1/2}\ \overline{u}^n_{j+1/2}\delta u^{\alpha,\star}_{j+1/2} + \rho^n_{j-1/2}\ \overline{u}^n_{j-1/2}\delta u^{\alpha,\star}_{j-1/2}\right)\Delta x^n_j \\ & + \frac{\nu}{2}\left(\rho^n_{j+1/2}\overline{u}^n_{j,j+1/2}\delta u^{\alpha,\star}_{j+1/2} + \rho^n_{j-1/2}\overline{u}^n_{j,j-1/2}\delta u^{\alpha,\star}_{j-1/2}\right)\Delta x^n_j \\ & - \frac{\nu}{2}\left(\rho^n_{j+1/2}\overline{u}^n_{j,j+1/2} + \rho^n_{j-1/2}\overline{u}^n_{j,j-1/2}\right)\delta u^{\alpha n+1}_j \Delta x^n_j. \end{aligned} \tag{6.A.3}$$

We emphasis that the discretization of the source terms in Eq. (6.A.3) is more complex than the one in [Ena07] p.128. We found it necessary for our multi-D proofs, in particular the entropy and AP behaviour. However, if we consider only the barotropic case (no energy equation and $p^\alpha_j(\rho^\alpha_j)$) we recover the same discretization of the conservations laws as in [Ena07] pp114–121 except for the definition of the mean velocity (called $\overline{u}^n_{j+1/2}$ in this work and $\tilde{u}$ in [Ena07]) which is slightly different.

The associated Riemann solver in 1D reads:

$$p^{\alpha,\star}_{j+1/2} = p^{\alpha,n}_j - \rho^{\alpha,n}_j c^{\alpha,n}_j (u^{\alpha,\star,n}_{j+1/2} - u^{\alpha,n}_{j+1/2}) - \frac{\nu}{2}\rho^n_{j+1/2}\Delta x^n_j \delta u^{\alpha,\star}_{j+1/2}, \tag{6.A.4}$$

and

$$p^{\alpha,\star}_{j+1/2} = p^{\alpha,n}_{j+1} + \rho^{\alpha,n}_{j+1} c^{\alpha,n}_{j+1} (u^{\alpha,\star,n}_{j+1/2} - u^{\alpha,n}_{j+1}) - \frac{\nu}{2}\rho^n_{j+1/2}\Delta x^n_{j+1}\delta u^{\alpha,\star}_{j+1/2}. \tag{6.A.5}$$

This solver is very similar to those proposed by Enaux [Ena07]. It combines the acoustic Godunov approximation to the usual *trick* of getting well-balanced scheme by incorporating the source terms into the solver. There are different ways to incorporate the source terms into the solver, the more grounded theoretically being described in [Gos13, DB16], leading to the above expression. In 1D and in the barotropic approximation, both conservation laws discretization and Riemann solver are very close to those in [Ena07], and all the proofs (including asymptotic behaviour) are the same.

Using the framework of well-balanced schemes leads however to a different discretization of the source terms (also different from what is proposed in [Ena07]). Then the system (6.A.1)-(6.A.3)

reads

$$m_j^\alpha \frac{\tau_j^{\alpha n+1} - \tau_j^{\alpha n}}{\Delta t} = u_{j+1/2}^{\alpha,\star} - u_{j-1/2}^{\alpha,\star}, \tag{6.A.6}$$

$$m_j^\alpha \frac{u_j^{\alpha n+1} - u_j^{\alpha n}}{\Delta t} = -\left(p_{j+1/2}^{\alpha,\star} - p_{j-1/2}^{\alpha,\star}\right) - \frac{\nu}{2}(\rho_{j-1/2}^n \delta u_{j-1/2}^{\alpha,\star}{}^{n+1} + \rho_{j+1/2}^n \delta u_{j+1/2}^{\alpha,\star}{}^{n+1})\Delta x_j^n, \tag{6.A.7}$$

$$m_j^\alpha \frac{E_j^{\alpha n+1} - E_j^{\alpha n}}{\Delta t} = -\left(p_{j+1/2}^{\alpha,\star} u_{j+1/2}^{\alpha,\star} - p_{j-1/2}^{\alpha,\star} u_{j-1/2}^{\alpha,\star}\right) - \frac{\nu}{2}\left(\rho_{j+1/2}^n\, \overline{u}_{j+1/2}^n \delta u_{j+1/2}^{\alpha,\star} + \rho_{j-1/2}^n\, \overline{u}_{j-1/2}^n \delta u_{j-1/2}^{\alpha,\star}\right)\Delta x_j^n. \tag{6.A.8}$$

The multi-D counterpart for this scheme should be (following [Fra12])

$$m_j^\alpha \frac{\tau_j^{\alpha n+1} - \tau_j^{\alpha n}}{\Delta t} = \sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n}, \tag{6.A.9}$$

$$m_j^\alpha \frac{\mathbf{u}_j^{\alpha n+1} - \mathbf{u}_j^{\alpha n}}{\Delta t} = -\sum_r \mathbf{F}_{jr}^{\alpha,n} - \sum_r \nu \rho_r^n B_{jr}^n \delta \mathbf{u}_r^{\alpha n}, \tag{6.A.10}$$

$$m_j^\alpha \frac{E_j^{\alpha n+1} - E_j^{\alpha n}}{\Delta t} = -\sum_r \mathbf{F}_{jr}^{\alpha,n} \cdot \mathbf{u}_r^{\alpha n} - \sum_r \nu \rho_r^n\, \overline{\mathbf{u}}_r^{nT} B_{jr}^n \delta \mathbf{u}_r^{\alpha n}. \tag{6.A.11}$$

However, we were not able to prove the asymptotic preserving property for this scheme. Moreover, applying it to the test problem 6-6.3, we found that the scheme (6.A.9)–(6.A.11) produce oscillations on $\delta \mathbf{u}_j$ which are not observed with the scheme (6.39)–(6.41) proposed in this work.

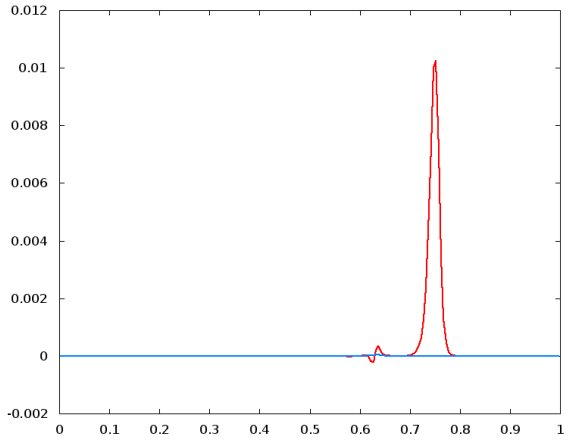

Figure 6.11 – "Sod shock tube". $\nu = 10^6$. Comparison of the $\delta \mathbf{u}_j$ obtained for the AP scheme (6.39)–(6.43) (blue) and for the *well-balanced* scheme (6.A.9)–(6.A.11) (red) at time $t = 1.4$ for a 200×3 grid.

Appendix 6.B Technical proofs

6-2.1 Proof of Property 3

Proof. $\forall \nu \geq 0$, $(\mathbf{u}_r^{\alpha,\nu}, \mathbf{u}_r^{\beta,\nu})$ is the unique solution of

$$\begin{pmatrix} A_r^\alpha + \nu\rho_r B_r & -\nu\rho_r B_r \\ -\nu\rho_r B_r & A_r^\beta + \nu\rho_r B_r \end{pmatrix} \begin{pmatrix} \mathbf{u}_r^{\alpha,\nu} \\ \mathbf{u}_r^{\beta,\nu} \end{pmatrix} = \mathbf{b}_r, \quad \text{with} \quad \mathbf{b}_r := \begin{pmatrix} \sum_j \mathbf{C}_{jr} p_j^\alpha \\ \sum_j \mathbf{C}_{jr} p_j^\beta \end{pmatrix}.$$

So, since $\mathbf{b}_r$ is independent of ν, one has

$$\forall \nu \geq 0, \quad \left(\mathbb{A}_r^0 + \nu\rho_r \Delta_r\right) \mathbf{u}_r^\nu = \mathbb{A}_r^0 \mathbf{u}_r^0, \tag{6.B.1}$$

where

$$\mathbb{A}_r^0 := \begin{pmatrix} A_r^\alpha & 0 \\ 0 & A_r^\beta \end{pmatrix}, \quad \Delta_r := \begin{pmatrix} B_r & -B_r \\ -B_r & B_r \end{pmatrix} \quad \text{and} \quad \mathbf{u}_r^\nu := \begin{pmatrix} \mathbf{u}_r^{\alpha,\nu} \\ \mathbf{u}_r^{\beta,\nu} \end{pmatrix}.$$

Multiplying on the left by $\mathbf{u}_r^\nu$ yields $\mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^\nu + \nu\rho_r \mathbf{u}_r^{\nu T}\Delta_r\mathbf{u}_r^\nu = \mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^0$. Since B_r is a positive matrix Δ_r is also positive, and since $\nu\rho_r \geq 0$, one gets

$$\forall \nu \geq 0, \quad \mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^\nu \leq \mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^0.$$

Finally, $\mathbb{A}_r^0$ being symmetric and positive definite, the simple following Youngs inequality,

$$\mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^0 \leq \frac{1}{2}\mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^\nu + \frac{1}{2}\mathbf{u}_r^{0T}\mathbb{A}_r^0\mathbf{u}_r^0,$$

allows to prove (6.28).

The proof of (6.29) follows the same way. Multiplying (6.B.1) on the left by $\mathbf{u}_r^\nu$, one has

$$\forall \nu \geq 0, \quad \nu\rho_r \mathbf{u}_r^{\nu T}\Delta_r\mathbf{u}_r^\nu + \mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^\nu = \mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^0.$$

Then, using the same Youngs inequality, one gets after a few arrangements

$$\forall \nu \geq 0, \quad \nu\rho_r \mathbf{u}_r^{\nu T}\Delta_r\mathbf{u}_r^\nu + \frac{1}{2}\mathbf{u}_r^{\nu T}\mathbb{A}_r^0\mathbf{u}_r^\nu \leq \frac{1}{2}\mathbf{u}_r^{0T}\mathbb{A}_r^0\mathbf{u}_r^0,$$

which yields to (6.29) since $\mathbb{A}_r^0$ is positive.

The third inequality is a bit more difficult to establish. Let us introduce the quadratic form $\mathsf{J}_\mathbf{v}^\nu := \frac{1}{2}\mathbf{v}^T\left(\mathbb{A}_r^0 + \nu\rho_r\Delta_r\right)\mathbf{v} - \mathbf{b}_r \cdot \mathbf{v}$. So, since $\mathbf{u}_r^\nu$ is the unique solution of the linear system, one has

$$\forall \nu \geq 0, \forall \mathbf{v}, \quad \mathsf{J}_{\mathbf{u}_r^\nu}^\nu \leq \mathsf{J}_\mathbf{v}^\nu.$$

In the particular case $\mathbf{v} = \mathbf{u}_r^0$, one gets $\mathsf{J}_{\mathbf{u}_r^\nu}^\nu \leq \mathsf{J}_{\mathbf{u}_r^0}^\nu$. It is then easy to check that

$$\mathsf{J}_{\mathbf{u}_r^0}^\nu = \frac{1}{2}\mathbf{u}_r^{0T}\left(\mathbb{A}_r^0 + \nu\rho_r\Delta_r\right)\mathbf{u}_r^0 - \mathbf{b}_r \cdot \mathbf{u}_r^0 = \mathsf{J}_{\mathbf{u}_r^0}^0 + \frac{\nu\rho_r}{2}\mathbf{u}_r^{0T}\Delta_r\mathbf{u}_r^0.$$

So, one has established a first inequality

$$\mathsf{J}_{\mathbf{u}_r^\nu}^\nu \leq \mathsf{J}_{\mathbf{u}_r^0}^0 + \frac{\nu\rho_r}{2}\mathbf{u}_r^{0T}\Delta_r\mathbf{u}_r^0. \tag{6.B.2}$$

Similarly, since $\mathbf{u}_r^0$ is the unique solution of the linear system in the case $\nu = 0$, one has $\mathsf{J}_{\mathbf{u}_r^0}^0 \leq \mathsf{J}_{\mathbf{u}_r^\nu}^0$, which can be written as

$$\mathsf{J}_{\mathbf{u}_r^0}^0 \leq \mathsf{J}_{\mathbf{u}_r^\nu}^\nu - \frac{\nu\rho_r}{2}\mathbf{u}_r^{\nu T}\Delta_r\mathbf{u}_r^\nu.$$

This actually gives a lower bound to $\mathsf{J}_{\mathbf{u}_r^\nu}^\nu$ which combined with its upper bound (6.B.2) yields

$$\mathsf{J}_{\mathbf{u}_r^0}^0 + \frac{\nu\rho_r}{2}\mathbf{u}_r^{\nu T}\Delta_r\mathbf{u}_r^\nu \leq \mathsf{J}_{\mathbf{u}_r^0}^0 + \frac{\nu\rho_r}{2}\mathbf{u}_r^{0T}\Delta_r\mathbf{u}_r^0.$$

Since $\nu\rho_r$ is positive, elementary calculations allow to write (6.30). ■

6-2.2 Proof of Property 4 (Conservation)

Proof. Conservations of mass and volume for each fluid are obvious since the associated balance equations are unchanged with regard to the mono-fluid schemes (see for instance [DM05, CDDL09, MABO07, Mai11]).

Summing momenta equations in (6.24) for both fluids gives

$$m_j^\alpha d_t \mathbf{u}_j^\alpha + m_j^\beta d_t \mathbf{u}_j^\beta = -\sum_r \mathbf{F}_{jr}^\alpha - \sum_r \mathbf{F}_{jr}^\beta - \sum_r \nu \rho_r B_{jr} (\delta \mathbf{u}_j^\alpha + \delta \mathbf{u}_j^\beta).$$

Recalling that by definition, $\delta\phi^\alpha + \delta\phi_j^\beta = 0$, one has

$$m_j^\alpha d_t \mathbf{u}_j^\alpha + m_j^\beta d_t \mathbf{u}_j^\beta = -\sum_r \mathbf{F}_{jr}^\alpha - \sum_r \mathbf{F}_{jr}^\beta.$$

The conservativity proof is ended in a standard way. One now sums these equations over the cells which gives

$$\sum_j m_j^\alpha d_t \mathbf{u}_j^\alpha + \sum_j m_j^\beta d_t \mathbf{u}_j^\beta = -\sum_j \sum_{r \in \mathcal{R}_j} \mathbf{F}_{jr}^\alpha - \sum_j \sum_{r \in \mathcal{R}_j} \mathbf{F}_{jr}^\beta,$$

that we rewrite

$$\sum_j m_j^\alpha d_t \mathbf{u}_j^\alpha + \sum_j m_j^\beta d_t \mathbf{u}_j^\beta = -\sum_r \sum_{j \in \mathcal{J}_r} \mathbf{F}_{jr}^\alpha - \sum_r \sum_{j \in \mathcal{J}_r} \mathbf{F}_{jr}^\beta.$$

This proves that momenta sum is conserved using (6.26) and recalling that cell masses are Lagrangian.

Conservation of total energies sum is obtained in the exact same way. ■

6-2.3 Proof of Property 5 (Entropy)

Let us establish a simple technical Lemma that will be useful in the following and to demonstrate Property 5.

Lemma 2. *Let M denote a symmetric matrix of $\mathbb{R}^{d\times d}$, then*

$$\forall \mathbf{v}, \mathbf{w} \in \mathbb{R}^d, \quad \mathbf{v}^T M \mathbf{v} - \mathbf{w}^T M (\mathbf{v} - \mathbf{w}) = \frac{1}{2}\mathbf{v}^T M \mathbf{v} + \frac{1}{2}\mathbf{w}^T M \mathbf{w} + \frac{1}{2}(\mathbf{w} - \mathbf{v})^T M (\mathbf{w} - \mathbf{v}).$$

Proof. Let $\xi := \mathbf{v}^T M \mathbf{v} - \mathbf{w}^T M (\mathbf{v} - \mathbf{w})$. Obviously, one has

$$\xi = \mathbf{v}^T M \mathbf{v} + \mathbf{w}^T M \mathbf{w} - \mathbf{w}^T M \mathbf{v}.$$

Since M is symmetric, one has $-2\mathbf{w}^T M \mathbf{v} = (\mathbf{v} - \mathbf{w})^T M (\mathbf{v} - \mathbf{w}) - \mathbf{v}^T M \mathbf{v} - \mathbf{w}^T M \mathbf{w}$. Injecting this equality in the expression of ξ ends the demonstration. ■

Corollary 1. *Let M denote a symmetric and positive matrix of $\mathbb{R}^{d\times d}$, then*

$$\forall \mathbf{v}, \mathbf{w} \in \mathbb{R}^d, \quad \mathbf{v}^T M \mathbf{v} - \mathbf{w}^T M (\mathbf{v} - \mathbf{w}) \geq \frac{1}{2}\mathbf{v}^T M \mathbf{v} + \frac{1}{2}\mathbf{w}^T M \mathbf{w}.$$

Proof. This is a direct consequence of Lemma 2, since M is a positive matrix. ■

We can now give the proof of Property 5.

Proof of Property 5. Gibbs formula reads $Td\eta = de + pd\tau$, so that one has

$$T_j^\alpha d_t \eta_j^\alpha = d_t e_j^\alpha + p_j^\alpha d_t \tau_j^\alpha,$$

which rewrites also

$$m_j^\alpha T_j^\alpha d_t \eta_j^\alpha = m_j^\alpha d_t E_j^\alpha - \mathbf{u}_j^\alpha \cdot m_j^\alpha d_t \mathbf{u}_j^\alpha + p_j^\alpha m_j^\alpha d_t \tau_j^\alpha.$$

Using (6.27), one gets

$$\begin{aligned} m_j^\alpha T_j^\alpha d_t \eta_j^\alpha = &- \sum_r \mathbf{C}_{jr} p_j^\alpha \cdot \mathbf{u}_r^\alpha + \sum_r \mathbf{u}_r^{\alpha T} A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) + \nu \sum_r \rho_r^\beta \ \delta\mathbf{u}_r^{\alpha T} B_{jr} \delta\mathbf{u}_r^\alpha \\ &- \nu \sum_r \rho_r^\beta \ \delta\mathbf{u}_j^{\alpha T} B_{jr} (\delta\mathbf{u}_r^\alpha - \delta\mathbf{u}_j^\alpha) + \nu \sum_r \rho_r \ \mathbf{u}_j^{\alpha T} B_{jr} (\delta\mathbf{u}_r^\alpha - \delta\mathbf{u}_j^\alpha) \\ &+ \mathbf{u}_j^\alpha \cdot \left(\sum_r A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) + \nu \sum_r \rho_r B_{jr} \left(\delta\mathbf{u}_r^\alpha - \delta\mathbf{u}_j^\alpha \right) \right) + \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r^\alpha p_j^\alpha, \end{aligned}$$

which simplifies as

$$m_j^\alpha T_j^\alpha d_t \eta_j^\alpha = \sum_r (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha)^T A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) + \nu \sum_r \rho_r^\beta \delta\mathbf{u}_r^{\alpha T} B_{jr} \delta\mathbf{u}_r^\alpha - \nu \sum_r \rho_r^\beta \delta\mathbf{u}_j^{\alpha T} B_{jr} (\delta\mathbf{u}_r^\alpha - \delta\mathbf{u}_j^\alpha).$$

Since B_{jr} matrices are symmetric and positive, one can apply Corollary 1 to obtain

$$m_j^\alpha T_j^\alpha d_t \eta_j^\alpha \geq \sum_r (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha)^T A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) + \frac{1}{2} \nu \sum_r \rho_r^\beta \delta\mathbf{u}_r^{\alpha T} B_{jr} \delta\mathbf{u}_r^\alpha + \frac{1}{2} \nu \sum_r \rho_r^\beta \delta\mathbf{u}_j^{\alpha T} B_{jr} \delta\mathbf{u}_j^\alpha.$$

Matrix A_{jr}^α being positive, one finally has

$$m_j^\alpha T_j^\alpha d_t \eta_j^\alpha \geq \frac{1}{2} \nu \sum_r \rho_r^\beta \delta\mathbf{u}_r^{\alpha T} B_{jr} \delta\mathbf{u}_r^\alpha + \frac{1}{2} \nu \sum_r \rho_r^\beta \delta\mathbf{u}_j^{\alpha T} B_{jr} \delta\mathbf{u}_j^\alpha,$$

which is positive. ■

6-2.4 Formal derivation of the asymptotic scheme

Formal derivation. Let $\alpha \in \{f_1, f_2\}$, β denoting the other fluid. Let us introduce $\epsilon := \nu^{-1}$. One rewrites (6.27) as

$$m_j^\alpha d_t \tau_j^\alpha = \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r^\alpha, \tag{6.B.3}$$

$$d_t m_j^\alpha = 0,$$

$$m_j^\alpha d_t \mathbf{u}_j^\alpha = \sum_r A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) - \frac{1}{\epsilon} \sum_r \rho_r B_{jr} (\delta\mathbf{u}_j^\alpha - \delta\mathbf{u}_r^\alpha), \tag{6.B.4}$$

$$\begin{aligned} m_j^\alpha d_t E_j^\alpha = &- \sum_r \mathbf{C}_{jr} p_j^\alpha \cdot \mathbf{u}_r^\alpha + \sum_r \mathbf{u}_r^{\alpha T} \ A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j^\alpha) + \frac{1}{\epsilon} \sum_r \rho_r^\beta (\delta\mathbf{u}_r^\alpha)^T B_{jr} \delta\mathbf{u}_r^\alpha \\ &+ \frac{1}{\epsilon} \sum_r (\rho_r \mathbf{u}_j^{\alpha T} - \rho_r^\beta \delta\mathbf{u}_j^{\alpha T}) B_{jr} (\delta\mathbf{u}_r^\alpha - \delta\mathbf{u}_j^\alpha), \end{aligned} \tag{6.B.5}$$

and

$$\sum_j A_{jr}^\alpha \mathbf{u}_r^\alpha + \sum_j \frac{1}{\epsilon} \rho_r B_{jr} \delta\mathbf{u}_r^\alpha = \sum_j A_{jr}^\alpha \mathbf{u}_j^\alpha + \sum_j \mathbf{C}_{jr} p_j^\alpha. \tag{6.B.6}$$

Following the analysis of the asymptotic model, we perform an Hilbert expansion.

The first information one gets is from equation (6.B.6) which reads

$$\sum_j A_{jr}^{\alpha,0}\mathbf{u}_r^{\alpha,0} + \sum_j \frac{1}{\epsilon}\rho_r^0 B_{jr}\delta\mathbf{u}_r^{\alpha,0} + \sum_j \rho_r^0 B_{jr}\delta\mathbf{u}_r^{\alpha,1} + \sum_j \rho_r^1 B_{jr}\delta\mathbf{u}_r^{\alpha,0} = \sum_j A_{jr}^{\alpha,0}\mathbf{u}_j^{\alpha,0} + \sum_j \mathbf{C}_{jr}p_j^{\alpha,0} + \mathbf{O}(\epsilon).$$

So that multiplying this equation by ϵ leads to $\rho_r^0(\sum_j B_{jr})\delta\mathbf{u}_r^{\alpha,0} = \mathbf{0}$. That is

$$\delta\mathbf{u}_r^{\alpha,0} = \mathbf{0}, \tag{6.B.7}$$

since $\sum_j B_{jr}$ is symmetric positive definite and $\rho_r = \rho_r^0 + O(\epsilon) > 0$ so that $\rho_r^0 > 0$ when $\epsilon \to 0$. One gets volume conservation equation (6.32).

Now, the momentum equation (6.B.4) is considered, using (6.B.7), one has

$$m_j^\alpha d_t\mathbf{u}_j^{\alpha,0} = \sum_r A_{jr}^{\alpha,0}(\mathbf{u}_r^{\alpha,0} - \mathbf{u}_j^{\alpha,0}) - \frac{1}{\epsilon}\sum_r \rho_r^0 B_{jr}\delta\mathbf{u}_j^{\alpha,0} - \sum_r \rho_r^1 B_{jr}\delta\mathbf{u}_j^{\alpha,0} - \sum_r \rho_r^0 B_{jr}(\delta\mathbf{u}_j^{\alpha,1} - \delta\mathbf{u}_r^{\alpha,1}) + \mathbf{O}(\epsilon),$$

which gives

$$\delta\mathbf{u}_j^{\alpha,0} = \mathbf{0}. \tag{6.B.8}$$

Using, (6.B.8) and (6.B.7), one defines $\mathbf{u}_j^0 := \mathbf{u}_j^{\alpha,0} = \mathbf{u}_j^{\beta,0}$ and $\mathbf{u}_r^0 := \mathbf{u}_r^{\alpha,0} = \mathbf{u}_r^{\beta,0}$.

So, Hilbert expansions of equations (6.B.3), (6.B.4) and (6.B.5) simplify as

$$\begin{aligned} m_j^\alpha d_t\tau_j^{\alpha,0} &= \sum_r \mathbf{C}_{jr}\cdot\mathbf{u}_r^0, \\ m_j^\alpha d_t\mathbf{u}_j^0 &= \sum_r A_{jr}^{\alpha,0}(\mathbf{u}_r^0 - \mathbf{u}_j^0) - \sum_r \rho_r^0 B_{jr}(\delta\mathbf{u}_j^{\alpha,1} - \delta\mathbf{u}_r^{\alpha,1}), \qquad (6.B.9) \\ m_j^\alpha d_t E_j^{\alpha,0} &= -\sum_r (\mathbf{C}_{jr}p_j^{\alpha,0} - A_{jr}^{\alpha,0}(\mathbf{u}_r^0 - \mathbf{u}_j^0))\cdot\mathbf{u}_r^0 \\ &\quad + \sum_r \rho_r^0 \mathbf{u}_j^{\alpha,0^T} B_{jr}(\delta\mathbf{u}_r^{\alpha,1} - \delta\mathbf{u}_j^{\alpha,1}). \qquad (6.B.10) \end{aligned}$$

Our aim is now to evaluate the term $\sum_r \rho_r^0 \mathbf{u}_j^{\alpha,0^T} B_{jr}(\delta\mathbf{u}_r^{\alpha,1} - \delta\mathbf{u}_j^{\alpha,1})$. To do so, we divide momentum equation (6.B.9) by $\rho_j^\alpha (> 0)$, which gives

$$V_j d_t\mathbf{u}_j^0 = \frac{1}{\rho_j^\alpha}\sum_r A_{jr}^{\alpha,0}(\mathbf{u}_r^0 - \mathbf{u}_j^0) - \sum_r \frac{\rho_r^0}{\rho_j^\alpha} B_{jr}(\delta\mathbf{u}_j^{\alpha,1} - \delta\mathbf{u}_r^{\alpha,1}).$$

The same relation can be written for fluid β. The difference of these two equations reads, recalling that $\delta\phi^\alpha = -\delta\phi^\beta$,

$$\mathbf{0} = \sum_r \delta\left(\frac{A_{jr}^0}{\rho_j}\right)^\alpha (\mathbf{u}_r^0 - \mathbf{u}_j^0) - \frac{\rho_j}{\rho_j^\alpha \rho_j^\beta}\sum_r \rho_r^0 B_{jr}(\delta\mathbf{u}_j^{\alpha,1} - \delta\mathbf{u}_r^{\alpha,1}). \tag{6.B.11}$$

Injecting this relation in (6.B.10) gives the limit scheme total energy balance equation (6.33). The momentum equation (6.31) is obtained in the same way or by simply summing equations (6.B.9) for both fluids α and β. ■

6-2.5 Proof or Property 7 (Limit scheme consitency)

Proof. Consistency for volume, mass and momentum is a direct consequence of Property 6, it remains to show the consistency for total energy.

We rewrite equation (6.5) using a more convenient form

$$\rho^\alpha D_t E^\alpha = -\nabla\cdot(p^\alpha + p^\beta)\mathbf{u} + p^\beta \nabla\cdot\mathbf{u} + \frac{\rho^\beta}{\rho}\nabla(p^\alpha+p^\beta)\cdot\mathbf{u}.$$

As a starting point we recall (6.34) for fluid α

$$m_j^\alpha d_t E_j^\alpha = -\sum_r \mathbf{C}_j p_j^\alpha\cdot\mathbf{u}_r + \sum_r \mathbf{u}_r^T A_{jr}^\alpha(\mathbf{u}_r-\mathbf{u}_j) - \frac{\rho_j^\alpha\rho_j^\beta}{\rho_j}\sum_r \mathbf{u}_j^T\delta\left(\frac{A_{jr}}{\rho_j}\right)^\alpha(\mathbf{u}_r-\mathbf{u}_j),$$

that we rewrite

$$\begin{aligned} m_j^\alpha d_t E_j^\alpha = &-\sum_r \mathbf{C}_j(p_j^\alpha+p_j^\beta)\cdot\mathbf{u}_r + \sum_r \mathbf{u}_r^T(A_{jr}^\alpha+A_{jr}^\beta)(\mathbf{u}_r-\mathbf{u}_j)\\ &+\sum_r \mathbf{C}_j p_j^\beta\cdot\mathbf{u}_r - \sum_r \mathbf{u}_r^T A_{jr}^\beta(\mathbf{u}_r-\mathbf{u}_j) - \frac{\rho_j^\alpha\rho_j^\beta}{\rho_j}\sum_r \mathbf{u}_j^T\left(\frac{A_{jr}^\alpha}{\rho_j^\alpha}-\frac{A_{jr}^\beta}{\rho_j^\beta}\right)(\mathbf{u}_r-\mathbf{u}_j). \end{aligned}$$

Simple algebraic manipulations on the later term allow to write

$$\begin{aligned} m_j^\alpha d_t E_j^\alpha = &-\sum_r \mathbf{C}_j(p_j^\alpha+p_j^\beta)\cdot\mathbf{u}_r + \sum_r \mathbf{u}_r^T(A_{jr}^\alpha+A_{jr}^\beta)(\mathbf{u}_r-\mathbf{u}_j)\\ &+\sum_r \mathbf{C}_j p_j^\beta\cdot\mathbf{u}_r - \sum_r (\mathbf{u}_r-\mathbf{u}_j)^T A_{jr}^\beta(\mathbf{u}_r-\mathbf{u}_j) - \frac{\rho_j^\beta}{\rho_j}\mathbf{u}_j^T\sum_r\left(A_{jr}^\alpha+A_{jr}^\beta\right)(\mathbf{u}_r-\mathbf{u}_j). \end{aligned}$$

— According to Property 6 the term

$$\frac{1}{V_j}\left(-\sum_r \mathbf{C}_j(p_j^\alpha+p_j^\beta)\cdot\mathbf{u}_r + \sum_r \mathbf{u}_r^T(A_{jr}^\alpha+A_{jr}^\beta)(\mathbf{u}_r-\mathbf{u}_j)\right),$$

is weakly consistent with $\left(-\nabla\cdot(p^\alpha+p^\beta)\mathbf{u}\right)\big|_{\mathbf{x}_j}$.

— Also since $\frac{1}{V_j}(\sum_r \mathbf{C}_j\cdot\mathbf{u}_r)$ is weakly consistent with $\nabla\cdot\mathbf{u}$,

$$\frac{1}{V_j}\left(p_j^\beta\sum_r \mathbf{C}_j\cdot\mathbf{u}_r\right) \approx \left(p^\beta\nabla\cdot\mathbf{u}\right)\Big|_{\mathbf{x}_j}.$$

— Now, since $\sum_r \mathbf{C}_{jr} = \mathbf{0}$, one has

$$-\sum_r \mathbf{F}_{jr} = \sum_r (A_{jr}^\alpha + A_{jr}^\beta)(\mathbf{u}_r-\mathbf{u}_j),$$

so that Property 6 implies that

$$\frac{1}{V_j}\left(-\frac{\rho_j^\beta}{\rho_j}\mathbf{u}_j^T\sum_r\left(A_{jr}^\alpha+A_{jr}^\beta\right)(\mathbf{u}_r-\mathbf{u}_j)\right) \approx \left(\frac{\rho^\beta}{\rho}\nabla(p^\alpha+p^\beta)\cdot\mathbf{u}\right)\Big|_{\mathbf{x}_j}.$$

To conclude, it remains to prove for the remaining term

$$\frac{1}{V_j}\left(-\sum_r (\mathbf{u}_r-\mathbf{u}_j)^T A_{jr}^{\beta}(\mathbf{u}_r-\mathbf{u}_j)\right)\approx 0.$$

Let ζ^α denote its limit:

$$\frac{1}{V_j}\left(-\sum_r (\mathbf{u}_r-\mathbf{u}_j)^T A_{jr}^{\beta}(\mathbf{u}_r-\mathbf{u}_j)\right)\underset{V_j\to 0}{\longrightarrow}\zeta^\alpha.$$

We have shown

$$\rho_j^\alpha d_t E_j^\alpha \approx \left(-\nabla\cdot(p^\alpha+p^\beta)\mathbf{u}+p^\beta\nabla\cdot\mathbf{u}+\frac{\rho^\beta}{\rho}\nabla(p^\alpha+p^\beta)\cdot\mathbf{u}\right)\Big|_{\mathbf{x}_j}+\zeta^\alpha.$$

Since the same result holds for fluid β, simple calculations lead to

$$\rho_j^\alpha d_t E_j^\alpha+\rho_j^\beta d_t E_j^\beta=\rho_j d_t E_j\approx\left(-\nabla\cdot(p^\alpha+p^\beta)\mathbf{u}\right)\Big|_{\mathbf{x}_j}+\zeta^\alpha+\zeta^\beta.$$

According to Property 6

$$\rho_j d_t E_j\approx\left(-\nabla\cdot(p^\alpha+p^\beta)\mathbf{u}\right)\Big|_{\mathbf{x}_j},$$

so that $\zeta^\alpha+\zeta^\beta\approx 0$.

Actually, one has

$$\frac{1}{V_j}\left(\sum_r (\mathbf{u}_r-\mathbf{u}_j)^T A_{jr}^{\beta}(\mathbf{u}_r-\mathbf{u}_j)\right)+\frac{1}{V_j}\left(\sum_r (\mathbf{u}_r-\mathbf{u}_j)^T A_{jr}^{\alpha}(\mathbf{u}_r-\mathbf{u}_j)\right)\to 0,$$

since A_{jr}^α and A_{jr}^β are positive matrices, one has finally

$$\frac{1}{V_j}\left(\sum_r (\mathbf{u}_r-\mathbf{u}_j)^T A_{jr}^{\alpha}(\mathbf{u}_r-\mathbf{u}_j)\right)\to 0\quad\text{and}\quad\frac{1}{V_j}\left(\sum_r (\mathbf{u}_r-\mathbf{u}_j)^T A_{jr}^{\beta}(\mathbf{u}_r-\mathbf{u}_j)\right)\to 0,$$

which ends the proof. ■

6-2.6 Proof of Lemma 1 (Internal energy variation)

Proof. Rewriting $e_j^{\alpha n+1}=-\frac{1}{2}\|\mathbf{u}_j^{\alpha n+1}\|^2+E_j^{\alpha n+1}$ and using (6.46), one gets after a few arrangements

$$\begin{aligned}e_j^{\alpha n+1}=&\frac{1}{2}\left\|\mathbf{u}_j^{\alpha n}\right\|^2-\frac{1}{2}\left\|\mathbf{u}_j^{\alpha n+1}\right\|^2\\&+e_j^{\alpha n}-\frac{\Delta t}{m_j^\alpha}\left(\sum_r p_j^{\alpha n}\mathbf{C}_{jr}^n\cdot\mathbf{u}_r^{\alpha n}+\sum_r \mathbf{u}_r^{\alpha nT}A_{jr}^{\alpha\,n}(\mathbf{u}_j^{\alpha n}-\mathbf{u}_r^{\alpha n})\right)\\&+\nu\frac{\Delta t}{m_j^\alpha}\left\{\sum_r \rho_r^{\beta n}\;\delta\mathbf{u}_r^{\alpha nT}B_{jr}^n\delta\mathbf{u}_r^{\alpha n}-\sum_r \rho_r^{\beta n}\;\delta\mathbf{u}_r^{\alpha n+1^T}B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n}-\delta\mathbf{u}_j^{\alpha n+1}\right)\right.\\&\left.+\sum_r \rho_r^n\;\mathbf{u}_j^{n+1^T}B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n}-\delta\mathbf{u}_j^{\alpha n+1}\right)\right\}.\end{aligned}\tag{6.B.12}$$

As a first step one estimates kinetic energy variation

$$-\Delta\mathcal{K}_j^\alpha := \frac{1}{2}\|\mathbf{u}_j^{\alpha n}\|^2 - \frac{1}{2}\|\mathbf{u}_j^{\alpha n+1}\|^2 = \frac{\mathbf{u}_j^{\alpha n}+\mathbf{u}_j^{\alpha n+1}}{2}\cdot\left(\mathbf{u}_j^{\alpha n}-\mathbf{u}_j^{\alpha n+1}\right),$$

which rewrites using (6.45)

$$\begin{aligned}-\Delta\mathcal{K}_j^\alpha = \left(\mathbf{u}_j^{\alpha n} - \frac{\Delta t}{2m_j^\alpha}\left[\sum_r A_{jr}^{\alpha\,n}(\mathbf{u}_j^{\alpha n}-\mathbf{u}_r^{\alpha n}) + \nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_j^{\alpha n+1}-\delta\mathbf{u}_r^{\alpha n})\right]\right)\\ \cdot\frac{\Delta t}{m_j^\alpha}\left[\sum_r A_{jr}^{\alpha\,n}(\mathbf{u}_j^{\alpha n}-\mathbf{u}_r^{\alpha n}) + \nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_j^{\alpha n+1}-\delta\mathbf{u}_r^{\alpha n})\right],\end{aligned}$$

that is

$$\begin{aligned}-\Delta\mathcal{K}_j^\alpha =& \frac{\Delta t}{m_j^\alpha}\left(\sum_r \mathbf{u}_j^{\alpha nT} A_{jr}^{\alpha\,n}(\mathbf{u}_j^{\alpha n}-\mathbf{u}_r^{\alpha n}) + \nu\sum_r \mathbf{u}_j^{\alpha nT}\rho_r^n B_{jr}^n(\delta\mathbf{u}_j^{\alpha n+1}-\delta\mathbf{u}_r^{\alpha n})\right)\\ &-\frac{\Delta t^2}{2m_j^{\alpha 2}}\left(\sum_r A_{jr}^{\alpha\,n}(\mathbf{u}_j^{\alpha n}-\mathbf{u}_r^{\alpha n}) + \nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_j^{\alpha n+1}-\delta\mathbf{u}_r^{\alpha n})\right)^2.\end{aligned}$$

So, one has

$$\begin{aligned}e_j^{\alpha n+1} = e_j^{\alpha n} + \frac{\Delta t}{m_j^\alpha}\Bigg\{&\sum_r (\mathbf{u}_j^{\alpha n}-\mathbf{u}_r^{\alpha n})^T A_{jr}^{\alpha\,n}(\mathbf{u}_j^{\alpha n}-\mathbf{u}_r^{\alpha n}) - \sum_r p_j^{\alpha n}\mathbf{C}_{jr}^n\cdot\mathbf{u}_r^{\alpha n}\\ &-\frac{\Delta t}{2m_j^\alpha}\left(\sum_r A_{jr}^{\alpha\,n}(\mathbf{u}_j^{\alpha n}-\mathbf{u}_r^{\alpha n}) + \nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_j^{\alpha n+1}-\delta\mathbf{u}_r^{\alpha n})\right)^2\Bigg\}\\ &+\nu\frac{\Delta t}{m_j^\alpha}\left[\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n\delta\mathbf{u}_r^{\alpha n} + \sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_j^{\alpha n+1^T} B_{jr}^n(\delta\mathbf{u}_j^{\alpha n+1}-\delta\mathbf{u}_r^{\alpha n})\right]\\ &+\nu\frac{\Delta t}{m_j^\alpha}\sum_r \rho_r^n\ (\mathbf{u}_j^{\alpha n}-\mathbf{u}_j^{\alpha n+1})^T B_{jr}^n(\delta\mathbf{u}_j^{\alpha n+1}-\delta\mathbf{u}_r^{\alpha n}),\end{aligned}$$

which using (6.43) is nothing but (6.47). ■

6-2.7 Proof of Property 10 (Entropy)

Proof. Let $U = \left(\tau, \mathbf{u}^T, E\right)^T$ and let η be the entropy of the fluid. Gibbs formula reads $Td\eta = de + pd\tau$. Following [Maz07, Des10a], we estimate the entropy change, by means of a third-order Taylor expansion, due to the proposed scheme:

$$\begin{aligned}\eta(U_j^{\alpha n+1}) - \eta(U_j^{\alpha n}) = (U_j^{\alpha n+1}-U_j^{\alpha n})^T\left.\frac{\partial\eta}{\partial U}\right|_{U_j^{\alpha n}}\\ +\frac{1}{2}(U_j^{\alpha n+1}-U_j^{\alpha n})^T\left.\frac{\partial^2\eta}{\partial U^2}\right|_{U_j^{\alpha n}}(U_j^{\alpha n+1}-U_j^{\alpha n}) + \mathcal{O}\left((U_j^{\alpha n+1}-U_j^{\alpha n})^3\right).\end{aligned}$$

One has $\left.\frac{\partial\eta}{\partial U}\right|_{U_j^{\alpha n}} = \frac{1}{T_j^{\alpha n}}(p_j^{\alpha n}, -\mathbf{u}_j^{\alpha n}, 1)^T$ and the variable change $V = (p, -\mathbf{u}, \eta)^T$ reads

$$\begin{aligned}(U_j^{\alpha n+1}-U_j^{\alpha n})^T\left.\frac{\partial^2\eta}{\partial U^2}\right|_{U_j^{\alpha n}}(U_j^{\alpha n+1}-U_j^{\alpha n}) = (V_j^{\alpha n+1}-V_j^{\alpha n})^T\left.\frac{\partial^2\eta}{\partial V^2}\right|_{V_j^{\alpha n}}(V_j^{\alpha n+1}-V_j^{\alpha n})\\ +\mathcal{O}\left((U_j^{\alpha n+1}-U_j^{\alpha n})^3\right),\end{aligned}$$

where, see [Des01, Maz07] for instance,

$$\left.\frac{\partial^2\eta}{\partial V^2}\right|_{V_j^{\alpha n}} = -\frac{1}{T_j^{\alpha n}}\begin{pmatrix} \left((\rho c)_j^{\alpha n}\right)^{-2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Let $O_1 := (U_j^{\alpha n+1} - U_j^{\alpha n})^T \left.\frac{\partial \eta}{\partial U}\right|_{U_j^{\alpha n}}$, using (6.39), (6.45) and (6.46), one gets

$$\begin{aligned} O_1 = \frac{1}{T_j^{\alpha n}}\frac{\Delta t}{m_j^\alpha}\Bigg\{ & p_j^{\alpha n}\sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} \\ & - \mathbf{u}_j^{\alpha nT}\left(\sum_r A_{jr}^{\alpha,n}(\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) + \nu\sum_r \rho_r^n B_{jr}^n(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1})\right) \\ & - \sum_r \mathbf{C}_{jr}^n p_j^{\alpha n}\cdot\mathbf{u}_r^{\alpha n} + \sum_r \mathbf{u}_r^{\alpha nT} A_{jr}^{\alpha,n}(\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) \\ & + \nu\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} - \nu\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha n+1^T} B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right) \\ & + \nu\sum_r \rho_r^n\ \mathbf{u}_j^{\alpha n+1^T} B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right)\Bigg\}, \end{aligned}$$

which simplifies as

$$\begin{aligned} O_1 = \frac{1}{T_j^{\alpha n}}\frac{\Delta t}{m_j^\alpha}\Bigg\{ & \sum_r (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n})^T A_{jr}^{\alpha,n}(\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) \\ & + \nu\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} - \nu\sum_r \rho_r^{\beta n}\ \delta\mathbf{u}_r^{\alpha n+1^T} B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right)\Bigg\} \\ & + \frac{1}{T_j^{\alpha n}}\frac{\Delta t}{m_j^\alpha}\left\{\nu\sum_r \rho_r^n\ (\mathbf{u}_j^{\alpha n+1} - \mathbf{u}_j^{\alpha n})^T B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right)\right\}. \end{aligned}$$

Now using Lemma 2, one gets

$$\begin{aligned} O_1 = & \frac{1}{T_j^{\alpha n}}\frac{\Delta t}{m_j^\alpha}\left\{\frac{1}{2}\nu\sum_r \rho_r^\beta\ \delta\mathbf{u}_r^{\alpha nT} B_{jr}^n \delta\mathbf{u}_r^{\alpha n} + \frac{1}{2}\nu\sum_r \rho_r^\beta\ \delta\mathbf{u}_j^{\alpha n+1^T} B_{jr}^n \delta\mathbf{u}_j^{\alpha n+1}\right\} \\ & + \frac{1}{T_j^{\alpha n}}\frac{\Delta t}{m_j^\alpha}\left\{\sum_r (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n})^T A_{jr}^{\alpha,n}(\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) + \nu\sum_r \rho_r^{\beta n}\ \left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right)^T B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right)\right\} \\ & + \frac{1}{T_j^{\alpha n}}\frac{\Delta t}{m_j^\alpha}\left\{\nu\sum_r \rho_r^n\ (\mathbf{u}_j^{\alpha n+1} - \mathbf{u}_j^{\alpha n})^T B_{jr}^n\left(\delta\mathbf{u}_r^{\alpha n} - \delta\mathbf{u}_j^{\alpha n+1}\right)\right\}. \end{aligned}$$

Observe that later term is second-order in time, so one retrieves as expected the entropy production of the continuous in time scheme established in Property 5 page 133.

One now focuses on the second-order term of the entropy variation

$$O_2 := \frac{1}{2}(V_j^{\alpha n+1} - V_j^{\alpha n})^T \left.\frac{\partial^2\eta}{\partial V^2}\right|_{V_j^{\alpha n}} (V_j^{\alpha n+1} - V_j^{\alpha n}),$$

which rewrites

$$O_2 = \frac{1}{2}(\Delta\mathbf{\Psi})^T \begin{pmatrix} \left((\rho c)_j^{\alpha n}\right)^{-2} & 0 \\ 0 & 1 \end{pmatrix}\Delta\mathbf{\Psi}, \quad \text{with} \quad \Delta\mathbf{\Psi} = \begin{pmatrix} p_j^{\alpha n+1} - p_j^{\alpha n} \\ -\mathbf{u}_j^{\alpha n+1} + \mathbf{u}_j^{\alpha n} \end{pmatrix}.$$

One has to estimate $p_j^{\alpha n+1} - p_j^{\alpha n}$. Assuming that the equation of state $p : (\tau, e) \to p(\tau, e)$ is regular enough, one has

$$p_j^{\alpha n+1} - p_j^{\alpha n} = (\tau_j^{\alpha n+1} - \tau_j^{\alpha n}) \left.\frac{\partial p}{\partial \tau}\right|_{jn} + (e_j^{\alpha n+1} - e_j^{\alpha n}) \left.\frac{\partial p}{\partial e}\right|_{jn} + O(\Delta t^2).$$

Using (6.39) and (6.47) and keeping only first-order terms, one has

$$\begin{aligned} p_j^{\alpha n+1} - p_j^{\alpha n} = {} & \frac{\Delta t}{m_j^\alpha} \left(\sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} \right) \left.\frac{\partial p}{\partial \tau}\right|_{jn} \\ & + \frac{\Delta t}{m_j^\alpha} \left\{ \left(\sum_r (\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n})^T A_{jr}^{\alpha\, n} (\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n}) - \sum_r p_j^{\alpha n} \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} \right) \right. \\ & + \nu \left(\sum_r \rho_r^{\beta n}\ \delta \mathbf{u}_r^{\alpha n T} B_{jr}^n \delta \mathbf{u}_r^{\alpha n} + \sum_r \rho_r^{\beta n}\ \delta \mathbf{u}_j^{\alpha n+1 T} B_{jr}^n (\delta \mathbf{u}_j^{\alpha n+1} - \delta \mathbf{u}_r^{\alpha n}) \right) \\ & \left. + \nu \left(\sum_r \rho_r^n\ (\mathbf{u}_j^{\alpha n} - \mathbf{u}_j^{\alpha n+1})^T B_{jr}^n (\delta \mathbf{u}_j^{\alpha n+1} - \delta \mathbf{u}_r^{\alpha n}) \right) \right\} \left.\frac{\partial p}{\partial e}\right|_{jn} + O(\Delta t^2). \end{aligned}$$

Then, using (6.45), one gets

$$\begin{aligned} O_2 = -\frac{1}{T_j^{\alpha n}} \frac{\Delta t^2}{2 m_j^{\alpha 2}} \Bigg[& \left\{ \left(\sum_r \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} \right) \left.\frac{\partial p}{\partial \tau}\right|_{jn} \right. \\ & + \left(\sum_r (\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n})^T A_{jr}^{\alpha\, n} (\mathbf{u}_j^{\alpha n} - \mathbf{u}_r^{\alpha n}) - \sum_r p_j^{\alpha n} \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} \right. \\ & + \nu \sum_r \rho_r^{\beta n}\ \delta \mathbf{u}_r^{\alpha n T} B_{jr}^n \delta \mathbf{u}_r^{\alpha n} + \nu \sum_r \rho_r^{\beta n}\ \delta \mathbf{u}_j^{\alpha n+1 T} B_{jr}^n (\delta \mathbf{u}_j^{\alpha n+1} - \delta \mathbf{u}_r^{\alpha n}) \\ & \left. + \nu \sum_r \rho_r^n\ (\mathbf{u}_j^{\alpha n} - \mathbf{u}_j^{\alpha n+1})^T B_{jr}^n (\delta \mathbf{u}_j^{\alpha n+1} - \delta \mathbf{u}_r^{\alpha n}) \right) \left.\frac{\partial p}{\partial e}\right|_{jn} \Bigg\}^2 \left((\rho c)_j^{\alpha n} \right)^{-2} \\ & \left. + \left\{ \sum_r A_{jr}^{\alpha, n} (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) + \nu \sum_r \rho_r^n B_{jr}^n (\delta \mathbf{u}_r^{\alpha n} - \delta \mathbf{u}_j^{\alpha n+1}) \right\}^2 \right] + O(\Delta t^3). \end{aligned}$$

Finally, putting all the pieces together, one has

$$\begin{aligned} & \eta(U_j^{\alpha n+1}) - \eta(U_j^{\alpha n}) \\ & \quad = \frac{1}{T_j^{\alpha n}} \frac{\Delta t}{m_j^\alpha} \left\{ \frac{1}{2} \nu \sum_r \rho_r^\beta\ \delta \mathbf{u}_r^{\alpha n T} B_{jr}^n \delta \mathbf{u}_r^{\alpha n} + \frac{1}{2} \nu \sum_r \rho_r^\beta\ \delta \mathbf{u}_j^{\alpha n+1 T} B_{jr}^n \delta \mathbf{u}_j^{\alpha n+1} \right\} \\ & \qquad + \frac{1}{T_j^{\alpha n}} \frac{\Delta t}{m_j^\alpha} \left(a - \frac{\Delta t}{m_j^\alpha} (b + c) + O(\Delta t^2) \right), \end{aligned}$$

with $a \geq 0$ and $b \geq 0$.

Thus it remains to study the positiveness of $a - \frac{\Delta t}{m_j^\alpha}(b + c) + O(\Delta t^2)$. There are two possibilities.

6-2.7.0.1 Case $a > 0$ In that case, there obviously exists $\Delta t > 0$ such that

$$T_j^{\alpha n} m_j^\alpha \frac{\eta(U_j^{\alpha n+1}) - \eta(U_j^{\alpha n})}{\Delta t} \geq \frac{1}{2} \nu \sum_r \rho_r^\beta\ \delta \mathbf{u}_r^{\alpha n T} B_{jr}^n \delta \mathbf{u}_r^{\alpha n} + \frac{1}{2} \nu \sum_r \rho_r^\beta\ \delta \mathbf{u}_j^{\alpha n+1 T} B_{jr}^n \delta \mathbf{u}_j^{\alpha n+1}.$$

6-2.7.0.2 Case $a = 0$ If $a = 0$, one has

$$\sum_r (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n})^T A_{jr}^{\alpha,n} (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) + \nu \sum_r \rho_r^{\beta n} \left(\delta \mathbf{u}_r^{\alpha n} - \delta \mathbf{u}_j^{\alpha n+1}\right)^T B_{jr}^n \left(\delta \mathbf{u}_r^{\alpha n} - \delta \mathbf{u}_j^{\alpha n+1}\right) = 0.$$

Since $A_{jr}^{\alpha,n}$ and B_{jr}^n are positive matrices, all the terms in the sum are zeros. Let us first focus on $(\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n})^T A_{jr}^{\alpha,n} (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) = 0$ terms. Two cases occur. In case of Eucclhyd scheme, $A_{jr}^{\alpha,n}$ is positive definite so that one has $\mathbf{u}_r^{\alpha n} = \mathbf{u}_j^{\alpha n}$. For Glace scheme

$$(\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n})^T A_{jr}^{\alpha,n} (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) = \frac{(\rho c)_j^{\alpha n}}{\|\mathbf{C}_{jr}^n\|} \left\| \mathbf{C}_{jr}^n \cdot (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) \right\|^2 = 0.$$

So, for both schemes, one has $\mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} = \mathbf{C}_{jr}^n \cdot \mathbf{u}_j^{\alpha n}$ and $A_{jr}^{\alpha,n} (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) = \mathbf{0}$. Recalling that $\sum_r \mathbf{C}_{jr}^n = \mathbf{0}$, one also has $\sum_r p_j^{\alpha n} \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} = 0$.

One now analyzes $\left(\delta \mathbf{u}_r^{\alpha n} - \delta \mathbf{u}_j^{\alpha n+1}\right)^T B_{jr}^n \left(\delta \mathbf{u}_r^{\alpha n} - \delta \mathbf{u}_j^{\alpha n+1}\right) = 0$. Since B_{jr}^n are positive definite, this implies $\delta \mathbf{u}_r^{\alpha n} - \delta \mathbf{u}_j^{\alpha n+1} = \mathbf{0}$.

Finally, if $a = 0$, one has

$$\begin{aligned} T_j^{\alpha n} m_j^\alpha \frac{\eta(U_j^{\alpha n+1}) - \eta(U_j^{\alpha n})}{\Delta t} &= \nu \sum_r \rho_r^\beta \ \delta \mathbf{u}_r^{\alpha n T} B_{jr}^n \delta \mathbf{u}_r^{\alpha n} \\ &\quad - \frac{\Delta t}{2 m_j^\alpha} \left(\nu \sum_r \rho_r^{\beta n} \ \delta \mathbf{u}_r^{\alpha n T} B_{jr}^n \delta \mathbf{u}_r^{\alpha n} \right)^2 \left. \frac{\partial p}{\partial e} \right|_{jn}^2 \left((\rho c)_j^{\alpha n} \right)^{-2} + O(\Delta t^2). \end{aligned}$$

Before enunciating the result, one should remark that in the general case, one has

$$\eta(U_j^{\alpha n+1}) - \eta(U_j^{\alpha n}) = \frac{1}{T_j^{\alpha n}} \frac{\Delta t}{m_j^\alpha} \left((a + a_\nu) - \frac{\Delta t}{m_j^\alpha} (b + c) + O(\Delta t^2) \right),$$

with $a \geq 0$, $a_\nu \geq 0$ and $b \geq 0$. Again, one has two alternatives $a + a_\nu > 0$ or $a + a_\nu = 0$. In the first case, there exists Δt such that $\eta(U_j^{\alpha n+1}) - \eta(U_j^{\alpha n}) > 0$. In the second case, one has $a = a_\nu = 0$ so as previously, $a = 0 \implies \mathbf{C}_{jr}^n \cdot \mathbf{u}_r^{\alpha n} = \mathbf{C}_{jr}^n \cdot \mathbf{u}_j^{\alpha n}$, $A_{jr}^{\alpha,n} (\mathbf{u}_r^{\alpha n} - \mathbf{u}_j^{\alpha n}) = \mathbf{0}$ and $\delta \mathbf{u}_r^{\alpha n} - \delta \mathbf{u}_j^{\alpha n+1} = \mathbf{0}$. Also, since $a_\nu = 0$ and since B_{jr}^n is positive definite one has $\delta \mathbf{u}_r^{\alpha n} = \mathbf{0}$. Therefore the scheme (6.39)–(6.41) gives $U_j^{\alpha n+1} = U_j^{\alpha n}$ and finally one has $\forall \Delta t > 0$, $\eta(U_j^{\alpha n+1}) = \eta(U_j^{\alpha n})$. ■

Appendix 6.C Asymptotic behaviour of $\delta \mathbf{u}_r^{\alpha,\nu}$

In this section we discuss the asymptotic behavior suggested by inequality (6.29) established in Property 3:

$$\left(\mathbf{u}_r^{\alpha,\nu} - \mathbf{u}_r^{\beta,\nu}\right)^T B_r \left(\mathbf{u}_r^{\alpha,\nu} - \mathbf{u}_r^{\beta,\nu}\right) \leq \frac{1}{2\nu\rho_r} \left(\mathbf{u}_r^{\alpha,0 T} A_r^\alpha \mathbf{u}_r^{\alpha,0} + \mathbf{u}_r^{\beta,0 T} A_r^\beta \mathbf{u}_r^{\beta,0} \right).$$

As a first remark, it is obvious that this inequality is not optimal: for instance, since the schemes are Galilean invariant, the right hand side could be replaced by

$$\inf_{\mathbf{v} \in \mathbb{R}^d} \frac{1}{2\nu\rho_r} \left((\mathbf{u}_r^{\alpha,0} - \mathbf{v})^T A_r^\alpha (\mathbf{u}_r^{\alpha,0} - \mathbf{v}) + (\mathbf{u}_r^{\beta,0} - \mathbf{v})^T A_r^\beta (\mathbf{u}_r^{\beta,0} - \mathbf{v}) \right),$$

but this is not a major issue since we are discussing the asymptotic behavior of $\delta \mathbf{u}_r^{\alpha,\nu}$.

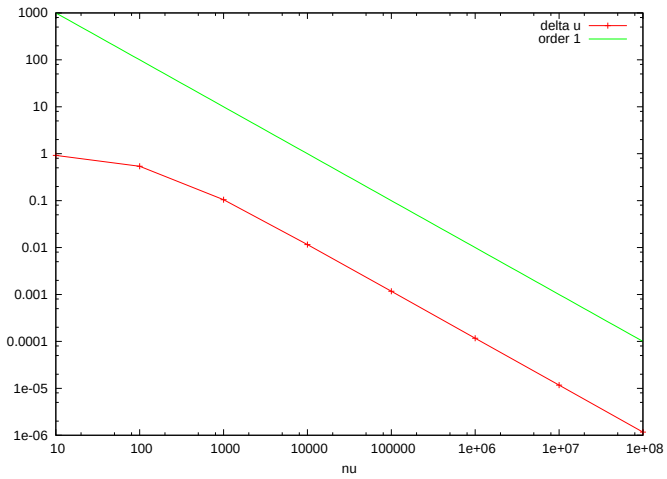

Figure 6.12 – $\|\delta \mathbf{u}_r^{\alpha,\nu}\|$ according to ν. One observes a $\mathcal{O}(\nu^{-1})$ behavior.

We test numerically the behavior of $\|\delta \mathbf{u}_r^{\alpha,\nu}\|_{B_r}$ using the following test. The domain is $\Omega =]0,1[\times]0,0.1[$. For both fluids, we set the initial condition $\rho = 1$ and $p = 1$ and $\gamma = 1.4$. Initial velocities are $\left(\pm\frac{1}{2}\sin(\pi x), 0\right)^T$ so that $\delta \mathbf{u}\left(\frac{1}{2}, y\right) = 1$. We use a 200×3 mesh and compute $\delta \mathbf{u}_r$ value at some node such that $x = \frac{1}{2}$ for values of ν in $\{10, 10^2, ..., 10^8\}$.

Figure 6.12 shows a first-order convergence to 0, while (6.29) implies $\mathcal{O}(\nu^{-1/2})$. From this observation, we made other tests and we always observed $\mathcal{O}(\nu^{-1})$. If this does not prove that (6.29) is not optimal, it might indicate that it could be improved. However, improving (6.29) does not seem to be easy. Moreover, the aim of Property 3 is to provide *a priori* estimates that proves some stability of $\mathbf{u}_r^{\alpha,\nu}$ according to ν. In that view, (6.29) is satisfactory.

Appendix 6.D Asymptotic behaviour of the reference scheme

We shall discuss here properties of the scheme (6.22)–(6.23). Actually, conservation, consistency and entropy stability are obtained exactly as for the AP scheme (6.24)–(6.26).

Following the formal derivation of Paragraph 6-4.2.4, we perform an Hilbert expansion of the scheme variables with regard to $\epsilon = \nu^{-1}$.

One rewrites (6.22)–(6.23) as

$$m_j^\alpha d_t \tau_j^{\alpha,0} = \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r^{\alpha,0} + \mathcal{O}(\epsilon), \tag{6.D.1}$$

$$m_j^\alpha d_t \mathbf{u}_j^{\alpha,0} = -\sum_r \mathbf{F}_{jr}^{\alpha,0} - \frac{1}{\epsilon}\sum_r \rho_r^0 B_{jr} \delta \mathbf{u}_j^{\alpha,0} - \sum_r \rho_r^0 B_{jr} \delta \mathbf{u}_j^{\alpha,1} + \mathcal{O}(\epsilon), \tag{6.D.2}$$

$$m_j^\alpha d_t E_j^{\alpha,0} = -\sum_r \mathbf{F}_{jr}^{\alpha,0} \cdot \mathbf{u}_r^{\alpha,0} + \frac{1}{\epsilon}\sum_r \rho_r^0 \overline{\mathbf{u}}_{jr}^{0\,T} B_{jr} \delta \mathbf{u}_j^{\alpha,0} + \sum_r \rho_r^0 \overline{\mathbf{u}}_{jr}^{0\,T} B_{jr} \delta \mathbf{u}^{\alpha,1} + \mathcal{O}(\epsilon), \tag{6.D.3}$$

with

$$\mathbf{F}_{jr}^{\alpha,0} = \mathbf{C}_{jr} p_j^{\alpha,0} - A_{jr}^{\alpha,0}(\mathbf{u}_r^{\alpha,0} - \mathbf{u}_j^{\alpha,0}) + \mathcal{O}(\epsilon), \quad \text{and} \tag{6.D.4}$$

$$\sum_j \mathbf{F}_{jr}^{\alpha,0} = \mathcal{O}(\epsilon). \tag{6.D.5}$$

It comes from 6.D.2 that the limit $\epsilon \to 0$ gives $\delta \mathbf{u}_j^{\alpha,0} = \mathbf{0}$ so that one sets $\mathbf{u}_j^0 := \mathbf{u}_j^{\alpha,0} = \mathbf{u}_j^{\beta,0} = \overline{\mathbf{u}}_{jr}^0$. One gets from (6.D.2) that for each fluid α,

$$d_t \mathbf{u}_j^0 = -\frac{1}{m_j^\alpha} \sum_r \mathbf{F}_{jr}^{\alpha,0} - \frac{1}{m_j^\alpha} \sum_r \rho_r^0 B_{jr} \delta \mathbf{u}_j^{\alpha,1}.$$

The difference of both equalities yields

$$\underbrace{\delta \left(d_t \mathbf{u}_j^0 \right)^\alpha}_{\mathbf{0}} = -\frac{1}{m_j^\alpha} \sum_r \mathbf{F}_{jr}^{\alpha,0} + \frac{1}{m_j^\beta} \sum_r \mathbf{F}_{jr}^{\beta,0} - \frac{1}{m_j^\alpha} \sum_r \rho_r^0 B_{jr} \delta \mathbf{u}_j^{\alpha,1} + \frac{1}{m_j^\beta} \sum_r \rho_r^0 B_{jr} \delta \mathbf{u}_j^{\beta,1},$$

which rewrites, since $\delta \phi^\beta = -\delta \phi^\alpha$

$$\left(\frac{1}{m_j^\alpha} + \frac{1}{m_j^\beta} \right) \sum_r \rho_r^0 B_{jr} \delta \mathbf{u}_j^{\alpha,1} = -\frac{1}{m_j^\alpha} \sum_r \mathbf{F}_{jr}^{\alpha,0} + \frac{1}{m_j^\beta} \sum_r \mathbf{F}_{jr}^{\beta,0}.$$

Thus recalling that $\rho_j = \rho_j^\alpha + \rho_j^\beta$,

$$\sum_r \rho_r^0 B_{jr} \delta \mathbf{u}_j^{\alpha,1} = -\frac{\rho_j^\beta}{\rho_j} \sum_r \mathbf{F}_{jr}^{\alpha,0} + \frac{\rho_j^\alpha}{\rho_j} \sum_r \mathbf{F}_{jr}^{\beta,0}. \tag{6.D.6}$$

So, summing (6.D.2) for both fluids, one gets

$$\left(m_j^\alpha + m_j^\beta \right) d_t \mathbf{u}_j^0 = -\sum_r \left(\mathbf{F}_{jr}^{\alpha,0} - \mathbf{F}_{jr}^{\beta,0} \right).$$

Injecting (6.D.6) into (6.D.3) and considering the limit $\epsilon \to 0$, yields

$$m_j^\alpha d_t E_j^{\alpha,0} = -\sum_r \mathbf{F}_{jr}^{\alpha,0} \cdot \mathbf{u}_r^{\alpha,0} - \frac{\rho_j^\beta}{\rho_j} \mathbf{u}_j^0 \cdot \sum_r \mathbf{F}_{jr}^{\alpha,0} + \frac{\rho_j^\alpha}{\rho_j} \mathbf{u}_j^0 \cdot \sum_r \mathbf{F}_{jr}^{\beta,0}.$$

Then one can write the limit scheme

$$m_j^\alpha d_t \tau_j^\alpha = \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r^\alpha, \tag{6.D.7}$$

$$\left(m_j^\alpha + m_j^\beta \right) d_t \mathbf{u}_j = -\sum_r \left(\mathbf{F}_{jr}^\alpha - \mathbf{F}_{jr}^\beta \right), \tag{6.D.8}$$

$$m_j^\alpha d_t E_j^\alpha = -\sum_r \mathbf{F}_{jr}^\alpha \cdot \mathbf{u}_r^\alpha - \frac{\rho_j^\beta}{\rho_j} \mathbf{u}_j \cdot \sum_r \mathbf{F}_{jr}^\alpha + \frac{\rho_j^\alpha}{\rho_j} \mathbf{u}_j \cdot \sum_r \mathbf{F}_{jr}^\beta, \tag{6.D.9}$$

with

$$\mathbf{F}_{jr}^\alpha = \mathbf{C}_{jr} p_j^\alpha - A_{jr}^\alpha (\mathbf{u}_r^\alpha - \mathbf{u}_j), \quad \text{and} \tag{6.D.10}$$

$$\sum_j \mathbf{F}_{jr}^\alpha = \mathbf{0}. \tag{6.D.11}$$

On one hand, for each fluid α, one recognizes the classical fluxes definition (6.19)–(6.20) or the frictionless case (monofluid). So, according to B. Després [Des10b], for each fluid one has the following consistency results

$$\frac{1}{V_j} \sum_r \mathbf{C}_{jr} \cdot \mathbf{u}_r^\alpha \approx \nabla \cdot \mathbf{u}^\alpha,$$

$$\frac{1}{V_j} \sum_r \mathbf{F}_{jr}^\alpha \approx \nabla p^\alpha.$$

On the other hand, solving (6.D.10)–(6.D.11), one obtains the nodal velocity for each fluid

$$\mathbf{u}_r^\alpha = A_r^{\alpha\,-1}\left(\sum_j \mathbf{C}_{jr} p_j^\alpha + \sum_j A_{jr}^\alpha \mathbf{u}_j\right), \quad \text{with } A_r^\alpha := \sum_j A_{jr}^\alpha,$$

which allows to compute the difference of the nodal velocities of both fluids

$$\delta\mathbf{u}_r^\alpha = A_r^{\alpha\,-1}\sum_j \mathbf{C}_{jr} p_j^\alpha - A_r^{\beta\,-1}\sum_j \mathbf{C}_{jr} p_j^\beta + A_r^{\alpha\,-1}\sum_j A_{jr}^\alpha \mathbf{u}_j - A_r^{\beta\,-1}\sum_j A_{jr}^\beta \mathbf{u}_j.$$

Unlike the AP scheme case, one has this time $\delta\mathbf{u}_r^\alpha \neq \mathbf{0}$ for a given grid, which is not uniformly consistent with the equation (6.16). This result implies that the scheme does not respect the diffusion regime corresponding to (6.15). It can be easily checked in writing the semi-discrete evolution of the specific volume difference between the fluids

$$m_j^\alpha d_t \tau_j^{\alpha,0} - m_j^\alpha d_t \tau_j^{\beta,0} = \sum_r \mathbf{C}_{jr} \cdot \delta\mathbf{u}_r^{\alpha,0} + O(\epsilon). \tag{6.D.12}$$

The right hand side of this equation should be $O(\epsilon)$ as for the AP scheme, but is $O(1)$ in this case. It explains the over-diffusive behaviour of this scheme when $\nu >> 1$.

Conclusions and perspectives

Conclusions

In this document, the Trefftz discontinuous Galerkin (TDG) method applied to transport models has been studied and analyzed. In particular, a special attention has been devoted to the P_N reduced model of the transport equation. The transport equation is challenging to solve numerically because it may involve, among other, a diffusion limit and boundary layers. The goal of this work was to obtained asymptotic preserving and well-balanced schemes to capture both of these phenomena with reasonable computational time. This document has shown that the TDG method naturally leads to well-balanced and asymptotic preserving schemes.

In particular, the well-balanced property of the scheme has been established in Chapter 2 together with the TDG formulation for general Friedrichs systems.

Additionally, an asymptotic study of the method has been performed in the Chapter 3 for the P_1 model in 1D. Taking advantage of the one dimensional framework, it has been shown that the TDG method recovers the diffusion limit at least for a particular choice of basis functions. Then, the convergence of the scheme and the asymptotic preserving property have been numerically confirmed.

In Chapter 4, the TDG method has been studied and analyzed in the general case of the two dimensional P_N model. After recalling the derivation of the P_N model, some of its properties were given. Concerning the TDG method, two important results were provided in this chapter

(i) *Construction of the basis functions.* Stationary and time dependent basis functions have been constructed. In particular, polynomial and exponential stationary solutions have been derived. Due to the well-balanced property, the exponential solutions lead to very efficient schemes to capture boundary layers as illustrated later in Chapter 5.

(ii) *High order convergence of the scheme.* High order convergence has been proven in the stationary case, mainly through the study of the approximation properties of the basis functions. Even if this approximation result may not be optimal for the case $N > 1$, a well known advantage of the TDG method has been recovered: to obtain high order convergence, the TDG method uses (at least asymptotically) less basis functions than the standard DG method. The example given in Table 4.1 is a good illustration of this property for the case $N = 1$.

Finally, numerical results for the two dimensional P_1 and P_3 models were provided in Chapter 5. The asymptotic preserving property of the TDG method has been illustrated both for the P_1 and P_3 models. Moreover, it has been shown that the TDG method outperforms the standard DG method for some numerical tests with boundary layers, using less degrees of freedom for a better accuracy. The main drawback of the TDG method is that it may lead to ill-conditioning systems when considering too many basis functions per cell or in some asymptotic regimes. Such behaviors have also been numerically illustrated in Chapter 5.

Perspectives

A first perspective is to develop good preconditioners to deal with the ill-conditioning systems of the TDG method. This could be particularly useful when considering, for example, stationary and time dependent basis functions.

It could also be interesting to extend the TDG method to the discrete ordinate method (S_N model) which is the other popular approximation of the transport equation. Since the S_N model can be written under the form of a Friedrichs system, the general formulation given in Chapter 2 can be used. It remains to construct the basis functions.

Another possibility would be to applied the TDG method directly to the transport equation using, for example, Case's and Birkhoff's solutions [BA69, BA70, Cas60].

The formulation of the TDG method given in Chapter 2 can be easily generalized to the three dimensional case. The only additional difficulty concerns the construction of the basis functions. For the P_N model, the basis functions can be constructed as in Chapter 4, starting from a one dimensional solution and then applying a rotation. Note however that the three dimensional rotation is not as simple as in the two dimensional case [BFB97, IR96, PH07]. Note also that the choice of directions may be tricky since it is not possible to get equi-distributed directions on the sphere.

Finally, an interesting perspective concerns the extension of the TDG method to non linear models. Of course this brings new difficulties, such as the construction of the basis functions or the discretization of the non linearity.

Appendices

Appendix A

Spherical harmonics

We recall some definitions and properties of the spherical harmonics and adopt the presentation given in [Her16].

A.1 Legendre functions

The spherical harmonics are based on the Legendre functions P_k^l which read

$$P_k^l(\mu) = \begin{cases} \frac{1}{2^k k!}(1-\mu^2)^{l/2}\frac{d^{k+l}}{d\mu^{k+l}}((\mu^2-1)^k), & l \geq 0, \\ (-1)^l \frac{(k+l)!}{(k-l)!} P_k^{-l}(\mu), & l < 0. \end{cases} \tag{A.1}$$

The Legendre polynomials satisfy the orthogonal relations

$$\frac{1}{2}\int_{-1}^{1} P_k^0 d\mu = \delta_k^0, \quad \frac{1}{2}\int_{-1}^{1} P_k^l P_m^l d\mu = \frac{1}{(a_k^l)^2}\delta_k^m,$$

where a_k^l is the normalization factor

$$a_k^l = \sqrt{(2k+1)\frac{(k-l)!}{(k+l)!}}.$$

They also satisfy the following recursion relations which are fundamentals to derive the P_N model

$$\begin{cases} \sqrt{1-\mu^2}P_k^m = \frac{1}{2k+1}\Big(P_{k+1}^{m+1} - P_{k-1}^{m+1}\Big), \\ \sqrt{1-\mu^2}P_k^m = \frac{1}{2k+1}\Big(-(k-m+1)(k-m+2)P_{k+1}^{m-1} + (k+m-1)(k+m)P_{k-1}^{m-1}\Big), \\ \mu P_k^m = \frac{1}{2k+1}\Big((k-m+1)P_{k+1}^m + (k+m)P_{k-1}^m\Big). \end{cases}$$

A.2 Spherical harmonics

The complex valued spherical harmonics read

$$Y_k^l(\psi,\phi) := Y_k^l(\mathbf{\Omega}) := (-1)^l a_k^l P_k^l(\cos\phi)e^{il\psi}, \quad |l| \leq k, \tag{A.2}$$

The real valued spherical harmonics $Y_{k,l}$ are defined from the complex valued spherical harmonics Y_k^l as follow:

$$\begin{cases} Y_{k,l}(\mathbf{\Omega}) = Y_k^l(\mathbf{\Omega}) = a_k^l P_k^l(\cos\phi), & l = 0, \\ Y_{k,l}(\mathbf{\Omega}) = \dfrac{(-1)^l}{\sqrt{2}}\big(Y_k^l(\mathbf{\Omega}) + \bar{Y}_k^l(\mathbf{\Omega})\big) = a_k^l\sqrt{2}\cos(l\psi)P_k^l(\cos\phi), & 0 < l \leq k, \\ Y_{k,l}(\mathbf{\Omega}) = \dfrac{i}{\sqrt{2}}\big(Y_k^l(\mathbf{\Omega}) - \bar{Y}_k^l(\mathbf{\Omega})\big) = a_k^{|l|}\sqrt{2}\sin(|l|\psi)P_k^{|l|}(\cos\phi), & -k \leq l < 0. \end{cases} \tag{A.3}$$

In particular, the real valued spherical harmonics satisfy the relations

$$\frac{1}{4\pi}\int_{S^2} Y_{k,l}\,d\psi d\mu = \delta_k^0\delta_l^0, \quad \frac{1}{4\pi}\int_{S^2} Y_{k,l}Y_{m,n}\,d\psi d\mu = \delta_k^m\delta_l^n.$$

Moreover, they also satisfy the following recursion relations

$$\begin{cases} \sin\phi\cos\psi Y_{k,m} &= \varepsilon^m(A_k^m Y_{k+1,m+1} - B_k^m Y_{k-1,m+1}) - \zeta^m(C_k^m Y_{k+1,m-1} - D_k^m Y_{k-1,m-1}), \\ \sin\phi\sin\psi Y_{k,m} &= \eta^m(A_k^m Y_{k+1,-m-1} - B_k^m Y_{k-1,-m-1}) + \phi^m(C_k^m Y_{k+1,-m+1} - D_k^m Y_{k-1,-m+1}), \\ \cos\phi Y_{k,m} &= E_k^m Y_{k+1,m} + F_{k,m} Y_{k-1,m}, \end{cases} \tag{A.4}$$

where all the coefficients are given in Table A.1 and by

$$\begin{cases} A_k^m = \sqrt{\dfrac{(k+m+1)(k+m+2)}{(2k+1)(2k+3)}}, & B_k^m = \sqrt{\dfrac{(k-m-1)(k-m)}{(2k-1)(2k+1)}}, \\ C_k^m = \sqrt{\dfrac{(k-m+1)(k-m+2)}{(2k+1)(2k+3)}}, & D_k^m = \sqrt{\dfrac{(k+m-1)(k+m)}{(2k-1)(2k+1)}}, \\ E_k^m = \sqrt{\dfrac{(k-m+1)(k+m+1)}{(2k+1)(2k+3)}}, & F_k^m = \sqrt{\dfrac{(k-m)(k+m)}{(2k-1)(2k+1)}}. \end{cases}$$

	$m < -1$	$m = -1$	$m = 0$	$m = 1$	$m > 1$
ε^m	$-\frac{1}{2}$	0	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
ζ^m	$-\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$
η^m	$-\frac{1}{2}$	$-\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
ϕ^m	$-\frac{1}{2}$	$-\frac{1}{2}$	0	0	$\frac{1}{2}$

Table A.1 – Coefficients of the equations (A.4)

Appendix B

Technical results for the P_N model

In this appendix, we give some technical proofs from Chapter 4.

B.1 Polynomial solutions for a simplified second order equation

B.1.1 Proof of Proposition 4.37

In this section, we prove Proposition 4.37. We recall that Proposition 4.37 reads

Proposition. *Assume $u \in C^{n+1}(\Omega)$ is solution to (4.44). Then, the double sum Taylor expansion in (4.45) can be recast as a simple sum with only zero or first order derivatives with respect to y. Therefore (4.45) can be written*

$$\begin{aligned} u(\mathbf{x}) = \beta_0^0(\mathbf{x})u(\mathbf{x}_0) + \sum_{k=1}^{n} \left[\beta_k^k(\mathbf{x})\partial_x^k u(\mathbf{x}_0) + \beta_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y u(\mathbf{x}_0)\right] \\ + \sum_{p=0}^{n+1} \gamma_{n+1}^p(\mathbf{x})\partial_x^p \partial_y^{n+1-p} u(\mathbf{x}_s), \quad \forall \mathbf{x} \in \Omega, \end{aligned} \tag{B.1}$$

where $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1-s)x_0 + sx$ and $y_s = (1-s)y_0 + sy$.

We will need the following lemma.

Lemma B.1. *Assume hypotheses of Proposition 4.37 are satisfied. Then for all $0 \leq l \leq n-2$ one has the identity*

$$\begin{aligned} &\sum_{p=0}^{l} \gamma_l^p(\mathbf{x})\partial_x^p \partial_y^{l-p} u(\mathbf{x}_0) + \sum_{p=0}^{l+2} \alpha_{l+2}^p(\mathbf{x})\partial_x^p \partial_y^{l+2-p} u(\mathbf{x}_0) = \\ &\sum_{p=0}^{l} \alpha_l^p(\mathbf{x})\partial_x^p \partial_y^{l-p} u(\mathbf{x}_0) + \beta_{l+2}^{l+2}(\mathbf{x})\partial_x^{l+2} u(\mathbf{x}_0) + \beta_{l+2}^{l+1}(\mathbf{x})\partial_x^{l+1}\partial_y u(\mathbf{x}_0). \end{aligned} \tag{B.2}$$

Proof. Let $l \in \mathbb{N}$, $0 \leq l \leq n-2$. For $l_1 \in \mathbb{Z}$, $-1 \leq l_1 \leq l-1$, we define the function

$$\begin{aligned} f(l_1) = &\sum_{p=0}^{l_1} \alpha_l^p(\mathbf{x})\partial_x^p \partial_y^{l-p} u(\mathbf{x}_0) + \sum_{p=l_1+1}^{l} \gamma_l^p(\mathbf{x})\partial_x^p \partial_y^{l-p} u(\mathbf{x}_0) + \sum_{p=l_1+3}^{l+2} \alpha_{l+2}^p(\mathbf{x})\partial_x^p \partial_y^{l+2-p} u(\mathbf{x}_0) \\ &+ \beta_{l+2}^{l_1+2}(\mathbf{x})\partial_x^{l_1+2}\partial_y^{l-l_1} u(\mathbf{x}_0) + \beta_{l+2}^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l+1-l_1} u(\mathbf{x}_0), \end{aligned} \tag{B.3}$$

where we use the convention $\sum_{p=a}^{b} = 0$ for $a, b \in \mathbb{Z}$ and $b < a$. First, we show $f(l_1) = f(l_1+1)$ for $-1 \leq l_1 \leq l-1$. Because u is solution to the equation (B.2) one notices

$$\beta_{l+2}^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l+1-l_1}u(\mathbf{x}_0) = \Big(-\partial_x^{l_1+3}\partial_y^{l-l_1-1} + \omega\beta_{l+2}^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}\Big)u(\mathbf{x}_0). \tag{B.4}$$

Now we consider the definition of the function f (B.3) and we study the difference $f(l_1+1)-f(l_1)$. After easy simplifications on the elements which appear both in $f(l_1)$ and $f(l_1+1)$ one finds

$$\begin{aligned} f(l_1+1) - f(l_1) =& \alpha_l^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}u(\mathbf{x}_0) - \gamma_l^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}u(\mathbf{x}_0) \\ & - \alpha_{l+2}^{l_1+3}(\mathbf{x})\partial_x^{l_1+3}\partial_y^{l-l_1-1}u(\mathbf{x}_0) + \beta_{l+2}^{l_1+3}(\mathbf{x})\partial_x^{l_1+3}\partial_y^{l-l_1-1}u(\mathbf{x}_0) \\ & - \beta_{l+2}^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l+1-l_1}u(\mathbf{x}_0). \end{aligned}$$

Using the equality (B.4) to reformulate the fifth term in the right hand side, one gets

$$\begin{aligned} f(l_1+1) - f(l_1) =& \alpha_l^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}u(\mathbf{x}_0) - \gamma_l^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}u(\mathbf{x}_0) \\ & - \alpha_{l+2}^{l_1+3}(\mathbf{x})\partial_x^{l_1+3}\partial_y^{l-l_1-1}u(\mathbf{x}_0) + \beta_{l+2}^{l_1+3}(\mathbf{x})\partial_x^{l_1+3}\partial_y^{l-l_1-1}u(\mathbf{x}_0) \\ & + \beta_{l+2}^{l_1+1}(\mathbf{x})\Big(\partial_x^{l_1+3}\partial_y^{l-l_1-1} - \omega\partial_x^{l_1+1}\partial_y^{l-l_1-1}\Big)u(\mathbf{x}_0). \end{aligned}$$

Ordering the terms with respect to the derivatives gives

$$\begin{aligned} f(l_1+1) - f(l_1) =& \Big(\alpha_l^{l_1+1}(\mathbf{x}) - \gamma_l^{l_1+1}(\mathbf{x}) - \omega\beta_{l+2}^{l_1+1}(\mathbf{x})\Big)\partial_x^{l_1+1}\partial_y^{l-l_1-1}u(\mathbf{x}_0) \\ & + \Big(-\alpha_{l+2}^{l_1+3}(\mathbf{x}) + \beta_{l+2}^{l_1+1}(\mathbf{x}) + \beta_{l+2}^{l_1+3}(\mathbf{x})\Big)\partial_x^{l_1+3}\partial_y^{l-l_1-1}u(\mathbf{x}_0). \end{aligned}$$

Using the definitions (4.46) and (4.47) one finds $\alpha_l^{l_1+1}(\mathbf{x}) - \gamma_l^{l_1+1}(\mathbf{x}) - \omega\beta_{l+2}^{l_1+1}(\mathbf{x}) = 0$ and $\beta_{l+2}^{l_1+3}(\mathbf{x}) - \alpha_{l+2}^{l_1+3}(\mathbf{x}) + \beta_{l+2}^{l_1+1}(\mathbf{x}) = 0$. Therefore, one has $f(l_1+1)-f(l_1) = 0$ for all $-1 \leq l_1 \leq l-1$. One deduces $f(-1) = f(l)$ which can be written

$$\begin{aligned} &\sum_{p=0}^{l}\gamma_l^p(\mathbf{x})\partial_x^p\partial_y^{l-p}u(\mathbf{x}_0) + \sum_{p=2}^{l+2}\alpha_{l+2}^p(\mathbf{x})\partial_x^p\partial_y^{l+2-p}u(\mathbf{x}_0) + \beta_{l+2}^0(\mathbf{x})\partial_y^{l+2}u(\mathbf{x}_0) + \beta_{l+2}^1(\mathbf{x})\partial_x\partial_y^{l+1}u(\mathbf{x}_0) = \\ &\quad \sum_{p=0}^{l}\alpha_l^p(\mathbf{x})\partial_x^p\partial_y^{l-p}u(\mathbf{x}_0) + \beta_{l+2}^{l+2}(\mathbf{x})\partial_x^{l+2}u(\mathbf{x}_0) + \beta_{l+2}^{l+1}(\mathbf{x})\partial_x^{l+1}\partial_y u(\mathbf{x}_0). \end{aligned}$$

Noticing from (4.79) that $\alpha_{l+2}^0(\mathbf{x}) = \beta_{l+2}^0(\mathbf{x})$ and $\alpha_{l+2}^1(\mathbf{x}) = \beta_{l+2}^1(\mathbf{x})$, one incorporates the two corresponding terms in the second sum so one finds the equality (B.2). It ends the proof. ■

We can now prove the Proposition 4.37.

Proof of Proposition 4.37. We start from the Taylor expansion (4.45). From the equation (4.46) one deduces $\alpha_n^p(\mathbf{x}) = \gamma_n^p(\mathbf{x})$ and $\alpha_{n-1}^p(\mathbf{x}) = \gamma_{n-1}^p(\mathbf{x})$. Therefore

$$\begin{aligned} u(\mathbf{x}) =& \sum_{k=0}^{n-2}\sum_{p=0}^{k}\gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}u(\mathbf{x}_0) + \sum_{p=0}^{n-1}\alpha_{n-1}^p(\mathbf{x})\partial_x^p\partial_y^{n-1-p}u(\mathbf{x}_0) \\ & + \sum_{p=0}^{n}\alpha_n^p(\mathbf{x})\partial_x^p\partial_y^{n-p}u(\mathbf{x}_0) + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s). \end{aligned}$$

One can recursively use the equality (B.2) from $l = n-2$ to $l = 0$. More precisely, rearranging the first sum one has

$$u(\mathbf{x}) = \sum_{k=0}^{n-3}\sum_{p=0}^{k}\gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}u(\mathbf{x}_0) + \sum_{p=0}^{n-1}\alpha_{n-1}^p(\mathbf{x})\partial_x^p\partial_y^{n-1-p}u(\mathbf{x}_0) + \left(\sum_{p=0}^{n-2}\gamma_{n-2}^p(\mathbf{x})\partial_x^p\partial_y^{n-2-p}u(\mathbf{x}_0) + \sum_{p=0}^{n}\alpha_n^p(\mathbf{x})\partial_x^p\partial_y^{n-p}u(\mathbf{x}_0)\right) + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s).$$

One can reformulate the terms between parenthesis using (B.2) with the index correspondance $n-2 = l$. One finds

$$u(\mathbf{x}) = \sum_{k=0}^{n-3}\sum_{p=0}^{k}\gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}u(\mathbf{x}_0) + \sum_{p=0}^{n-1}\alpha_{n-1}^p(\mathbf{x})\partial_x^p\partial_y^{n-1-p}u(\mathbf{x}_0) + \sum_{p=0}^{n-2}\alpha_{n-2}^p(\mathbf{x})\partial_x^p\partial_y^{n-2-p}u(\mathbf{x}_0) + \left[\beta_n^n(\mathbf{x})\partial_x^n u(\mathbf{x}_0) + \beta_n^{n-1}(\mathbf{x})\partial_x^{n-1}\partial_y u(\mathbf{x}_0)\right] + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s). \tag{B.5}$$

And one can now recursively repeat this simple operation using the equality (B.2) for $l = n-3$, $\ldots$, to $l = 0$. One finally gets the formula (B.5) where the first line is written for $n = 2$, the term $[\cdot]$ becomes a sum and the last term remains unchanged

$$u(\mathbf{x}) = 0 + \sum_{p=0}^{1}\alpha_1^p(\mathbf{x})\partial_x^p\partial_y^{1-p}u(\mathbf{x}_0) + \alpha_0^0(\mathbf{x})u(\mathbf{x}_0) + \sum_{k=2}^{n}\left[\beta_k^k(\mathbf{x})\partial_x^k u(\mathbf{x}_0) + \beta_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y u(\mathbf{x}_0)\right] + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}u(\mathbf{x}_s).$$

That is

$$u(\mathbf{x}) = \alpha_0^0(\mathbf{x})u(\mathbf{x}_0) + \alpha_1^1(\mathbf{x})\partial_x u(\mathbf{x}_0) + \alpha_1^0(\mathbf{x})\partial_y u(\mathbf{x}_0) + \sum_{k=2}^{n}\left[\beta_k^k(\mathbf{x})\partial_x^k u(\mathbf{x}_0) + \beta_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y u(\mathbf{x}_0)\right] + \sum_{p=0}^{n+1}\partial_x^p\partial_y^{n+1-p}\gamma_{n+1}^p(\mathbf{x})u(\mathbf{x}_s).$$

Noticing from (4.79) $\alpha_0^0(\mathbf{x}) = \beta_0^0(\mathbf{x})$, $\alpha_1^0(\mathbf{x}) = \beta_1^0(\mathbf{x})$, $\alpha_1^1(\mathbf{x}) = \beta_1^1(\mathbf{x})$ one finds the expression (B.1). The proof is complete. ■

B.1.2 Proof of Proposition 4.38

In this section, we prove Proposition 4.38. We recall that Proposition 4.38 reads

Proposition (Limit of the coefficients $\beta_k^k(\mathbf{x})$ and $\beta_k^{k-1}(\mathbf{x})$)**.** *Assume $\omega = 0$ and $0 \le k \le n$. The coefficients β_k^k and β_k^{k-1} are harmonic polynomials when $\omega = 0$. More precisely one has*

$$\boldsymbol{\beta}(\mathbf{x}) \underset{\omega\to 0}{\to} \mathbf{q}(\mathbf{x}).$$

Proof. One has

$$\Big((x-x_0) + i(y-y_0)\Big)^k = \sum_{l=0}^{k} C_k^{k-l} i^l (x-x_0)^{k-l}(y-y_0)^l.$$

Therefore

$$\Re\Big((x-x_0)+i(y-y_0)\Big)^k = \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} C_k^{k-2l}(-1)^l (x-x_0)^{k-2l}(y-y_0)^{2l},$$
$$\Im\Big((x-x_0)+i(y-y_0)\Big)^k = \sum_{l=0}^{\lfloor \frac{k-1}{2} \rfloor} C_k^{k-2l-1}(-1)^l (x-x_0)^{k-2l-1}(y-y_0)^{2l+1}.$$

That is from the definition of the coefficients γ_k^p

$$\Re\Big((x-x_0)+i(y-y_0)\Big)^k = k! \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} (-1)^l \gamma_k^{k-2l},$$
$$\Im\Big((x-x_0)+i(y-y_0)\Big)^k = k! \sum_{l=0}^{\lfloor \frac{k-1}{2} \rfloor} (-1)^l \gamma_k^{k-2l-1}.$$

When $\omega = 0$ one has $\alpha_k^p = \gamma_k^p$. Therefore, from the recurrence formula (4.47) one deduces

$$\beta_k^k(\mathbf{x}) = \frac{1}{k!}\Re\Big((x-x_0)+i(y-y_0)\Big)^k, \quad \beta_k^{k-1}(\mathbf{x}) = \frac{1}{k!}\Im\Big((x-x_0)+i(y-y_0)\Big)^k, \quad 0 \le k \le n.$$

This completes the proof. ■

B.2 Polynomial solutions to the P_N model

B.2.1 Proof of Proposition 4.45

In this section, we prove Proposition 4.45. We recall that Proposition 4.45 reads

Proposition (A first simplification of the Taylor expansion)**.** *Let* $\mathbf{u}(\mathbf{x}) \in C^{n+2}(\Omega)$ *be a solution to* (4.59)*. The beginning of the Taylor expansion on the vectorial function* $\mathbf{u}(\mathbf{x})$ *can be recast as a Taylor expansion on the vectorial function* $\mathbf{u}_{|_1^1}(\mathbf{x}_0)$

$$\mathbf{u}(\mathbf{x}) = \sum_{k=0}^{n}\sum_{p=0}^{k} \Gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}\mathbf{u}_{|_1^1}(\mathbf{x}_0) + \boldsymbol{\xi}(\mathbf{x}),$$
$$\boldsymbol{\xi}(\mathbf{x}) = \sum_{p=0}^{n+1} \Gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}\mathbf{u}(\mathbf{x}_0) + \sum_{p=0}^{n+2} \gamma_{n+2}^p(\mathbf{x}) I_m \partial_x^p\partial_y^{n+2-p}\mathbf{u}(\mathbf{x}_s),$$

where $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1-s)x_0 + sx$ *and* $y_s = (1-s)y_0 + sy$.

Proof. For $l \in \mathbb{Z}$, $-1 \le l \le n$ we define the function

$$f(l) = \sum_{k=0}^{l}\sum_{p=0}^{k} \Gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}\mathbf{u}_{|_1^1}(\mathbf{x}_0) + \sum_{p=0}^{l+1} \Gamma_{l+1}^p(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\mathbf{u}(\mathbf{x}_0) + \sum_{k=l+2}^{n+1}\sum_{p=0}^{k} \gamma_k^p(\mathbf{x}) I_m \partial_x^p\partial_y^{k-p}\mathbf{u}(\mathbf{x}_0), \tag{B.6}$$

where we use the convention $\sum_{p=a}^{b} = 0$ for $a, b \in \mathbb{Z}$ and $b < a$. First we show $f(l) = f(l+1)$ for $-1 \leq l \leq n-1$. Because $\mathbf{u}$ is solution to (4.59) and from the definition of the matrix $R_{|_m^2}^{-1}$ one has

$$\mathbf{u}_{|_m^2}(\mathbf{x}_0) = -R_{|_m^2}^{-1}\Big(A_1\partial_x + A_2\partial_y\Big)\mathbf{u}(\mathbf{x}_0).$$

Therefore

$$\begin{aligned}\sum_{p=0}^{l+1}\Gamma_{l+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\mathbf{u}_{|_m^2}(\mathbf{x}_0) &= -\sum_{p=0}^{l+1}\Gamma_{l+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}R_{|_m^2}^{-1}\Big(A_1\partial_x + A_2\partial_y\Big)\mathbf{u}(\mathbf{x}_0),\\ &= -\sum_{p=1}^{l+2}\Gamma_{l+1}^{p-1}(\mathbf{x})\partial_x^p\partial_y^{l+2-p}R_{|_m^2}^{-1}A_1\mathbf{u}(\mathbf{x}_0) - \sum_{p=0}^{l+1}\Gamma_{l+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+2-p}R_{|_m^2}^{-1}A_2\mathbf{u}(\mathbf{x}_0).\end{aligned}$$

And therefore since $\Gamma_k^{-1} = \Gamma_k^{k+1} = 0$

$$\sum_{p=0}^{l+1}\Gamma_{l+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\mathbf{u}_{|_m^2}(\mathbf{x}_0) = -\sum_{p=0}^{l+2}\Big(\Gamma_{l+1}^{p-1}(\mathbf{x})R_{|_m^2}^{-1}A_1 + \Gamma_{l+1}^{p}(\mathbf{x})R_{|_m^2}^{-1}A_2\Big)\partial_x^p\partial_y^{l+2-p}\mathbf{u}(\mathbf{x}_0). \quad \text{(B.7)}$$

Now, we consider the definition of the function f (B.6) and we study the difference $f(l+1)-f(l)$. After easy simplifications on the elements which appear both in $f(l)$ and $f(l+1)$ one finds

$$\begin{aligned}f(l+1) - f(l) =& \sum_{p=0}^{l+1}\Gamma_{l+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\mathbf{u}_{|_1^1}(\mathbf{x}_0) + \sum_{p=0}^{l+2}\Gamma_{l+2}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+2-p}\mathbf{u}(\mathbf{x}_0) - \sum_{p=0}^{l+1}\Gamma_{l+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\mathbf{u}(\mathbf{x}_0)\\ &- \sum_{p=0}^{l+2}\gamma_{l+2}^{p}(\mathbf{x})I_m\partial_x^p\partial_y^{l+2-p}\mathbf{u}(\mathbf{x}_0).\end{aligned}$$

Regrouping the terms one gets

$$f(l+1)-f(l) = \sum_{p=0}^{l+1}\Gamma_{l+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\Big(\mathbf{u}_{|_1^1}(\mathbf{x}_0)-\mathbf{u}(\mathbf{x}_0)\Big)+\sum_{p=0}^{l+2}\Big(\Gamma_{l+2}^{p}(\mathbf{x})-\gamma_{l+2}^{p}(\mathbf{x})I_m\Big)\partial_x^p\partial_y^{l+2-p}\mathbf{u}(\mathbf{x}_0).$$

That is using $\mathbf{u}_{|_m^2} = \mathbf{u} - \mathbf{u}_{|_1^1}$ and the definition of the coefficients Γ_k^p (4.41)

$$f(l+1)-f(l) = -\sum_{p=0}^{l+1}\Gamma_{l+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\mathbf{u}_{|_m^2}(\mathbf{x}_0)+\sum_{p=0}^{l+2}\Big(-\Gamma_{l+1}^{p-1}(\mathbf{x})R_{|_m^2}^{-1}A_1-\Gamma_{l+1}^{p}(\mathbf{x})R_{|_m^2}^{-1}A_2\Big)\partial_x^p\partial_y^{l+2-p}\mathbf{u}(\mathbf{x}_0).$$

Using (B.7) one finally finds $f(l+1) - f(l) = 0$ for all $-1 \leq l \leq n-1$. Therefore, one gets $f(-1) = f(n)$. That is using $\Gamma_0^0 = \gamma_0^0$ and the definition (B.6) of the function f

$$\sum_{k=0}^{n+1}\sum_{p=0}^{k}\gamma_k^p(\mathbf{x})I_m\partial_x^p\partial_y^{k-p}\mathbf{u}(\mathbf{x}_0) = \sum_{k=0}^{n}\sum_{p=0}^{k}\Gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}\mathbf{u}_{|_1^1}(\mathbf{x}_0)+\sum_{p=0}^{n+1}\Gamma_{n+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\mathbf{u}(\mathbf{x}_0). \quad \text{(B.8)}$$

We consider now the Taylor expansion of the function $\mathbf{u}(\mathbf{x})$

$$\mathbf{u}(\mathbf{x}) = \sum_{k=0}^{n+1}\sum_{p=0}^{k}\gamma_k^p(\mathbf{x})I_m\partial_x^p\partial_y^{k-p}\mathbf{u}(\mathbf{x}_0) + \sum_{p=0}^{n+2}\gamma_{n+2}^{p}(\mathbf{x})I_m\partial_x^p\partial_y^{n+2-p}\mathbf{u}(\mathbf{x}_s).$$

Using (B.8) one finally gets

$$\mathbf{u}(\mathbf{x}) = \sum_{k=0}^{n}\sum_{p=0}^{k}\Gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}\mathbf{u}_{|_1^1}(\mathbf{x}_0)+\sum_{p=0}^{n+1}\Gamma_{n+1}^{p}(\mathbf{x})\partial_x^p\partial_y^{l+1-p}\mathbf{u}(\mathbf{x}_0)+\sum_{p=0}^{n+2}\gamma_{n+2}^{p}(\mathbf{x})I_m\partial_x^p\partial_y^{n+2-p}\mathbf{u}(\mathbf{x}_s).$$

This completes the proof. ■

B.3 Convergence of the scheme

B.3.1 Proof of Proposition 4.54

In this section, we prove Proposition 4.54. We recall that Proposition 4.54 reads

Proposition. *Assume* $\mathbf{u}_e(\mathbf{x}) \in C^{n+1}(\Omega)$ *is solution to* (4.73). *Then, the double sum Taylor expansion* (4.75) *can be recast as a simple sum with only zero or first order derivatives with respect to* y

$$\begin{aligned}\mathbf{u}_e(\mathbf{x}) = L_0^0(\mathbf{x})\mathbf{u}_e(\mathbf{x}_0) + \sum_{k=1}^{n}\Big[L_k^k(\mathbf{x})\partial_x^k\mathbf{u}_e(\mathbf{x}_0) + L_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y\mathbf{u}_e(\mathbf{x}_0)\Big]\\ + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}\mathbf{u}_e(\mathbf{x}_s), \quad \forall \mathbf{x}\in\Omega,\end{aligned} \tag{B.9}$$

where $\mathbf{x}_s = (x_s, y_s)^T$, $x_s = (1-s)x_0 + sx$ *and* $y_s = (1-s)y_0 + sy$.

Lemma B.2. *Assume hypotheses of Proposition* 4.54 *are satisfied. Then for all* $0 \le l \le n-2$ *one has the identity*

$$\begin{aligned}\sum_{p=0}^{l}\gamma_l^p(\mathbf{x})\partial_x^p\partial_y^{l-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=0}^{l+2}K_{l+2}^p(\mathbf{x})\partial_x^p\partial_y^{l+2-p}\mathbf{u}_e(\mathbf{x}_0) =\\ \sum_{p=0}^{l}K_l^p(\mathbf{x})\partial_x^p\partial_y^{l-p}\mathbf{u}_e(\mathbf{x}_0) + L_{l+2}^{l+2}\partial_x^{l+2}\mathbf{u}_e(\mathbf{x}_0) + L_{l+2}^{l+1}(\mathbf{x})\partial_x^{l+1}\partial_y\mathbf{u}_e(\mathbf{x}_0).\end{aligned} \tag{B.10}$$

Proof. Let $l\in\mathbb{N}$, $0\le l\le n-2$. For $l_1\in\mathbb{Z}$, $-1\le l_1\le l-1$, we define the function

$$\begin{aligned}f(l_1) =& \sum_{p=0}^{l_1}K_l^p(\mathbf{x})\partial_x^p\partial_y^{l-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=l_1+1}^{l}\gamma_l^p(\mathbf{x})\partial_x^p\partial_y^{l-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=l_1+3}^{l+2}K_{l+2}^p(\mathbf{x})\partial_x^p\partial_y^{l+2-p}\mathbf{u}_e(\mathbf{x}_0)\\ &+\Big(K_{l+2}^{l_1+2}(\mathbf{x}) - L_{l+2}^{l_1}(\mathbf{x})(BB^T)^{-1}(AA^T)\Big)\partial_x^{l_1+2}\partial_y^{l-l_1}\mathbf{u}_e(\mathbf{x}_0) + L_{l+2}^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l+1-l_1}\mathbf{u}_e(\mathbf{x}_0),\end{aligned} \tag{B.11}$$

where we use the convention $\sum_{p=a}^{b}=0$ for $a,b\in\mathbb{Z}$ and $b<a$. First, we show $f(l_1)=f(l_1+1)$ for $-1\le l_1\le l-1$. Because $\mathbf{u}_e$ is solution to the equation (4.73), it statisfies (4.74) and one notices

$$\begin{aligned}L_{l+2}^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l+1-l_1}\mathbf{u}_e(\mathbf{x}_0) =& L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}\Big(-AA^T\partial_x^{l_1+3}\partial_y^{l-l_1-1}\\ &-(AB^T+BA^T)\partial_x^{l_1+2}\partial_y^{l-l_1}+\sigma_t R_1\partial_x^{l_1+1}\partial_y^{l-l_1-1}\Big)\mathbf{u}_e(\mathbf{x}_0).\end{aligned} \tag{B.12}$$

Now, we consider the definition of the function f (B.11) and we study the difference $f(l_1+1)-f(l_1)$. After simplifications on the elements which appear both in $f(l_1)$ and $f(l_1+1)$, one finds

$$\begin{aligned}f(l_1+1)-f(l_1) =& K_l^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0) - \gamma_l^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0)\\ &- K_{l+2}^{l_1+3}(\mathbf{x})\partial_x^{l_1+3}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0)\\ &+\Big(K_{l+2}^{l_1+3}(\mathbf{x}) - L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}AA^T\Big)\partial_x^{l_1+3}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0)\\ &+\Big(L_{l+2}^{l_1+2}(\mathbf{x}) - [K_{l+2}^{l_1+2}(\mathbf{x}) - L_{l+2}^{l_1}(\mathbf{x})(BB^T)^{-1}AA^T]\Big)\partial_x^{l_1+2}\partial_y^{l-l_1}\mathbf{u}_e(\mathbf{x}_0)\\ &- L_{l+2}^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l+1-l_1}\mathbf{u}_e(\mathbf{x}_0).\end{aligned}$$

Simplifying the term $K_{l+2}^{l_1+3}(\mathbf{x})\partial_x^{l_1+3}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0)$ and using the equality (B.12) to reformulate the last term in the right hand side one gets

$$\begin{aligned}
f(l_1+1)-f(l_1) =& K_l^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0) - \gamma_l^{l_1+1}(\mathbf{x})\partial_x^{l_1+1}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0)\\
&- \Big(L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}(AA^T)\Big)\partial_x^{l_1+3}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0)\\
&+ \Big(L_{l+2}^{l_1+2}(\mathbf{x}) - [K_{l+2}^{l_1+2}(\mathbf{x}) - L_{l+2}^{l_1}(\mathbf{x})(BB^T)^{-1}AA^T]\Big)\partial_x^{l_1+2}\partial_y^{l-l_1}\mathbf{u}_e(\mathbf{x}_0)\\
&- L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}\Big(- (AB^T+BA^T)\partial_x^{l_1+2}\partial_y^{l-l_1}\\
&- AA^T\partial_x^{l_1+3}\partial_y^{l-l_1-1} + \sigma_t R_1\partial_x^{l_1+1}\partial_y^{l-l_1-1}\Big)\mathbf{u}_e(\mathbf{x}_0).
\end{aligned}$$

Simplifying the term $\Big(L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}(AA^T)\Big)\partial_x^{l_1+3}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0)$ and ordering the terms with respect to the derivatives gives

$$\begin{aligned}
f(l_1+1)-f(l_1) =& \Big(K_l^{l_1+1}(\mathbf{x}) - \gamma_l^{l_1+1}(\mathbf{x}) - \sigma_t L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}R_1\Big)\partial_x^{l_1+1}\partial_y^{l-l_1-1}\mathbf{u}_e(\mathbf{x}_0)\\
&+ \Big(L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}(AB^T+BA^T) + L_{l+2}^{l_1+2}\\
&- [K_{l+2}^{l_1+2}(\mathbf{x}) - L_{l+2}^{l_1}(\mathbf{x})(BB^T)^{-1}AA^T]\Big)\partial_x^{l_1+2}\partial_y^{l-l_1}\mathbf{u}_e(\mathbf{x}_0).
\end{aligned}$$

Using the Definition (4.76), one finds $K_l^{l_1+1}(\mathbf{x}) - \gamma_l^{l_1+1}(\mathbf{x}) - \sigma_t L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}R_1 = 0$. With the Definition (4.77) one gets (since $l_1 < l$) $L_{l+2}^{l_1+2}(\mathbf{x}) := K_{l+2}^{l_1+2}(\mathbf{x}) - L_{l+2}^{l_1+1}(\mathbf{x})(BB^T)^{-1}(AB^T + BA^T) - L_{l+2}^{l_1}(\mathbf{x})(BB^T)^{-1}AA^T$. Therefore, one has $f(l_1+1)-f(l_1)=0$ for all $-1 \le l_1 \le l-1$. One deduces $f(-1) = f(l)$ which can be written

$$\begin{aligned}
&\sum_{p=0}^{l}\gamma_l^p(\mathbf{x})\partial_x^p\partial_y^{l-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=2}^{l+2}K_{l+2}^p(\mathbf{x})\partial_x^p\partial_y^{l+2-p}\mathbf{u}_e(\mathbf{x}_0) + L_{l+2}^0(\mathbf{x})\partial_y^{l+2}\mathbf{u}_e(\mathbf{x}_0)\\
&+ \Big(K_{l+2}^1(\mathbf{x}) - L_{l+2}^{-1}(\mathbf{x})(BB^T)^{-1}(AA^T)\Big)\partial_x\partial_y^{l+1}\mathbf{u}_e(\mathbf{x}_0) =\\
&\sum_{p=0}^{l}K_l^p(\mathbf{x})\partial_x^p\partial_y^{l-p}\mathbf{u}_e(\mathbf{x}_0) + \Big(K_{l+2}^{l+2}(\mathbf{x}) - L_{l+2}^{l}(\mathbf{x})(BB^T)^{-1}(AA^T)\Big)\partial_x^{l+2}\mathbf{u}_e(\mathbf{x}_0) + L_{l+2}^{l+1}(\mathbf{x})\partial_x^{l+1}\partial_y\mathbf{u}_e(\mathbf{x}_0).
\end{aligned} \tag{B.13}$$

By definition $L_{l+2}^{-1} = 0$ and one notices from (4.79) that $K_{l+2}^0(\mathbf{x}) = L_{l+2}^0(\mathbf{x})$. One can therefore incorporate the two terms K_{l+2}^0 and K_{l+2}^1 in the second sum on the left hand side of (B.13). Moreover from (4.78), one has $L_{l+2}^{l+2} = K_{l+2}^{l+2}(\mathbf{x}) - L_{l+2}^{l}(\mathbf{x})(BB^T)^{-1}AA^T$. Using this equality on the right hand side of (B.13) completes the proof. ■

We can now give the proof of Proposition 4.54.

Proof of Proposition 4.54. We start from the Taylor expansion (4.75). From the Definition (4.76) one has $K_n^p(\mathbf{x}) = \gamma_n^p(\mathbf{x})$ and $K_{n-1}^p(\mathbf{x}) = \gamma_{n-1}^p(\mathbf{x})$. Therefore

$$\begin{aligned}
\mathbf{u}_e(\mathbf{x}) =& \sum_{k=0}^{n-2}\sum_{p=0}^{k}\gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=0}^{n-1}K_{n-1}^p(\mathbf{x})\partial_x^p\partial_y^{n-1-p}\mathbf{u}_e(\mathbf{x}_0)\\
&+ \sum_{p=0}^{n}K_n^p(\mathbf{x})\partial_x^p\partial_y^{n-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}\mathbf{u}_e(\mathbf{x}_s).
\end{aligned}$$

One can recursively use the equality (B.10) from $l = n-2$ to $l = 0$. More precisely, rearranging the first sum one has

$$\begin{aligned}\mathbf{u}_e(\mathbf{x}) =& \sum_{k=0}^{n-3}\sum_{p=0}^{k}\gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=0}^{n-1}K_{n-1}^p(\mathbf{x})\partial_x^p\partial_y^{n-1-p}\mathbf{u}_e(\mathbf{x}_0)\\ &+\left(\sum_{p=0}^{n-2}\gamma_{n-2}^p(\mathbf{x})\partial_x^p\partial_y^{n-2-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=0}^{n}K_n^p(\mathbf{x})\partial_x^p\partial_y^{n-p}\mathbf{u}_e(\mathbf{x}_0)\right) + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}\mathbf{u}_e(\mathbf{x}_s).\end{aligned}$$

One can reformulate the terms between parenthesis using (B.10) with the index correspondence $l = n-2$. One finds

$$\begin{aligned}\mathbf{u}_e(\mathbf{x}) =& \sum_{k=0}^{n-3}\sum_{p=0}^{k}\gamma_k^p(\mathbf{x})\partial_x^p\partial_y^{k-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=0}^{n-1}K_{n-1}^p(\mathbf{x})\partial_x^p\partial_y^{n-1-p}\mathbf{u}_e(\mathbf{x}_0) + \sum_{p=0}^{n-2}K_{n-2}^p(\mathbf{x})\partial_x^p\partial_y^{n-2-p}\mathbf{u}_e(\mathbf{x}_0)\\ &+\left[L_n^n(\mathbf{x})\partial_x^n\mathbf{u}_e(\mathbf{x}_0) + L_n^{n-1}(\mathbf{x})\partial_x^{n-1}\partial_y\mathbf{u}_e(\mathbf{x}_0)\right] + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}\mathbf{u}_e(\mathbf{x}_s).\end{aligned} \tag{B.14}$$

And one can now recursively repeat this simple operation using the equality (B.10) for $l = n-3$, …, to $l = 0$. One finally gets the formula (B.14) where the first line is written for $n = 2$, the term $[\cdot]$ becomes a sum and the last term remains unchanged

$$\begin{aligned}\mathbf{u}_e(\mathbf{x}) =& 0 + \sum_{p=0}^{1}K_1^p(\mathbf{x})\partial_x^p\partial_y^{1-p}\mathbf{u}_e(\mathbf{x}_0) + K_0^0(\mathbf{x})\mathbf{u}_e(\mathbf{x}_0)\\ &+\sum_{k=2}^{n}\left[L_k^k(\mathbf{x})\partial_x^k\mathbf{u}_e(\mathbf{x}_0) + L_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y\mathbf{u}_e(\mathbf{x}_0)\right] + \sum_{p=0}^{n+1}\gamma_{n+1}^p(\mathbf{x})\partial_x^p\partial_y^{n+1-p}\mathbf{u}_e(\mathbf{x}_s).\end{aligned}$$

That is

$$\begin{aligned}\mathbf{u}_e(\mathbf{x}) =& K_0^0(\mathbf{x})\mathbf{u}_e(\mathbf{x}_0) + K_1^1(\mathbf{x})\partial_x\mathbf{u}_e(\mathbf{x}_0) + K_1^0(\mathbf{x})\partial_y\mathbf{u}_e(\mathbf{x}_0)\\ &+\sum_{k=2}^{n}\left[L_k^k(\mathbf{x})\partial_x^k\mathbf{u}_e(\mathbf{x}_0) + L_k^{k-1}(\mathbf{x})\partial_x^{k-1}\partial_y\mathbf{u}_e(\mathbf{x}_0)\right] + \sum_{p=0}^{n+1}\partial_x^p\partial_y^{n+1-p}\gamma_{n+1}^p(\mathbf{x})\mathbf{u}_e(\mathbf{x}_s).\end{aligned}$$

Noticing from (4.79) $K_0^0(\mathbf{x}) = L_0^0(\mathbf{x})$, $K_1^0(\mathbf{x}) = L_1^0(\mathbf{x})$, $K_1^1(\mathbf{x}) = L_1^1(\mathbf{x})$, one finds the expression (4.80). This completes the proof. ■

Appendix C

Discontinuous Galerkin method using adjoint solutions as basis functions

In the beginning of this work, adjoint solutions to the model were used as basis functions. The idea was to easily adapt the ultra weak formalism [CD98] to transport problemsCC. However, as this document has made clear, it is much more efficient to use direct solutions to the problem. We give here some numerical examples which were performed with adjoint basis functions. The adjoint P_1 model reads

$$\begin{cases} \varepsilon \partial_t p + \dfrac{c}{\sqrt{3}} \partial_x v = \varepsilon \sigma_a(x) p, \\ \varepsilon \partial_t v + \dfrac{c}{\sqrt{3}} \partial_x p = \sigma_t(x) v, \end{cases} \tag{C.1}$$

C.1 Asymptotic study in one dimension

As in section 3-1.2, one can study the asymptotic behavior of the scheme for the one dimensional hyperbolic heat equation. We consider the following two adjoint solutions

$$\mathbf{e}_{k,1}(t,x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_{k,2}(t,x) = \begin{pmatrix} \frac{\sqrt{3}\sigma_s}{c\varepsilon}(x - x_k) \\ 1 \end{pmatrix}, \tag{C.2}$$

where x_k is the abscissa of the center of the cell k. Instead of using direct solutions to the P_1 model, one can use the adjoint solutions (C.2). One gets the following finite difference scheme

$$\begin{cases} \varepsilon \dfrac{p_k^{n+1} - p_k^n}{\Delta t} + \dfrac{c}{2\sqrt{3}h}[-p_{k+1} + 2p_k - p_{k-1} + (1+a)(v_{k+1} - v_{k-1})]^{n+1} = 0, \\ \varepsilon(1 + \dfrac{a^2}{3}) \dfrac{v_k^{n+1} - v_k^n}{\Delta t} + \dfrac{c}{2\sqrt{3}h} \left[a^2(v_{k+1} + 2v_k + v_{k-1}) + (-v_{k+1} + 2v_k - v_{k-1}) \right. \\ \left. + (1-a)(p_{k+1} - p_{k-1})\right]^{n+1} = -\dfrac{\sigma_s}{\varepsilon} v_k^{n+1}, \end{cases} \tag{C.3}$$

with $a = \frac{\sqrt{3}\sigma_s \Delta x}{2c\varepsilon}$. This scheme is very similar to the scheme (3.14) and can be obtained from (3.14) simply by replacing the coefficient a by its inverse. Using Hilbert expansion, one can show that this scheme is asymptotic preserving when $\varepsilon \to 0$ for some average values for the variable p.

Proposition C.1. *Using Hilbert expansion in the limit $\varepsilon \to 0$, the scheme (C.3) admits the*

following limits $\forall k$

$$\begin{cases} (v^0_{k+1} + v^0_k)^{n+1} = 0, \\ (\dfrac{v^1_{k+1} + 2v^1_k + v^1_{k-1}}{4})^{n+1} = \dfrac{c}{\sqrt{3}\sigma_s}(\dfrac{p^0_{k+1} - p^0_{k-1}}{2h})^{n+1}, \\ \dfrac{(\bar{p}^0_k)^{n+1} - (\bar{p}^0_k)^n}{\Delta t} - \dfrac{c^2}{3\sigma_s}\left(\dfrac{p^0_{k+2} - 2p^0_k + p^0_{k-2}}{4h^2}\right)^{n+1} = 0, \end{cases} \tag{C.4}$$

with $\bar{p}^0_k$ *a mean value of* p^0_k *define as* $\bar{p}^0_k = (\frac{2}{3}p^0_{k+2} + 4p^0_{k+1} + \frac{20}{3}p^0_k + 4p^0_{k-1} + \frac{2}{3}p^0_{k-2})/16$.

Comparing with the model limit (3.2), the scheme is asymptotic preserving for the variable p but not for the variable v. Actually, the variable v in the limit scheme is consistent with the opposite of the model limit. This is confirmed by the numerical test from section 3-1.3.2 in Figure C.1.

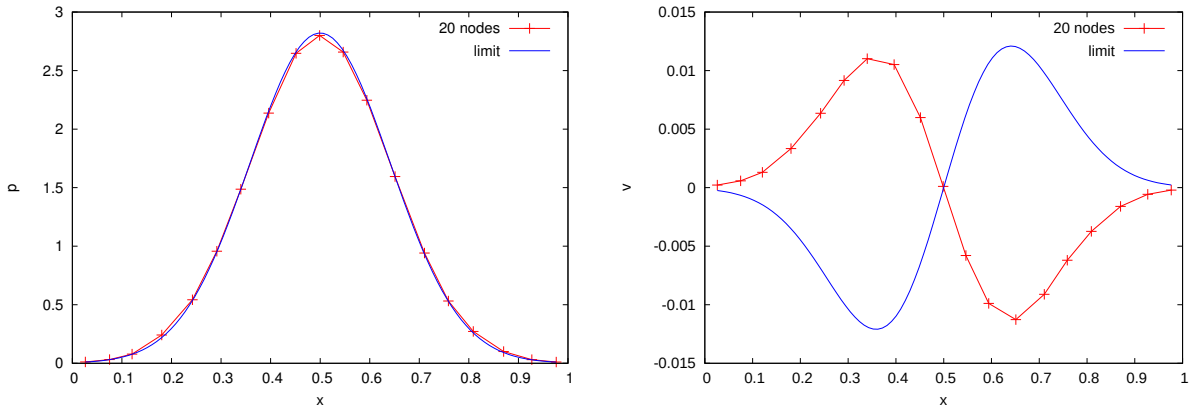

Figure C.1 – Numerical solution obtained for the variable p (on the left) and v (on the right) with the numerical scheme (C.3) with $\varepsilon = 0.001$. Random mesh with 20 nodes and $dt = 0.01/20$. Good accuracy illustrate the AP properties of the scheme for the first variable.

C.2 Order study in one and two dimensions

Consider the time dependent P_1 model in one dimension (3.1). To apply the adjoint method, one first needs to find solutions to the adjoint model (C.1).

Proposition C.2. *Assume* $c \neq 0, \sigma_a \neq 0$. *The adjoint the one dimensional* P_1 *model* (C.1) *admits the following four solutions*

$$\begin{aligned} \mathbf{v}^{\pm}_1(x) &= \begin{pmatrix} \mp\sqrt{\sigma_t} \\ \sqrt{\varepsilon\sigma_a} \end{pmatrix} e^{\mp\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}x}, \\ \mathbf{v}^{\pm}_2(t,x) &= \begin{pmatrix} -\frac{c}{\varepsilon}(\varepsilon\sigma_a - \sigma_t) \pm \sqrt{\frac{3\sigma_a\sigma_t}{\varepsilon}}(\varepsilon\sigma_a + \sigma_t)x + 2\frac{c}{\varepsilon}\sigma_a\sigma_t t \\ -\sqrt{3}\sigma_a(\varepsilon\sigma_a + \sigma_t)x \mp 2c\sigma_a\sqrt{\frac{\sigma_a\sigma_t}{\varepsilon}}t \end{pmatrix} e^{\mp\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}x}. \end{aligned} \tag{C.5}$$

For the two dimensional P_1 model one can also construct adjoint solutions.

Proposition C.3. *Assume* $\mathbf{d}_k = (\cos(\phi_k), \sin(\phi_k))^T \in \mathbb{R}^2, c \neq 0$. *The functions*

$$\mathbf{v}_i = \begin{pmatrix} \sqrt{\sigma_t} \\ -\sqrt{\varepsilon\sigma_a}\mathbf{d}_i \end{pmatrix} e^{-\frac{1}{c}\sqrt{3\varepsilon\sigma_a\sigma_t}\mathbf{d}_i^T\mathbf{x}}, \tag{C.6}$$

are solutions of the adjoint model associated to the two dimensional P_1 *system with constant coefficients* σ_a, σ_T.

We can now reproduce the test cases from sections 3-1.3.1 and 5-4.1 using respectively the adjoint solutions (C.5) and (C.6) as basis functions. The Figures C.2 and C.3 show that, with adjoint basis functions, one can not increase the order of the method for the standard L^2 norm.

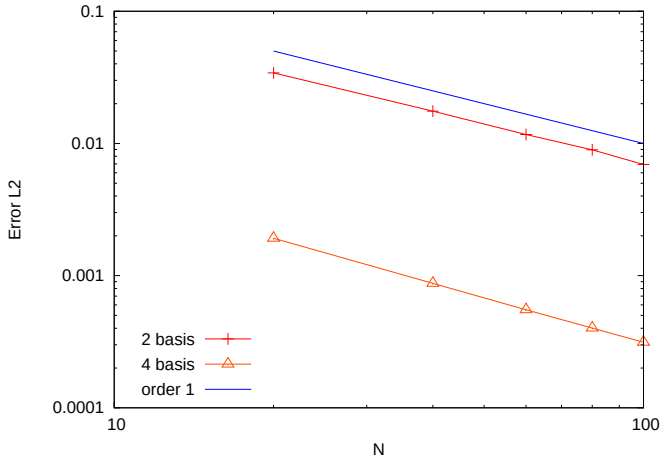

Figure C.2 – Numerical test from section 3-1.3.1. Study of the L^2 error in logarithmic scale using adjoint solution as basis functions for temporal one dimensional model.

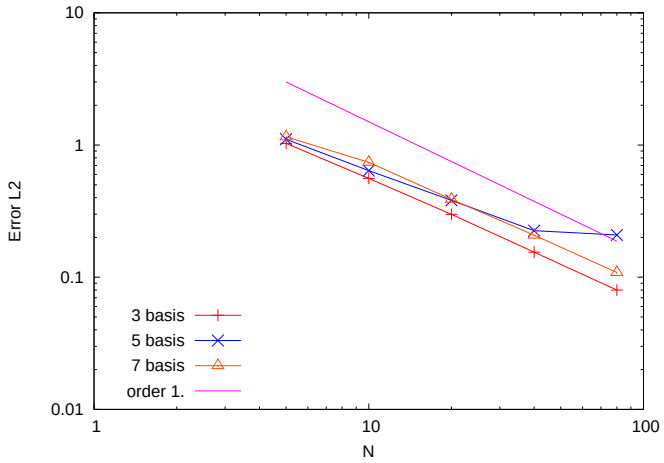

Figure C.3 – Numerical test from section 5-4.1. Study of the L^2 error in logarithmic scale using adjoint solution as basis functions for stationary two dimensional model.

Bibliography

[ABDG15] Gregoire Allaire, Xavier Blanc, Bruno Despres, and François Golse. *Transport et diffusion*. Cours de l'école Polytechnique, 2015.

[ACCG14] A. Ambroso, C. Chalons, Frédéric. Coquel, and T. Galié. Interface model coupling via prescribed local flux balance. *ESAIM: M2AN*, 48(3):895–918, April 2014.

[AD14] R. Abgrall and S. Dallet. *An Asymptotic Preserving Scheme for the Barotropic Baer-Nunziato Model*, pages 749–757. Springer International Publishing, Cham, 2014.

[AH08] Yazid Abdelaziz and Abdelmadjid Hamouine. A survey of the extended finite element. *Computers & Structures*, 86(11):1141 – 1151, 2008.

[And96] Joel Anderson. A secular equation for the eigenvalues of a diagonal matrix perturbation. *Linear Algebra and its Applications*, 246(0):49 – 70, 1996.

[Arn82] Douglas N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982.

[BA69] G. Birkhoff and I. Abu-Shumays. Harmonic solutions of transport equations. *J. Math. Anal. Appl.*, 28:211–221, 1969.

[BA70] G. Birkhoff and I.K. Abu-Shumays. Exact analytic solutions of transport equations. *J. Math. Anal. Appl.*, 32:468–481, 1970.

[BB99] T. Belytschko and T. Black. Elastic crack growth in finite elements with minimal remeshing. *International Journal for Numerical Methods in Engineering*, 45(5):601–620, 1999.

[BDDM11] S. Brull, P. Degond, F. Deluzet, and A. Mouton. Asymptotic-preserving scheme for a bi-fluid Euler-Lorentz model. *Kinetic and Related Models*, 4(4):991–1023, December 2011.

[BDF12] C. Buet, B. Després, and E. Franck. Design of asymptotic preserving finite volume schemes for the hyperbolic heat equation on unstructured meshes. *Numerische Mathematik*, 122(2):227–278, 2012.

[BDF15] Christophe Buet, Bruno Després, and Emmanuel Franck. Asymptotic preserving schemes on distorted meshes for Friedrichs systems with stiff relaxation: application to angular models in linear transport. *J. Sci. Comput.*, 62(2):371–398, 2015.

[BDFL16] Christophe Buet, Bruno Després, Emmanuel Franck, and Thomas Leroy. Proof of uniform convergence for a cell-centered AP discretization of the hyperbolic heat equation on general meshes. *Mathematics of Computation*, September 2016.

[BDM18] Christophe Buet, Bruno Després, and Guillaume Morel. Trefftz discontinuous Galerkin method for Friedrichs systems with linear relaxation: application to the P_1 model. *Computational Methods in Applied Mathematics*, 2018.

[Ben92] David J. Benson. Computational methods in lagrangian and eulerian hydrocodes. *Computer Methods in Applied Mechanics and Engineering*, 99(2):235 – 394, 1992.

[BFB97] A. Miguel Blanco, Michell Florez, and M Bermejo. Evaluation of the rotation matrices in the basis of real spherical harmonics. 419:19–27, 12 1997.

[BH01] Thomas A. Brunner and James Paul Holloway. One-dimensional riemann solvers and the maximum entropy closure. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 69(5):543 – 566, 2001.

[BM96] I. Babuska and J. M. Melenk. The partition of unity method. *International Journal of Numerical Methods in Engineering*, 40:727–758, 1996.

[BM08] Annalisa Buffa and Peter Monk. Error estimates for the ultra weak variational formulation of the helmholtz equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 42(6):925–940, 8 2008.

[BN86] M. R. Baer and J. W. Nunziato. A two-phase mixture theory for the deflagration-to-detonation transition (DDT) in reactive granular materials. *Int. J. Multiphase Flow*, 12(6):861–889, 1986.

[BPV03] Ramaz Botchorishvili, Benoit Perthame, and Alexis Vasseur. Equilibrium schemes for scalar conservation laws with stiff sources. *Mathematics of Computation*, 72(241):131–157, 2003.

[Bre03] Susanne C. Brenner. Poincaré–Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1):306–324, 2003.

[Bru02] Thomas A. Brunner. Form of approximate radiation transport. *Sandia report*, 2002.

[BT11] Christophe Berthon and Rodolphe Turpault. Asymptotic preserving HLL schemes. *Numer. Methods Partial Differ. Equations*, 27(6):1396–1422, 2011.

[Cas60] K.M. Case. Elementary solutions of the transport equation and their applications. *Ann. Phys.*, 9:1–23, 1960.

[CBS98] E. J. Caramana, D. E. Burton, and M. J. Shashkov. The construction of compatible hydrodynamics algorithms utilizing conservation of total energy. *J. Comput. Phys.*, 146(1):227–262, 1998.

[CD98] Olivier Cessenat and Bruno Despres. Application of an ultra weak variational formulation of elliptic pdes to the two-dimensional helmholtz problem. *SIAM J. Numer. Anal.*, 35(1):255–299, February 1998.

[CDDL09] G. Carré, S. Del Pino, B. Després, and E. Labourasse. A cell-centered lagrangian hydrodynamics scheme on general unstructured meshes in arbitrary dimension. *Journal of Computational Physics*, 228:5160–5183, 2009.

[CDV07] Pierre Crispel, Pierre Degond, and Marie-Hélène Vignal. An asymptotic preserving scheme for the two-fluid Euler–Poisson model in the quasineutral limit. *Journal of Computational Physics*, 223(1):208–234, 2007.

[CDW99] Phillip Colella, Milo R Dorr, and Daniel D Wake. A conservative finite difference method for the numerical solution of plasma fluid equations. *Journal of Computational Physics*, 149(1):168–193, 1999.

[CGHS02] F. Coquel, T. Gallouët, J.-M. Hérard, and N. Seguin. Closure laws for a two-fluid two-pressure model. *C. R. Acad. Sci. Paris*, Ser. I 334:927–932, 2002.

[Cha50] S. Chandrasekhar. Radiative transfer. (International Series of Monographs on Physics) Oxford: Clarendon Press; London: Oxford University Press. XIV, 394 p. (1950)., 1950.

[CHSN13] F. Coquel, J.-M. Hérard, K. Saleh, and Seguin N. A robust entropy-satisfying finite volume scheme for the isentropic Baer-Nunziato model. *ESAIM: M2AN*, 48(1):165 – 206, January-February 2013.

[CKS99] Bernardo Cockburn, George E. Karniadakis, and Chi-Wang Shu. The development of discontinuous galerkin methods, 1999.

[CL95] P. Cargo and A. Leroux. Un schéma équilibre adapté au modèle d'atmosphère avec termes de gravité. *Comptes rendus de l'Académie des sciences*, 318:73–76, 1995.

[CLO08] D.A. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Undergraduate Texts in Mathematics. Springer New York, 2008.

[CS05] B. Cheng and A. J. Scannapieco. Buoyancy-drag mix model obtained by multifluid interpenetration equations. *Physics Letters E*, 72:1–5, 2005.

[CS12] C.H. Chang and A.K. Stagg. A compatible Lagrangian hydrodynamic scheme for multicomponent flows with mixing. *J. Comput. Phys.*, 231(11):4279 – 4294, 2012.

[CZ97] Olgierd C. Zienkiewicz. Trefftz type approximation and the generalized finite element method - history and development. 4, 01 1997.

[DB16] Bruno Després and Christophe Buet. The structure of well-balanced schemes for Friedrichs systems with linear relaxation. *Applied Mathematics and Computation*, 272:440–459, 2016.

[Des01] B. Després. Lagrangian systems of conservation laws. *Numer. Math.*, 89(1):99–134, July 2001.

[Des10a] B. Després. *Lois de Conservations Eulériennes, Lagrangiennes et Méthodes Numériques*. Number 68 in Math. Appl. Springer, Berlin, 2010.

[Des10b] B. Després. Weak consistency of the cell-centered Lagrangian GLACE scheme on general meshes in any dimension. *Computer Methods in Applied Mechanics and Engineering*, 199:2669–2679, 2010.

[DF15] Vít Dolejší and Miloslav Feistauer. *Discontinuous Galerkin method. Analysis and applications to compressible flow.* Cham: Springer, 2015.

[DM05] B. Després and C. Mazeran. Lagrangian gas dynamics in two dimensions and Lagrangian systems. *Arch. Rational Mech. Anal.*, 178:327–372, 2005.

[DMP98] A. Decoster, P. A. Markowich, and B. Perthame. *Modeling of collisions*. Gauthier-Villars, 1998.

[DPE11] D.A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Mathématiques et Applications. Springer Berlin Heidelberg, 2011.

[DX13] Feng Dai and Yuan Xu. *Approximation theory and harmonic analysis on spheres and balls.* New York, NY: Springer, 2013.

[EG06] Alexandre Ern and Jean-Luc Guermond. Discontinuous galerkin methods for friedrichs' systems. i. general theory. *SIAM J. Numerical Analysis*, 44(2):753–778, 2006.

[Ena07] C. Enaux. *Analyse mathématique et numérique d'un modèle multifluide multivitesse pour l'interpénétration de fluides miscibles*. PhD thesis, École Centrale de Paris, 2007.

[FB10] Thomas-Peter Fries and Ted Belytschko. The extended/generalized finite element method: An overview of the method and its applications. *International Journal for Numerical Methods in Engineering*, 84(3):253–304, 2010.

[FHK16] Martin Frank, Cory Hauck, and Kerstin Küpper. Convergence of filtered spherical harmonic equations for radiation transport. *Commun. Math. Sci.*, 14(5):1443–1465, 2016.

[Fle77] Wendell Fleming. Functions of several variables. 2nd ed. Undergraduate Texts in Mathematics. New York - Heidelberg - Berlin: Springer-Verlag. XI, 411 p. DM 41.00; $ 18.10 (1977)., 1977.

[FM16] E. Franck and L. S. Mendoza. Finite volume scheme with local high order discretization of the hydrostatic equilibrium for the euler equations with external forces. *Journal of Scientific Computing*, pages 1–41, 2016.

[FR00] Richard S. Falk and Gerard R. Richter. Explicit finite element methods for linear hyperbolic systems. In Bernardo Cockburn, George E. Karniadakis, and Chi-Wang Shu, editors, *Discontinuous Galerkin Methods*, pages 209–219, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[Fra12] Emmanuel Franck. *Design and numerical analysis of asymptotic preserving schemes on unstructured meshes. Application to the linear transport and Friedrichs systems.* Theses, Université Pierre et Marie Curie - Paris VI, October 2012.

[Fra14] E. Franck. Modified Finite Volume Nodal Scheme for Euler Equations with Gravity and Friction. In *Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects*, volume 77 of *Springer Proceedings in Mathematics & Statistics*, pages 285–292. Springer International Publishing, 2014.

[Fri58] K. O. Friedrichs. Symmetric positive linear differential equations. *Communications on Pure and Applied Mathematics*, 11(3):333–418, 1958.

[GA07] Pablo Gamallo and R.J. Astley. A comparison of two trefftz-type methods: The ultraweak variational formulation and the least-squares method, for solving shortwave 2-d helmholtz problems. 71:406 – 432, 07 2007.

[Gab06] G. Gabard. Discontinuous galerkin methods with plane waves for the displacement-based acoustic equation. *International Journal for Numerical Methods in Engineering*, 66:549–569, 2006.

[Gab07] G. Gabard. Discontinuous galerkin methods with plane waves for time-harmonic problems. *Journal of Computational Physics*, 225(2):1961–1984, August 2007.

[Gab09] G. Gabard. Exact integration of polynomial-exponential products with application to wave-based numerical methods. *Communications in Numerical Methods in Engineering*, 25(3):237–246, 2009.

[Gal03] G. Gallice. Positive and Entropy Stable Godunov-type Schemes for Gas Dynamics and MHD Equations in Lagrangian or Eulerian Coordinates. *Numerische Mathematik*, 94(4):673–713, 2003.

[GH16] C.Kristopher Garrett and Cory D. Hauck. On the eigenstructure of spherical harmonic equations for radiative transport. *Comput. Math. Appl.*, 72(2):264–270, 2016.

[GHP09] Claude J. Gittelson, Ralf Hiptmair, and Ilaria Perugia. Plane wave discontinuous Galerkin methods: Analysis of the *h*-version. *ESAIM, Math. Model. Numer. Anal.*, 43(2):297–331, 2009.

[GL96] J. M. Greenberg and A. Y. Leroux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM Journal on Numerical Analysis*, 33(1):1–16, 1996.

[Gos13] Laurent Gosse. *Computing qualitatively correct approximations of balance laws. Exponential-fit, well-balanced and asymptotic-preserving.* Milano: Springer, 2013.

[GS00] Michael Griebel and Marc Alexander Schweitzer. A particle-partition of unity method for the solution of elliptic, parabolic, and hyperbolic pdes. *SIAM Journal on Scientific Computing*, 22(3):853–890, 2000.

[GT02] Laurent Gosse and Giuseppe Toscani. An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations. *C. R., Math., Acad. Sci. Paris*, 334(4):337–342, 2002.

[HBH+03] Michael Heroux, Roscoe Bartlett, Vicki Howle Robert Hoekstra, Jonathan Hu, Tamara Kolda, Richard Lehoucq, Kevin Long, Roger Pawlowski, Eric Phipps, Andrew Salinger, Heidi Thornquist, Ray Tuminaro, James Willenbring, and Alan Williams. An Overview of Trilinos. Technical Report SAND2003-2927, Sandia National Laboratories, 2003.

[Her16] F. Hermeline. A discretization of the multigroup P_N radiative transfer equation on general meshes. *J. Comput. Phys.*, 313:549–582, 2016.

[HM10] Cory Hauck and Ryan McClarren. Positive P_N closures. *SIAM J. Sci. Comput.*, 32(5):2603–2626, 2010.

[HMK02] Tomi Huttunen, Peter Monk, and Jari P. Kaipio. Computational aspects of the ultra-weak variational formulation. *J. Comput. Phys.*, 182(1):27–46, 2002.

[HMM07] T. Huttunen, M. Malinen, and P. Monk. Solving maxwell's equations using the ultra weak variational formulation. *Journal of Computational Physics*, 223(2):731 – 758, 2007.

[HMP11] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version. *SIAM J. Numer. Anal.*, 49(1):264–284, 2011.

[HMP16a] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods: exponential convergence of the hp-version. *Found. Comput. Math.*, 16(3):637–675, 2016.

[HMP16b] Ralf Hiptmair, Andrea Moiola, and Ilaria Perugia. A survey of trefftz methods for the helmholtz equation. In *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, volume 114, pages 237–278. Springer, 2016.

[HW07] Jan S. Hesthaven and Tim Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2007.

[IG13] Lise-Marie Imbert-Gérard. *Mathematical and numerical problems of some wave phenomena appearing in magnetic plasmas*. Theses, Université Pierre et Marie Curie - Paris VI, September 2013.

[IG15a] Lise-Marie Imbert-Gérard. Interpolation properties of generalized plane waves. *Numer. Math.*, 131(4):683–711, 2015.

[IG15b] Lise-Marie Imbert-Gérard. Well-posedness and generalized plane waves simulations of a 2D mode conversion model. *J. Comput. Phys.*, 303:105–124, 2015.

[IGD14] Lise-Marie Imbert-Gérard and Bruno Després. A generalized plane-wave numerical method for smooth nonconstant coefficients. *IMA J. Numer. Anal.*, 34(3):1072–1103, 2014.

[IR96] J. Ivanic and K. Ruedenberg. Rotation matrices for real spherical harmonics. direct determination by recursion. *Journal of Physical Chemistry*, 100(15), Apr 1996.

[Jin04] Shi Jin. A steady-state capturing method for hyperbolic systems with geometrical source terms. In Naoufel Ben Abdallah, Irene M. Gamba, Christian Ringhofer, Anton Arnold, Robert T. Glassey, Pierre Degond, and C. David Levermore, editors, *Transport in Transition Regimes*, pages 177–183, New York, NY, 2004. Springer New York.

[Jin10] S. Jin. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. In *Lecture notes for summer school on methods and models of kinetic theory (M&MKT)*, pages 177–216. Porto Ercole (Grosseto, Italy), 2010.

[JL91] Shi Jin and David Levermore. The discrete-ordinate method in diffusive regimes. 20:413–439, 10 1991.

[JL96] Shi Jin and C.David Levermore. Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. *J. Comput. Phys.*, 126(2):449–467, art. no. 0149, 1996.

[JTH09] Shi Jin, Min Tang, and Houde Han. A uniformly second order numerical method for the one-dimensional discrete-ordinate transport equation and its diffusion limit with interface. *Netw. Heterog. Media*, 4(1):35–65, 2009.

[KK95] Eisuke Kita and Norio Kamiya. Trefftz method: an overview. *Advances in Engineering Software*, 24(1):3 – 12, 1995.

[Klu08] G. Kluth. *Analyse mathématique et numérique de systèmes hyperélastiques et introduction de la plasticité*. PhD thesis, Université Paris VI, 2008.

[KMPS16] Fritz Kretzschmar, Andrea Moiola, Ilaria Perugia, and Sascha M. Schnepp. A priori error analysis of space–time trefftz discontinuous galerkin methods for wave problems. *IMA Journal of Numerical Analysis*, 36(4):1599, 2016.

[LeV98] Randall J. LeVeque. Balancing source terms and flux gradients in high-resolution godunov methods: The quasi-steady wave-propagation algorithm. *Journal of Computational Physics*, 146(1):346 – 365, 1998.

[Li08] Z.C. Li. *Trefftz and Collocation Methods*. WIT Press, 2008.

[LMM87] Edward W Larsen, J.E Morel, and Warren F Miller. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *Journal of Computational Physics*, 69(2):283 – 324, 1987.

[Lou05] R. Loubère. Validation test case suite for compressible hydrodynamics computation. Technical report, Los Alamos National Laboratory, 2005.

[LR74] P. Lesaint and P. A. Raviart. On a finite element method for solving the neutron transport equation. *Publications mathématiques et informatique de Rennes*, (S4):1–40, 1974.

[LRv95] R. Lowrie, P. Roe, and B. van Leer. A space-time discontinuous Galerkin method for the time-accurate numerical solution of hyperbolic conservation laws. 1995.

[Luo13] T. Luostari. *Non polynomial approximation methods in accoustics and elasticity*. PhD thesis, University of Eastern Finland, 2013.

[MABO07] P.-H. Maire, R. Abgrall, J. Breil, and J Ovadia. A Cell-Centered Lagrangian Scheme for Two-Dimensional Compressible Flow Problems. *SIAM J. Sci. Comput.*, pages 1781–1824, 2007.

[Mai11] P.-H. Maire. *Contribution à la modélisation numérique de la Fusion par Confinement Inertiel*. Habilitation à diriger des recherches, Université de Bordeaux, 2011.

[Mau03] Edward A.W. Maunder. Trefftz in translation. *Comput. Assist. Mech. Eng. Sci.*, 10(4):545–563, 2003.

[Maz07] C. Mazeran. *Sur la structure mathématique et l'approximation numérique de l'hydrodynamique lagrangienne bidimensionnelle*. PhD thesis, Université de Bordeaux, 2007.

[MB96] J.M. Melenk and Ivo Babuška. The partition of unity finite element method: Basic theory and applications. *Comput. Methods Appl. Mech. Eng.*, 139(1-4):289–314, 1996.

[MDB99] Nicolas Moës, John Dolbow, and Ted Belytschko. A finite element method for crack growth without remeshing. *International Journal for Numerical Methods in Engineering*, 46(1):131–150, 1999.

[MDBCF16] Victor Michel-Dansac, Christophe Berthon, Stéphane Clain, and Françoise Foucher. A well-balanced scheme for the shallow-water equations with topography. *Computers & Mathematics with Applications*, 72(3):568 – 593, 2016.

[MH10] Ryan G. McClarren and Cory D. Hauck. Robust and accurate filtered spherical harmonics expansions for radiative transfer. *J. Comput. Phys.*, 229(16):5597–5614, 2010.

[Moi11] Andrea Moiola. *Trefftz-discontinuous Galerkin methods for time-harmonic wave problems.* PhD thesis, ETH Zürich, 2011.

[MR05] Peter Monk and Gerard R. Richter. A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media. *J. Sci. Comput.*, 22-23:443–477, 2005.

[Ols12] Gordon L. Olson. Alternate closures for radiation transport using Legendre polynomials in 1D and spherical harmonics in 2D. *J. Comput. Phys.*, 231(7):2786–2793, 2012.

[PH07] Didier Pinchon and Philip E. Hoggan. Rotation matrices for real spherical harmonics: general rotations of atomic orbitals in space-fixed axes. *J. Phys. A, Math. Theor.*, 40(7):1597–1610, 2007.

[PLM18] S. Del Pino, E. Labourasse, and G. Morel. An asymptotic preserving multidimensional ale method for a system of two compressible flows coupled with friction. *Journal of Computational Physics*, 363:268 – 301, 2018.

[PvHVD07] B. Pluymers, B. van Hal, D. Vandepitte, and W. Desmet. Trefftz-based methods for time-harmonic acoustics. *Archives of Computational Methods in Engineering*, 14(4):343–381, Dec 2007.

[Qin05] Q.H. Qin. Trefftz finite element method and its applications. *Appl. Mech.Rev*, 58:316 – 337, 2005.

[RARO13] David Radice, Ernazar Abdikamalov, Luciano Rezzolla, and Christian D. Ott. A new spherical harmonics scheme for multi-dimensional radiation transport. I: Static matter configurations. *J. Comput. Phys.*, 242:648–669, 2013.

[RGK12] J.C. Ragusa, J.-L. Guermond, and G. Kanschat. A robust S_N-DG-approximation for radiation transport in optically thick and diffusive regimes. *J. Comput. Phys.*, 231(4):1947–1962, 2012.

[RH73] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. 1973.

[SA99] R. Saurel and R. Abgrall. A multiphase godunov method for compressible multifluid and multiphase flows. *Journal of Computational Physics*, 150(2):425 – 467, 1999.

[SBC00] T. Strouboulis, I. Babuška, and K. Copps. The design and analysis of the generalized finite element method. *Computer Methods in Applied Mechanics and Engineering*, 181(1):43 – 69, 2000.

[SC02] A. J. Scannapieco and B. Cheng. A multifluid interpenetration mix model. *Physics Letters A*, 299:49–64, 2002.

[SCB00] T. Strouboulis, K. Copps, and I. Babuška. The generalized finite element method: an example of its implementation and illustration of its performance. *International Journal for Numerical Methods in Engineering*, 47(8):1401–1417, 2000.

[Sen14] Rémi Sentis. *Mathematical Models and Methods for Plasma Physics, Volume 1: Fluid Models.* Springer Science & Business Media, 2014.

[SFL11] Matthias Schäfer, Martin Frank, and C.David Levermore. Diffusive corrections to P_N approximations. *Multiscale Model. Simul.*, 9(1):1–28, 2011.

[SO96] Bingjing Su and Gordon L. Olson. Benchmark results for the non-equilibrium marshak diffusion problem. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 56(3):337 – 351, 1996.

[Tan09] Min Tang. A uniform first-order method for the discrete ordinate transport equation with interfaces in X,Y-geometry. *J. Comput. Math.*, 27(6):764–786, 2009.