Thesis presented to obtain the grade of Doctor at the

Université de Bordeaux

and the

RWTH Aachen University

by **Teddy Pichard**

Specialty: Applied mathematics and scientific computing

Mathematical Modelling for dose deposition in photontherapy

Date of the defense: 4th November 2016

Before the jury composed of:

Benoît PERTHAME	Professor, Univ. Paris 6	Reviewer
Cory HAUCK	Researcher, ORNL	Reviewer
Bruno DESPRÉS	Professor, Univ. Paris 6	
Michael HERTY	Professor, RWTH Aachen	
Axel KLAR	Professor, TU Kaiserslautern	
Bruno DUBROCA	Expert-Senior, CEA	Advisor
Martin FRANK	Professor, RWTH Aachen	Advisor
Denise AREGBA-DRIOLLET	Maître de conférence, INP Bordeaux	
Stéphane BRULL	Maître de conférence, INP Bordeaux	Advisor

Abstract Radiation therapy is one of the most common types of cancer treatments. It consists in irradiating the patient with beams of energetic particles (typically photons). Such particles are transported through the medium and interact with it. Especially, during such interactions, a part of the energy of the transported is deposited in the medium, this is the so-called dose, responsible for the biological effect of the radiation.

The aim of the present thesis is to develop numerical method for dose computation that are competitive in terms of computational cost and accuracy compared to reference method such as the statistical Monte Carlo methods or the empirical superposition-convolution methods.

The motion of such particles is studied through a system of linear transport equations at the kinetic level with a special consideration on the conservation of mass, momentum and energy.

Computational costs required to solve directly such systems is typically higher than available in medical center. In order to reduce those costs, the moment method is used. This consists in averaging the transport equation over one of the variables. However, such a method leads to a system of equations with more unknowns than equation. An entropy minimization procedure is used to close this system, leading to the so-called M_N models. The moments extraction preserves the major properties of the kinetic system such as hyperbolicity, entropy decay and realizability (existence of a positive solution). However, computing numerically the M_N closure may also be computationally costly for the application in medical physics, and furthermore it is valid only under condition, called realizability condition, on the unknowns. The realizability domain, *i.e.* the domain of validity of the M_N model, is studied. Based on these results, approximations of the first order entropy-based closures, *i.e.* the M_1 and the M_2 closures, are developped for 3D problems which require lower computational costs to compute.

The resulting moment equation are non-linear and valid under realizability condition. Standard numerical schemes for moment equations are constrained by stability conditions which happen to be very restrictive when the medium contains low density regions. Numerical approaches adapted to moment equations are developped. The non-linearity is treated by using a relaxation method originally developped for hyperbolic systems of equations. Then inconditionally stable schemes are proposed to treat the problem of restrictive stability conditions. A first explicit scheme based on the method of characteristics is proposed for hyperbolic equations. A second numerical scheme with implicit non-linear flux terms is proposed. Those schemes preserve the realizability property and they are competitive in terms of computational costs compared to reference approaches.

Résumé La radiothérapie est l'une des familles de traitements de cancer les plus utilisés. Ces traitements consistent en l'irradiation du patient par des faisceaux de particules énergétiques (le plus couramment des photons). Ces particules sont transportées à travers le milieu et interagissent avec. En particulier, à travers ces interactions, une partie de l'énergie de ces particules est déposée dans le milieu, cette énergie déposée est appelée la dose, et les effets biologiques des radiations sont souvent considérés comme une conséquence directe de l'énergie déposée.

L'objectif de la présente thèse est de développer des méthodes numériques pour le calcul de dose qui sont compétitives en terme de coûts numériques et précision comparé à des méthodes de référence comme les méthodes statistiques Monte-Carlo ou les méthodes empiriques de superposition-convolution.

Le transport de photons et d'électrons est étudié à travers un système d'équations transport linéaire au niveau cinétique en prenant en considération des propriétés de conservations de masse, de quantité de mouvement et d'énergie.

Le coût numérique pour résoudre directement ces équations cinétiques est généralement trop élevé pour être utilisé dans les centres médicaux. Afin de réduire ces coûts numériques, la méthode aux moments est utilisée. Cette méthode consiste à moyenner les équations de transport par rapport à l'une des variables. Cependant cette méthode mène à un système d'équations avec plus d'inconnues que d'équations. Une procédure de minimisation d'entropie est utilisée pour fermer ce système, menant aux modèles M_N. L'extraction de moments préserve les principales propriétés du système cinétique sous-jacent, notamment l'hyperbolicité, la dissipation d'entropie et la réalisabilité (existence d'une solution positive). Cependant, calculer numériquement la fermeture M_N peut également être coûteuse au niveau numérique. De plus, cette fermeture n'est valide que si l'inconnue satisfait la condition dite de réalisabilité. Dans un premier temps, le domaine de validité des modèles M_N est étudié, puis, à partir de ces résultats, des approximations des fermetures entropiques d'ordre un et deux, à savoir la fermeture M_1 et M_2, sont développées pour des problèmes 3D et dont le calcul nécessite un faible coût numérique.

Les équations aux moments obtenues sont non-linéaires et valide sous condition de réalisabilité. Les méthodes numériques standards pour ces équations sont conxtraintes par des conditions de stabilité qui, en pratique, sont très restrictives lorsque le milieu contient des zones sous denses. Des approches numériques adaptées aux équations aux moments sont développées. La non-linéarité est traitée en utilisant une méthode de relaxation, à l'origine développée pour les systèmes d'équations hyperboliques. Ensuite des schémas numériques inconditionnellement stables sont proposés pour pallier le problème des conditions de stabilité restrictives. Un premier schéma explicite basé sur la méthode des caractéristiques est proposé. Un second schéma avec des termes de flux non linéaires implicites est proposé. Ces schémas préservent

la propriété de réalisabilité et sont compétitifs en terme de coûts de calcul comparé à des méthodes de référence.

Inhaltsangabe Strahlentherapie ist eine der häufigsten Arten von Krebsbehandlungen. Das Verfahren besteht darin, den Patienten mit Strahlen energetischer Pertikel (typischerweise Photonen) zu bestrahlen. Solche Partikel propagieren durch das Medium und interagieren damit. Insbesondere wird bei solchen Interaktion ein Teil der Energie der Partikel im Medium hinterlegt, dies ist die sogenannte Dosis, die für die biologische Wirkung der Strahlung verantwortlich ist.

Ziel der vorliegenden Arbeit ist es, eine numerische Methode zur Dosisberechnung zu entwickeln, die im Vergleich zu Referenzmethoden wie die statistischen Monte-Carlo-Methoden oder die empirischen Superposition-Convolution-Methoden, hinsichtlich Rechenkosten und Genauigkeit konkurrenzfähig ist.

Diese Partikelbewegung wird durch ein lineares Transportgleichungensystem auf kinetischer Ebene untersucht, mit besondere Aufmerksamkeit auf die Bewahrung von Masse, Impuls und Energie.

Die erforderliche Rechenleistung, um ein solches System direkt zu lösen, ist in der Regel höher als die Rechenleistung, die in medizinischen Zentren verfügbar sind. Um diese Kosten zu senken, wird die Momentmethode benutzt. Diese besteht darin, die Transportgleichung über eine der Variablen zu mitteln. Ein solches Verfahren führt jedoch zu einem Gleichungssystem mit mehr Unbekannten als Gleichungen. Ein Entropie-Minimierungsverfahren wird verwendet, um dieses System zu schließen, was zu den sogenannten M_N-Modellen führt. Die Momentextraktion bewahrt die Haupteigenschaften des kinetische System wie Hyperbolizität, Entropiezerfall und Realisierbarkeit (Existenz einer positiven Lösung). Allerdings kann die numerische Berechnung der M_N Schließung auch zu kostspielig für die Anwendung in der medizinischen Physik sein. Außerdem ist es gültig nur unter Bedingung auf die Unbekannten, genannt Realisierbarkeitbedingung. Die Realisierbarkeitsdomäne, d.h. die Gültigkeitsdomäne der M_N-Modellen, wird untersucht. Basierend auf diesen Ergebnissen werden Approximationen der Entropie-basierten Schließungen für den ersten Ordnungen, d.h. die M_1 und die M_2 Schließungen für 3D-Probleme entwickelt, die niedrieger Berechnungskosten erfordern im Vergleich zu Minimierungverfaren Algorithmus.

Die resultierende Momentengleichungen sind nichtlinear und gültig unter Realisierbarkeitsbedingunge. Standard numerische Schemata für Momentgleichungen sind durch Stabilitätsbedingungen eingeschränkt, die sehr restriktiv sind, wenn das Medium Bereiche mit geringe Dichte enthält. Numerische Methoden angepasst für Momentgleichungen werden entwickelt. Die Nichtlinearität ist durch ein Relaxationsverfahren behandelt, das ursprünglich für hyperbolische Gleichungssysteme entwickelt war. Darüber hinaus werden bedingungslose Schemata entwickelt, um das Problem der restriktiven Stabilitätsbedingungen zu behandeln. Für hyperbolische Gleichungen wird ein erstes explizites Schema entwickelt, das auf der Charakteristikmethode basiert ist. Ein zweites numerisches Schemata mit impliziten nichtlinearen Flusstermen wird entwickelt. Diese Methoden bewahren die Realisierbarkeitseigenschaft und sind hinsichtlich der Rechenkosten gegenüber Referenzmethode wettbewerbsfähig.

Contents

Acknowledgement

Firstly, I would like to thank Benoît Perthame for the interesting discussion about the radiobiology perspectives of this work and Cory Hauck for the discussions about the moment models. Also I would like to thank them for accepting to review the present thesis.

I also thank Bruno Després for our exchange on hyperbolic equations and, together with Axel Klar and Michael Herty, for accepting to be part of the jury (even with such a short notice for Michael Herty). I would like to express my gratitude to Denise Aregba-Driollet for her valuable and always precise comments about relaxation methods which lead to the developpement of the numerical schemes in Part III.

Many thanks to my three advisors for their patience with me and my writings and for their involvement in this work: Bruno Dubroca who was always present and opened to scientific discussions, Martin Frank for his welcomeness in Aachen, his scientific contributions and for offering me the opportunity of teaching in Aachen, and Stéphane Brull who pushed me to attend to conferences and to meet the scientific community.

This work is also the result of a collaboration with the radiotherapy group (subgroup of the fusion group) in CELIA, Gabriele Birindelli, Jerome Caron, Bruno Dubroca, Jean-Luc Feugeas, Philippe Nicolai, Jonathan Page; and the MathCCES gruppe in Aachen, especially Graham Alldredge whose contribution lead to writing the paper on the approximated M_2 closure, Martin Frank and Kerstin Küpper. I acknowledge Kerstin Küpper, Gabriele Birindelli, Jerome Caron and Jonathan Page for performing the Monte Carlo simulations which were used as reference results for the 2D simulations in Part III.

I would like to thank Ed Larsen for the discussions during his stay in Aachen and at the conference in Nashville about Acuros code and the expansion in order of scattering which I understood the importance lately. I also thank Oscar Lopez Pouso for his interest in my work and the interesting discussion about the radiation therapy and kinetic equations in general during the ICTT in Taormina.

Last but not least, I would like to thank the two groups in CELIA and in MathCCES, and all my friends and family for their support during this intensive three year period.

General introduction

Together with surgery and chemotherapy, radiation therapy is one of the most common types of cancer treatment.

Definition 0.1 *[10]* ***Radiation therapy:*** *The use of high-energy radiation from X-rays, gamma rays, neutrons, protons, and other sources to kill cancer cells and shrink tumors. Radiations may come from a machine outside the body (external-beam radiation therapy), or it may come from radioactive material placed in the body near cancer cells (internal radiation therapy or brachytherapy). Also called irradiation and radiotherapy.*

Radiations can be seen as beams of energetic particles travelling through a medium and interacting with it. Therefore particle transport models are the basis of the numerical methods applied in this field. As a first approach, one often considers that the biological impact of the radiations on the cells is a function of the energy deposited by the radiations (or equivalently the particles) in the medium. This deposited energy is called the dose.

In the last decades, the recent advances and the developement of new techniques in the field of radiation therapy lead to an enhanced need for new algorithms and numerical methods for dose computation. These treatment improvements consist either in better adapting the treatments to each patient or in better controling the dose delivered. For instance, the image guided radiotherapy (IGRT, see *e.g.* [17]) is an online adapted radiation therapy (ART, see *e.g.* [15]) technique in development which purpose is to take into account potential movements of the patient in the dose computations; or the intensity modulated radiation therapy (IMRT, see *e.g.* [5, 3]) which consists in adapting the radiations by modulating the intensity of the source.

Such sophisticated techniques require adapted numerical methods and classical numerical techniques may be inappropriate for those emerging problems. Especially such techniques typically require fast and accurate methods of dose computation in order to adapt and optimize previously computed treatment plannings. Most of the numerical approaches available present drawbacks, the majors of which being:

- The high computational costs. As described below, certain methods typically require more computational power than available in medical centers.

- The inaccuracy for particular applications. The results obtained with those method can be trusted only for some applications for which they are known to be accurate.

Existing numerical methods

Common numerical techniques for the computations of dose deposition and/or used in the field of transport theory are listed below.

Superposition-convolution method

The superposition method (see *e.g.* [18]) is an empirical method based on the Fermi-Eyges theory ([7]). The primary (never scattered) and the secondary particles are considered seperately.

The dose is defined as a convolution of a primary particles energy fluence, *i.e.* the quantity of energy travelling in the medium from a source, and a kernel modeling the quantity of energy deposited per unit energy fluence (see *e.g.* [9]).

Such algorithms are very fast and are accurate in homogeneous weakly collisional media, *e.g.* when modelling photon beams in water. However they are unadapted for dose computations in media containing strong heterogeneities, when the deflection of the particles have an non-negligible impact on the dose distribution.

Discrete ordinate methods

The discrete ordinate methods (see *e.g.* [14]) are deterministic methods based on a kinetic model (as the one presented in the next chapter). They consist in discretizing directly such an equation in all the variables. However the dimension of those variables is relatively high, *e.g.* the variables in the models presented in Chapter 1 evolve in a six dimensional space (three of space, one of energy and two of direction). Discretizing directly such high dimensional equations requires a consequent amount of data storage, and are also generally time consuming.

Monte Carlo algorithms

The Monte Carlo solvers are probabilistic algorithms. The radiations are seen as particles transported through a medium. The result is averaged over a large number of samples. In practice, a particle is injected on the boundary of the medium and its motion is computed (or approximated) based on the physics of the modeled interactions. This operation is repeated a large number of times and the final result is averaged over all these samples. The length of the free flight and the effect of the interactions are governed by random processes. Such random processes introduce noise in the results. They are used in this manuscript to obtain reference results as they present a good accuracy for the present applications. However, similarily to discrete ordinate methods, they typically require a computational time too long for medical applications. Furthermore, they are not adapted to numerical optimization and can therefore not be used for some of the applications presented in this manuscript. The reader is referred to [11, 19, 1, 2, 20] for applications of Monte Carlo methods in medical physics and to [12, 4] for further applications.

Recent advances lead to constructing more time-efficient probabilistic methods referred to as fast Monte Carlo algorithms (see *e.g.* [22] and references therein) and/or to reduce the noise in the Monte Carlo results, although those methods are not as accurate as the original Monte Carlo methods.

An alternative deterministic approach

A recent alternative uses similar techniques as the approach developed in this manuscript. The Acuros® ([21]) code is a deterministic method based on the decomposition in order of scattering (see *e.g* [13, 6, 8]). It uses the method of moments to solve the transport equations of multiply-scattered particles.

This method was originally developed for applications in neutrons transport ([16]) and was afterward adapted for medical applications. This code is already used in medical center and is shown to be time-efficient and accurate.

The present approach

The present approach is based on the physics of the transport and the collisions of the considered particles. As for the discrete ordinate methods, it is based on a kinetic description of this physics and it uses deterministic numerical methods.

However, in order to reduce the computational costs due to the high-dimensionality of the kinetic equations, the method of moments, *i.e.* a model reduction technique, is used. Solving moment equations requires considerably less computational power than solving a kinetic equation, although several difficulties emerge when using this method. The moment equations have more unknowns than equations, so they require a closure, *i.e.* an additional equation such that the number of equations equals the number of unknowns. For this problem, a closure based on a entropy minimization procedure is chosen, *i.e.* the M_N closure. This choice presents desirable properties both on the mathematical and physical level. However, the M_N closure is defined under condition afterward called realizability, which needs to be carefully taken into account when developing numerical schemes for moment equations.

This manuscript is organized as follow: in the Part I, *i.e.* Chapter 1, the physics of the transport and collisions of photons and electrons in the field of radiotherapy is presented. In Part II, the method of moments is presented: First in Chapter 2 the moment equations are computed. Then in Chapter 3, the realizability condition linking the kinetic model to the angular moment models is studied. Finally, the constructions of some closures are described in Chapter 4. In Part III, numerical methods based on the moment equations are presented. Numerical schemes adapted to M_N models are developed in Chapter 5. Chapter 6 deals with a numerical approach for dose optimization.

Bibliography

[1] J. Barò, J. Sempau, J.M. Fernández-Varea, and F. Salvat. PENELOPE: An algorithm for Monte Carlo simulation of the penetration and energy loss of electrons in matter. *Nuclear instruments and methods*, 100:31–46, 1995.

[2] J. Barò, J. Sempau, J.M. Fernández-Varea, and F. Salvat. PENELOPE: An algorithm and computer code for Monte Carlo simulation of electron-photon shower. *Ciemat technical report*, pages 31–46, 1996.

[3] Memorial Sloan-Kettering Cancer Center. *A practical guide to intensity-modulated radiation therapy.* Medical Physics Publishing, 2003.

[4] CERN. *Geant4 User's Guide for Application Developers*, 2015.

[5] International commission on radiation units and measurements. Prescribing, recording, and reporting photon-beam intensity-modulated radiationtherapy (IMRT). Technical Report 1, 2010.

[6] T. K. Das and Ó. López Pouso. New insights into the numerical solution of the Boltzmann transport equation for photons. *Kin. Rel. Mod.*, 7(3):433–461, 2014.

[7] L. Eyges. Multiple scattering with energy loss. *Phys. Rev.*, 74:1534–1535, Nov 1948.

[8] H. Hensel, R. Iza-Teran, and N. Siedow. Deterministic model for dose calculation in photon radiotherapy. *Phys. Med. Biol.*, 51:675–693, 2006.

[9] K. R. Hogstrom, M. D. Mills, and P. R. Almond. Electron beam dose calculations. *Phys. Med. Biol.*, 26(3):445–459, 1981.

[10] National Cancer Institute. NCI dictionary of cancer terms.

[11] I. Kawrakow, E. Mainegra-Hing, D. W. O. Rogers, F. Tessier, and B. R. B. Walters. *The EGSnrc Code System: Monte Carlo Simulation of Electron and Photon Transport*, 2013.

[12] Los Alamos National Laboratory. *MCNP - A general Monte Carlo N-particle transport code, Version 5*, 2003.

[13] E. W. Larsen. Solution of neutron transport problems in L_1^*. *Commun. Pur. Appl. Math.*, 28(6):729–746, 1975.

[14] E. E. Lewis and W. F. Miller. *Computational methods of neutron transport.* American nuclear society, 1993.

[15] X. Allen Li, editor. *Adaptative radiation therapy.* CRC press, 2011.

[16] D. S. Lucas, H. D. Gougar, T. Wareing, G. Failla, J. McGhee, D. A. Barnett, and I. Davis. Comparison of the 3-D deterministic neutron transport code Attila® to measure data, MCNP and MCNPX for the advanced test reactor. Technical report, Idaho National Laboratory, 2005.

[17] T. R. Mackie, J. Kapatoes, K. Ruchala, G. Olivera W. Lu, C. Wu, L. Forrest, W. Tome, J. Welsh, R. Jeraj, P. Harari, P. Reckwerdt, B. Paliwal, M. Ritter, H. Keller, J. Fowler, and M. Mehta. Image guidance for precise conformal radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.*, 56(1):89–105, 2003.

[18] P. Mayles, A. Nahum, and J.C. Rosenwald, editors. *Handbook of radiotherapy physics: Theory and practice.* Taylor & Francis, 2007.

[19] F. Salvat, J. M. Fernández-Varea, and J. Sempau. *PENELOPE-2011: A code system for Monte Carlo simulation of electron and photon transport*, 2011.

[20] E. Spezi and G. Lewis. An overview of Monte Carlo treatment planning for radiotherapy. *Radiat. Prot. Dos.*, 131(1):123–129, 2008.

[21] T.A. Wareing, J.M. McGhee, Y. Archambault, and S. Thompson. Acuros XB® advanced dose calculation for the Eclipse$^{\text{TM}}$ treatement planning system. *Clinical perspectives*, 2010.

[22] C. Zankowski, M. Laitinen, and H. Neuenschwander. Fast electron Monte Carlo for Eclipse$^{\text{TM}}$. Technical report, Varian Medical System.

Part I

Physical description

Chapter 1

Models

1.1 Introduction

This first chapter is devoted to the mathematical and physical description of the transport of particles in the field of external radiotherapy.

Radiotherapy is a type of cancer treatment based on radiations which can be seen as beams of energetic particles. Most commonly beams of photons are used. But electrons can potentially be used for near-skin cancers. Heavier particles, such as protons or other hadrons, can also be prescribed in particular cases requiring highly accurate treatements, *e.g.* for eye cancers. Those particles are transported through a medium and collide with other particles. The particles involved in such collisions may exchange energy. In particular, a part of the energy of the transported particles may be transfered to particles constituting the medium (atoms or molecules). Modelling this transfer of energy is important as the biological effects of radiations are assumed to be a function of this energy transfered to the medium, so-called dose.

Mathematical models for particle transport are organized in a hierarchy, where each representation models the particles motion at a different scale and retains some properties of the previous scale. The first of those is the molecular description. The movements and the interactions of every particle are modeled. However, such models become difficult to use when considering a very large number of particles. Therefore molecular models are often unadapted to observe mean behaviour, *e.g.* to compute the density or the mean velocity. The second scale in the hierarchy are the kinetic models. For some problems, this type of model can be obtained from a molecular one in a mean field regime, *i.e.* by considering a large number of particles, by using a Bogolioubov-Born-Green-Kirkwood-Yvon (BBGKY, see *e.g.* [6, 8]) hierarchy from the Liouville equation. A third scale consists of angular moment models. They are obtained from a moment extraction, *i.e.* an integration over a direction variable $\Omega \in S^2$, of the kinetic equation and will be studied in the next chapter. One last scale is the hydrodynamic or fluid models. Typically, hydrodynamic models are obtained from a moment extraction, *i.e.* here an integration over a velocity variable $v \in \mathbb{R}^3$, of a kinetic model in a certain regime.

This part is devoted to presenting a state-of-the-art kinetic model (also described *e.g.* in [17, 13, 12, 26]). Under hypothesis, it models the physics of the transport of photons and electrons when considering that the movements of the particles are modified by Compton (see Subsection 1.3.1 or [11, 14, 18]), Mott (see Subsection 1.3.2 or [24]) and Møller effects (see Subsection 1.3.3 or [18, 14]) which are the predominant effects in the range of energy spectra studied in this manuscript. This model can easily be extended by taking into account more types of physical interactions, such as Bremsstrahlung effect (see *e.g.* [21]) or pair production (see *e.g.* [25]), but, as a first approach, only Compton, Mott and Møller effects were considered in order to simplify the notations and computations. This model retains some of the basic properties (conservation of particles, momentum and energy) of an underlying molecular model.

The chapter is organized as follow. First, assumptions arising from physics, and required for the mathematical modeling are presented. Then, some basic properties of the molecular model are recalled, with a special focus on conservation of particles, momentum and energy. In a third part, a kinetic model is presented. This model is based on linear Boltzmann collision terms and some approximations of such a collision term are recalled.

In this chapter, and in the rest of the manuscript, the formulae and equations presented are always non dimensionalized. Especially the energies ϵ are always normalized by the energy $m_e c^2$ of electrons at rest, the momenta p by $m_e c$ and the velocities v by c the celerity of light.

1.2 Assumptions

Before describing the model, some assumptions are made. They arise from the physics of the studied phenomena and are necessary to derive the kinetic model in Section 1.4.

1.2.1 Non-alteration of the medium

The quantity of particles that are transported (photons and electrons) is very low compared to the quantity of atoms composing the medium. This leads to the following two assumptions.

The first assumption is related to the motion of the transported particles.

Assumption 1.1 *The collisions involving two transported particles are assumed to be so rare that they have a negligible impact on the macroscopic motion.*
Consequence: *The collisions involving two or more transported particles are neglected. Only the collisions involving a transported particles and a particle from the background medium are considered.*

The second assumption is related to the medium itself.

Assumption 1.2 *The effect of those collisions on the medium is assumed to be negligible. It is not alterate.*
Consequence: *The transported particles have no effect on the background*

medium.

Those hypothesis will lead to consider linear models at the kinetic level, in Section 1.4.

1.2.2 Time independency

The transported particles (photons γ and electrons e) are relativistic in the sense that they have a non-negligible velocity compared to the speed of light in vacuum c. We assume the following.

Assumption 1.3 *The velocity of the medium is assumed to be negligible compared to the speed of light in vacuum c.*
Consequence: *The particles of the medium are assumed to be fixed, and the medium can be represented by a distribution of particles that is assumed to be a given data.*

At this point, one does not need to study the motion of the medium, but only the motion of electrons and photons through the medium.

Assumption 1.4 *The flux of injected particles in the domain is assumed to be constant. Furthermore, the time required for the flow of transported particles to reach a steady state is assumed to be negligible compared to the time of irradiation. Therefore this flow is alway assumed to be at steady state.*
Consequence: *In this description, the time appears only as a parameter. The transport model of electrons and photons can be assumed to be steady and the time is therefore removed from the model described below.*

1.2.3 Medium composition

In practice the collisions studied always involve one transported particle and the background medium. The composition of this medium impacts on the nature and the effects of the collisions. In order to simplify the kinetic model, the following assumption is made.

Assumption 1.5 *The effect of a collision involving a transported particle and any atom or molecule of the medium, i.e. the deflection angle and the energy loss, is assumed to be identical to the effect of a collision with a water molecule.*
Consequence: *The characteristics of the collisions, i.e. the deflection angle and the energy loss (modeled in the next sections by the so-called cross sections σ), do not depend on the composition of the medium. However, the quantity of those collisions is assumed to depend linearily of the relative density $\rho(x)$ of the medium at point x compared to the density of water.*

In the model describe below, the composition of the medium only affects physical parameters (the cross sections σ) that are assumed to be given data in this manuscript and it does not affect the model itself. Therefore this assumption is not required, it only simplifies the problem and the notations.

1.3 Molecular description

Before describing the considered kinetic model, the physics of the interactions are described at the molecular level. Especially, the variations of density, of momentum and of energy due to the collisions are computed for each collision. Those basic features of the collisions are focused on in order to translate them and exhibit them at the kinetic level in the next section.

The particles travel in straight line until colliding with a molecule of the medium. The transported particles are characterized by their type, *i.e.* an electron e or a photon γ, their position $x \in \mathbb{R}^3$, their energy $\epsilon\mathbb{R}^+$ and their direction of flight $\Omega \in S^2$. The momentum $\mathbf{p}$ of a particle is given by

$$\mathbf{p}(\epsilon, \Omega) = p(\epsilon)\Omega, \tag{1.1}$$

and the norm of the relativistic momentum of photons and electrons are funtions of their energy ϵ given by

$$p_\gamma(\epsilon) = \epsilon, \qquad p_e(\epsilon) = \sqrt{\epsilon(\epsilon + 2)}. \tag{1.2}$$

The following subsections describe the interactions considered in this manuscript, *i.e.* Compton, Mott and Møller effect which are predominant in the energy range studied in this manuscript. The physics presented here can be extended by considering more types of collision that are non-negligible in certain energy ranges, such as Bremmstrahlung effect (see *e.g.* [21]) or pair production (see *e.g.* [25]).

1.3.1 Compton effect

Compton effect is an ionizing collision involving an incoming photon. The incoming photon transfers part of its energy to an electron bound to a molecule of the medium. The photon also transfers some energy to the molecule in order to break the link between the electron and the molecule, which results in a detachment of the electron from the molecule. In practice, this energy is the binding energy or ionization energy ϵ_B. Therefore, the electron is considered as a transported particle after collision. Fig. 1.1 represents schematically this interaction.

Before collision, the bound electron is assumed to have a negligible energy $\epsilon \approx 0$ and has therefore a negligible momentum

$$p_e(0) = 0. \tag{1.3}$$

Compton effect satisfies the following conservation properties.

Property 1.1 *(a)* ***Quantity of particles:***
The quantity of transported photons γ before collision equals the one after collision, and equals the quantity of transported electron e after collision.

(b) ***Energy:***

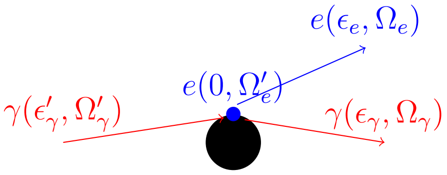
$e(\epsilon_e, \Omega_e)$
$e(0, \Omega'_e)$
$\gamma(\epsilon'_\gamma, \Omega'_\gamma)$
$\gamma(\epsilon_\gamma, \Omega_\gamma)$

Figure 1.1: Schematic representation of Compton's collision. The black ball represents an atom, the red arrows are the photon before and after scattering, the blue point and the blue arrow are the electron before (bound electron) and after scattering.

The total energy is preserved during collision. This means

$$\epsilon'_\gamma + 0 = \epsilon_\gamma + \epsilon_e + \epsilon_B. \tag{1.4}$$

Here the superscript $'$ *refers to the precollisional state and* ϵ_B *is the binding or ionization energy.*

(c) ***Momentum:***
The total momentum is preserved during collision. This means

$$\begin{aligned} \mathbf{p}_\gamma(\epsilon'_\gamma, \Omega'_\gamma) + \mathbf{p_e}(0, \Omega'_e) &= \mathbf{p}_\gamma(\epsilon'_\gamma, \Omega'_\gamma) + 0_{\mathbb{R}^3} \\ &= \mathbf{p_e}(\epsilon_e, \Omega_e) + \mathbf{p}_\gamma(\epsilon_\gamma, \Omega_\gamma). \end{aligned} \tag{1.5}$$

1.3.2 Mott effect

Mott effect is a Coulombian elastic scattering. This effect is an elastic deflection of an electron, *i.e.* an electron passes near an atom core and is deflected without losing any energy. Fig. 1.2 represents schematically this interaction. Mott effect satisfies the following conservation properties.

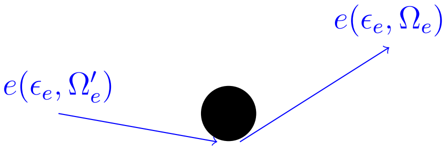

Figure 1.2: Schematic representation of Mott's collision. The black ball represents an atom and blue arrows represent an electron before and after interaction.

Property 1.2 *(a)* ***Quantity of particles:***
The quantity of transported electrons e before and after collision are equal.

(b) ***Energy:***
The energy of the electron before and after collision are equal

$$\epsilon'_e = \epsilon_e. \tag{1.6}$$

(c) ***Momentum:***
The momentum is not preserved.

1.3.3 Møller effect

Møller effect is an ionizing interaction involving an incoming electron. This electron transfers a part of its energy to an electron bound to a molecule and to the molecule itself so that the bound electron escapes from the atomic shell. Fig. 1.3 represents schematically this effect.

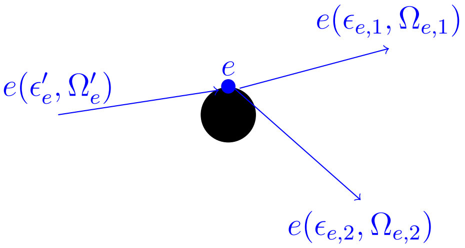
$e(\epsilon_{e,1}, \Omega_{e,1})$
e
$e(\epsilon'_e, \Omega'_e)$
$e(\epsilon_{e,2}, \Omega_{e,2})$

Figure 1.3: Schematic representation of Møller's collision. The black ball represents an atom, the blue point is the bound electron before scattering and the blue arrows are the transported electrons before and after scattering.

The indifferentiation principle states that the two outgoing electrons are indistinguishable. After collision one can not determine which one was bound and which one was transported before collision. It is although convenient to differentiate them by their energy.

Definition 1.1 *The higher energetic outgoing electron is called "primary electron", the lower energetic one is called "secondary electron".*

Møller effect satisfy the following conservation properties.

Property 1.3 (a) ***Quantity of particles:***
The quantity of transported electron e before collision equals the quantity of primary and of secondary electron after collision.

(b) ***Energy:***
The total energy is preserved during collision. This means

$$\epsilon'_e + 0 = \epsilon_{e,1} + \epsilon_{e,2} + \epsilon_B. \tag{1.7}$$

(c) ***Momentum:***
The total momentum is preserved during collision. This means

$$\mathbf{p_e}(\epsilon'_e, \Omega'_e) + 0_{\mathbb{R}^3} = \mathbf{p_e}(\epsilon_{e,1}, \Omega_{e,1}) + \mathbf{p_e}(\epsilon_{e,2}, \Omega_{e,2}). \tag{1.8}$$

Here the indices $e, 1$ and $e, 2$ refer respectively to the primary and secondary electrons after collision.

1.4 Kinetic model

The following kinetic description is a state-of-the-art model in the field of radiotherapy used for the computation of deposited doses. It offers an accurate description of the physical phenomena and is usable for applications.

1.4.1 Kinetic equation

The motion of the particles composing the medium (atoms and bound electrons) and of transported particles is differently modeled.

Assumption 1.6 *In this manuscript, the transported particles and the particles of the medium are considered as two seperated families of particles.*

Especially, the transported electrons and the bound electrons are two different families of particles, although electrons may switch from one family to the other.

One may evoke Assumption 1.2 to neglect the motion of the particles composing the medium and therefore study only the motion of the transported particles.

The motion of transported photons and electrons are modeled by their fluence ψ_γ and ψ_e depending on position x in a compact set $Z \subset \mathbb{R}^3$, on energy ϵ assumed to be bounded in the interval $[\epsilon_{\text{min}}, \epsilon_{\text{max}}] \subset \mathbb{R}^+$ (with $\epsilon_{\text{min}} > \epsilon_B$) and on direction of flight Ω on the unit sphere S^2. The quantity dN_α of particles of type α in a spatial neighboordhood dx around x having an energy in a neighbourhood $d\epsilon$ around ϵ and travelling in a solid angle $d\Omega$ around Ω is given by

$$dN_\alpha = \psi_\alpha(\epsilon, x, \Omega) d\epsilon dx d\Omega. \tag{1.9}$$

The direction of flight is often written in spherical coordinates

$$\Omega = (\mu, \sqrt{1-\mu^2}\cos\phi, \sqrt{1-\mu^2}\sin\phi)^T,$$

where $\mu \in [-1, 1]$ and $\phi \in [0, 2\pi[$. Under this notation $d\Omega = d\mu d\cos\phi$.

The density n_α, the macroscopic momentum $\mathbf{q}_\alpha$ and the macroscopic energy E_α of particles α at point $x \in Z$ can be defined from the fluence ψ_α

$$\begin{pmatrix} n_\alpha(x) \\ \mathbf{q}_\alpha(x) \\ E_\alpha(x) \end{pmatrix} = \int_{\epsilon_{\text{min}}}^{\epsilon_{\text{max}}} \int_{S^2} \begin{pmatrix} 1 \\ \mathbf{p}_\alpha(\epsilon, \Omega) \\ \epsilon \end{pmatrix} \psi_\alpha(\epsilon, x, \Omega) d\Omega d\epsilon. \tag{1.10}$$

The fluences of the transported photons and electrons in radiotherapy satisfy the following system of equations (see *e.g.* [5, 7, 17])

$$\begin{aligned}
\Omega.\nabla_x\psi_\gamma(\epsilon,x,\Omega) &= \rho(x)\left(Q_{\gamma\to\gamma}(\psi_\gamma)+Q_{e\to\gamma}(\psi_e)\right)(\epsilon,x,\Omega), && \text{(1.11a)}\\
\Omega.\nabla_x\psi_e(\epsilon,x,\Omega) &= \rho(x)\left(Q_{e\to e}(\psi_e)+Q_{\gamma\to e}(\psi_\gamma)\right)(\epsilon,x,\Omega). && \text{(1.11b)}
\end{aligned}$$

This system is composed of one equation for the photons and one for the electrons. The left-hand side of those equations is a free transport term. It is time-independent according to Assumption 1.4. The right-hand side is composed of collision operators.

Since only collision involving one transported particle and one particle of the medium are considered, the quantity of collisions can be assumed to be proportional to the quantity of particles of the medium. As only collisions with atoms and bound electrons are considered, one can assume that this quantity of particles of the medium is proportional to the relative density ρ of the medium compared to the density of water (water is chosen as it is the main component of a human body). In this study, the density ρ is chosen to be in the interval $[10^{-3}, 2]$. This interval includes the densities of liquid water ($\rho = 1$), main component of a human body, of air ($\rho = 10^{-3}$), and of bones ($\rho \approx 1.8$), corresponding to the highest density in a human body.

The collision operator $Q_{\alpha\to\beta}(\psi_\alpha)$ represents the variations of the fluence ψ_β due to collisions involving incoming particles α. Similarily the quantity of those collisions can be assumed to be proportional to ψ_α. This implies that $Q_{\alpha\to\beta}$ is a linear operator. Those collision operators are defined in the following subsections.

> **Remark 1.1** *The system (1.11) represents only the motion of the transported particles. For Compton and Møller collisions, the bound electron before scattering is not transported. It receives sufficient energy to be transported after collision. Therefore the kinetic model proposed in this section is not conservative because particles (i.e. density n_α and also momentum $\mathbf{q}_\alpha$ and energy E_α) are created in the system.*

The next subsection describes a collision operator representing interactions with a fixed background medium. A special focus is made to translate the conservation properties 1.1-1.3 from the molecular level (Section 1.3) to the kinetic level. Those properties needs to be considered when constructing numerical method for dose computaions.

1.4.2 Linear Boltzmann (LB) collision operator

As a first approach, the collisions can be modeled by linear Boltzmann gain $G_{\alpha\to\beta}(\psi_\alpha)$ and loss terms $P_\alpha(\psi_\alpha)$ (see *e.g.* the derivation of linear Boltzmann equations in [5]).

Loss term

A loss term $P_\alpha(\psi_\alpha)$ represents the loss of particles of type α due to collisions involving incoming particle α. This loss corresponds to the particles that are removed from the system during collision, *i.e.* the particles at precollisional state. It is given by

$$P_\alpha(\psi_\alpha)(\epsilon, x, \Omega) = \sigma_{T,\alpha}(\epsilon)\psi_\alpha(\epsilon, x, \Omega), \tag{1.12}$$

where the total cross section $\sigma_{T,\alpha}$ is a positive function which quantifies the interactions of a particle α, *i.e.* the quantity

$$\rho(x)\sigma_{T,\alpha}(\epsilon)\psi_\alpha(\epsilon, x, \Omega)d\epsilon dx d\Omega$$

is the quantity of particles α in a neighborhood $(d\epsilon, dx, d\Omega)$ around (ϵ, x, Ω) colliding with the medium.

Gain term

A gain term $G_{\alpha\to\beta}(\psi_\alpha)(\epsilon, x, \Omega)$ represents the gain of particles of type β at state (ϵ, x, Ω) due to collisions involving a incoming particle β. This gain term corresponds to the particles that are created in the system during collision, *i.e.* the particles at postcollisional state. It has the form

$$G_{\alpha\to\beta}(\psi_\alpha)(\epsilon, x, \Omega) = \int_\epsilon^{\epsilon_{\max}} \int_{S^2} \sigma_{\alpha\to\beta}(\epsilon', \epsilon, \Omega'.\Omega)\psi_\alpha(\epsilon', x, \Omega')d\Omega' d\epsilon'. \tag{1.13}$$

This gain term is an integral over all possible precollisional state (ϵ', Ω') of the fluence ψ_α of the incident particles α multiplied by a differential cross section $\sigma_{\alpha\to\beta}$. The differential cross section $\sigma_{\alpha\to\beta}(\epsilon', \epsilon, \Omega'.\Omega)$ is a non-negative function such that the following quantity

$$\rho(x)\sigma_{\alpha\to\beta}(\epsilon', \epsilon, \Omega'.\Omega)\psi_\alpha(\epsilon', x, \Omega')d\epsilon' d\Omega' d\Omega dx$$

is the quantity of particles β created in a neighborhood $(d\epsilon, dx, d\Omega)$ around (ϵ, x, Ω) by a collision involving a particles α in a precollisional state in the neighborhood $(d\epsilon', dx, d\Omega')$ around (ϵ', x, Ω').

Remark 1.2 *The density, the macroscopic momentum and the macroscopic energy created, repsectively removed, in the system for each effect is defined by replacing ψ_α by $G_{\alpha\to\beta}(\psi_\alpha)$, respectively $P_\alpha(\psi_\alpha)$, in (1.10), i.e. the quantity of particles, the momentum and the energy of the particles α removed from the system is*

$$\int_{\epsilon_{\min}}^{\epsilon_{\max}} \int_{S^2} \begin{pmatrix} 1 \\ \mathbf{p}_\alpha(\epsilon, \Omega) \\ \epsilon \end{pmatrix} P_\alpha(\psi_\alpha)(\epsilon, x, \Omega)d\Omega d\epsilon,$$

and the quantities of particles, momentum and energy of the particles β created by collisions involving incoming particles of type α is

$$\int_{\epsilon_{\min}}^{\epsilon_{\max}} \int_{S^2} \begin{pmatrix} 1 \\ \mathbf{p}_\beta(\epsilon, \Omega) \\ \epsilon \end{pmatrix} G_{\alpha\to\beta}(\psi_\alpha)(\epsilon, x, \Omega)d\Omega d\epsilon.$$

For Compton, Mott and Møller interactions, the collision operators read

$$\begin{aligned}
Q_{\gamma\to\gamma}(\psi_\gamma) &= G_{\gamma\to\gamma}(\psi_\gamma) - P_\gamma(\psi_\gamma) &&= G_{C,\gamma}(\psi_\gamma) \; -P_C(\psi_\gamma), && (1.14a)\\
Q_{\gamma\to e}(\psi_\gamma) &= G_{\gamma\to e}(\psi_\gamma) &&= G_{C,e}(\psi_\gamma), && (1.14b)\\
Q_{e\to\gamma}(\psi_e) &= G_{e\to\gamma}(\psi_e) &&= 0, && (1.14c)\\
Q_{e\to e}(\psi_e) &= G_{e\to e}(\psi_e) - P_e(\psi_e) &&= G_{M,1}(\psi_e) + G_{M,2}(\psi_e) + G_{Mott}(\psi_e) \\
& && \quad -P_M(\psi_e) - P_{Mott}(\psi_e), && (1.14d)
\end{aligned}$$

where the subscript C refers to Compton scattering, M to Møller and $Mott$ to Mott. Therefore the indices C, γ and C, e refer respectively to the outgoing photon and electron of Compton effect. Similarily, the indices $M, 1$ and $M, 2$ refer respectively to the primary (more energetic) and secondary electron of Møller effect. Using these notations, the cross sections can be rewritten

$$\begin{aligned}
\sigma_{\gamma\to\gamma} &= \sigma_{C,\gamma}, \qquad \sigma_{T,\gamma} = \sigma_{T,C}, \qquad \sigma_{\gamma\to e} = \sigma_{C,e},\\
\sigma_{e\to e} &= \sigma_{M,1} + \sigma_{M,2} + \sigma_{Mott}, \qquad \sigma_{T,e} = \sigma_{T,M} + \sigma_{T,Mott}.
\end{aligned}$$

The conservation properties 1.1-1.3 from the molecular description leads to impose conditions on the cross sections at the kinetic level. Those conditions needs to be fulfilled for the physics of the collision to be correctly modeled at the kinetic level.

Compton scattering

Compton differential cross section for photons reads (see [11, 14, 18] and references therein)

$$\begin{aligned}
\sigma_{C,\gamma}(\epsilon', \epsilon, \mu) &= \frac{r_e^2}{2}\left(\frac{\epsilon}{\epsilon'}\right)^2 \left(\frac{\epsilon}{\epsilon'} + \frac{\epsilon'}{\epsilon} - (1-\mu^2)\right)\\
&\quad * \;\; \delta\left(\epsilon - \frac{\epsilon'}{1+\epsilon'(1-\mu)}\right), \qquad (1.16)
\end{aligned}$$

where r_e is the radius of the electron. The Dirac distribution δ in this formula represents the fact that the energy ϵ of the photon after collision is a given function of its energy ϵ' before collision and of the deflection cosangle μ.

The conservation of particles (Property 1.1(a) at the molecular) is described at the kinetic level by

Proposition 1.1 *The quantity of photons removed, created, and the quantity* *of electrons created by Compton effect are equal at the kinetic level iff*

$$\begin{aligned}\sigma_{T,C}(\epsilon) &= 2\pi\int_{\epsilon_B}^{\epsilon}\int_{-1}^{+1}\sigma_{C,\gamma}(\epsilon,\epsilon',\mu)d\mu d\epsilon' \\ &= 2\pi\int_{\epsilon_B}^{\epsilon}\int_{-1}^{+1}\sigma_{C,e}(\epsilon,\epsilon',\mu)d\mu d\epsilon'. \qquad (1.17)\end{aligned}$$

Proof According to Property 1.1, the total quantity of photons created, lost and the total quantity of electrons created by Compton effect at point x simply are equal. According to Remark 1.2, this reads

$$\begin{aligned}\int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}G_{C,\gamma}(\psi_\gamma)(\epsilon,x,\Omega)d\epsilon d\Omega &= \int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}G_{C,e}(\psi_\gamma)(\epsilon,x,\Omega)d\epsilon d\Omega \\ &= \int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}P_C(\psi_\gamma)(\epsilon,x,\Omega)d\epsilon d\Omega.\end{aligned}$$

Using Fubini theorem leads to

$$\begin{aligned}&\int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}G_{C,\gamma}(\psi_\gamma)(\epsilon,x,\Omega)d\epsilon d\Omega \\ =&\int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}\int_{\epsilon}^{\epsilon_{max}}\int_{S^2}\sigma_{C,\gamma}(\epsilon',\epsilon,\Omega'.\Omega)\psi_\gamma(\epsilon',x,\Omega')d\epsilon' d\Omega' d\epsilon d\Omega \\ =&\int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}\int_{\epsilon_B}^{\epsilon}\int_{S^2}\sigma_{C,\gamma}(\epsilon,\epsilon',\Omega.\Omega')d\epsilon' d\Omega'\psi_\gamma(\epsilon,x,\Omega)d\epsilon d\Omega,\end{aligned}$$

Similar computations with $G_{C,e}$ leads to

$$\begin{aligned}&\int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}\left(\int_{\epsilon_B}^{\epsilon}\int_{S^2}\sigma_{C,\gamma}(\epsilon,\epsilon',\Omega.\Omega')d\epsilon' d\Omega'\right)\psi_\gamma(\epsilon,x,\Omega)d\epsilon d\Omega \\ =&\int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}\left(\int_{\epsilon_B}^{\epsilon}\int_{S^2}\sigma_{C,e}(\epsilon,\epsilon',\Omega.\Omega')d\epsilon' d\Omega'\right)\psi_\gamma(\epsilon,x,\Omega)d\epsilon d\Omega \\ =&\int_{\epsilon_B}^{\epsilon_{max}}\int_{S^2}\sigma_{T,C}(\epsilon)\psi_\gamma(\epsilon,x,\Omega)d\epsilon d\Omega,\end{aligned}$$

which is satisfied for all possible ψ_γ. Therefore one obtains the result in a weak sense. Computations show that the functions defined in (1.17) are $C^\infty([\epsilon_B,\epsilon_{max}])$. Therefore this equality also holds in a strong sense. □

The conservation of momentum (Property 1.1(c) at the molecular level) is described at the kinetic level by

Proposition 1.2 *The macroscopic momentum in the system is preserved by Compton effect at the kinetic level iff*

$$\begin{aligned}\mathbf{p}_\gamma(\epsilon,\Omega)\sigma_{T,C}(\epsilon)-\int_{\epsilon_B}^{\epsilon}\int_{S^2} &\quad [\mathbf{p}_\gamma(\epsilon',\Omega')\sigma_{C,\gamma}(\epsilon,\epsilon',\Omega.\Omega') \qquad (1.18)\\ &+\quad \mathbf{p}_\gamma(\epsilon',\Omega')\sigma_{C,e}(\epsilon,\epsilon',\Omega.\Omega')]\quad d\Omega' d\epsilon' = 0_{\mathbb{R}^3}.\end{aligned}$$

The proof is identical to the one of Proposition 1.1.

The differential cross sections for photons and electrons are related to each others. Indeed according to Property 1.1, there is as many photons γ created at state $(\epsilon_\gamma, x, \Omega_\gamma)$ as electrons at state $(\epsilon_e, x, \Omega_e)$ from photons at state (ϵ', x, Ω'), *i.e.*

$$\begin{aligned} & \sigma_{C,\gamma}(\epsilon', \epsilon_\gamma, \Omega'.\Omega_\gamma) \quad d\epsilon' \quad d\epsilon_\gamma \quad d\Omega' \quad d\Omega_\gamma \quad dx \\ = & \sigma_{C,e}(\epsilon', \epsilon_e, \Omega'.\Omega_e) \quad d\epsilon' \quad d\epsilon_e \quad d\Omega' \quad d\Omega_e \quad dx. \end{aligned} \tag{1.19}$$

Therefore, the differential cross sections are related to each others. Using (1.4) and (1.5), one finds that

$$\epsilon_e = \epsilon'_\gamma - \epsilon_\gamma - \epsilon_B, \quad \Omega_e = \frac{\mathbf{p}_\gamma(\epsilon'_\gamma, \Omega'_\gamma) - \mathbf{p}_\gamma(\epsilon_\gamma, \Omega_\gamma)}{p_e(\epsilon_e)}, \tag{1.20a}$$

$$\epsilon_\gamma = \epsilon'_\gamma - \epsilon_e - \epsilon_B, \quad \Omega_\gamma = \frac{\mathbf{p}_\gamma(\epsilon'_\gamma, \Omega'_\gamma) - \mathbf{p_e}(\epsilon_e, \Omega_e)}{p_\gamma(\epsilon_\gamma)}. \tag{1.20b}$$

Using (1.19) and (1.20) together leads to the following property.

Property 1.4 *The differential cross section for electrons and photons satisfy*

$$\begin{aligned} \sigma_{C,e}(\epsilon'_\gamma, \epsilon_e, \mu_e) \quad = \quad & \sigma_{C,\gamma}\left(\epsilon', \epsilon' - \epsilon - \epsilon_B, \frac{p_\gamma(\epsilon') - p_e(\epsilon)\mu}{p_\gamma(\epsilon' - \epsilon - \epsilon_B)}\right) \\ & * \quad \frac{p_e(\epsilon)}{p_\gamma(\epsilon' - \epsilon - \epsilon_B)}. \end{aligned} \tag{1.21}$$

In this study, $\epsilon' > \epsilon > \epsilon_B$ and one may verify that the deflection cosangle

$$\frac{p_\gamma(\epsilon') - p_\gamma(\epsilon)\mu}{p_e(\epsilon' - \epsilon - \epsilon_B)}$$

in this formula is always in $[-1, +1]$.

Remark 1.3 *In practice, the differential and total cross sections are defined such that the conservation properties (1.17) and (1.18) are satisfied (see e.g. [11]). One may also verify that those propositions are always satisfied when the differential cross section for electron is defined from (1.21).*

In the litterature, it is often preferred to work with analytical formula of the differential and total cross sections (see e.g. [17]). However, those properties need to be kept in mind when constructing numerical schemes for (1.11).

Mott scattering

Since Mott effect is an elastic scattering effect, the energy of the electron before and after scattering is the same. Therefore the differential cross section can be written

$$\sigma_{Mott}(\epsilon', \epsilon, \mu) = \tilde{\sigma}_{Mott}(\epsilon, \mu)\delta(\epsilon' - \epsilon), \tag{1.22a}$$

where $\tilde{\sigma}_{Mott}$ reads (see [17, 24] and references therein)

$$\begin{aligned}\tilde{\sigma}_{Mott}(\epsilon,\mu) &= \left[\frac{Zr_e(1+\epsilon)}{4\left(\epsilon(\epsilon+2)\right)\left(1+2\eta(\epsilon)-\mu\right)}\right]^2 \\ &\quad * \left[1-\frac{\epsilon(\epsilon+2)}{(1+\epsilon)^2}\frac{1-\mu}{2}\right], \qquad (1.22b)\\ \eta(\epsilon) &= \frac{\pi\alpha^2 Z^{\frac{2}{3}}}{\epsilon(\epsilon+2)}, \qquad (1.22c)\end{aligned}$$

Z is the equivalent atomic number of water, and $\alpha \approx \frac{1}{137}$.

Similarily as for Compton scattering, the conservation Properties 1.2 can be translated at the kinetic level by

Proposition 1.3 *The quantity of particles and the energy in the system is preserved by Mott effect at the kinetic level iff*

$$\sigma_{T,Mott}(\epsilon) = 2\pi\int_{-1}^{+1}\tilde{\sigma}_{Mott}(\epsilon,\mu)d\mu. \qquad (1.23)$$

The proof is identical to the one of Proposition 1.1.

Møller scattering

Møller differential cross section for primary electrons reads (see [18, 14] and references therein)

$$\begin{aligned}\sigma_{M,1}(\epsilon',\epsilon,\mu) &= \sigma_M(\epsilon',\epsilon,\mu)\mathbf{1}_{[\frac{\epsilon'-\epsilon_B}{2},\epsilon'-\epsilon_B]}(\epsilon), \qquad (1.24a)\\ \sigma_M(\epsilon',\epsilon,\mu) &= \left(\frac{r_e(\epsilon'+1)}{\epsilon'(\epsilon'+2)}\right)^2\left[\frac{1}{W(\epsilon',\epsilon)^2}+\frac{1}{(\epsilon'-W(\epsilon',\epsilon))^2}\right. \qquad (1.24b)\\ &\quad \left.-\frac{2\epsilon'+1}{(\epsilon'+1)^2}\frac{1}{W(\epsilon',\epsilon)\left(\epsilon'-W(\epsilon',\epsilon)\right)}+\frac{1}{(\epsilon'+1)^2}\right]\\ &\quad * \delta\left(W(\epsilon',\epsilon)-Q(\epsilon',\epsilon,\mu)\right).\end{aligned}$$

with W the energy transfered to the bound electron and Q is the recoil energy

$$W(\epsilon',\epsilon) = \epsilon'-\epsilon, \quad Q(\epsilon',\epsilon,\mu) = \sqrt{p_e(\epsilon')^2+p_e(\epsilon)^2-2p_e(\epsilon')p_e(\epsilon)\mu+1}-1.$$

The conservation of particles (Property 1.3(a) at the molecular) is described at the kinetic level by

Proposition 1.4 *The quantity of electrons removed, and of primary and sec-*

ondary electrons created by Møller effect are equal at the kinetic level iff

$$\begin{aligned}\sigma_{T,M}(\epsilon) &= 2\pi\int_{\epsilon_B}^{\epsilon}\int_{-1}^{+1}\sigma_{M,1}(\epsilon,\epsilon',\mu)d\mu d\epsilon' \\ &= 2\pi\int_{\epsilon_B}^{\epsilon}\int_{-1}^{+1}\sigma_{M,2}(\epsilon,\epsilon',\mu)d\mu d\epsilon'.\end{aligned} \tag{1.25}$$

The proof is identical to the one of Proposition 1.1.

The conservation of momentum (Property 1.3(c) at the molecular level) is described at the kinetic level by

Proposition 1.5 *The macroscopic momentum in the system is preserved by Møller effect at the kinetic level iff*

$$\begin{aligned}0_{\mathbb{R}^3} &= \mathbf{p_e}(\epsilon,\Omega)\sigma_{T,M}(\epsilon) \\ &\quad -\int_{\epsilon_B}^{\epsilon}\int_{S^2}\mathbf{p_e}(\epsilon',\Omega')(\sigma_{M,1}+\sigma_{M,2})(\epsilon,\epsilon',\Omega.\Omega')d\Omega' d\epsilon'.\end{aligned} \tag{1.26}$$

The proof is identical to the one of Proposition 1.1.

Similarily as for Compton effect, the differential cross section for primary and secondary electrons are related to each others. Indeed according to Property 1.3, there is as many primary electrons at state $(\epsilon_{e,1},x,\Omega_{e,1})$ as secondary electrons at state $(\epsilon_{e,2},x,\Omega_{e,2})$ from electrons at state (ϵ',x,Ω'), *i.e.*

$$\begin{aligned}&\sigma_{M,1}(\epsilon',\epsilon_{e,1},\Omega'.\Omega_{e,1}) \quad d\epsilon' \quad d\epsilon_{e,1} \quad d\Omega' \quad d\Omega_{e,1} \quad dx \\ =\ &\sigma_{M,2}(\epsilon',\epsilon_{e,2},\Omega'.\Omega_{e,2}) \quad d\epsilon' \quad d\epsilon_{e,2} \quad d\Omega' \quad d\Omega_{e,2} \quad dx.\end{aligned} \tag{1.27}$$

Therefore, the differential cross section are related to each others. Using (1.7) and (1.8), one finds that

$$\epsilon_{e,2} = \epsilon'_e - \epsilon_{e,1} - \epsilon_B, \qquad \Omega_{e,2} = \frac{\mathbf{p_e}(\epsilon'_e,\Omega'_e) - \mathbf{p_e}(\epsilon_{e,1},\Omega_{e,1})}{p_e(\epsilon_{e,2})}. \tag{1.28}$$

Using (1.27) and (1.28) together leads to the following property.

Property 1.5 *The differential cross sections for Møller effect satisfy*

$$\begin{aligned}\sigma_{M,2}(\epsilon',\epsilon,\mu) &= \sigma_{M,1}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B,\frac{p_e(\epsilon')-p_e(\epsilon)\mu}{p_e(\epsilon'-\epsilon-\epsilon_B)}\right) \\ &\quad * \frac{p_e(\epsilon)}{p_e(\epsilon'-\epsilon-\epsilon_B)}.\end{aligned} \tag{1.29}$$

In this study, $\epsilon' > \epsilon > \epsilon_B$*, one may verify that the deflection cosangle*

$$\frac{p_e(\epsilon')-p_e(\epsilon)\mu}{p_e(\epsilon'-\epsilon-\epsilon_B)}$$

in this formula is always in $[-1,+1]$.

Remark 1.4 *One may verify that*

$$\sigma_{M,2}(\epsilon',\epsilon,\mu) \quad = \quad \sigma_M(\epsilon',\epsilon,\mu)\mathbf{1}_{[\epsilon_B,\frac{\epsilon'-\epsilon_B}{2}]}(\epsilon).$$

According to the indifferentiation principle, one can not determine which of the outgoing electrons was the incoming one. Originally a single differential cross section was written for both primary and secondary electrons. Those differential cross sections were afterward differentiated by defining the bounds of integration $\left[\frac{\epsilon'-\epsilon_B}{2},\epsilon'\right]$ *for primary electrons and* $\left[\epsilon_B,\frac{\epsilon'-\epsilon_B}{2}\right]$ *for secondary electrons.*

Some of the differential cross sections (Møller and Mott collisions) present high gradients. When constructing numerical schemes, such high gradients commonly leads to stiff terms at the numerical level. The following two sections present methods to remove this stiffness directly at the continuous level.

1.4.3 Continuous slowing-down approximation (CSDA)

On the theoretical level, the linear Boltzmann collision operator described in the previous subsection is appropriate. Although, for numerical applications, discretizing directly the linear Boltzmann gain terms can be difficult. In particular, Møller cross section σ_M is known to be very peaked along the line $\epsilon' = \epsilon$ (see Fig. 1.4). Typically, discretizations of the gain term require the energy grid to be fine enough to capture this peak, which leads to requiring a significant numer of energy cells.

Instead it is often preferred to apply the continuous slowing-down approximation (see *e.g.* [20]).

Suppose a linear Boltzmann collision term

$$Q(\psi) = (G - P)(\psi),$$

with loss and gain terms of the form (1.12) and (1.13) characterized by a differential cross section σ peaked along the line $\epsilon' = \epsilon$, and its associated total cross section σ_T.

Formally, the continuous slowing-down approximation consists in replacing the peaked gain term and its associated loss term by an energy derivative and an elastic linear Boltzmann term, *i.e.*

$$Q(\psi) \approx Q_{CSD}(\psi) \quad := \quad \partial_\epsilon(S\psi) + (G_{el} - P_{el})(\psi). \tag{1.30}$$

where the stopping power S, the differential σ_{el} and total cross section $\sigma_{T,el}$ of the elastic collision operator $(G_{el} - P_{el})$ are related to the original cross sections σ and σ_T through

$$S(\epsilon) \quad = \quad \int_{\epsilon_{min}}^{\epsilon} \int_{S^2} (\epsilon - \epsilon')\sigma(\epsilon,\epsilon',\Omega'.\Omega)d\Omega' d\epsilon', \tag{1.31}$$

$$\sigma_{el}(\epsilon',\epsilon,\Omega'.\Omega) \quad = \quad \int_{\epsilon_{min}}^{\epsilon} \sigma(\epsilon,\epsilon',\Omega'.\Omega)d\epsilon' \quad \delta(\epsilon'-\epsilon), \tag{1.32}$$

$$\sigma_{T,el}(\epsilon) \quad = \quad \int_{\epsilon_{min}}^{\epsilon} \int_{S^2} \sigma_{el}(\epsilon,\epsilon',\Omega.\Omega')d\Omega' d\epsilon'. \tag{1.33}$$

Based on its definition (1.31), the stopping power S is a positive function of the energy ϵ. The ϵ-derivative represents a continuous loss of energy of the particles. Therefore, S characterizes the capacity of the system to slow particles down. The deflection phenomenum is represented by an elastic linear Boltzmann operator.

Møller cross section for primary electrons is peaked along the line $\epsilon' = \epsilon$. This peak can be observed on Fig. 1.4 through the quantity

$$\sigma^0_{M,1}(\epsilon', \epsilon) = 2\pi \int_{-1}^{+1} \sigma_{M,1}(\epsilon', \epsilon, \mu) d\mu.$$

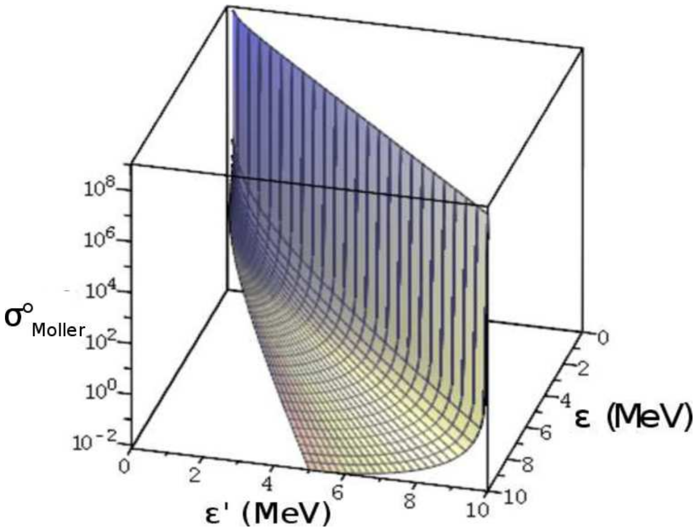

Figure 1.4: Representation of $\sigma^0_{M,1}$ as a function of ϵ' and ϵ.

Using the CSDA (1.30) leads to write the following approximation

$$(G_{M,1} - P_M)(\psi_e)(\epsilon, x, \Omega) \approx \left[\partial_\epsilon(S_M \psi_e) + (G_{M,1,el} - P_{M,el})(\psi_e)\right](\epsilon, x, \Omega),$$

where S_M, $G_{M,1,el}$ and $P_{M,el}$ are respectively the stopping power, the elastic gain term of primary electrons and the loss term after application of the CSDA (1.30) to Møller collision.

In practice, the angular deflection due to Møller effect is negligible compared to the one due to Mott effect. This leads to consider the linear Boltzmann continuous slowing-down (LBCSD) operator

$$\boxed{\begin{aligned} Q_{e\to e}(\psi_e) &\approx Q_{LBCSD}(\psi_e) \\ Q_{LBCSD}(\psi_e) &:= \partial_\epsilon(S\psi_e) + (G + G_{el} - P_{el})(\psi_e), \\ S = S_M, \qquad G &= G_{M,2}, \qquad G_{el} = G_{Mott}, \qquad P_{el} = P_{Mott}. \end{aligned} \qquad (1.34)}$$

Remark 1.5 *This approximation is convenient for numerical purposes. Although, at the theoretical level, this approximation leads to violate the conservation of quantity of particles, momentum and energy (Property 1.5 and Propositions 1.4 and 1.5). Indeed, the approximation* (1.34) *consists in approximating the distribution*

$$\sigma_{M,1} \approx S_M(\epsilon)\delta'(\epsilon' - \epsilon),$$

without modifying the distribution $\sigma_{M,2}$. Since the term $\partial_\epsilon(S\psi)$ corresponds now to both terms of loss and gain of primary electrons, one can now only deduce, at the kinetic level, the equality of quantity of primary electrons created and of electrons removed

$$\begin{aligned}\int_{\epsilon_B}^{\epsilon_{max}} \int_{S^2} \partial_\epsilon(S\psi_e)(\epsilon, x, \Omega) d\epsilon d\Omega = \int_{S^2} \quad & [S(\epsilon_{max}) \quad \psi_e(\epsilon_{max}, x, \Omega) \\ - \quad & S(\epsilon_B) \quad \psi_e(\epsilon_B, x, \Omega)]\, d\Omega,\end{aligned}$$

where $\psi_e(\epsilon_{max}, x, \Omega)$ is assumed to be zero. This corresponds to requiring that the particles have a energy bounded by $\epsilon_{\max}$. The other term in $\psi_e(\epsilon_B, x, \Omega)$ corresponds to the quantity of particles leaving the system, i.e. particles of energy below the threshold ϵ_B are absorbed by the medium and are therefore not transported anymore.

Remark 1.6 *By considering ϵ as a "numerical time", the CSDA* (1.30) *makes the following operator appear out of the collision operator* (1.34)

$$-\partial_\epsilon(S\psi_e) + \Omega.\nabla_x\psi_e,$$

which can be studied similarily as an hyperbolic operator.

1.4.4 Fokker-Planck (FP) approximation

Similarily to Møller cross section, Mott cross section is forward-peaked, *i.e.* peaked along the line $\Omega'.\Omega = 1$. This peak is represented on Fig. 1.5 through the $\tilde{\sigma}_{Mott}(\epsilon, \mu)$.

The Fokker-Planck approximation for forward-peaked collisions with small energy losses was rigourously derivated in [29] through the following theorem.

Theorem 1.1 ([29]) *Consider the linear Boltzmann operator*

$$\begin{aligned} Q(\psi)(\epsilon, x, \Omega) &= \int_{\epsilon}^{\epsilon_{max}} \int_{S^2} \sigma(\epsilon', \epsilon, \Omega'.\Omega)\psi(\epsilon', x, \Omega')d\Omega' d\epsilon' - \sigma_T(\epsilon)\psi(\epsilon, x, \Omega), \\ \sigma(\epsilon', \epsilon, \mu) &= \frac{1}{s}\hat{\sigma}\left(\epsilon', \frac{\epsilon' - \epsilon}{e}, \frac{1-\mu}{m}\right), \\ \sigma_T(\epsilon) &= \frac{2\pi}{s} \int_{\epsilon_{min}}^{\epsilon} \int_{-1}^{+1} \hat{\sigma}\left(\epsilon, \frac{\epsilon - \epsilon'}{e}, \frac{1-\mu}{m}\right) d\mu d\epsilon', \end{aligned}$$

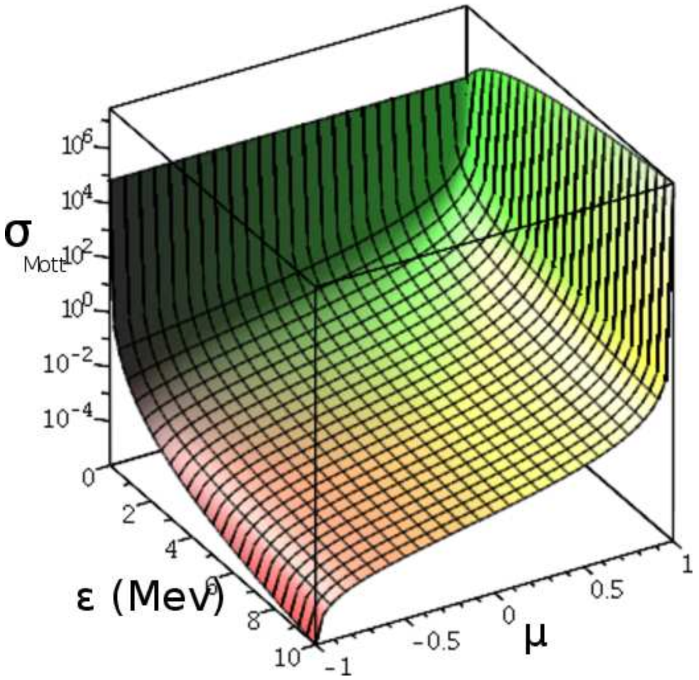

Figure 1.5: Representation of $\tilde{\sigma}_{Mott}$ as a function of ϵ and μ.

where s, e and m are asymptotic parameters characterizing the "peakness" of the cross section and tending to 0.

If $\hat{\sigma} = O(1)$ then

$$Q(\psi)(\epsilon, x, \Omega) = Q_{FP}(\psi) + O\left(\frac{e^2}{s}\right) + O\left(\frac{m^2}{s}\right) + O\left(\frac{em}{s}\right),$$

where the Fokker-Planck (FP) operator reads

$$Q_{FP}(\psi) = \partial_\epsilon (S(\epsilon)\psi)(\epsilon, x, \Omega) + T(\epsilon)\Delta_\Omega \psi(\epsilon, x, \Omega), \tag{1.35a}$$

$$S(\epsilon) = \int_{\epsilon_{min}}^{\epsilon} \int_{S^2} (\epsilon - \epsilon')\sigma(\epsilon, \epsilon', \Omega'.\Omega) d\Omega' d\epsilon', \tag{1.35b}$$

$$T(\epsilon) = \int_{\epsilon_{min}}^{\epsilon} \int_{S^2} (1 - \Omega'.\Omega)\sigma(\epsilon, \epsilon', \Omega'.\Omega) d\Omega' d\epsilon', \tag{1.35c}$$

$$\Delta_\Omega \psi(\epsilon, x, \Omega) = \left[\partial_\mu \left((1-\mu^2)\partial_\mu \psi\right) + \frac{1}{1-\mu^2}\partial_\phi^2 \psi\right](\epsilon, x, \Omega), \tag{1.35d}$$

and the variable μ and ϕ are such that

$$\Omega = (\mu, \sqrt{1-\mu^2}\cos\phi, \sqrt{1-\mu^2}\sin\phi)^T.$$

The function T is called "transport coefficient". Based on the positivity of the differential cross sections σ, the transport coefficient T is a positive function of ϵ.

Applying Theorem 1.1 leads to write the following approximation

$$(G_{Mott} + G_{M,1} - P_{Mott} - P_M)(\psi_e)(\epsilon, x, \Omega) \approx (\partial_\epsilon(S\psi_e) + T\Delta_\Omega\psi_e)(\epsilon, x, \Omega),$$
$$S \approx S_M, \qquad T \approx T_{Mott}.$$

In practice, $S_{Mott} = 0$. This leads to approximate the LBCSD operator by a Fokker-Planck with a linear Boltzmann gain term (LBFP) operator

$$\boxed{\begin{aligned} &Q_{e\to e}(\psi_e) \approx Q_{LBFP}(\psi_e) \\ &Q_{LBFP}(\psi_e) := \partial_\epsilon(S\psi_e) + T\Delta_\Omega\psi_e + G(\psi_e), \\ &S = S_M, \qquad T = T_{Mott}, \qquad G = G_{M,2}. \end{aligned} \qquad (1.36)}$$

In the rest of the manuscript, $S = S_M$ and $T = T_{Mott}$.

> **Remark 1.7** *By considering ϵ as a "numerical time", the Fokker-Planck operator* (1.35a) *is backward parabolic in Ω. Therefore a kinetic equation with such a collision operator can be well-posed only if the ϵ derivative is read in the "backward" direction, from a maximum energy ϵ_{max} to a minimum one ϵ_{min}. In this manuscript, the considered kinetic equation are always studied so.*
>
> *Physically, it is sensefull to study the system in the backward direction in energy since the particles (and the system) only lose energy in the medium.*

If the approximation of LB operator into a LBCSD operator (previous section) is commonly accepted in the litterature, the accuracy of the LBFP approximation is more discussable (see *e.g.* comparison in [27, 26]).

> **Remark 1.8** *For the numerical applications, of the Part III, the stopping power S and the cross sections $\sigma_{\gamma\to\gamma}$, $\sigma_{e\to e}$ are obtained from tabulations of the ones used in the Monte Carlo code PENELOPE ([14, 2, 3]) that were obtained by E. Olbrant ([26]). The total cross sections sigma$_{T,\gamma}$ and $\sigma_{T,e}$ and Compton's cross section for electrons are obtained from the formulae* (1.17), (1.23), (1.25) *and* (1.21).

1.5 Well-posedness of the kinetic equations

Consider the system (1.11) where the spatial domain $Z \subset \mathbb{R}^3$ is compact. First, initial and boundary conditions for (1.11) are defined. Then the well-posedness of the resulting problem is studied.

1.5.1 Initial-boundary conditions

According to Remark 1.7, the natural initial condition consists in fixing the value of ψ_e and ψ_γ at energy $\epsilon = \epsilon_{max}$. The considered particles are assumed to have a bounded energy ϵ below $\epsilon_{\max}$. Therefore in this manuscript, the following initial condition is always fixed

$$\psi_\gamma(\epsilon_{max}, x, \Omega) = \psi_e(\epsilon_{max}, x, \Omega) = 0, \quad \forall (x, \Omega) \in Z \times S^2. \tag{1.37}$$

In order to formulate boundary conditions, one needs to define in- and outgoing boundaries Γ^- and Γ^+

$$\begin{aligned} \Gamma^- &= \Big\{ (x, \Omega) \in \partial Z \times S^2, \quad \text{s.t.} \quad \Omega.n(x) < 0 \Big\}, & \text{(1.38a)} \\ \Gamma^+ &= \Big\{ (x, \Omega) \in \partial Z \times S^2, \quad \text{s.t.} \quad \Omega.n(x) \geq 0 \Big\}. & \text{(1.38b)} \end{aligned}$$

A natural boundary condition for (1.11) consists in fixing the flux of particles coming inside the medium (see *e.g.* [5, 10]), *i.e.*

$$\left. \begin{aligned} \psi_\gamma(\epsilon, x, \Omega) &= \psi^b_\gamma(\epsilon, x, \Omega) \\ \psi_e(\epsilon, x, \Omega) &= \psi^b_e(\epsilon, x, \Omega) \end{aligned} \right\} \quad \forall (\epsilon, x, \Omega) \in [\epsilon_{min}, \epsilon_{max}] \times \Gamma^-. \tag{1.39}$$

1.5.2 Preliminaries

The following notations will be used in this chapter and in Chapter 6.

Notation 1.1 *Choose a positive weight function* $0 < h \in C^1([\epsilon_{\min}, \epsilon_{\max}])$ *and define the following weighted inner product*

$$\begin{aligned} (\psi, \lambda)^h_i &= \int_{\epsilon_{\min}}^{\epsilon_{\max}} \int_Z \int_{S^2} h(\epsilon)\psi(\epsilon, x, \Omega)\lambda(\epsilon, x, \Omega) d\Omega dx d\epsilon, \\ (\psi, \lambda)^h_\epsilon &= \int_Z \int_{S^2} h(\epsilon)\psi(\epsilon, x, \Omega)\lambda(\epsilon, x, \Omega) d\Omega dx, \\ (\psi^b, \eta)^h_{b_-} &= \int_{\epsilon_{\min}}^{\epsilon_{\max}} \int_{\Gamma^-} h(\epsilon)|\Omega.n(x)|\psi^b(\epsilon, x, \Omega)\eta(\epsilon, x, \Omega) d\Omega dx d\epsilon, \\ (\psi^b, \eta)^h_{b_+} &= \int_{\epsilon_{\min}}^{\epsilon_{\max}} \int_{\Gamma^+} h(\epsilon)|\Omega.n(x)|\psi^b(\epsilon, x, \Omega)\eta(\epsilon, x, \Omega) d\Omega dx d\epsilon, \\ (\psi^b, \eta)^h_b &= (\psi^b, \eta)^h_{b_-} + (\psi^b, \eta)^h_{b_+}, \end{aligned}$$

and, by extension, write the inner product with the weight function $h = 1$ *and the norms associated to these inner products*

$$\begin{aligned} &\textit{for } s = i,\ b_-,\ b_+,\ b, \\ &\qquad (\psi, \lambda)_s = (\psi, \lambda)^1_s, \qquad \|\psi\|^h_s = (\psi, \psi)^h_s, \qquad \|\psi\|_s = (\psi, \psi)_s. \end{aligned}$$

The following proofs of well-posedness are widely inspired of those proposed in [31] (see also [10, 15, 16, 1]). Those proofs are written here as they provide (slightly) different constraints on the physical parameters compared to [31] and they provide an understanding of the considered kinetic equations that will be exploited for dose optimization in Chapter 6.

1.5.3 Non scattered particles

Non-zero boundary conditions introduce difficulties when proving the existence and uniqueness of a solution to (1.11). One method to circumvent such difficulties consists in exploiting the linearity of the kinetic equations. First, the equation satisfied by the fluence of the particles that have never scattered is studied. Secondly, the equation satisfied by the secondary particles, *i.e.* those that have scattered at least one, is studied. Such equations have zero incoming fluxes.

The following two lemmas prove the existence of solutions to kinetic equations with non-zero boundary conditions corresponding to the non-scattered particles.

Lemma 1.1 ([31]) *Consider the equation*

$$\left\{\begin{array}{rclll} \Omega.\nabla_x\psi + \rho P(\psi) & = & q & \textit{in} & [\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2, \\ \psi|_{\Gamma^-} & = & \psi^b & \textit{on} & [\epsilon_{\min},\epsilon_{\max}]\times \Gamma^-, \\ \psi(\epsilon_{\max},x,\Omega) & = & 0 & \textit{in} & Z\times S^2, \end{array}\right. \tag{1.40}$$

where P is a generic Boltzmann loss term of the form (1.12) *characterized by a total cross section $0<\sigma_T\in L^\infty([\epsilon_{\min},\epsilon_{\max}])$.*

Suppose

$$0\le\psi^b\in L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-),\qquad 0\le q\in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2)$$

then (1.40) *has a unique solution ψ satisfying*

$$0\le\psi\in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2),\qquad \Omega.\nabla_x\psi\in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2),$$
$$0\le\psi|_{[\epsilon_{\min},\epsilon_{\max}]\times\Gamma^+}\in L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^+).$$

Proof Fix $\Omega\in S^2$ and choose a point $x_0\in\partial Z$. Along the line x_0 directed by Ω, the problem (1.40) turns into a first order linear ordinary differential equation (ODE) which is therefore well-posed. The unique solution to Problem (1.40) is

$$\begin{array}{rcl} \psi_0(\epsilon,x,\Omega) & = & \exp\left(-\int_0^{L(x,x_0(x,\Omega))}\rho(x+\Omega s)ds\sigma_T(\epsilon)\right)\psi^b(\epsilon,x_0(x,\Omega),\Omega) \\ & + & \int_0^{L(x,x_0(x,\Omega))}\exp\left(-\int_0^{L(x,y)}\rho(x+\Omega s)ds\sigma_T(\epsilon)\right)q(\epsilon,y,\Omega)dy, \\ L(x,y) & = & |x-y|, \end{array}$$

where $x_0(x,\Omega)$ is the point of intersection between the boundary ∂Z and the line passing through x and directed by Ω. See further discussion on (1.40) in [31]. □

One obtains a similar result with the following equation.

Lemma 1.2 ([31]) *Consider the equation*

$$\left\{\begin{array}{rcll} -\rho\partial_\epsilon(S\psi)+\Omega.\nabla_x\psi+\rho P(\psi) & = & q & in \quad [\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2, \\ \psi|_{\Gamma^-} & = & \psi^b & on \quad [\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-, \\ \psi(\epsilon_{\max},x,\Omega) & = & 0 & in \quad Z\times S^2, \end{array}\right. \tag{1.41}$$

where P is a generic Boltzmann loss term of the form (1.12) characterized by a total cross section $0<\sigma_T\in L^\infty([\epsilon_{\min},\epsilon_{\max}])$ and the stopping power $0<S\in C^1([\epsilon_{\min},\epsilon_{max}])$.

Suppose

$$0\le\psi^b\in L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-),\qquad 0\le q\in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2),$$

then (1.41) has a unique solution ψ satisfying

$$0\le\psi\in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2),$$
$$-\rho\partial(S\psi)+\Omega.\nabla_x\psi\in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2),$$
$$0\le\psi|_{[\epsilon_{\min},\epsilon_{\max}]\times\Gamma^+}\in L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^+),\quad 0\le\psi(\epsilon_{\min},.,.)\in L^2(Z\times S^2).$$

Proof This result can be obtained by adapting the proof of the Lemma 1.1 and using the method of characteristics. The reader is referred to [31] for further discussion on the problems (1.40) and (1.41) with non-zero boundary conditions. □

1.5.4 Well-posedness of scalar kinetic equations

In the following, the well-posedness of the kinetic system (1.11) with a zero inital condition (1.37) and a given boundary condition of the form (1.39) is proven when considering a LB collision term (1.14d), a LBCSD collision term (1.34) or a LBFP collision term (1.36) for the electron collisions. For this purpose, the following three lemmas are given. Each provides the well-posed of a kinetic equation with one of those collision operators.

Those propositions are proven using variational approaches. For this purpose, the following generalization of Lax-Milgram theorem is recalled.

Theorem 1.2 (Lions-Lax-Milgram)

Let H_1 be two Hilbert spaces and H_2 a normed space. Consider a bilinear form B on $H_1\times H_2$ and a linear form l on H_2.

Suppose the operator l is bounded, B is bounded and coercive in the sense

$$\textbf{\textit{Boundedness of l}}\qquad |l(\lambda)| \le C_1\,\|\lambda\|_{H_2}, \tag{1.42}$$

$$\textbf{\textit{Boundedness of B}}\qquad |B(\psi,\lambda)| \le C_2\,\|\psi\|_{H_1}\,\|\lambda\|_{H_2}, \tag{1.43}$$

$$\textbf{\textit{Coercivity of B}}\quad \forall\lambda\in H_2,\ \sup_{\|\psi\|_{H_1}\le1}|B(\psi,\lambda)| \ge C_3\,\|\lambda\|_{H_2}, \tag{1.44}$$

for some scalars $C_1 < \infty$, $C_2 < \infty$ and $C_3 > 0$.

Then there exists a (potentially non-unique) solution $\psi \in H_1$ to the problem

$$\forall \lambda \in H_2, \qquad B(\psi, \lambda) = l(\lambda). \tag{1.45}$$

The first equation considered is a linear Boltzmann equation.

Proposition 1.6 (Well-posedness of a linear Boltzmann equation) *Consider the problem*

$$\left\{ \begin{array}{rcll} \Omega.\nabla_x \psi + \rho(P-G)(\psi) & = & q & \textit{in} \quad [\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2, \\ \psi|_{\Gamma^-} & = & \psi^b & \textit{on} \quad [\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-, \\ \psi(\epsilon_{\max}, x, \Omega) & = & 0 & \textit{for} \quad (x, \Omega) \in Z \times S^2. \end{array} \right. \tag{1.46}$$

where G and P are linear Boltzmann gain and loss terms of the form (1.13) and (1.12).

Suppose that there exists a positive weight function

$$0 < h_{\min} \le h \in C^0([\epsilon_{\min}, \epsilon_{\max}]),$$

such that, together with density ρ and the cross sections σ and σ_T, it satisfies

$$0 \le \rho\sigma < +\infty, \qquad 0 < \alpha \le \rho\sigma_T < +\infty \tag{1.47a}$$

$$0 < \alpha \le \rho(x) \left(\sigma_T(\epsilon) - \int_{\epsilon}^{\epsilon_{\max}} \int_{S^2} \sigma(\epsilon, \epsilon', \Omega.\Omega') d\Omega' d\epsilon' \right) \tag{1.47b}$$

$$0 < \alpha \le \rho(x) \left(h(\epsilon)\sigma_T(\epsilon) - \int_{\epsilon_{\min}}^{\epsilon} \int_{S^2} h(\epsilon')\sigma(\epsilon, \epsilon', \Omega.\Omega') d\Omega' d\epsilon' \right) \tag{1.47c}$$

for some positive scalar $\alpha > 0$. Suppose furthermore that

$$\psi^b \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-), \qquad q \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2),$$

then the problem (1.46) has a unique solution ψ satisfying

$$\psi \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2), \qquad \Omega.\nabla_x \psi \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2),$$
$$\psi|_{[\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^+} \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^+).$$

Proof First, we exploit the linearity of the kinetic equation (1.46).

Consider the two problems

$$\left\{ \begin{array}{rcll} \Omega.\nabla_x \psi_0 + \rho P(\psi_0) & = & q & \text{in} \quad [\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2, \\ \psi_0|_{\Gamma^-} & = & \psi^b & \text{on} \quad [\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-, \\ \psi(\epsilon_{\max}, x, \Omega) & = & 0 & \text{in} \quad Z \times S^2. \end{array} \right. \tag{1.48}$$

and

$$\left\{\begin{array}{rcll}\Omega.\nabla_x\psi_s+\rho(P-G)(\psi_s) & = & G(\psi_0) & \text{in} \quad [\epsilon_{\text{min}},\epsilon_{\text{max}}]\times Z\times S^2,\\ \psi_s|_{\Gamma^-} & = & 0 & \text{on} \quad [\epsilon_{\text{min}},\epsilon_{\text{max}}]\times\Gamma^-,\\ \psi_s(\epsilon_{\text{max}},x,\Omega) & = & 0 & \text{in} \quad Z\times S^2.\end{array}\right. \tag{1.49}$$

As Problem (1.46) is linear, proving the existence of a unique solution to (1.48) and to (1.49) provides the existence of a unique solution to (1.46) which has the form

$$\psi=\psi_0+\psi_s$$

where ψ_0 solves (1.48) and ψ_s solves (1.49). This decomposition corresponds to the first order expansion in order of scattering ([19, 9, 17]), ψ_0 is the fluence of the particles that have never scattered and ψ_s the fluence of the other particles.

The well-posedness of (1.48) follows from Lemma 1.1. Due to hypothesis (1.47), G sends $L^2([\epsilon_{\text{min}},\epsilon_{\text{max}}]\times Z\times S^2)$ into itself, therefore $G(\psi_0)\in L^2([\epsilon_{\text{min}},\epsilon_{\text{max}}]\times Z\times S^2)$.

Well-posedness of (1.49)**:** Choose a regular positive weight function $0<h\in C^0([\epsilon_{\text{min}},\epsilon_{\text{max}}])$. In the spirit of [31], define the following inner products

$$(\psi,\lambda)_{H_1} = (\psi,\lambda)_i+(\Omega.\nabla_x\psi,\Omega.\nabla_x\lambda)_i+(\psi,\lambda)_{b_+}\,, \tag{1.50a}$$

$$(\psi,\lambda)_{H_2} = (\psi,\lambda)_i+(\psi,\lambda)_{b_+}\,, \tag{1.50b}$$

and the spaces

$$\begin{array}{rcl}H_1 & := & \left\{\psi\in L^2([\epsilon_{\text{min}},\epsilon_{\text{max}}]\times Z\times S^2),\ \text{s.t.}\ \|\psi\|_{H_1}<\infty \text{ and } \|\psi\|_{b_-}=0\right\},\\ H_2 & := & \left\{\psi\in L^2([\epsilon_{\text{min}},\epsilon_{\text{max}}]\times Z\times S^2),\ \text{s.t.}\ \|\psi\|_{H_2}<\infty \text{ and } \|\psi\|_{b_-}=0\right\}.\end{array}$$

The set H_1 with the inner product $(,)_{H_1}$ is a Hilbert space, see [31] for a proof of completeness of H_1.

Define the bilinear form B and the linear form l as

$$\begin{array}{rcl}B(\psi,\lambda) & = & (\Omega.\nabla_x\psi,\lambda)_i^h+(\psi,\lambda)_{b_+}^h+(\rho(P-G)(\psi),\lambda)_i^h\,,\\ l(\lambda) & = & (q,\lambda)_i^h\,.\end{array}$$

Firstly, the operator l is bounded using Cauchy-Schwartz inequalities

$$|l(\lambda)|\leq\|q\|_i\|\lambda\|_i\leq C_1\|\lambda\|_{H_2},$$

and is therefore continuous in the sense (1.42).

Secondly, the operator B is proven to be bounded in the sense (1.43). Using Cauchy-Schwartz inequality leads to

$$\begin{array}{rcl}|\,(\Omega.\nabla_x\psi,\lambda)_i^h\,| & \leq & \|\Omega.\nabla_x\psi\|_i^h\|\lambda\|_i^h\leq\|h\|_\infty\|\psi\|_{H_1}\|\lambda\|_{H_2},\\ |(\rho(P-G)(\psi),\lambda)_i^h| & \leq & \alpha\|\psi\|_i\|\lambda\|_i\leq\alpha\|\psi\|_{H_1}\|\lambda\|_{H_2},\end{array}$$

for some positive scalar

$$\begin{aligned}\alpha \quad \geq \quad & \max_{x\in Z}\rho(x)(\max_{\epsilon\in[\epsilon_{\min},\epsilon_{\max}]} h(\epsilon)\sigma_T(\epsilon) \\ & + \max_{\epsilon\in[\epsilon_{\min},\epsilon_{\max}]} \max\left(h(\epsilon)\int_{\epsilon}^{\epsilon_{\max}}\int_{S^2}\sigma(\epsilon',\epsilon,\Omega'.\Omega)d\Omega' d\epsilon',\right. \\ & \left.\int_{\epsilon_{\min}}^{\epsilon}\int_{S^2} h(\epsilon')\sigma(\epsilon,\epsilon',\Omega.\Omega')d\Omega' d\epsilon'\right)).\end{aligned}$$

Therefore one obtains

$$|B(\psi,\lambda)| \leq C_2\|\psi\|_{H_1}\|\lambda\|_{H_2},$$

with $0 < C_2 < \infty$.

Thirdly, B is proven to be coercive in the sense (1.44). One remarks that $H_2 \subset H_1$, and therefore one may use $\psi \in H_1$ as a test function. Using an integration by parts and Green theorem, one obtains

$$(\Omega.\nabla_x\psi,\psi)_i^h \quad = \quad \int_Z\int_{S^2}\int_{\epsilon_{\min}}^{\epsilon_{\max}} h(\epsilon)\Omega.\nabla_x\left(\frac{\psi^2}{2}\right)d\epsilon d\Omega dx = \frac{(\|\psi\|_{b_+}^h)^2}{2}.$$

Using Fubini theorem leads to

$$\begin{aligned}(G(\psi),\psi) = \int_Z\int_{S^2}\int_{\epsilon_{\min}}^{\epsilon_{\max}}\int_{S^2}\int_{\epsilon}^{\epsilon_{\max}} \quad & h(\epsilon)\sigma(\epsilon',\epsilon,\Omega'.\Omega)\psi(\epsilon',x,\Omega') \\ & \psi(\epsilon,x,\Omega)d\epsilon' d\Omega' d\epsilon d\Omega dx,\end{aligned}$$

then a Cauchy-Schwartz inequality provides

$$(G(\psi),\psi) \quad \leq \quad \left(\int_Z\int_{\epsilon_{\min}}^{\epsilon_{\max}}\int_{S^2} f_1(\epsilon)\psi(\epsilon,x,\Omega)^2 d\epsilon d\Omega dx\right)^{\frac{1}{2}} \tag{1.51}$$

$$* \left(\int_Z\int_{\epsilon_{\min}}^{\epsilon_{\max}}\int_{S^2} f_2(\epsilon)\psi(\epsilon,x,\Omega)^2 d\epsilon d\Omega dx\right)^{\frac{1}{2}},$$

$$f_1(\epsilon) \quad = \quad h(\epsilon)\int_{S^2}\int_{\epsilon}^{\epsilon_{\max}}\sigma(\epsilon',\epsilon,\Omega'.\Omega)d\epsilon' d\Omega', \tag{1.52}$$

$$f_2(\epsilon) \quad = \quad \int_{S^2}\int_{\epsilon_{\min}}^{\epsilon} h(\epsilon')\sigma(\epsilon,\epsilon',\Omega.\Omega')d\epsilon' d\Omega'. \tag{1.53}$$

Together with hypothesis (1.47), this leads to

$$B(\psi,\psi) \geq C_3\|\psi\|_{H_2}^2, \tag{1.54}$$

for some constant C_3 satisfying

$$0 < C_3 \quad \leq \quad \min\left(\frac{1}{2}, \min_{\epsilon\in[\epsilon_{\min},\epsilon_{max}]}\sigma_T(\epsilon)-f_1(\epsilon), \min_{\epsilon\in[\epsilon_{\min},\epsilon_{max}]}\sigma_T(\epsilon)-f_2(\epsilon)\right).$$

Therefore (1.44) holds. Thus Lions-Lax-Milgram theorem 1.2 provides the existence of a solution $\psi \in H_1$.

Suppose $\psi \in H_1$ and $\psi' \in H_1$ are two solutions to (1.46), then

$$B(\psi, \lambda) = l(\lambda) = B(\psi', \lambda),$$

and therefore, by bilinearity of B, one has

$$B(\psi - \psi', \lambda) = 0. \tag{1.55}$$

According to the definition of H_1 and H_2, and Sobolev embedding theorem, one has $H_1 \subset H_2$, and therefore $\psi - \psi' \in H_2$. Replacing $\lambda = \psi - \psi'$ in (1.55) reads to

$$B(\psi - \psi', \psi - \psi') = 0.$$

According to (1.54), this leads to $\|\psi - \psi'\|_{H_2} = 0$ which is enough to provide the uniqueness of the solution in H_1. □

Remark 1.9

- *An alternative proof was written using the theory of evolution equation (semi-groups theory, see e.g. [32, 4]) in [10, 31, 15] under the condition (1.47) with $h = 1$. Such a condition is too constraining for the present application. However, this method provides the non-negativity of the solution $0 \leq \psi$ under the additional non-negativity conditions on the sources*

$$0 \leq \psi^b, \qquad 0 \leq q.$$

- *The physical meaning of condition (1.47c) is that the medium absorbes the quantity h, i.e. if (1.47c) is satisfied with $h(\epsilon) = 1$ the medium absorbes particles. In practice, due to Compton and Møller effects, there is creation of particles and therefore those conditions are not satisfied. However, some quantities are absorbed by the medium. Indeed, according to the physics described by the model (1.11), a part of the energy of the particles is absorbed by the medium, and none is created. This idea will be used in the following and it corresponds to choosing $h(\epsilon) = \epsilon$ in (1.47c).*

- *The condition (1.47) can be sharpened by the studying the equation satisfied by the quantity $h_2(\epsilon)\psi$ for some positive weight function $0 < h_2 \in C^0([\epsilon_{\min}, \epsilon_{\max}])$. This idea was used in [31] with the choice $h_2(\epsilon) = \exp(C\epsilon)$ with some constant C. Remark that, when chosing $h_2(\epsilon) = \epsilon$, the quantity $\epsilon\psi$ corresponds to the specific intensity in the field of radiative transfer (see e.g. [28, 23]), and one may rewrite (1.11) with this quantity for unknown.*

Similarily, the following lemma provides the well-posedness of the kinetic equation for the electons when using the CSDA.

Proposition 1.7 (Well-posedness of a LBCSD equation) *Consider the problem*

$$\left\{\begin{array}{rcll} -\rho\partial_\epsilon(S\psi) & + & \Omega.\nabla_x\psi & +\rho(P_{el}-G_{el}-G)(\psi)=q \\ & & & \textit{in} \quad [\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2, \\ \psi|_{\Gamma^-} & = & \psi_b & \textit{on} \quad [\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-, \\ \psi(\epsilon_{\max},x,\Omega) & = & 0 & \textit{for} \quad (x,\Omega)\in Z\times S^2, \end{array}\right. \tag{1.56}$$

where P_{el} and G_{el} are linear Boltzmann loss and gain terms of the form (1.12) and (1.13) with cross sections of the form (1.32) and (1.33) and G is a linear Boltzmann gain term of the form (1.13).

Suppose that there exists a positive function $0 < h \in C^1([\epsilon_{\min},\epsilon_{\max}])$ such that together with the density ρ, the cross sections σ_{el}, $\sigma_{T,el}$, σ and the stopping power S, it satisfies

$$0 \le \rho\sigma_{el} < +\infty, \qquad 0 < \alpha \le \rho\sigma_{T,el} < +\infty, \qquad 0 \le \rho\sigma < +\infty, \tag{1.57a}$$

$$0 < \alpha \le S \in C^1([\epsilon_{\min},\epsilon_{\max}]), \tag{1.57b}$$

$$0 < \alpha \le \rho\frac{S(\epsilon)h'(\epsilon)-h(\epsilon)S'(\epsilon)}{2} - h(\epsilon)\int_\epsilon^{\epsilon_{\max}}\int_{S^2}\sigma(\epsilon',\epsilon,\Omega'.\Omega)d\Omega'd\epsilon', \tag{1.57c}$$

$$0 < \alpha \le \rho\frac{S(\epsilon)h'(\epsilon)-h(\epsilon)S'(\epsilon)}{2} - \int_\epsilon^{\epsilon_{\max}}\int_{S^2}h(\epsilon')\sigma(\epsilon,\epsilon',\Omega.\Omega')d\Omega'd\epsilon', \tag{1.57d}$$

for some positive scalar $0 < \alpha$. Suppose furthermore that

$$\psi^b \in L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-), \qquad q \in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2),$$

then the problem (1.56) has a unique solution ψ satisfying

$$\psi \in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2),$$
$$-\rho\partial_\epsilon(S\psi)+\Omega.\nabla_x\psi \in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2),$$
$$\psi|_{[\epsilon_{\min},\epsilon_{\max}]\times\Gamma^+} \in L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^+), \qquad \psi(\epsilon_{\min},.,.) \in L^2(Z\times S^2).$$

Proof One obtains this results by adapting the proof of Proposition 1.6 to the present equation.

Using the expansion

$$\psi = \psi_0 + \psi_s$$

where ψ_0 and ψ_s solve the following problems

$$\left\{\begin{array}{rcll} -\rho\partial_\epsilon(S\psi_0)+\Omega.\nabla_x\psi_0+\rho P_{el}(\psi) &=& q, & \text{in} \quad [\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2 \\ \psi_0 &=& \psi^b & \text{on} \quad [\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-, \\ \psi_0(\epsilon_{\max},x,\Omega) &=& 0 & \text{for} \quad (x,\Omega)\in Z\times S^2, \end{array}\right. \tag{1.58a}$$

$$\left\{\begin{array}{rcl} -\rho\partial_\epsilon(S\psi_s) &+& \Omega.\nabla_x\psi_s \\ &+& \rho(P_{el}-G_{el}-G)(\psi_s)=q+\rho\left(G_{el}(\psi_0)+G(\psi_0)\right), \\ && \text{in} \quad [\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2 \\ \psi_s &=& 0 \quad \text{on} \quad [\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-, \\ \psi_s(\epsilon_{\max},x,\Omega) &=& 0 \quad \text{for} \quad (x,\Omega)\in Z\times S^2. \end{array}\right. \tag{1.58b}$$

Using Lemma 1.2, there exists a unique solution ψ_0 to the problem (1.58a) in $L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2)$.

As in the previous proof, one rewrites the inner products

$$\begin{array}{rcl} (\psi,\lambda)_{H_1} &=& (\psi,\lambda)_i+(\psi,\lambda)_{b_+}+(\psi,\lambda)_{\epsilon_{\min}} \\ && +(\Omega.\nabla_x\psi,\Omega.\nabla_x\lambda)_i+(\partial_\epsilon\psi,\partial_\epsilon\lambda)_i, \\ (\psi,\lambda)_{H_2} &=& (\psi,\lambda)_i+(\psi,\lambda)_{b_+}+(\psi,\lambda)_{\epsilon_{\min}}, \end{array}$$

and the Hilbert spaces

$$\begin{array}{rcl} H_1 &:=& \Big\{\psi\in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2), \quad \text{s.t.} \quad \|\psi\|_{H_1}<\infty, \ \|\psi\|_{b_-}=0 \\ && \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{and } \|\psi\|_{\epsilon_{\max}}=0\Big\}, \\ H_2 &:=& \Big\{\psi\in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2), \quad \text{s.t.} \quad \|\psi\|_{H_2}<\infty, \ \|\psi\|_{b_-}=0 \\ && \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{and } \|\psi\|_{\epsilon_{\max}}=0\Big\}. \end{array}$$

Choose the bilinear form

$$B(\psi,\lambda)=(-\rho\partial_\epsilon(S\psi)+\Omega.\nabla_x\psi+\rho(P_{el}-G_{el}-G)(\psi),\lambda)_i^h.$$

Adapting the computations of the proof of Proposition 1.6 leads to

$$\begin{array}{rcl} (-\rho\partial_\epsilon(S\psi),\lambda)_i^h &\leq& \alpha\|\partial_\epsilon\psi\|_i\|\lambda\|_i, \\ (\Omega.\nabla_x\psi,\lambda)_i^h &\leq& \|h\|_\infty\|\Omega.\nabla_x\psi\|_i\|\lambda\|_i, \\ ((P_{el}-G_{el}-G)(\psi),\lambda)_i^h &\leq& \beta\|\psi\|_i\|\lambda\|_i, \\ |B(\psi,\lambda)| &\leq& C_1\|\psi\|_{H_1}\|\lambda\|_{H_2}, \end{array}$$

with

$$\begin{array}{rcl} C_1 &=& \min(1,\ \alpha,\ \beta), \\ \alpha &\geq& \max\limits_{x\in Z}\rho(x) \max\limits_{\epsilon\in[\epsilon_{\min},\epsilon_{\max}]} |h(\epsilon)S(\epsilon)|, \\ \beta &\geq& \max\limits_{x\in Z}\rho(x) \max\limits_{\epsilon\in[\epsilon_{\min},\epsilon_{\max}]} \left[|h(\epsilon)S'(\epsilon)| + \max(f_1(\epsilon), f_2(\epsilon))\right], \\ f_1(\epsilon) &=& h(\epsilon)\displaystyle\int_{S^2}\int_{\epsilon}^{\epsilon_{\max}} \sigma(\epsilon',\epsilon,\Omega'.\Omega)d\epsilon' d\Omega', \\ f_2(\epsilon) &=& \displaystyle\int_{S^2}\int_{\epsilon_{\min}}^{\epsilon} h(\epsilon')\sigma(\epsilon,\epsilon',\Omega.\Omega')d\epsilon' d\Omega'. \end{array}$$

For the coercivity of B, using a product rule and an integration by part leads to write

$$\begin{array}{rcl} (-\rho\partial_\epsilon(S\psi),\psi)_i^h &=& (-S', h\psi^2)_i + (-\partial_\epsilon(\psi), hS\psi)_i, \\ (-\rho\partial_\epsilon(S\psi),\psi)_i^h &=& (-\partial_\epsilon(Sh\psi^2))_i + (S\psi, \partial_\epsilon(h\psi))_i \\ &=& (-\partial_\epsilon(Sh\psi^2))_i + (S\psi^2, h')_i + (Sh\psi, \partial_\epsilon(\psi))_i, \end{array}$$

and therefore

$$\begin{array}{rcl} (-\rho\partial_\epsilon(S\psi),\psi)_i^h &=& \frac{1}{2}\Big[(-\rho(Sh\psi^2))_{\epsilon_{\max}} - (-\rho(Sh\psi^2))_{\epsilon_{\min}} \\ && +(\psi^2, \rho(Sh' - hS'))_i\Big] \\ &=& \frac{1}{2}\left[(\rho(Sh\psi^2))_{\epsilon_{\min}} + (\psi^2, \rho(Sh' - hS'))_i\right]. \end{array} \quad (1.59)$$

Using a Cauchy-Schwartz inequality as in (1.51) leads to

$$(\rho G(\psi),\psi)_i^h \leq \int_Z \rho(x)\int_{S^2}\int_{\epsilon_{\min}}^{\epsilon_{\max}} \max(f_1(\epsilon), f_2(\epsilon))\psi^2(\epsilon,x,\Omega)d\epsilon d\Omega dx. \quad (1.60)$$

Writig together (1.59) and (1.60) leads to the coercivity of B under the condition (1.57). The rest of the proof is identical to the one of Proposition 1.6. □

Finally this can also be proven when using the Fokker-Planck equation with linear Boltzmann gain term.

Proposition 1.8 (Well-posedness of a LBFP equation)

Consider the problem

$$\left\{\begin{array}{rclll} -\rho\partial_\epsilon(S\psi) &+& \Omega.\nabla_x\psi & -\rho T\Delta_\Omega\psi - \rho G(\psi) = q & \\ &&& \textit{in} \quad [\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2, & \\ \psi|_{\Gamma^-} &=& \psi_b & \textit{on} \quad [\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-, & \\ \psi(\epsilon_{\max},x,\Omega) &=& 0 & \textit{for} \quad (x,\Omega)\in Z\times S^2, & \end{array}\right. \quad (1.61)$$

where G is a linear Boltzmann gain term of the form (1.13).

Suppose that the density ρ, the cross section σ, the stopping power S satisfy

(1.57) *and furthermore that the transport coefficient T satisfy*

$$0 < T \in C^0([\epsilon_{\min}, \epsilon_{\max}]), \tag{1.62}$$

and that

$$\psi^b \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-), \qquad q \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2),$$

then the problem (1.61) *has a unique solution ψ satisfying*

$$\begin{gathered} \psi \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2), \\ -\rho\partial_\epsilon(S\psi) + \Omega.\nabla_x\psi - \rho T \Delta_\Omega \psi \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2), \\ \psi|_{[\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^+} \in L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^+), \qquad \psi(\epsilon_{\min}, ., .) \in L^2(Z \times S^2). \end{gathered}$$

Proof This result can be obtained by adapting the proof of Propositions 1.6 and 1.7 . □

Using the theory of evolution equations, *i.e.* here the Hille-Yosida theorem (see *e.g.* [32, 4]), one may obtain a better regularity of the solution and its non-negativity under the conditions

$$0 \le \psi^b, \qquad 0 \le q.$$

In the rest of the manuscript, the non-negativity of the solution is always assumed.

1.5.5 Well-posedness of the system of transport equations

Those three lemma lead to the existence of a solution to the problem (1.11).

Theorem 1.3 *Consider the problem* (1.11) *with the initial-boundary conditions*

$$\begin{cases} \psi_\gamma|_{\Gamma^-} &= \psi_{\gamma b} \ge 0, \qquad \psi_e|_{\Gamma^-} = \psi_{eb} \ge 0, \\ \psi_\gamma(\epsilon_{\max}, x, \Omega) &= \psi_e(\epsilon_{\max}, x, \Omega) = 0. \end{cases} \tag{1.63}$$

and where the electrons collision operator $Q_{e\to e}$ is the LB collision term (1.14d) *or the LBCSD one* (1.34) *or the LBFP one* (1.36).

Then this problem has a unique solution (ψ_γ, ψ_e) satisfying

$$\begin{aligned} (\psi_\gamma, \psi_e) &\in (L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2))^2, \\ \Omega.\nabla_x\psi_\gamma - \rho Q_{\gamma\to\gamma}(\psi_\gamma) &\in L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2), \\ \Omega.\nabla_x\psi_e - \rho(Q_{e\to e}(\psi_e) + Q_{\gamma\to e}(\psi_\gamma)) &\in L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2), \\ (\psi_\gamma, \psi_e)|_{[\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^+} &\in (L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^+))^2, \\ \psi_e(\epsilon_{\max}, ., .) &\in L^2(Z \times S^2). \end{aligned}$$

Proof One first remarks that the equations for photons and electrons are decoupled. One can first solve the equation for the photons to obtain ψ_γ and then use it to compute $G_{\gamma\to e}(\psi_\gamma)$ which is simply a source term in the electron equation.

The equation for the photons has the form (1.46) and under the constraints (1.47) on the cross sections $\sigma_{\gamma\to\gamma}$ and $\sigma_{T,\gamma}$, one obtains the existence of a unique solution $\psi_\gamma \in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2)$ from Proposition 1.6.

One verifies that $G_{\gamma\to e}(\psi_\gamma) \in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2)$ as long as $\psi_\gamma \in L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2)$.

When using a LB collision term (1.14d), under the constraints (1.47) on the cross sections, one may use Proposition 1.6 to obtain the result.

When using a LBCSD collision term (1.34), under the constraints (1.57a) and (1.57b) on the cross sections and the stopping power satisfy, one may use Proposition 1.7 to obtain the result.

When using a LBFP collision term (1.36), under the constraints (1.57a), (1.57b) and (1.62) on the cross sections, the stopping power and the transport coefficient, one may use proposition 1.8 to obtain the result. □

1.5.6 Deposited dose

The function of interest in medical physics is the quantity of energy deposited per unit of mass, locally in space, by the system of transported particles, so-called dose $D(x)$. The biological impact, *e.g.* for the linear-quadratic model ([30]) the proportion of cells surviving to the treatement, is assumed to be a known function of this dose. This quantity of energy transfered to the medium is the quantity of energy lost by the system of particles. Based on the collision operators, one can obtain the dose by computing a global balance of energy. It reads

$$\begin{aligned} D(x) = \int_{\epsilon_{min}}^{\epsilon_{max}}\int_{S^2} -\epsilon \quad & [Q_{\gamma\to\gamma}(\psi_\gamma) + Q_{\gamma\to e}(\psi_\gamma) \\ & +Q_{e\to\gamma}(\psi_e) + Q_{e\to e}(\psi_e)] \quad (\epsilon, x, \Omega) d\Omega d\epsilon. \end{aligned}$$

When using a linear Boltzmann collision operator for the collisions $e \to e$, using (1.17) and (1.25) leads to simplify the dose into

$$\begin{aligned} D(x) = & \int_{\epsilon_{min}}^{\epsilon_{max}}\int_{S^2}\Big[\int_{\epsilon_B}^{\epsilon}\int_{S^2} (\epsilon-\epsilon')\sigma_{C,\gamma}(\epsilon,\epsilon',\Omega.\Omega') \\ & + \qquad \epsilon'\sigma_{C,e}(\epsilon,\epsilon',\Omega.\Omega')\ d\Omega' d\epsilon'\Big]\psi_\gamma(\epsilon,x,\Omega) \\ & + \Big[\int_{\epsilon_B}^{\epsilon}\int_{S^2} (\epsilon-\epsilon')\sigma_{M,1}(\epsilon,\epsilon',\Omega.\Omega') \\ & + \qquad \epsilon'\sigma_{M,2}(\epsilon,\epsilon',\Omega.\Omega')\ d\Omega' d\epsilon'\Big]\psi_e(\epsilon,x,\Omega) d\Omega d\epsilon. \end{aligned} \tag{1.64}$$

In the formula (1.64), one can identify the stopping power S defined in (1.35b), and, therefore, when considering a LBCSD operator (1.34) or LBFP operator (1.35a), the formula for the dose is the same.

Another method to compute the dose consists in computing the energetic balance in a spatial cell C_h centered in x of radius h. Then having h tends to zero provides another definition of the quantity of energy deposited at point x. The flux of energy on the boundary ∂C_h is $\epsilon\Omega\psi$, then using the divergence theorem leads to

$$\begin{aligned} D(x) &= \lim_{h\to 0}\frac{1}{\text{Mass}(C_h)}\int_{\epsilon_{min}}^{\epsilon_{max}}\oint_{\partial C_h}\epsilon(\Omega.n)\psi(x,\epsilon,\Omega)dS(x)d\Omega d\epsilon \\ &= \lim_{h\to 0}\frac{1}{\text{Mass}(C_h)}\int_{\epsilon_{min}}^{\epsilon_{max}}\int_{S^2}\int_{C_h}\epsilon\Omega.\nabla_x\psi(\epsilon,x,\Omega)dxd\Omega d\epsilon \\ &= \int_{\epsilon_{min}}^{\epsilon_{max}}\int_{S^2}\frac{\epsilon}{\rho(x)}\Omega.\nabla_x\psi(\epsilon,x,\Omega)d\Omega d\epsilon, \end{aligned}$$

where $n(x)$ is the outgoing normal to ∂C_h at point x, $dS(x)$ is the Lebesgue measure on the boundary of the cell ∂C_h (*i.e.* on rectangles) and Mass(C_h) is the mass of particles of the medium contained in C_h, *i.e.*

$$\text{Mass}(C_h) = \int_{C_h}\rho(x)dx.$$

Remark that both definitions of the dose are equivalent according to (1.11).

In order to compare the dose obtained with different methods, it is often convenient to normalize the dose with the maximum dose, so-called Percentage Depth Dose (PDD)

$$PDD(x) = \frac{D(x)}{\max(D)}.$$

1.5.7 Reduction to 1D problems

As a first approach, it is often easier and more convenient to start studying one dimensional problems instead of three dimensional ones. One dimensional problems correspond here to the transport of particles in slab geometry. In 1D, (1.11) turns into

$$\begin{aligned} \mu\partial_x\psi_\gamma(\epsilon,x,\mu) &= \rho(x)\left(Q_{\gamma\to\gamma}(\psi_\gamma)+Q_{e\to\gamma}(\psi_e)\right)(\epsilon,x,\mu), && \text{(1.65a)} \\ \mu\partial_x\psi_e(\epsilon,x,\mu) &= \rho(x)\left(Q_{e\to e}(\psi_e)+Q_{\gamma\to e}(\psi_\gamma)\right)(\epsilon,x,\mu), && \text{(1.65b)} \end{aligned}$$

with the collision operator in (1.14) and where the gain terms proposed in (1.13) in 1D are replaced by

$$\begin{aligned} G_{\alpha\to\beta}(\psi_\alpha)(\epsilon,x,\mu) &= \int_{\epsilon}^{\epsilon_{max}}\int_{-1}^{+1}\sigma^{1D}_{\alpha\to\beta}(\epsilon',\epsilon,\mu',\mu)\psi_\alpha(\epsilon',x,\mu')d\mu' d\epsilon', \\ \sigma^{1D}_{\alpha\to\beta}(\epsilon',\epsilon,\mu',\mu) &= \int_{\phi\in[0,2\pi]}\sigma_{\alpha\to\beta}(\epsilon',\epsilon,\mu'\mu+\sqrt{1-\mu'^2}\sqrt{1-\mu^2}\cos\phi)d(\cos\phi), \end{aligned}$$

the loss term are unchanged (*i.e.* given by (1.12)). The differential and total cross sections are defined in Subsections 1.4.2. The approximations of the linear Boltzmann operator in 1D leads to write the LBCSD (1.34) and the LBFP (1.35a) operators

$$\begin{aligned} Q_{LBCSD}(\psi)(\epsilon,x,\mu) &= \left[\partial_\epsilon(S\psi)+(G_{el}-P_{el})(\psi)\right](\epsilon,x,\mu), && \text{(1.65c)} \\ Q_{LBFP}(\psi)(\epsilon,x,\mu) &= \left[\partial_\epsilon(S\psi)+\partial_\mu\left((1-\mu^2)\partial_\mu\psi\right)\right](\epsilon,x,\mu). && \text{(1.65d)} \end{aligned}$$

1.6 Conclusion

The aim of the present thesis is to propose a numerical approach for computing the dose (1.64) and eventually to optimize this dose.

The model (1.11) with the different collision operators is the fundation of the numerical methods for dose computation and optimization in this manuscript.

However, discretizing directly such an equation leads to numerical methods (typically discrete ordinate methods, see *e.g.* [22]) that require very large computational ressources due to the high dimensionality of (1.11).

In order to reduce these computational costs, the method of moments is introduced in the next part. This method provides lower dimensional models. Part III presents numerical methods adapted to such moment models which are shown to require less computational power than some reference methods.

Bibliography

[1] R. Barnard, M. Frank, and M. Herty. Optimal radiotherapy treatment planning using minimum entropy models. *Appl. Math. Comput.*, 219(5):2668 – 2679, 2012.

[2] J. Barò, J. Sempau, J. M. Fernández-Varea, and F. Salvat. PENELOPE: An algorithm for Monte Carlo simulation of the penetration and energy loss of electrons in matter. *Nuclear instruments and methods*, 100:31–46, 1995.

[3] J. Barò, J. Sempau, J. M. Fernández-Varea, and F. Salvat. PENELOPE: An algorithm and computer code for Monte Carlo simulation of electron-photon shower. *Ciemat technical report*, pages 31–46, 1996.

[4] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations.* Springer, 2011.

[5] C. Cercignani. *The Boltzmann equation and its applications.* Springer, 1988.

[6] C. Cercignani. *Ludwig Boltzmann, The man who trusted atoms.* Oxford, 1998.

[7] C. Cercignani. *The relativistic Boltzmann equation: Theory and applications.* Birkäuser, 2002.

[8] C. Cercignani, R. Illner, and M. Pulvirenti. *The mathematical theory of dilute gases.* Springer, 1994.

[9] T. K. Das and Ó. López Pouso. New insights into the numerical solution of the Boltzmann transport equation for photons. *Kin. Rel. Mod.*, 7(3):433–461, 2014.

[10] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology: Volume 6, Evolution problems II.* Springer, 2000.

[11] C. M. Davisson and R. D. Evans. Gamma-ray absorption coefficients. *Rev. Mod. Phys.*, 1952.

[12] R. Duclous. *Modélisation et simulation numérique multi-échelle du transport cinétique électronique.* PhD thesis, Université de Bordeaux, 2009.

[13] R. Duclous, B. Dubroca, and M. Frank. A deterministic partial differential equation model for dose calculation in electron radiotherapy. *Phys. Med. Biol.*, 55:3843–3857, 2010.

[14] J. Sempau F. Salvat, J. M. Fernández-Varea. *PENELOPE-2011: A code system for Monte Carlo simulation of electron and photon transport*, 2011.

[15] M. Frank, M. Herty, and A. N. Sandjo. Optimal radiotherapy treatment planning governed by kinetic equations. *Math. Mod. Meth. Appl. S.*, 20(04):661–678, 2010.

[16] M. Frank, M. Herty, and M. Schäfer. Optimal treatment planning in radiotherapy based on Boltzmann transport calculations. *Math. Mod. Meth. Appl. S.*, 18(04):573–592, 2008.

[17] H. Hensel, R. Iza-Teran, and N. Siedow. Deterministic model for dose calculation in photon radiotherapy. *Phys. Med. Biol.*, 51:675–693, 2006.

[18] I. Kawrakow, E. Mainegra-Hing, D. W. O. Rogers, F. Tessier, and B. R. B. Walters. *The EGSnrc code system: Monte Carlo simulation of electron and photon transport*, 2013.

[19] E. W. Larsen. Solution of neutron transport problems in L_1^*. *Commun. Pur. Appl. Math.*, 28(6):729–746, 1975.

[20] E. W. Larsen, M. M. Miften, B. A. Fraass, and I. A. Bruinvis. Electron dose calculations using the method of moments. *Med. Phys.*, 24(1):111–125, 1997.

[21] T. Leroy, R. Duclous, B. Dubroca, V.T. Tikhonchuk, S. Brull, A. Decoster, M. Lobet, and L. Gremillet. Deterministic and stochastic kinetic descriptions of electron-ion Bremsstrahlung: From thermal to non-thermal regimes. *submitted*, 2016.

[22] E. E. Lewis and W. F. Miller. *Computational methods of neutron transport.* American nuclear society, 1993.

[23] D. Mihalas and B. W. Mihalas. *Foundations of radiation hydrodynamics.* Dover publications, inc, 1984.

[24] N. F. Mott and H. S. W. Massey. *The theory of atomic collisions.* Oxford, 1965.

[25] J. W. Motz, H. A. Olsen, and H. W. Koch. Pair production by photons. *Rev. Mod. Phys.*, 41:581–639, Oct 1969.

[26] E. Olbrant. *Models and numerical methods for time- and energy-dependent particle transport.* PhD thesis, Rheinisch-Westfälische Technische Hochschule, 2012.

[27] E. Olbrant and M. Frank. Generalized Fokker–Planck theory for electron and photon transport in biological tissues: Application to radiotherapy. *Comput. Math. Method. M.*, 11(4):313–339, 2010. PMID: 20924856.

[28] G. C. Pomraning. *The equations of radiation hydrodynamics.* Pergamon Press, 1973.

[29] G. C. Pomraning. The Fokker-Planck operator as an asymptotic limit. *Math. Mod. Meth. Appl. S.*, 2(1):21–36, 1991.

[30] C. P. South, M. Partridge, and P. M. Evans. A theoretical framework for prescribing radiotherapy dose distributions using patient-specific biological information. *Med. Phys.*, 35(10):4599–4611, 2008.

[31] J. Tervo, P. Kokkonen, M. Frank, and M. Herty. On existence of L^2-solutions of coupled Boltzmann continuous slowing down transport equation system. *arxive*, 2016.

[32] K. Yosida. *Functional analysis.* Springer, 1995.

Part II

Angular moments

Chapter 2

Moment models

2.1 Introduction

The method of moments is used in a very large of physics. It was first used for linear transport equations in astrophysics ([10, 11]) and in radiative transfer ([12], see also *e.g.* [35, 14, 30, 24, 8, 6] for more recent applications). It has been afterwhile applied in a large range of fields of physics, such as in fluid dynamics, particularily for rarefied gas dynamics (see *e.g.* [20, 31, 28, 34, 21]), in quantum mechanics with application to the study of the semi-conductors (see *e.g.* [3, 36, 25]) and in plasma physics with application for the inertial or magnetic confinement fusion (see *e.g.* [32, 33, 22, 23, 16]).

The method of moments is a type of model reduction of kinetic equations. The purpose of this method is to reduce of the number of degrees of freedom, *i.e.* the number of variables. In practice, it consists in studying some integrals, afterward called moments, of the unknowns ψ_γ and ψ_e instead of the fluences themselves. As those moments depend on less variables than the fluences, their computation typically requires less numerical efforts. However the method of moments can also be used as a numerical method to discretize the angular dependencies Ω, *i.e.* when studying high order moments (see *e.g.* [9, 27, 2, 1, 26, 7]).

Moment models are applied in such a diverse range of physical problems because they preserve the major properties of the underlying kinetic models. In particular, the derivation of the moment models described in this manuscript focuses on hyperbolicity, positivity and entropy dissipation.

Although, several issues emerge when deriving those models. In this chapter, the derivation of the angular moment models is presented and illustrated through the extraction of the moments of the kinetic equation (1.11). In the next two chapters, two of the problems emerging from the moment extraction are studied.

In order to simplify some notations and to exhibit properties of the moment method in this chapter, the following kinetic equations and their moments are also studied

$$\textbf{in 1D:} \qquad \partial_t\psi+\partial_x \; F(\psi) = C(\psi), \qquad F(\psi) = \mu\psi \tag{2.1a}$$

$$\textbf{in 3D:} \qquad \partial_t\psi+\nabla_x.F(\psi) = C(\psi), \qquad F(\psi) = \Omega\psi \tag{2.1b}$$

and $C(\psi)$ is a collision operator which satisfies certain properties that will be specified later.

2.2 Angular moment extraction

First, the following notations are defined.

2.2.1 In 1D

In order to simplify the writing and/or the computations, 1D problems are often studied.

Notation 2.1 *The integral of a function ψ over all cosangles $\mu \in [-1,+1]$ is denoted*

$$\langle \psi \rangle = \int_{-1}^{+1} \psi(\mu) d\mu.$$

Moments, in 1D, are defined by

Definition 2.1 *(a)* ***Moment of order*** **i*:***
The angular moment ψ^i of order i of a function ψ is given by

$$\psi^i \quad = \quad \left\langle \mu^i \psi \right\rangle.$$

(b) ***Vector of moments:***
Suppose an independent family of polynomials of the components of $\mu \in [-1,+1]$ ranged into a vector $\mathbf{m}(\mu)$. The vector $\boldsymbol{\psi}$ of moments of a function ψ according to $\mathbf{m}$ is

$$\boldsymbol{\psi} \quad = \quad \langle \mathbf{m}\psi \rangle.$$

The vectors, and especially the moments vectors are written in bold.

Example 2.1 *Consider $\psi \in L^1([-1,1])$, then*

$$\psi^0 = \int_{-1}^{+1} \psi(\mu) d\mu, \qquad \psi^1 = \int_{-1}^{+1} \mu\psi(\mu) d\mu, \qquad \psi^2 = \int_{-1}^{+1} \mu^2\psi(\mu) d\mu.$$

Chose for instance $\mathbf{m}(\mu) = (1,\ \mu,\ \mu^2)$ then

$$\boldsymbol{\psi} = \int_{-1}^{+1} \mathbf{m}(\mu)\psi(\mu) d\mu = (\psi^0,\ \psi^1,\ \psi^2).$$

2.2.2 In 3D

For 3D problems, the following notations are used.

Notation 2.2 *For convenience, the integral of a function ψ over all directions $\Omega \in S^2$ is denoted*

$$\langle \psi \rangle = \int_{S^2} \psi(\Omega) d\Omega,$$

as in 1D. In case of ambiguity, the integral notation is used.

In 3D, moments are commonly ranged into tensors or vectors.

Definition 2.2 *(a)* ***Moment of order* i *(tensorial form):***
The angular moment ψ^i of order i of a function ψ is given by

$$\psi^i \quad = \quad \left\langle \Omega^{\otimes i} \psi \right\rangle, \quad \Omega^{\otimes i} = \underbrace{\Omega \otimes \cdots \otimes \Omega}_{i \text{ times}},$$

where $\otimes$ refers to the tensor product, and $\Omega^{\otimes i}$ is the i-th power of Ω according to the tensor product. The moments under tensorial form are always written with a superscript. The components of a moment ψ^i is denoted by adding subscripts, e.g.

$$\psi^2_{i,j} = \langle \Omega_i \Omega_j \psi \rangle .$$

(b) ***Vector of moments:***
Suppose an independent family of polynomials of the components of $\Omega \in S^2$ ranged into a vector $\mathbf{m}(\Omega)$. The vector $\boldsymbol{\psi}$ of moments of a function ψ according to $\mathbf{m}$ is

$$\boldsymbol{\psi} \quad = \quad \langle \mathbf{m} \psi \rangle .$$

When $\mathbf{m}$ is a vector composed of all monomials of degree i, one can pass from one notation to the other by simply reodering the components of $\boldsymbol{\psi}$ into tensorial form.

Example 2.2 *Consider $\psi \in L^1(S^2)$, then*

$$\begin{aligned}
\psi^0 &= \int_{S^2} \psi(\Omega) d\Omega, \quad \psi^1_i = \int_{S^2} \Omega_i \psi(\Omega) d\Omega, \quad \psi^2_{i,j} = \int_{S^2} \Omega_i \Omega_j \psi(\Omega) d\Omega = \psi^2_{j,i} \\
\psi^1 &= \int_{S^2} \Omega \psi(\Omega) d\Omega = \left(\psi^1_1, \ \psi^1_2, \ \psi^1_3 \right)^T, \\
\psi^2 &= \int_{S^2} \Omega \otimes \Omega \psi(\Omega) d\Omega = \begin{pmatrix} \psi^2_{1,1} & \psi^2_{1,2} & \psi^2_{1,3} \\ \psi^2_{1,2} & \psi^2_{2,2} & \psi^2_{2,3} \\ \psi^2_{1,3} & \psi^2_{2,3} & \psi^2_{3,3} \end{pmatrix}.
\end{aligned}$$

Chose for instance

$$\mathbf{m}(\Omega) = (\Omega_1, \ \Omega_2, \ \Omega_3, \ \Omega_1^2, \ \Omega_1\Omega_2, \ \Omega_2^2, \ \Omega_1\Omega_3, \ \Omega_2\Omega_3, \ \Omega_3^2),$$

then

$$\boldsymbol{\psi} = \int_{S^2} \mathbf{m}(\Omega)\psi(\Omega)d\Omega = (\psi^1_1,\ \psi^1_2,\ \psi^1_3,\ \psi^2_{1,1},\ \psi^2_{1,2},\ \psi^2_{2,2},\ \psi^2_{1,3},\ \psi^2_{2,3},\ \psi^2_{3,3}).$$

Both notations have advantages that are exploited in the next chapters.

The tensorial notation typically provides a physical meaning to the moments, *e.g.*. by analogy with fluid model, ψ^0 corresponds to a density of particles and ψ^1 to a flux. Although this notation is redundant, because the components of a moment tensor are not independent, *e.g.* the matrix ψ^2 is symmetric $\psi^2_{i,j} = \psi^2_{j,i}$.

The vectorial notation is typically more compact and does not present redundancy as long as the vector $\mathbf{m}$ is composed of independent polynomials.

The purpose of the use of moments in this manuscript is to reduce the computational costs compared to direct discretization of a kinetic equation. This can be illustrated by the following simple computation.

Example 2.3 *Suppose that a fluence $\psi(\epsilon, x, \Omega)$ is discretized with $N_\epsilon = 100$ cells in energy, $N_x = 500$ cells in position and $N_\Omega = 128$ cells in direction. As a comparison, consider a system of four moments (corresponding e.g. to M_1 model in 3D, see below Chapter 4 Section 4.5) with the same number of cells in energy and position. Comparing the degrees of freedom for the kinetic and moment systems reads*

$$\begin{aligned} N_{kinetic} &= N_\epsilon \times N_x \times N_\Omega = 6.4 \times 10^6, \\ N_{moments} &= 4 \times N_\epsilon \times N_x = 2 \times 10^5 = \frac{N_{kinetic}}{32}. \end{aligned}$$

The kinetic system has more degrees of freedom than the moment system when the number of moments studied is lower than N_Ω.

2.3 Moment equations

One obtains equations for the moments ψ^i by extracting the moments of a kinetic equation. Those equations are computed here for 3D problems. The 1D moment equations can easily be computed using the same method. First the moments of the kinetic equations are computed (in Subsections 2.3.1 and 2.3.2), then the moments of the different parts of the collision operators are computed in Subsections 2.3.3 to 2.3.5 and gathered in Subsection 2.3.6.

2.3.1 Moments of the toy kinetic equation

Extracting the moments of (2.1) reads
under tensorial form

$$\partial_t \psi^i + \nabla_x . \psi^{i+1} = C^i, \qquad (2.2a)$$

or under vectorial form

$$\begin{aligned} \partial_t \boldsymbol{\psi} &+ \nabla_x . F = \mathbf{C}, \\ F &= \langle \Omega \otimes \mathbf{m}(\Omega) \psi(\epsilon, x, \Omega) \rangle . \end{aligned} \tag{2.2b}$$

Here C^i and $\mathbf{C}$ are the moments of the collision operator $C(\psi)$ under tensorial and vectorial form.

2.3.2 Moments of the electron and photon tranport equations

Extracting the moments of (1.11) reads
under tensorial form

$$\nabla_x . \psi_\gamma^{i+1}(x, \epsilon) = \rho(x) \left[Q^i_{\gamma \to \gamma}(\psi^i_\gamma) + Q^i_{e \to \gamma}(\psi^i_e) \right] (x, \epsilon), \tag{2.3a}$$

$$\nabla_x . \psi_e^{i+1}(x, \epsilon) = \rho(x) \left[Q^i_{e \to e}(\psi^i_e) + Q^i_{\gamma \to e}(\psi^i_\gamma) \right] (x, \epsilon), \tag{2.3b}$$

or under vectorial form

$$\nabla_x . F_\gamma(x, \epsilon) = \rho(x) \left[\mathbf{Q}_{\gamma \to \gamma}(\boldsymbol{\psi}_\gamma) + \mathbf{Q}_{e \to \gamma}(\boldsymbol{\psi}_e) \right] (x, \epsilon), \tag{2.3c}$$

$$\nabla_x . F_e(x, \epsilon) = \rho(x) \left[\mathbf{Q}_{e \to e}(\boldsymbol{\psi}_e) + \mathbf{Q}_{\gamma \to e}(\boldsymbol{\psi}_\gamma) \right] (x, \epsilon). \tag{2.3d}$$

In the next subsections, the moments of each part of the collision operators $Q_{\alpha \to \beta}$ are exhibited.

2.3.3 Moments of a linear Boltzmann loss term

Generically, the moments of a generic linear Boltzmann loss term P (1.12) based on a total cross section σ_T read

$$P^i(\psi^i)(\epsilon, x) = \sigma_T(\epsilon) \psi^i(\epsilon, x), \quad \mathbf{P}(\boldsymbol{\psi})(\epsilon, x) = \sigma_T(\epsilon) \boldsymbol{\psi}(\epsilon, x). \tag{2.4}$$

2.3.4 Moments of a linear Boltzmann gain term

Here the moments of a generic linear Boltzmann gain term (1.13) for inelastic and elastic scattering, *i.e.* based on generic cross sections σ and σ_{el}, are computed.

Inelastic gain term

Similarily, the moments of a generic linear Boltzmann gain term G based on a differential cross section σ read

$$\begin{aligned} G^0(\psi^0)(\epsilon,x) &= \int_\epsilon^{\epsilon_{max}} \sigma^0(\epsilon',\epsilon)\psi^0(\epsilon',x)d\epsilon', & (2.5a)\\ G^1(\psi^1)(\epsilon,x) &= \int_\epsilon^{\epsilon_{max}} \sigma^1(\epsilon',\epsilon)\psi^1(\epsilon',x)d\epsilon', & (2.5b)\\ G^2(\psi^2)(\epsilon,x) &= \int_\epsilon^{\epsilon_{max}} \left[\frac{\sigma^0-\sigma^2}{2}(\epsilon',\epsilon)\psi^0(x,\epsilon')Id\right.\\ &\qquad \left.+\frac{3\sigma^2-\sigma^0}{2}(\epsilon',\epsilon)\psi^2(x,\epsilon')\right]d\epsilon', & (2.5c)\\ \mathbf{G}(\boldsymbol{\psi})(\epsilon,x) &= \int_\epsilon^{\epsilon_{max}} s(\epsilon',\epsilon)\boldsymbol{\psi}(x,\epsilon')d\epsilon', & (2.5d) \end{aligned}$$

where σ^i are moments of σ according to the last variable

$$\sigma^i(\epsilon',\epsilon) = 2\pi\int_{-1}^{+1}\mu^i\sigma(\epsilon',\epsilon,\mu)d\mu, \qquad (2.5e)$$

and s is a matrix the components of which are linear combinations of the moments σ^i. See Appendix 2.A.1 for the computation of the matrix s.

Elastic gain term

For elastic effects (*e.g.* for the present problem, Mott effect (1.22) is elastic), the moments of an elastic linear Boltzmann gain term based on a generic cross section σ_{el} read

$$\begin{aligned} G^0_{el}(\psi^0)(\epsilon,x) &= \sigma^0_{el}(\epsilon)\psi^0(\epsilon,x), & (2.6a)\\ G^1_{el}(\psi^1)(\epsilon,x) &= \sigma^1_{el}(\epsilon)\psi^1(\epsilon,x), & (2.6b)\\ G^2_{el}(\psi^2)(\epsilon,x) &= \frac{\sigma^0_{el}-\sigma^2_{el}}{2}(\epsilon)\psi^0(x,\epsilon)Id+\frac{3\sigma^2_{el}-\sigma^0_{el}}{2}(\epsilon)\psi^2(x,\epsilon), & (2.6c)\\ \mathbf{G}_{el}(\boldsymbol{\psi})(\epsilon,x) &= s_{el}(\epsilon).\boldsymbol{\psi}(x,\epsilon). & (2.6d) \end{aligned}$$

2.3.5 Moments of a Fokker-Planck operator

The moments of a Fokker-Planck operator (1.35a) read

$$\begin{aligned} Q^0_{FP}(\psi^0)(\epsilon,x) &= \partial_\epsilon(S\psi^0)(\epsilon,x), & (2.7a)\\ Q^1_{FP}(\psi^1)(\epsilon,x) &= \partial_\epsilon(S\psi^1)(\epsilon,x)-2T(\epsilon)\psi^1(\epsilon,x), & (2.7b)\\ Q^2_{FP}(\psi^2)(\epsilon,x) &= \partial_\epsilon(S\psi^2)(\epsilon,x)+2T(\epsilon)\left[tr(\psi^2)Id-3\psi^2\right](\epsilon,x), & (2.7c)\\ Q^i_{FP}(\psi^i)(\epsilon,x) &= \partial_\epsilon(S\psi^2)(\epsilon,x)+T(\epsilon)(\Delta_\Omega\psi)^i(\epsilon,x), & (2.7d)\\ \mathbf{Q}_{FP}(\boldsymbol{\psi})(\epsilon,x) &= \partial_\epsilon(S\boldsymbol{\psi})(\epsilon,x)+T(\epsilon)M_{FP}\boldsymbol{\psi}(\epsilon,x), & (2.7e) \end{aligned}$$

where the matrix M_{FP} is computed in Appendix 2.A.2.

2.3.6 Moments of the full collision operator

The moments of the collision operators (1.14) for photon and electron transport described in Chapter 1 can be computed by gathering the moments of the collision operators defined in the previous subsections.

In tensorial form

The moments of the collision operators (1.14) under tensorial form read

$$\begin{aligned}
Q^i_{\gamma\to\gamma}(\psi^i_\gamma) &= [G^i_{C,\gamma} - P^i_C](\psi^i_\gamma), && \text{(2.8a)}\\
Q^i_{\gamma\to e}(\psi^i_\gamma) &= G^i_{C,e}(\psi^i_\gamma), && \text{(2.8b)}\\
Q^i_{e\to\gamma}(\psi^i_e) &= 0, && \text{(2.8c)}\\
Q^i_{e\to e}(\psi^i_e) &= [G^i_{M,1} + G^i_{M,2} + G^i_{Mott} - P^i_M - P^i_{Mott}](\psi^i_e), && \text{(2.8d)}
\end{aligned}$$

where $G^i_{C,\gamma}$, $G^i_{C,e}$, $G^i_{M,1}$ and $G^i_{M,2}$ have the form (2.5), G^i_{Mott} the form (2.6), and P^i_C, P^i_M and P^i_{Mott} have the form (2.4).
At the moment level, the CSDA leads to

$$Q^i_{e\to e}(\psi^i_e) \approx Q^i_{LBCSD}(\psi^i_e) = \partial_\epsilon(S\psi^i_e) + (G^i_{M,2} + G^i_{Mott} - P^i_{Mott})(\psi^i_e). \tag{2.9}$$

At the moment level, the FP approximation leads to

$$Q^i_{e\to e}(\psi^i_e) \approx Q^i_{LBFP}(\psi^i_e) = \partial_\epsilon(S\psi^i_e) + \left(T(\Delta_\Omega\psi_e)^i + G^i_{M,2}\right)(\psi^i_e). \tag{2.10}$$

In vectorial form

The moments of the collision operators (1.14) under vectorial form read

$$\begin{aligned}
\mathbf{Q}_{\gamma\to\gamma}(\boldsymbol{\psi}_\gamma) &= [\mathbf{G}_{\mathbf{C},\gamma} - \mathbf{P}_\mathbf{C}](\boldsymbol{\psi}_\gamma), && \text{(2.11a)}\\
\mathbf{Q}_{\gamma\to\mathbf{e}}(\boldsymbol{\psi}_\gamma) &= \mathbf{G}_{\mathbf{C},\mathbf{e}}(\boldsymbol{\psi}_\gamma), && \text{(2.11b)}\\
\mathbf{Q}_{\mathbf{e}\to\gamma}(\boldsymbol{\psi}_e) &= 0, && \text{(2.11c)}\\
\mathbf{Q}_{\mathbf{e}\to\mathbf{e}}(\boldsymbol{\psi}_e) &= [\mathbf{G}_{\mathbf{M},\mathbf{1}} + \mathbf{G}_{\mathbf{M},\mathbf{2}} + \mathbf{G}_{\mathbf{Mott}} - \mathbf{P}_\mathbf{M} - \mathbf{P}_{\mathbf{Mott}}](\boldsymbol{\psi}_e), && \text{(2.11d)}
\end{aligned}$$

where $\mathbf{G}_{\mathbf{C},\gamma}$, $\mathbf{G}_{\mathbf{C},\mathbf{e}}$, $\mathbf{G}_{\mathbf{M},\mathbf{1}}$ and $\mathbf{G}_{\mathbf{M},\mathbf{2}}$ have the form (2.5), $\mathbf{G}_{\mathbf{Mott}}$ the form (2.6), and $\mathbf{P}_\mathbf{C}$, $\mathbf{P}_\mathbf{M}$ and $\mathbf{P}_{\mathbf{Mott}}$ have the form (2.4).
At the moment level, the CSDA leads to

$$\mathbf{Q}_{\mathbf{e}\to\mathbf{e}}(\boldsymbol{\psi}_e) \approx \mathbf{Q}_{\mathbf{LBCSD}}(\boldsymbol{\psi}_e) = \partial_\epsilon(S\boldsymbol{\psi}_e) + (\mathbf{G}_{\mathbf{M},\mathbf{2}} + \mathbf{G}_{\mathbf{Mott}} - \mathbf{P}_{\mathbf{Mott}})(\boldsymbol{\psi}_e). \tag{2.12}$$

At the moment level, the FP approximation leads to

$$\mathbf{Q}_{\mathbf{e}\to\mathbf{e}}(\boldsymbol{\psi}_e) \approx \mathbf{Q}_{\mathbf{LBFP}}(\boldsymbol{\psi}_e) = \partial_\epsilon(S\boldsymbol{\psi}_e) + (\mathbf{Q}_{\mathbf{FP}} + \mathbf{G}_{\mathbf{M},\mathbf{2}})(\boldsymbol{\psi}_e). \tag{2.13}$$

The moments of a 1D function $\langle\mu^i\psi\rangle$ can be seen as the first component of a moment of a 3D function $\langle\Omega^i_1\psi\rangle$, and can be easily computed from (2.8) and (2.11).

Remark 2.1 *The moments of the LBCSD and the LBFP operators are identical up to order 1. Indeed, according to the definition (1.35c) of the transport coefficient T and the definition (2.5e) of the moments of the differential cross section σ^i, one has*

$$\begin{aligned} 0 &= \sigma_{T,Mott}(\epsilon) - \sigma^0_{Mott}(\epsilon), \\ 2T(\epsilon) &= \sigma_{T,Mott}(\epsilon) - \sigma^1_{Mott}(\epsilon). \end{aligned}$$

Injecting it in (2.10) reads

$$\begin{aligned} Q^0_{LBFP}(\psi^0_e) &= \partial_\epsilon(S\psi^0_e) + G^0_{M,2}(\psi^0_e) \\ &= \partial_\epsilon(S\psi^0_e) + (\sigma^0_{Mott} - \sigma_{T,Mott})\psi^0_e + G^0_{M,2}(\psi^0_e) = Q^0_{LBCSD}(\psi^0_e), \\ Q^0_{LBFP}(\psi^1_e) &= \partial_\epsilon(S\psi^1_e) - 2T\psi^1_e + G^1_{M,2}(\psi^1_e) \\ &= \partial_\epsilon(S\psi^0_e) + (\sigma^1_{Mott} - \sigma_{T,Mott})\psi^0_e + G^0_{M,2}(\psi^0_e) = Q^1_{LBCSD}(\psi^1_e). \end{aligned}$$

The moments of order higher than one of the LBCSD and the LBFP operators differ. Therefore moment models of order one can not differentiate a CSD operator from a FP one.

In the field of radiotherapy, the dose (1.64) is the function of interest. This dose can be obtained from the moments of ψ_e and ψ_γ. Indeed, the equation (1.64) can be rewritten

$$\begin{aligned} D(x) = \int_{\epsilon_{min}}^{\epsilon_{max}} & \left[\int_{\epsilon_B}^{\epsilon} (\epsilon-\epsilon')\sigma^0_{C,\gamma}(\epsilon,\epsilon') + \epsilon'\sigma^0_{C,e}(\epsilon,\epsilon')\, d\epsilon'\right]\psi^0_\gamma(\epsilon,x) \\ &+ \left[\int_{\epsilon_B}^{\epsilon} (\epsilon-\epsilon')\sigma^0_{M,1}(\epsilon,\epsilon') + \epsilon'\sigma^0_{M,2}(\epsilon,\epsilon')\, d\epsilon'\right]\psi^0_e(\epsilon,x)d\epsilon. \end{aligned} \tag{2.14}$$

Therefore, the angular moment models can be used to compute the dose.

2.4 Conservation properties at the moment level

Based on the Properties and Propositions 1.1 to 1.5 of Section 1.4.2 obtained at the kinetic level from the conservation properties of the underlying molecular model, one can easily deduce their equivalent after moment extraction.

2.4.1 Compton scattering

After moment extraction, Property 1.1 turns into the following proposition.

Proposition 2.1 *The quantity of photons removed, created, and the quantity of electrons created by Compton effect are equal at the moment level iff*

$$\sigma_{T,C}(\epsilon) = \int_{\epsilon_B}^{\epsilon} \sigma^0_{C,\gamma}(\epsilon,\epsilon')d\epsilon' = \int_{\epsilon_B}^{\epsilon} \sigma^0_{C,e}(\epsilon,\epsilon')d\epsilon'. \tag{2.15}$$

This follows from Proposition 1.1 (by extracting moments from (1.17)) and can also be obtained by reproducing the proof of Proposition 1.1.

Similarily the conservation of momentum (*i.e.* Proposition 1.2) at the moment level leads to:

Proposition 2.2 *The macroscopic momentum in the system is preserved by Compton effect at the moment level iff*

$$p_\gamma(\epsilon)\sigma_{T,C}(\epsilon) - \int_{\epsilon_B}^{\epsilon} [p_\gamma(\epsilon')\sigma^1_{C,\gamma}(\epsilon,\epsilon') + p_e(\epsilon')\sigma^1_{C,e}(\epsilon,\epsilon')]d\epsilon' = 0_{\mathbb{R}^3}. \tag{2.16}$$

This follows from Proposition 1.2.

After moment extraction, Property 1.4 turns into the following property

Property 2.1 *The moments of the differential cross sections for photons and electrons are related to each other through the following formulae*

$$\begin{aligned}
\sigma^0_{C,e}(\epsilon',\epsilon) &= \sigma^0_{C,\gamma}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B\right), && \text{(2.17a)}\\
\sigma^1_{C,e}(\epsilon',\epsilon) &= \frac{p_\gamma(\epsilon')}{p_e(\epsilon)}\sigma^0_{C,\gamma}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B\right) \\
&\quad -\frac{p_\gamma(\epsilon'-\epsilon-\epsilon_B)}{p_e(\epsilon)}\sigma^1_{C,\gamma}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B\right), && \text{(2.17b)}\\
\sigma^i_{C,e}(\epsilon',\epsilon) &= \sum_{j=0}^{i}\left(\begin{array}{c} i \\ j \end{array}\right)\left(\frac{p_\gamma(\epsilon')}{p_e(\epsilon)}\right)^j && \text{(2.17c)}\\
&\quad * \left(-\frac{p_\gamma(\epsilon'-\epsilon-\epsilon_B)}{p_e(\epsilon)}\right)^{(i-j)} \sigma^{i-j}_{C,\gamma}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B\right).
\end{aligned}$$

The first equality can also be interpreted as having equal quantity of outgoing photons of energy ϵ as of outgoing electrons of energy $\epsilon'-\epsilon-\epsilon_B$. The second and last equations are obtained from (1.21), by doing a change of variable

$$\mu" = \frac{p_\gamma(\epsilon') - \mu p_e(\epsilon'-\epsilon-\epsilon_B)}{p_\gamma(\epsilon)},$$

and then writing $\mu"^i$ as a function of μ.

2.4.2 Mott scattering

Similarily as for Compton scattering, the conservation Properties 1.2 can be interpreted at the moment level by the following proposition.

Proposition 2.3 *The quantity of particles and the energy in the system is preserved by Mott effect at the moment level iff*

$$\sigma_{T,Mott}(\epsilon) = \tilde{\sigma}^0_{Mott}(\epsilon). \tag{2.18}$$

This follows from Proposition 1.3 or by rewriting the proof of Proposition 1.3 at the moment level.

2.4.3 Møller scattering

As for Compton effect, at the moment level, Property 1.5 turns into the following property.

Property 2.2 *The moments of the differential cross sections for primary and secondary electrons are related to each other through the following formulae*

$$\sigma^0_{M,2}(\epsilon',\epsilon) = \sigma^0_{M,1}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B\right) \tag{2.19a}$$

$$\begin{aligned}\sigma^1_{M,2}(\epsilon',\epsilon) &= \frac{p_e(\epsilon')}{p_e(\epsilon)}\sigma^0_{M,1}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B\right)\\ &\quad -\frac{p_e(\epsilon'-\epsilon-\epsilon_B)}{p_e(\epsilon)}\sigma^1_{M,1}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B\right),\end{aligned} \tag{2.19b}$$

$$\begin{aligned}\sigma^i_{M,2}(\epsilon',\epsilon) &= \sum_{j=0}^{i}\left(\begin{array}{c} i\\ j\end{array}\right)\left(\frac{p_e(\epsilon')}{p_e(\epsilon)}\right)^j\\ &\quad *\left(-\frac{p_e(\epsilon'-\epsilon-\epsilon_B)}{p_e(\epsilon)}\right)^{(i-j)}\sigma^{i-j}_{M,2}\left(\epsilon',\epsilon'-\epsilon-\epsilon_B\right),\end{aligned} \tag{2.19c}$$

The conservation Properties 1.3(a) and 1.3(b) can be translated at the moment level by

Proposition 2.4 *The quantity of electrons removed, and of primary and secondary electrons created by Møller effect are equal at the moment level iff*

$$\sigma_{T,M}(\epsilon) = \int_{\epsilon_B}^{\epsilon}\sigma^0_{M,1}(\epsilon,\epsilon')d\epsilon' = \int_{\epsilon_B}^{\epsilon}\sigma^0_{M,2}(\epsilon,\epsilon')d\epsilon'. \tag{2.20}$$

The proof is identical to the one of Proposition 2.1.

Proposition 2.5 *The macroscopic momentum in the system is preserved by Møller effect at the moment level iff*

$$0_{\mathbb{R}^3} = p_e(\epsilon)\sigma_{T,M}(\epsilon) - \int_{\epsilon_B}^{\epsilon}p_e(\epsilon')(\sigma^1_{M,1}+\sigma^1_{M,2})(\epsilon,\epsilon')d\epsilon'. \tag{2.21}$$

The proof is identical to the one of Proposition 2.1.

2.5 Issues with moment models

The method of moments reduces the dimension of the space of the variables for the problem. Although several problems emerge from the moment extraction. Among the most important ones, the realizability and the closure problem are presented in the following subsections and are studied in detail in the next chapters.

2.5.1 Realizability property

In order to relate more closely the moment model to the associated kinetic model, one may want to reconstruct a fluence $\psi(\Omega)$ based on the solution $\boldsymbol{\psi}$ of a moment equation. At the kinetic level, a fluence ψ is necessarily non-negative because it corresponds to a density of particles in a certain variable space (see Definition (1.9)). Then, based on Definitions 2.2(a) and 2.2(b) of the moments, physically relevant solutions to moment equations such as (2.2) or (2.3) need to be the moments of a non-negative function ψ according to the Ω variable (or μ in 1D). This imposes a condition on the vector of moment afterward called realizability condition.

A vector $\boldsymbol{\psi}$ (or a set of tensors $(\psi^0, \psi^1, ..., \psi^i)$) is called realizable if there exists at least one non-negative distribution $\psi(\Omega)$ such that $\boldsymbol{\psi}$ is composed of moments of ψ. This theoretical definition is complicated to use for practical applications. Therefore practical characterization of the realizability property are seeked. The next chapter is devoted to this problem.

2.5.2 Closure problem

The moment system (2.3) does not have a unique solution. This can easily be observed because it has more unknowns $(\psi^0, \psi^1, ..., \psi^N)$ and ψ^{N+1} (or $\boldsymbol{\psi}$ and F in vectorial notation) than equations. Especially, the moments of the solution to the kinetic equation (1.11) are in the set of the solutions of the moment system (2.3).

In order to have a unique solution to the moment problem (2.3), one needs to close the system. This closure consists in adding relations in order to have as many unknowns as equations. Commonly, this closure consists in expressing ψ^{N+1} as a function of the other moments $(\psi^0, ..., \psi^N)$. This problem is commonly solved by reconstructing an ansatz ψ_R having $(\psi^0, ..., \psi^N)$ as moments of order 0 to N and then the closure relation is chosen to be

$$\psi^{N+1} = \left\langle \Omega^{\otimes N+1} \psi_R \right\rangle.$$

In order to relate the moment model to the underlying kinetic model, one may chose ψ_R to be non-negative, and therefore the closure problem is related to the realizability problem through the construction of this representing ansatz. Chapter 4 is devoted to this problem.

2.5.3 Boundary conditions

The moment extraction procedure of Section 2.2 is well-understood in the interior $x \in int(Z)$. However, defining properly boundary conditions for moment models corresponding to the underlying kinetic ones remains an open problem.

The kinetic boundary conditions impose incoming fluxes (see Section 1.5), *i.e.* ψ is fixed to ψ^b only on Γ^-. As moment extraction is an integration over the whole unit sphere S^2, this does not provide enough information to defined the whole flux $\mathbf{F}$ on the boundary for moment equations.

According to the theory of hyperbolic equations (see *e.g.* [13, 4, 15]), not all the flux $\mathbf{F}$ needs to be imposed on the boundary and one may define boundary conditions for moment equations based on this theory. However, it remains unclear how to relate such boundary conditions to the underlying kinetic ones in the general case.

Several approaches are based on the study of the half space problem (see *e.g.* [19, 5]). This problem was also adressed *e.g.* [29, 37, 18] in the particular case of linear closures (see the P_N closures in Subsection 4.7.1 below). Other methods, such as the partial moment method [38, 17], circumvent this issue.

The problem of boundary conditions for moment equations is not dealt with in this manuscript. In Part III, boundary conditions for the moment equations are imposed at the discrete level. They are assumed to approximate the underlying kinetic boundary conditions accurately enough for the present applications.

Appendix

2.A Moments of the collision operators

Integrating the collision operators over the angular variable Ω is not trivial. Computations are summarized here. Only the 3D computations are shown, the 1D ones can be deduced from them.

2.A.1 Moments of a linear Boltzmann gain term

Generically, a kinetic linear Boltzmann gain term can be written

$$G(\psi)(\Omega) = \int_{S^2} \sigma(\Omega'.\Omega)\psi(\Omega')d\Omega',$$

where the x and ϵ variables are removed because they can be seen as parameters in this equation.

Moment tensors

By using the change of variable $\Omega" = R^T\Omega$ where the rotation R is such that $\Omega' R = e_1$, the moments of order 0 and 1 of $G(\psi)$ in tensorial notation reads

$$\begin{aligned}
\mathbf{G}^0(\psi)(\Omega) &= \int_{S^2}\int_{S^2} \sigma(\Omega'.\Omega)\psi(\Omega')d\Omega' d\Omega \\
&= \int_{S^2}\int_{S^2} \sigma(\Omega'.\Omega)d\Omega\psi(\Omega')d\Omega' = \sigma^0\psi^0, \\
\mathbf{G}^1(\psi)(\Omega) &= \int_{S^2}\int_{S^2} \Omega\sigma(\Omega'.\Omega)\psi(\Omega')d\Omega' d\Omega \\
&= \int_{S^2}\int_{S^2} \Omega\sigma(\Omega'.\Omega)d\Omega\psi(\Omega')d\Omega' \\
&= \int_{S^2} R\int_{S^2} \Omega"\sigma(\Omega" e_1)d\Omega"\psi(\Omega')d\Omega' \\
&= \int_{S^2} Re_1\sigma^1\psi(\Omega')d\Omega' = \int_{S^2}\Omega'\sigma^1\psi(\Omega')d\Omega' = \sigma^1\psi^1,
\end{aligned}$$

where

$$\sigma^i = 2\pi\int_{-1}^{+1}\mu^i\sigma(\mu)d\mu.$$

The computation of the moment of order two in tensorial form of this gain term is presented (corresponding to (2.5c)). Then the computation in vectorial form is presented, and one can obtains the moment of any order i in tensorial form from it.

Multiplying $G(\psi)(\Omega)$ by $\Omega \otimes \Omega$ and integrating it over all $\Omega \in S^2$ yields

$$\begin{aligned}
\mathbf{G}^2(\psi)(\Omega) &= \int_{S^2} \Omega \otimes \Omega \int_{S^2} \sigma(\Omega'.\Omega)\psi(\Omega')d\Omega' d\Omega \\
&= \int_{S^2}\int_{S^2} \Omega \otimes \Omega \sigma(\Omega'.\Omega) d\Omega \psi(\Omega')d\Omega'.
\end{aligned}$$

By using the change of variable $\Omega" = R^T\Omega$ where a rotation R is such that $\Omega' R = e_1$, the moments of order 0 and 1 of $G(\psi)$ in tensorial notation reads

$$\begin{aligned}
&\int_{S^2} \Omega \otimes \Omega \sigma(\Omega'.\Omega)d\Omega = \int_{S^2} (R^T\Omega") \otimes (R^T\Omega")\sigma(e_1.\Omega")d\Omega" \\
=\ & R^T \int_{S^2} \Omega" \otimes \Omega" \sigma(e_1.\Omega")d\Omega" R \\
=\ & R^T \int_{S^2} \begin{pmatrix} \mu"^2 & \mu"\sqrt{1-\mu"^2}c & \mu"\sqrt{1-\mu"^2}s \\ \mu"\sqrt{1-\mu"^2}c & (1-\mu"^2)c^2 & (1-\mu"^2)cs \\ \mu"\sqrt{1-\mu"^2}s & (1-\mu"^2)cs & (1-\mu"^2)s^2 \end{pmatrix} \sigma(\mu")d\Omega" R, \\
=\ & 2\pi R^T \int_{-1}^{+1} \begin{pmatrix} \mu"^2 & 0 & 0 \\ 0 & \frac{1-\mu"^2}{2} & 0 \\ 0 & 0 & \frac{1-\mu"^2}{2} \end{pmatrix} \sigma(\mu")d\mu" R \\
=\ & R^T(\frac{\sigma^0-\sigma^2}{2} Id + \frac{3\sigma^2-\sigma^0}{2} e_1)R \\
=\ & \frac{\sigma^0-\sigma^2}{2} Id + \frac{3\sigma^2-\sigma^0}{2} \Omega \otimes \Omega,
\end{aligned}$$

where $c = \cos\phi"$ and $s = \sin\phi"$. This leads to (2.5c).

Moment vector

Multiplying $G(\psi)(\Omega)$ by a vector $\mathbf{m}(\Omega)$ and integrating it over $\Omega \in S^2$ yields

$$\begin{aligned}
\mathbf{G}(\boldsymbol{\psi})(\Omega) &= \int_{S^2} \mathbf{m}(\Omega) \int_{S^2} \sigma(\Omega'.\Omega)\psi(\Omega')d\Omega' d\Omega \\
&= \int_{S^2}\int_{S^2} \mathbf{m}(\Omega)\sigma(\Omega'.\Omega) d\Omega \psi(\Omega')d\Omega'
\end{aligned}$$

Now using the change of variable $\Omega" = R\Omega$ leads to

$$\int_{S^2} \mathbf{m}(\Omega)\sigma(\Omega'.\Omega)d\Omega = \int_{S^2} \mathbf{m}(R^T\Omega")\sigma(e_1.\Omega")d\Omega".$$

Suppose $\mathbf{m}(\Omega)$ is a linearily independent family generating the set of polynomials of degree N over the unit sphere S^2, then the vector $\mathbf{m}(R^T\Omega")$ can be written

$$\mathbf{m}(R^T\Omega") = A(\Omega")\mathbf{m}(\Omega'),$$

where the components of the matrix $A(\Omega")$ are linear combinations of the components of $\mathbf{m}(\Omega")$. This matrix can be computed for particular choices of $\mathbf{m}(\Omega)$, but there is, a priori, no simple formula general to any degree N.

Therefore one obtains

$$\begin{aligned} \mathbf{G}(\boldsymbol{\psi})(\Omega) &= s\boldsymbol{\psi}(\Omega), \\ s &= \int_{S^2} A(\Omega")\sigma(e_1.\Omega")d\Omega". \end{aligned}$$

The components of the moment of order i (in tensorial form) of a linear Boltzmann gain term are the same as the one of a moment vector. One can obtain the moments of this gain term from those computations.

Example 2.4 *For the moments according to* $\mathbf{m}(\Omega) = (1, \Omega)$*, the matrix* s *reads*

$$s = \begin{pmatrix} \sigma^0 & 0_{\mathbb{R}^3}^T \\ 0_{\mathbb{R}^3} & \sigma^1 Id \end{pmatrix}.$$

For the moments according to

$$\mathbf{m}(\Omega) = (\Omega_1,\ \Omega_2,\ \Omega_3,\ \Omega_1^2,\ \Omega_2^2,\ \Omega_3^2,\ \Omega_1\Omega_2,\ \Omega_1\Omega_3,\ \Omega_2\Omega_3),$$

the matrix s *reads*

$$s = \begin{pmatrix} \sigma^1 Id & 0_{\mathbb{R}^{3\times3}} & 0_{\mathbb{R}^{3\times3}} \\ 0_{\mathbb{R}^{3\times3}} & \dfrac{\sigma^0 - \sigma^2}{2} M_{\mathbb{R}^{3\times3}} + \dfrac{3\sigma^2 - \sigma^0}{2} Id & 0_{\mathbb{R}^{3\times3}} \\ 0_{\mathbb{R}^{3\times3}} & 0_{\mathbb{R}^{3\times3}} & \dfrac{3\sigma^2 - \sigma^0}{2} Id \end{pmatrix},$$

where

$$M_{\mathbb{R}^{3\times3}} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

2.A.2 Moments of a Fokker-Planck diffusion operator

Similarily the computation of the moments of the Fokker-Planck agular diffusion term is shown for moments under vectorial form and in the tensorial form for moments of order two.

The Fokker-Planck diffusion term reads

$$\Delta_\Omega \psi(\Omega) = \partial_\mu \left((1-\mu^2)\partial_\mu \psi(\Omega) \right) + \frac{1}{1-\mu^2}\partial_\phi^2 \psi(\Omega). \tag{2.22}$$

Moment vector

Multiplying (2.22) by $\mathbf{m}(\Omega)$, integrating it over all $\Omega \in S^2$ and using an integration by parts leads

$$\begin{aligned}\int_{S^2} \mathbf{m}(\Omega)\Delta_\Omega \psi(\Omega) d\Omega &= \int_{S^2} \mathbf{m}(\Omega) \left[\partial_\mu \left((1-\mu^2)\partial_\mu \psi(\Omega)\right) + \frac{1}{1-\mu^2}\partial_\phi^2 \psi(\Omega)\right] d\Omega \\ &= \int_{S^2} \left[\partial_\mu \left((1-\mu^2)\partial_\mu \mathbf{m}(\Omega)\right) + \frac{1}{1-\mu^2}\partial_\phi^2 \mathbf{m}(\Omega)\right] \psi(\Omega) d\Omega.\end{aligned}$$

Suppose $\mathbf{m}(\Omega)$ is a linearly independent family generating the set of polynomials of degree N over the unit sphere S^2. By exhibiting the value of

$$\partial_\mu \left((1-\mu^2)\partial_\mu \mathbf{m}(\Omega)\right) + \frac{1}{1-\mu^2}\partial_\phi^2 \mathbf{m}(\Omega),$$

one can easily show that this vector is only composed of linear combinations of $\mathbf{m}(\Omega)$. This leads to the matrix formulation

$$\left[\partial_\mu \left((1-\mu^2)\partial_\mu \mathbf{m}(\Omega)\right) + \frac{1}{1-\mu^2}\partial_\phi^2 \mathbf{m}(\Omega)\right] = M_{FP}\mathbf{m}(\Omega),$$

where the matrix M_{FP} can be computed for any choice of $\mathbf{m}(\Omega)$, *e.g.* when $\mathbf{m}(\Omega) = (1,\Omega)$

$$M_{FP}\mathbf{m}(\Omega) = \begin{pmatrix} 0 & 0_{\mathbb{R}^3} \\ 0_{\mathbb{R}^3} & -2Id \end{pmatrix} \mathbf{m}(\Omega),$$

and when $\mathbf{m}(\Omega) = (\Omega_1,\ \Omega_2,\ \Omega_3,\ \Omega_1^2,\ \Omega_2^2,\ \Omega_3^2,\ \Omega_1\Omega_2,\ \Omega_1\Omega_3,\ \Omega_2\Omega_3)$

$$\begin{aligned}M_{FP}\mathbf{m}(\Omega) &= \begin{pmatrix} -2Id & 0_{\mathbb{R}^{3\times3}} & 0_{\mathbb{R}^{3\times3}} \\ 0_{\mathbb{R}^{3\times3}} & 2M_{\mathbb{R}^{3\times3}} - 6Id & 0_{\mathbb{R}^{3\times3}} \\ 0_{\mathbb{R}^{3\times3}} & 0_{\mathbb{R}^{3\times3}} & -6Id \end{pmatrix} \mathbf{m}(\Omega), \\ M_{\mathbb{R}^{3\times3}} &= \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}\end{aligned}$$

Again the moments under tensorial form can be deduced from this notation.

Moment tensor

In the special case $N = 0$ and $N = 1$, the moment of (2.22) under tensorial notation reads

$$\begin{aligned}\int_{S^2} \Delta_\Omega \psi(\Omega) d\Omega &= 0, \\ \int_{S^2} \Omega\Delta_\Omega \psi(\Omega) d\Omega &= \int_{S^2} \left[\partial_\mu \left((1-\mu^2)\partial_\mu \Omega\right) + \frac{1}{1-\mu^2}\partial_\phi^2 \Omega\right] \psi(\Omega) d\Omega, \\ &= \int_{S^2} -2\Omega\psi(\Omega) d\Omega, \\ &= -2\psi^1.\end{aligned}$$

In the special case $N = 2$, the moment of (2.22) under tensorial notation reads

$$\begin{aligned}
\int_{S^2} \Omega \otimes \Omega \Delta_\Omega \psi(\Omega) d\Omega &= \int_{S^2} \left[\partial_\mu \left((1-\mu^2) \partial_\mu \Omega \otimes \Omega \right) + \frac{1}{1-\mu^2} \partial_\phi^2 \Omega \otimes \Omega \right] \psi(\Omega) d\Omega, \\
&= \int_{S^2} \left(2Id - 6\Omega \otimes \Omega \right) \psi(\Omega) d\Omega, \\
&= \int_{S^2} \left(2tr(\Omega \otimes \Omega) Id - 6\Omega \otimes \Omega \right) \psi(\Omega) d\Omega, \\
&= 2tr(\psi^2) Id - 6\psi^2.
\end{aligned}$$

Bibliography

[1] G. W. Alldredge, C. D. Hauck, D. P. O'Leary, and A. L. Tits. Adaptive change of basis in entropy-based moment closures for linear kinetic equations. *J. Comput. Phys.*, 74(4), february 2014.

[2] G. W. Alldredge, C. D. Hauck, and A. L. Tits. High-order entropy-based closures for linear transport in slab geometry II: A computational study of the optimization problem. *SIAM J. Sci. Comput.*, 34(4):361–391, 2012.

[3] A. M. Anile and S. Pennisi. Thermodynamic derivation of the hydrodynamical model for charge transport in semiconductors. *Phys. Rev. B*, 46:13186–13193, 1992.

[4] C. Bardos, A. Y. Leroux, and J. C. Nedelec. First order quasilinear equations with boundary conditions. *Commun. Part. Diff. Eq.*, 4(9):1017–1034, 1979.

[5] A. Bensoussan, J.-L. Lions, and G. C. Papanicolaou. Boundary layers and homogenization of transport processes. *Publ. RIMS, Kyoto Univ.*, 15:53–157, 1979.

[6] C. Berthon, P. Charrier, and B. Dubroca. An HLLC scheme to solve the M_1 model of radiative transfer in two space dimensions. *J. Sci. Comput.*, 31(3):347–389, 2007.

[7] Y. Bourgault, D. Broizat, and P.-E. Jabin. Convergence rate for the method of moments with linear closure relations. *Kin. Rel. Mod.*, 8(1):1–27, 2015.

[8] C. Buet and B. Despres. Asymptotic preserving and positive schemes for radiation hydrodynamics. *J. Comput. Phys.*, 215(2):717–740, July 2006.

[9] C. Canuto, M. Y. Hussaini, A. Quarteroni, and Th. A. Zang. *Spectral methods: Fundamental in single domains.* Springer, 2006.

[10] S. Chandrasekhar. On the radiative equilibrium of a stellar atmosphere. *Astrophys. J.*, 99:180 – 190, 1943.

[11] S. Chandrasekhar. On the radiative equilibrium of a stellar atmosphere X. *Astrophys. J.*, 103:351 – 370, 1946.

[12] S. Chandrasekhar. *Radiative transfer.* Dover publications, 1960.

[13] R. J. Diperna. Uniqueness of solutions to hyperbolic conservation laws. *Indiana Univ. Math. J.*, 28(1):137–188, 1979.

[14] B. Dubroca and J.-L. Feugeas. Hiérarchie des modèles aux moments pour le transfert radiatif. *C. R. Acad. Sci. Paris*, 329:915–920, 1999.

[15] B. Dubroca and G. Gallice. Resultats d'existence et d'unicite du problème mixte pour des systems hyperboliques de lois de conservation monodimensionnels. *Commun. Part. Diff. Eq.*, 15(1):59–80, 1990.

[16] R. Duclous. *Modélisation et simulation numérique multi-échelle du transport cinétique électronique.* PhD thesis, Université de Bordeaux 1, 2009.

[17] M. Frank. *Partial moment models for radiative transfer.* PhD thesis, Teschniche Universität Kaiserslautern, July 2005.

[18] B. D. Ganapol, C. T. Kelley, and G. C. Pomraning. Asymptotically exact boundary conditions for the P_N equations. *Nucl. Sci. Eng.*, 114, 1993.

[19] F. Golse, S. Jin, and C. D. Levermore. A domain decomposition analysis for a two-scale linear transport problem. *ESAIM-Math. Model. Num.*, 37(6):869–892, 2003.

[20] H. Grad. On the kinetic theory of rarefied gases. *Commun. Pur. Appl. Math.*, 2(4):331–407, 1949.

[21] C. Groth and J. McDonald. Towards physically realizable and hyperbolic moment closures for kinetic theory. *Cont. Mech. Therm.*, 21(6):467–493, 2009.

[22] S. Guisset, S. Brull, B. Dubroca, E. d'Humières, S. Karpov, and I. Potapenko. Asymptotic-preserving scheme for the M_1-Maxwell system in the quasi-neutral regime. *Commun. Comput. Phys.*, 19:301–328, 2016.

[23] S. Guisset, J.G. Moreau, R. Nuter, S. Brull, E. d'Humieres, B. Dubroca, and V.T. Tikhonchuk. Limits of the M_1 and M_2 angular moments models for kinetic plasma physics studies. *J. Phys. A: Math. Theor.*, 48, 2015.

[24] T. Hanawa and E. Audit. Reformulation of the M_1 model of radiative transfer. *J. Quant. Spectros. Radiat. Transfer*, 145:9 – 16, 2014.

[25] C. Hauck. *Entropy-based moment closures in semiconductor models.* PhD thesis, University of Maryland, 2006.

[26] C. Hauck and R. McClarren. Positive P_N closures. *SIAM J. Sci. Comput.*, 32(5):2603–2626, 2010.

[27] C. D. Hauck. High-order entropy-based closures for linear transport in slab geometry. *Commun. Math. Sci*, 9(1):187–205, 2011.

[28] C. D. Hauck, C. D. Levermore, and A. L. Tits. Convex duality and entropy-based moment closures: Characterizing degenerate densities. *SIAM J. Control Optim.*, 47(4):1977–2015, 2008.

[29] E. W. Larsen and G. C. Pomraning. The P_N theory as an asymptotic limit of transport theory in planar geometry I: Analysis. *Nucl. Sci. Eng.*, 109(49), 1991.

[30] C. D. Levermore. Relating Eddington factors to flux limiters. *J. Quant. Spectros. Radiat. Transfer*, 31:149–160, 1984.

[31] C. D. Levermore. Moment closure hierarchies for kinetic theories. *J. Stat. Phys.*, 83(5–6):1021–1065, 1996.

[32] J. Mallet, S. Brull, and B. Dubroca. An entropic scheme for an angular moment model for the classical Fokker-Planck-Landau equation of electrons. *Commun. Comput. Phys.*, 15(2):422–450, 2014.

[33] J. Mallet, S. Brull, and B. Dubroca. General moment system for plasma physics based on minimum entropy principle. *Kin. rel. mod.*, 8(3):533–558, 2015.

[34] J. Mcdonald and M. Torrilhon. Affordable robust moment closures for CFD based on the maximum-entropy hierarchy. *J. Comput. Phys.*, 251:500–523, 2013.

[35] G. C. Pomraning. *The equations of radiation hydrodynamics.* Pergamon Press, 1973.

[36] S. La Rosa, G. Mascali, and V. Romano. Exact maximum entropy closure of the hydrodynamical model for Si semiconductors: The 8-moment case. *SIAM J. Appl. Math.*, 70(3):710–734, 2009.

[37] R. P. Rulko, E. W. Larsen, and G. C. Pomraning. The P_N theory as an asymptotic limit of transport theory in planar geometry II: Numerical results. *Nucl. Sci. Eng.*, 109(76), 1991.

[38] R. Turpault, M. Frank, B. Dubroca, and A. Klar. Multigroup half space moment approximations to the radiative heat transfer equations. *J. Comput. Phys.*, 198(1):363 – 371, 2004.

Chapter 3

Realizability

3.1 Introduction

In order to provide deeper physical meaning to the solution of moment equation, one aims to relate the solution $\boldsymbol{\psi}$ of moment equation such as (2.3) to an underlying kinetic fluence ψ.

Among the properties that satisfy a kinetic fluence ψ, this chapter focuses on the positivity. Indeed, since ψ is a density in the space $Z \times [\epsilon_{\min}, \epsilon_{\max}] \times S^2$, a negative value of ψ is non-sense as it corresponds to a negative quantity of particles.

The necessity for a density to be positive is well understood. However it becomes more complicated, after extraction of moments, to determine whether a vector $\boldsymbol{\psi}$ is the moment vector of a non-negative function ψ. A vector is called realizable if it is the moment vector of a non-negative function $\psi(\Omega)$. The definition of the realizability property being difficult to use for practical applications, one typically seeks characterizations.

The study of the realizability property originally emerged in the end of the nineteenth century. This problem was mentioned by P. L. Chebyshev in 1873 ([8]) and was pushed forward by A. Markov ([21, 22]). However, it is T.-J. Stieltjes who gave the name "the Moment Problem" ([34]) to the problem of characterizing whether a vector $\mathbf{V}$ is realizable. This field of study was widely developed (see *e.g.* with [1, 2, 17, 9, 19] and references therein) after E. Artin solved the 17-th Hilbert problem ([3, 13])

Nowadays, the "moment problem" refers to the problem of characterizing the realizability of a vector $\mathbf{V}$ when the domain S of integration (for our purpose, $S = [-1, 1]$ in 1D, and $S = S^2$ in 3D) can be identified to a subset of $\mathbb{R}$. Among the moment problems, the problems of Hamburger ($S = \mathbb{R}$), Hausdorff ($S = [a, b]$), Stieltjes ($S = \mathbb{R}^+$) and Toeplitz ($S = S^1$) were solved and practical characterizations are available (see *e.g.* Subsection 3.3.2 and [1, 9, 18]).

Commonly when the number of moments is finite, *i.e.* the vector $\boldsymbol{\psi} \in \mathbb{R}^N$ with $N < \infty$, the problem is called "truncated moment problem". When the set of integration S is not restricted to one dimension, such as in the present 3D problem, the problem is called "the generalized moment problem". No practical characterization of the realizability property were provided in the general case where the set of integration S is not one dimensional and for arbitrarily large order N of moments. The moment problem on compact algebraic varieties (and semi-algebraic varieties), commonly called K-moment problem is widely studied ([12, 10, 20, 4]), but only partial results were obtained for moments on S^2. Some of them are recalled in this chapter and will be used in the next one to study the realizability of moment closures.

3.2 Preliminaries

The truncated moment problem, *i.e.* the problem of characterizing whether a vector $\mathbf{V}$ is a moment vector of a positive function ψ presents much more difficulties in 3D as 1D. Although similar methods are used for both problems. A generic workframe is given for both 1D and 3D problems. The generic variable of integration is written x and it belongs to a set S, *i.e.* when studying 1D problems

$$x = \mu \in [-1, 1] = S$$

and for 3D problems

$$x = \Omega \in S^2 = S.$$

The set of polynomials of degree less or equal to K is denoted $\mathbb{R}^K[X]$ in 1D, or $\mathbb{R}^K[X_1, X_2, X_3]$ in 3D.

3.2.1 The realizability domain

For writing purposes, the definition of the realizability property given here is slightly different compared to the one given in the introduction, *i.e.* the underlying function ψ is chosen to be strictly positive instead of non-negative. However, one can prove that those two problems are equivalent, *e.g.* with the work of [23, 5, 6, 7, 15, 31, 14].

The following notations and definitions are used

Notation 3.1 *The set of strictly positive L^1 functions over a set S is denoted $L^1(S)^+$, i.e. $f \in L^1(S)^+$ iff*

$$f \in L^1(S) \quad \textit{and} \quad \underset{S}{ess\,inf} f > 0.$$

The realizability property is commonly studied for moments under vectorial form (or under matricial form, see Subsection 3.2.2).

The truncated moment problem yields

$$\textit{Does there exist a function } \psi \in L^1(S)^+, \quad \textit{such that} \quad \mathbf{V} = \langle \mathbf{m}\psi \rangle ? \tag{3.1}$$

Remark that in 1D or 3D, only moments on the compact sets $[-1, 1]$ and S^2 are studied. Then having $\psi \in L^1(S)$ is sufficient for the existence of the moment according to any polynomial p. Indeed, on a compact S the polynomial p is bounded by $\min(p) \in \mathbb{R}$ and $\max(p) \in \mathbb{R}$ and so

$$\min(p) \langle \psi \rangle \leq \langle p\psi \rangle \leq \max(p) \langle \psi \rangle .$$

Therefore $\langle p\psi \rangle$ exists.

Definition 3.1 *The realizability domain $\mathcal{R}_{\mathbf{m}}$ associated to a vector of polynomials $\mathbf{m}$ over S is the set of all realizable moment vectors*

$$\mathcal{R}_{\mathbf{m}} := \left\{ \langle \mathbf{m}\psi \rangle, \quad \psi \in L^1(S)^+ \right\}. \tag{3.2}$$

We first prove the following proposition.

Proposition 3.1 *If $\mathbf{m}$ is a linearily independent family of polynomials over S, then the realizability domain $\mathcal{R}_{\mathbf{m}}$ is open.*

Proof Suppose $\mathbf{V} = \langle \mathbf{m}\psi \rangle \in \mathcal{R}_{\mathbf{m}}$ and define

$$\epsilon := \underset{S}{ess\ inf} \psi > 0.$$

Let us exhbibit a neighborhood of $\mathbf{V}$ included in $\mathcal{R}_{\mathbf{m}}$.

Define

$$\mathbf{V_i} = \langle \mathbf{m} m_i \rangle .$$

Simple computations show that the matrix $\langle \mathbf{m} \otimes \mathbf{m} \rangle$ is symmetric positive definite, and therefore non-singular. Therefore the family $(\mathbf{V}_i)_{\{i=1,...,Card(\mathbf{m})\}}$ of its column is a basis of $\mathbb{R}^{Card(\mathbf{m})}$.

Therefore for all $\alpha > 0$, the set

$$\mathcal{V} = \left\{ \mathbf{V} + \alpha \sum_i \lambda_i \mathbf{V_i}, \quad \lambda_i \in]-1, 1[, \quad i = 1, ..., Card(\mathbf{m}) \right\} \tag{3.3}$$

is a neighborhood of $\mathbf{V}$ in $\mathbb{R}^{Card(\mathbf{m})}$.

Choose *e.g.*

$$\alpha = \frac{\epsilon}{2Card(\mathbf{m})} > 0.$$

With this choice of α, one can easily show that

$$\alpha \sum_i \lambda_i \mathbf{m}_i(x) > -\epsilon, \quad \forall x \in S, \quad \forall \lambda_i \in]-1, 1[, \quad i = 1, ..., Card(\mathbf{m}).$$

Therefore, with this choice of α, each vector of the set (3.3) is represented by a function

$$\tilde{\psi} = \psi + \alpha \sum_i \lambda_i m_i \in L^1(S)$$

such that $\underset{S}{ess\ inf} \tilde{\psi} > \epsilon - \frac{\epsilon}{2} > 0$. Therefore $\tilde{\psi} \in L^1(S)^+$ and each point of $\mathcal{V}$ is included in $\mathcal{R}_{\mathbf{m}}$. For each point $\mathbf{V} \in \mathcal{R}_{\mathbf{m}}$, there exists a neighborhood of $\mathbf{V}$ included in $\mathcal{R}_{\mathbf{m}}$. □

In order to study the geometry of the realizability domain, and especially its boundary $\partial\mathcal{R}_{\mathbf{m}}$, one generally study at a slightly more general problem. By considering that $L^1(S)^+$ is a subset of $\mathcal{M}(S)$, the set of the Borel measures on S, one can generalize the notion of moments.

Example 3.1 *Suppose $S = [-1, 1]$ and $\mathbf{m}(\mu) = (1,\ \mu)$. Consider the sequence $(\psi_k)_{k\in\mathbb{N}}$ defined by*

$$\psi_k(\mu) = \frac{1}{k\sqrt{2\pi}} \exp\left(- \left(\frac{\mu - 1}{\sqrt{2}k} \right)^2 \right).$$

One can verify that ψ_k is bounded in $L^1(S)$ norm by 1. However the sequence ψ_k does not converge in $L^1(S)$, while the sequence of its moments

$$\int_{-1}^{+1} (1,\ \mu)\psi_k(\mu)d\mu \underset{k\to\infty}{\longrightarrow} (1,\ 1)$$

does converge in $\mathbb{R}^2$.

In the sense of measures, ψ_k converges to a Dirac measure in $\mu = 1$

$$\psi_k(\mu) \underset{k\to\infty}{\longrightarrow} \delta(\mu - 1),$$

and therefore the sequence of moments converges to the moments according to this measure

$$\int \mathbf{m}(\mu)\psi_k(\mu)d\mu \underset{k\to\infty}{\longrightarrow} \int_{-1}^{+1} \mathbf{m}(\mu)\delta(\mu - 1) = (1,\ 1).$$

This examples is a motivation to extend the notion of moments according to L^1 functions to moments according to measures.

Therefore one can generalize the truncated moment problem by looking for a representing measure γ instead of a positive function $\psi \in L^1(S)^+$.

The truncated moment problem (according to measures) yields

$$\text{Does there exist a measure } \gamma \text{ such that } \mathbf{V} = \int_S \mathbf{m} d\gamma(x)? \tag{3.4}$$

This notion of moments according to a measure is required in order to characterize the boundary of the realizability domain $\mathcal{R}_\mathbf{m}$.

In the next sections, the following set is studied.

Definition 3.2 *The set $\mathcal{R}^m_\mathbf{m}$ associated to a vector of polynomials $\mathbf{m}$ is defined by*

$$\begin{aligned} \mathcal{R}^m_\mathbf{m} \quad := \quad \Big\{ \mathbf{V} \in \mathbb{R}^{Card(\mathbf{m})}, \quad s.t. \quad \mathbf{V} = \lim_{k\to\infty} \langle \mathbf{m}\psi_k \rangle \\ \text{for some bounded sequence } \psi_k \text{ in } L^1(S)^+ \Big\}. \end{aligned} \tag{3.5}$$

Remark that each vector of $\mathcal{R}^m_\mathbf{m}$ is represented by some measure γ

$$\mathbf{V} \in \mathcal{R}^m_\mathbf{m} \quad \Rightarrow \quad \mathbf{V} = \int_S \mathbf{m}(x) d\gamma(x).$$

According to Definition 3.2 and to Proposition 3.1, the following relations hold

$$\mathcal{R}^m_{\mathbf{m}} = \overline{\mathcal{R}_{\mathbf{m}}} \cap \mathbb{R}^{Card(\mathbf{m})}, \qquad \mathcal{R}_{\mathbf{m}} = int(\mathcal{R}^m_{\mathbf{m}}), \tag{3.6}$$

where the superscript $\bar{E}$ refers to the closure of a set E and $int(E)$ refers to the interior of a set E.

Definition 3.1 of the realizability domain $\mathcal{R}_{\mathbf{m}}$ (and the definition of the set $\mathcal{R}^m_{\mathbf{m}}$) is difficult to apply to practical problems. Instead, one typically aims to characterize the realizability property by some numerical conditions easier to check. For this purpose, the following definition is introduced.

Definition 3.3 *A convex cone C is a set stable by positive combinations, i.e. iff*

$$\forall (\mathbf{V}_1, \mathbf{V}_2) \in C^2 \quad \forall (\alpha_1, \alpha_2) \in (\mathbb{R}^{*+})^2, \quad \alpha_1 \mathbf{V}_1 + \alpha_2 \mathbf{V}_2 \in C.$$

The following property offers a first characterization of realizability property.

Property 3.1 *The realizability domain $\mathcal{R}_{\mathbf{m}}$ and the set $\mathcal{R}^m_{\mathbf{m}}$ are convex cones.*

Proof Using Definition 3.1 of the realizability property, there exists $\psi_1 \in L^1(S)^+$ and $\psi_2 \in L^1(S)^+$ such that

$$\mathbf{V_1} = \langle \mathbf{m}\psi_1 \rangle, \qquad \mathbf{V_2} = \langle \mathbf{m}\psi_2 \rangle.$$

Therefore one obtains

$$\alpha_1 \mathbf{V}_1 + \alpha_2 \mathbf{V}_2 = \langle \mathbf{m} \left(\alpha_1 \psi_1 + \alpha_2 \psi_2 \right) \rangle,$$

where $\alpha_1 \psi_1 + \alpha_2 \psi_2 \in L^1(S)^+$.

Similar computations hold for $\mathcal{R}^m_{\mathbf{m}}$. □

Remark 3.1 *This property is commonly used when constructing numerical schemes for moment equations in order to prove that such schemes preserve the realizability property from one step to another.*

This characterization is convenient to construct realizable vector. However, in order to exhibit the realizability property of a vector $\mathbf{V}$, one still requires the knowledge of two realizable vectors $\mathbf{V}_1$ and $\mathbf{V}_2$, which is *a priori* not a simpler problem than the original one (3.1).

Strategy: In order to find practical characterization of the realizability property, first a necessary condition for a vector to be realizable (Subsection 3.2.2) is seeked. Then, under this condition, the existence of a representing measure (Sections 3.3 and 3.4) is proven by exhibiting one representing measure.

3.2.2 The Riesz functional

In the following, we seek necessary conditions for a vector to be realizable. Those are commonly easy to find. For numerical puprposes, we seek numerical inequality conditions (because they are easy to check), *i.e.* we seek a set of N functions $(f_i)_{i=1,...,N}$ such that

$$\mathbf{V} \in \mathcal{R}_{\mathbf{m}} \quad \Rightarrow \quad \forall i = 1, ..., N,\ f_i(\mathbf{V}) > 0.$$

For notation purposes, the Riesz functional ([27, 19, 12]) is used. This operator rearranges the terms of a vector $\mathbf{V}$.

Definition 3.4 (Riesz functional)
Consider a vector $\mathbf{m} \in (\mathbb{R}[X])^N$ *of* N *polynomials of* x*, and a vector* $\mathbf{V} \in \mathbb{R}^N$*.*

The Riesz functional $R_{\mathbf{V}}$ *associated to* $\mathbf{V}$ *sends any polynomial* $p = \boldsymbol{\lambda}\mathbf{m}$ *onto*

$$R_{\mathbf{V}}(p) = \boldsymbol{\lambda}\mathbf{V}. \tag{3.7}$$

Remark that the Riesz functional associated to $\mathbf{V}$ is a linear map from $Span(\mathbf{m})$ to $\mathbb{R}$.

If the vector $\mathbf{V} = \langle \mathbf{m}\psi \rangle$ is the vector of moments of a function $\psi \in L^1(S)^+$, then the Riesz functional of p is the moment of ψ according to p

$$R_{\mathbf{V}}(p) = \langle p\psi \rangle\,.$$

In the next sections, the Riesz functional is also applied to matrices of polynomials. The Riesz functional is simply applied to each component of such a matrix

$$R_{\mathbf{V}}(M)_{i,j} = R_{\mathbf{V}}(M_{i,j}).$$

Example 3.2 ***In 1D****, consider the vector* $\boldsymbol{\psi} = (\psi^0, \psi^1, \psi^2) \in \mathbb{R}^3$*, and the vector of monomials* $\mathbf{m}(\mu) = (1, \mu, \mu^2)$*. The Riesz function according to the vector* $\boldsymbol{\psi}$ *of the polynomial*

$$p(\mu) = 1 + 3\mu - \mu^2$$

reads

$$\begin{array}{rcccccc} R_{\boldsymbol{\psi}}(p) & = & R_{\boldsymbol{\psi}}(1) & + & 3R_{\boldsymbol{\psi}}(\mu) & - & R_{\boldsymbol{\psi}}(\mu^2) \\ & = & \psi^0 & + & 3\psi^1 & - & \psi^2 \quad . \end{array}$$

In 3D*, consider the vector*

$$\boldsymbol{\psi} = (\psi^0,\ \psi^1_1,\ \psi^1_2,\ \psi^1_3,\ \psi^2_{1,1},\ \psi^2_{1,2},\ \psi^2_{2,2},\ \psi^2_{1,3},\ \psi^2_{2,3},\ \psi^2_{3,3}) \in \mathbb{R}^{10},$$

and choose the vector of monomials $\mathbf{m}(\Omega) = (1, \Omega_1, \Omega_2, \Omega_3)$*. The Riesz func-*

tional according to ψ *of the matrix* $\mathbf{m}\otimes\mathbf{m}$ *reads*

$$\begin{aligned} R_\psi(\mathbf{m}\otimes\mathbf{m}) &= R_\psi\begin{pmatrix} 1 & \Omega_1 & \Omega_2 & \Omega_3 \\ \Omega_1 & \Omega_1^2 & \Omega_1\Omega_2 & \Omega_1\Omega_3 \\ \Omega_2 & \Omega_1\Omega_2 & \Omega_2^2 & \Omega_2\Omega_3 \\ \Omega_3 & \Omega_1\Omega_3 & \Omega_2\Omega_3 & \Omega_3^2 \end{pmatrix} = R_\psi\begin{pmatrix} 1 & \Omega^T \\ \Omega & \Omega\otimes\Omega \end{pmatrix} \\ &= \begin{pmatrix} \psi^0 & \psi^1_1 & \psi^1_2 & \psi^1_3 \\ \psi^1_1 & \psi^2_{1,1} & \psi^2_{1,2} & \psi^2_{1,3} \\ \psi^1_2 & \psi^2_{1,2} & \psi^2_{2,2} & \psi^2_{2,3} \\ \psi^1_3 & \psi^2_{1,3} & \psi^2_{2,3} & \psi^2_{3,3} \end{pmatrix} = \begin{pmatrix} \psi^0 & (\psi^1)^T \\ (\psi^1) & \psi^2 \end{pmatrix}. \end{aligned}$$

One can easily observe the following necessary condition.

Proposition 3.2 *A vector* $\mathbf{V}$ *is realizable only if the Riesz functional preserves positivity, i.e.*

$$\mathbf{V}\in\mathcal{R}_\mathbf{m} \quad\Rightarrow\quad (\forall p\in Span(\mathbf{m}), \quad s.t. \quad p>0, \quad then \quad R_\mathbf{V}(p)>0). \tag{3.8}$$

Proof Based on the definition of the realizability property

$$\mathbf{V}\in\mathcal{R}_\mathbf{m} \quad\Rightarrow\quad (\exists\psi\in L^1(S)^+, \quad \text{s.t.} \quad \mathbf{V}=\langle\mathbf{m}\psi\rangle\,).$$

Suppose $p=\boldsymbol{\lambda}\mathbf{m}>0$ is a positive polynomial, then the product $p\psi>0$ is positive and therefore the integral

$$R_\mathbf{V}(p)=\langle\psi p\rangle>0.$$

□

In Sections 3.3 and 3.4, the condition (3.8) is studied when $S=[-1,1]$ and $S=S^2$.

3.3 Moment realizability in 1D

This section is devoted to prove Theorems 3.1 and 3.2. This theorem was proven *e.g.* in [16, 1, 9] in a slightly different way than presented here. The present proof is devoted to introduce the 3D problem of the next section and the basics of the atomic closure presented in [24, 30, 29] and recalled in Chapter 4 Section 4.7.2.

The vector of polynomials $\mathbf{m}(\mu)$ is chosen to generate all the polynomials of degree less or equal to N. The vector of monomials of degree up to K is denoted

$$\mathbf{m}_K(\mu)=(1,\mu,...,\mu^K).$$

The vector $\mathbf{m}_N$ is composed of linearily indepent polynomials generating $\mathbb{R}^N[X]$. For simplicity, the following choice of $\mathbf{m}$ is made

$$\mathbf{m}=\mathbf{m}_N.$$

3.3.1 Positive polynomials on the interval $[-1, 1]$

This subsection is devoted to motivate the use of moment matrices.

The condition (3.8) requires knowledge about the set of positive polynomials on $[-1, 1]$ as the condition $R_{\mathbf{V}}(p) > 0$ needs to be satisfied for all positive polynomials p of degree less or equal to N.

First the following subset of positive polynomials of degree $2K$ on $[-1, 1]$ is defined.

Definition 3.5 *A polynomial $p \in \mathbb{R}^{2K}[X]$ of degree $2K$ is a sum of squares if it has the form*

$$p = \sum_{i=1}^{J} p_i^2, \quad \textit{for some } J \in \mathbb{N} \textit{ and some} \quad (p_i)_{\{i=1,\dots,J\}} \in (\mathbb{R}^K[X])^J.$$

The set of sum of squares of polynomials of degree K is denoted $\Sigma^{2K}[X]$.

Remark that a square of a polynomial $q = \boldsymbol{\lambda}\mathbf{m}$ for some $\boldsymbol{\lambda} \in \mathbb{R}^{Card(\mathbf{m})}$ can be written

$$q^2 = (\boldsymbol{\lambda}\mathbf{m})^2 = \boldsymbol{\lambda}^T(\mathbf{m} \otimes \mathbf{m})\boldsymbol{\lambda}.$$

Example 3.3 *Choose $\mathbf{m}(\mu) = (1, \ \mu)$. The polynomial*

$$p(\mu) = 6 + 6\mu + 2\mu^2$$

is a sum of squares, because it can be written e.g.

$$p(\mu) \ = \ (1+\mu)^2 + 1^2 + (2+\mu)^2 \ = \ \sum_{i=1}^{3} \boldsymbol{\lambda}_i^T \left(\mathbf{m}(\mu) \otimes \mathbf{m}(\mu)\right) \boldsymbol{\lambda}_i,$$

where the coefficients $\boldsymbol{\lambda}_i$ read

$$\boldsymbol{\lambda}_1 = (1, \ 1), \qquad \boldsymbol{\lambda}_2 = (1, \ 0), \qquad \boldsymbol{\lambda}_3 = (2, \ 1).$$

Obviously sums of squares are non-negative, and the following characterizations are a motivation to study moment matrices $R_{\mathbf{V}}(\mathbf{m} \otimes \mathbf{m})$.

Proposition 3.3 **([26, 28])** ***Even case:***

All polynomials $p \in \mathbb{R}^{2K}[X]$ of degree $2K$, strictly positive on $[-1, 1]$ have the form

$$p(\mu) = p_1(\mu) + (1 - \mu^2)p_2(\mu), \tag{3.9}$$

where $p_1 \in \Sigma^{2K}[X]$ and $p_2 \in \Sigma^{2K-2}[X]$ are sums of squares.

Proposition 3.4 ([26, 28]) ***Odd case:***
All polynomials $p \in \mathbb{R}^{2K+1}[X]$ of degree $2K+1$, strictly positive on $[-1,1]$ have the form

$$p(\mu) = (1-\mu)p_1(\mu) + (1+\mu)p_2(\mu), \tag{3.10}$$

where $p_1 \in \Sigma^{2K}[X]$ and $p_2 \in \Sigma^{2K}[X]$ are sums of squares.

The even and odd cases are subcases of both Putinar [26] and Schmüdgen [28] Positivstellensätze for univariate polynomials.

Propositions 3.3 and 3.4 can be rewritten under matricial form.

Even case:
A polynomial $p \in \mathbb{R}^{2K}[X]$ is strictly positive on $[-1,1]$ iff there exists coefficients $(\boldsymbol{\lambda}_{1,i})_{\{i=1,\ldots,J\}} \in \mathbb{R}^{K\times J}$ and $(\boldsymbol{\lambda}_{2,i})_{\{i=1,\ldots,J\}} \in \mathbb{R}^{K-1\times J}$ such that

$$p \;=\; \sum_{i=1}^{J} \left[\boldsymbol{\lambda_{1,i}}^T \left(\mathbf{m}_K \otimes \mathbf{m}_K\right) \boldsymbol{\lambda_{1,i}} \;+\; \boldsymbol{\lambda_{2,i}}^T \left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right) \boldsymbol{\lambda_{2,i}} \right].$$

Odd case:
A polynomial $p \in \mathbb{R}^{2K+1}[X]$ is strictly positive on $[-1,1]$ iff there exists coefficients
$(\boldsymbol{\lambda}_{1,i})_{\{i=1,\ldots,J\}} \in \mathbb{R}^{K\times J}$ and $(\boldsymbol{\lambda}_{2,i})_{\{i=1,\ldots,J\}} \in \mathbb{R}^{K\times J}$ such that

$$p \;=\; \sum_{i=1}^{J} \left[\boldsymbol{\lambda_{1,i}}^T \left((1-\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) \boldsymbol{\lambda_{1,i}} \;+\; \boldsymbol{\lambda_{2,i}}^T \left((1+\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) \boldsymbol{\lambda_{2,i}} \right].$$

Using Proposition 3.2, this leads to writing the following necessary conditions for a vector to be realizable.

Even case:
A vector $\mathbf{V} \in \mathcal{R}_{\mathbf{m}_{2K}}$ is realizable only if

$$R_{\mathbf{V}}\left(\mathbf{m}_K \otimes \mathbf{m}_K\right) \;>\; 0, \qquad R_{\mathbf{V}}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right) > 0.$$

Odd case:
A vector $\mathbf{V} \in \mathcal{R}_{\mathbf{m}_{2K+1}}$ is realizable only if

$$R_{\mathbf{V}}\left((1-\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) \;>\; 0, \qquad R_{\mathbf{V}}\left((1+\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) > 0.$$

In the next subsections, those conditions are shown to be sufficient for $\mathbf{V}$ to be realizable.

3.3.2 The truncated Hausdorff moment problem

The aim of this subsection is to exhibit the solution of the truncated Hausdorf moment problem in theorem 3.1 and 3.2, *i.e.* to give a necessary and sufficient condition on matrices to be realizable moment matrices. This results was proven *e.g.* in [9]. A proof of this theorem is provided in order to exhibit the particular representing measure (3.30) to realizable vector. This proof provides some understanding on potential representing measures for a moment vector $\mathbf{V}$, especially when $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}}$ is on the boundary of the realizability domain. Some of the ideas presented in this section will be generalized for the multi-D problem and will be used in the next chapter to construct a realizable closure. Furthermore, the representing measure exhibited in this subsection can be used to construct an atomic based closure which differs from the Kershaw closure as described in [16, 24, 30, 29].

The principle of the method consists in exhibiting a measure representing any vector $\mathbf{V}$ satisfying particular constraints. This measure is decomposed into a regular part, *i.e* the Lesbesgues measure multiplied by a positive scalar, and a discrete measure, *i.e.* a sum of Dirac measures. One also observes that the regular part vanishes on the boundary of the realizability domain and vectors on the boundary of the realizability domain can only be represented by a unique measure which is a sum of Dirac measures.

First the following lemmas characterize potential singularities of the moment matrices.

Lemma 3.1 (Even case) *Suppose* $\mathbf{V} \in \mathbb{R}^{2K}$ *such that*

$$R_{\mathbf{V}}\left(\mathbf{m}_K \otimes \mathbf{m}_K\right) \geq 0, \tag{3.13a}$$

$$R_{\mathbf{V}}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right) \geq 0, \tag{3.13b}$$

and at least one of those matrices is singular.

Even case 1: *In the case when*

$$rank\left(R_{\mathbf{V}}\left(\mathbf{m}_K \otimes \mathbf{m}_K\right)\right) \leq rank\left(R_{\mathbf{V}}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right)\right), \tag{3.14}$$

there exists a polynomial $p \in \mathbb{R}^J[X]$ *of degree* $J \leq K$ *with distinct real roots included in* $]-1,1[$*, such that*

$$R_{\mathbf{V}}(p^2) = 0. \tag{3.15}$$

Even case 2: *In the case when*

$$rank\left(R_{\mathbf{V}}\left(\mathbf{m}_K \otimes \mathbf{m}_K\right)\right) > rank\left(R_{\mathbf{V}}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right)\right), \tag{3.16}$$

there exists a polynomial $p \in \mathbb{R}^J[X]$ *of degree* $J \leq K-1$ *with distinct real roots included in* $]-1,1[$*, such that*

$$R_{\mathbf{V}}((1-\mu^2)p^2) = 0. \tag{3.17}$$

Lemma 3.2 (Odd case) *Suppose* $\mathbf{V} \in \mathbb{R}^{2K+1}$ *such that*

$$\begin{aligned} R_{\mathbf{V}}\left((1+\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) &\geq 0, && (3.18a)\\ R_{\mathbf{V}}\left((1-\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) &\geq 0, && (3.18b)\end{aligned}$$

and at least one of those matrices is singular.

Odd case 1: *In the case when*

$$rank\left(R_{\mathbf{V}}\left((1+\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right)\right) \geq rank\left(R_{\mathbf{V}}\left((1-\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right)\right), \qquad (3.19)$$

there exists a polynomial $p \in \mathbb{R}^J[X]$ *of degree* $J \leq K$ *with distinct real roots included in* $]-1,1[$*, such that*

$$R_{\mathbf{V}}((1-\mu)p^2) \;=\; 0. \qquad (3.20)$$

Odd case 2: *In the case when*

$$rank\left(R_{\mathbf{V}}\left((1+\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right)\right) \leq rank\left(R_{\mathbf{V}}\left((1-\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right)\right), \qquad (3.21)$$

there exists a polynomial $p \in \mathbb{R}^J[X]$ *of degree* $J \leq K$ *with distinct real roots included in* $]-1,1[$*, such that*

$$R_{\mathbf{V}}((1+\mu)p^2) \;=\; 0. \qquad (3.22)$$

Proof For the sake of brevity, the proof is written for the Even case 1. The result in the other cases can be obtained by similar computations.

According to the hypothesis of the lemma and to (3.14), the matrix (3.13a) is singular. Then there exists $\boldsymbol{\lambda} \in \mathbb{R}^K \backslash \{0_{\mathbb{R}^K}\}$ such that

$$\boldsymbol{\lambda}^T R_{\mathbf{V}}\left(\mathbf{m}_K \otimes \mathbf{m}_K\right)\boldsymbol{\lambda} = R_{\mathbf{V}}\left((\boldsymbol{\lambda}\mathbf{m}_K)^2\right) = 0. \qquad (3.23)$$

Define the polynomial $p = \boldsymbol{\lambda}\mathbf{m}_K \in \mathbb{R}^K[X]$. If several $\boldsymbol{\lambda} \in \mathbb{R}^K$ satisfy (3.23), we choose the one that provides a polynomial $p = \boldsymbol{\lambda}\mathbf{m}_K$ of lowest degree. In the following p is proven to have real distinct roots in $[-1,1]$.

p **is real rooted:** By contradiction, suppose p has a pair of complex roots, *i.e.* it has the form

$$p = (\mu^2 + 2b\mu + c^2)q \quad \text{with} \quad q \in \mathbb{R}^{K-2}[X], \quad (b,c) \in \mathbb{R}^2 \quad \text{s.t.} \quad b^2 < c.$$

Computing the Riesz functional of p^2 reads

$$\begin{aligned} 0 &= R_{\mathbf{V}}\left((\mu^2+2b\mu+c)^2q^2\right) = R_{\mathbf{V}}\left(\left[(\mu+b)^2 + \left(c-b^2\right)\right]^2 q^2\right)\\ &= R_{\mathbf{V}}\left((\mu+b)^4 q^2\right) + \kappa_1 R_{\mathbf{V}}\left((\mu+b)^2 q^2\right) + \kappa_2 R_{\mathbf{V}}\left(q^2\right), \qquad (3.24)\\ \kappa_1 &= 2\left(c-b^2\right) > 0, \quad \kappa_2 = \left(c-b^2\right)^2 > 0.\end{aligned}$$

Remark that each polynomial $(\mu+b)^4q^2$, $(\mu+b)^2q^2$ and q^2 is the square of a polynomial. Thus, according to (3.13a) each term in (3.24) is non-negative. Since their sum is zero, each term is zero. In particular $R_{\mathbf{V}}\left(q^2\right)=0$ which violates the hypothesis that p is the polynomial of lowest degree satisfying (3.23).

p **has distinct roots:** By contradiction, suppose that one of the roots of p is a double root, *i.e.* p has the form

$$p=\boldsymbol{\lambda}\mathbf{m}_K=(\mu-\mu_1)^2q\in\mathbb{R}^J[X].$$

The matrix $R_{\mathbf{V}}(\mathbf{m_K}\otimes\mathbf{m_K})$ is real symmetric positive semi-definite and therefore diagonalizable. If $\boldsymbol{\lambda}\in\mathbb{R}^K$ is such that

$$\boldsymbol{\lambda}^TR_{\mathbf{V}}(\mathbf{m_K}\otimes\mathbf{m_K})\boldsymbol{\lambda}=0,$$

then one actually has

$$R_{\mathbf{V}}(\mathbf{m_K}\otimes\mathbf{m_K})\boldsymbol{\lambda}=0_{\mathbb{R}^K}. \tag{3.25}$$

Multiplying (3.25) by the vector $\boldsymbol{\lambda}_2\in\mathbb{R}^K$ such that $\boldsymbol{\lambda}_2\mathbf{m}_K=q$ leads to write that

$$R_{\mathbf{V}}\left((\mathbf{m_K}\boldsymbol{\lambda}_2)(\mathbf{m_K}\boldsymbol{\lambda})\right)=R_{\mathbf{V}}\left(q^2(\mu-\mu_1)^2\right)=0.$$

Therefore the polynomial $p_2=q(\mu-\mu_1)$ satisfy (3.23) which violates the hypothesis that p is the polynomial of lowest degree satisfying (3.23)

The roots of p are included in $[-1,1]$: By contradiction, suppose that one of the root μ_1 of p is bigger than 1, *i.e.* p has the form

$$(\mu-\mu_1)q$$

with $q\in\mathbb{R}^{K-1}[X]$ and $\mu_1\in]+1,+\infty[$. Computing the Riesz functional of p^2 reads

$$\begin{aligned}
R_{\mathbf{V}}\left((\mu-\mu_1)^2q^2\right) &= \kappa_1R_{\mathbf{V}}\left(q^2\right)+\kappa_2R_{\mathbf{V}}\left((1-\mu^2)q^2\right)\\
&\quad+\kappa_3R_{\mathbf{V}}\left((1-\mu)^2q^2\right)+\kappa_4R_{\mathbf{V}}\left((1+\mu)^2q^2\right),\\
\kappa_1 &= (1-\mu_1)^2>0,\quad \kappa_2=|1-\mu_1|>0,\\
\kappa_3 &= \frac{|\mu_1|+\mu_1}{2}\geq 0,\quad \kappa_4=\frac{|\mu_1|-\mu_1}{2}\geq 0,
\end{aligned}$$

which leads again to $R_{\mathbf{V}}\left(q^2\right)=0$ and violates the hypothesis that p is the polynomial of lowest degree satisfying (3.23). □

Under the hypothesis of Lemmas 3.1 and 3.2, one can prove the existence of a representing measure for $\mathbf{V}$.

Lemma 3.3 (Even case) *If* $\mathbf{V} \in \mathbb{R}^{2K}$ *satisfies*

$$R_{\mathbf{V}}(\mathbf{m}_K \otimes \mathbf{m}_K) \geq 0, \qquad R_{\mathbf{V}}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right) \geq 0,$$

and such that one of those matrices is singular, then there exists a representing measure for $\mathbf{V}$*.*

Lemma 3.4 (Odd case) *If* $\mathbf{V} \in \mathbb{R}^{2K+1}$ *satisfies*

$$R_{\mathbf{V}}\left((1-\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) \geq 0, \qquad R_{\mathbf{V}}\left((1+\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) \geq 0,$$

and such that one of those matrices is singular, then there exists a representing measure for $\mathbf{V}$*.*

Proof The proof is again written for the even case when

$$\operatorname{rank}\left(R_{\mathbf{V}}(\mathbf{m}_K \otimes \mathbf{m}_K)\right) \leq \operatorname{rank}\left(R_{\mathbf{V}}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right)\right),$$

and the other cases can be obtained by reproducing those computations.

Lemma 3.1 provides the existence of a polynomial $p \in \mathbb{R}^J[X]$ of degree $J \leq K$ having J distinct roots $(\mu_i)_{\{i=1,\dots,J\}}$ in the interval $[-1,1]$ such that $R_{\mathbf{V}}(p^2) = 0$. Among the possible p, we choose one of lowest degree.

Suppose there exists a measure γ which moments are $\mathbf{V}$, then

$$\int_{-1}^{+1} p^2(\mu) d\gamma(\mu) = 0.$$

Since p^2 is non-negative, if such a measure γ exists, its support $Supp(\gamma)$, *i.e.* the intersection of all closed set the measure of which complements is zero, needs to be included in the set $Z(p)$ of zeros of p

$$Supp(\gamma) \subset Z(p). \tag{3.26}$$

The only measures having a union of a finite number of singletons μ_i for support have the form

$$\gamma = \sum_{i=1}^{J} \alpha_i \delta(\mu - \mu_i). \tag{3.27}$$

Therefore, the rest of the proof consists in proving that there exists a unique set of coefficients $\alpha_i > 0$ such that the measure (3.26) has $\mathbf{V}$ for moments.

The moments according to γ of the form (3.27) read

$$\int_{-1}^{+1} \mathbf{m}_{\mathbf{K}}(\mu) \otimes \mathbf{m}_{\mathbf{K}}(\mu) d\gamma(\mu) = \sum_{i=1}^{J} \alpha_i \mathbf{m}_{\mathbf{K}}(\mu_i) \otimes \mathbf{m}_{\mathbf{K}}(\mu_i).$$

Then γ is a representing measure for $\mathbf{V}$ if non-negative scalars α_i can be found such that

$$R_{\mathbf{V}}(\mathbf{m}_{\mathbf{K}} \otimes \mathbf{m}_{\mathbf{K}}) = \sum_{i=1}^{J} \alpha_i \mathbf{m}_{\mathbf{K}}(\mu_i) \otimes \mathbf{m}_{\mathbf{K}}(\mu_i).$$

Write $\mathbf{V}_J$ the truncation of the vector $\mathbf{V}$ up to the J-th component. By rearranging the terms composing the matrices $\mathbf{m}_{\mathbf{K}}(\mu_i) \otimes \mathbf{m}_{\mathbf{K}}(\mu_i)$ into vector and removing the redundant terms, one can rewrite this problem

$$M\boldsymbol{\alpha} = \mathbf{V}_J,$$

where M is a Vandermonde matrix, *i.e.* $M_{i,j} = \mu_i^{j-1}$ and $\boldsymbol{\alpha} \in \mathbb{R}^J$ is the desired vector of scalars. Vandermonde matrices are known to be invertible as long as the μ_i are distinct which is verified here according to Lemma 3.1. It remains to verify that the coefficients $\alpha_i \geq 0$ are positive.

Let us define the Lagrange polynomials

$$q_i(\mu) = \prod_{\substack{j=1 \\ j \neq i}}^{J} (\mu - \mu_j),$$

then computing the moment of γ according to q_i^2 reads

$$R_{\mathbf{V}}(q_i^2) = \alpha_i q_i(\mu_i)^2,$$

which is strictly positive according to the hypothesis that p is the polynomial of lowest degree such that $R_{\mathbf{V}}(p^2) = 0$. Since the zeros of p are distinct, $q_i(\mu_i)^2$ are strictly positive. Therefore the coefficients $\alpha_i > 0$, which leads to the result. □

This proof provides several information on the representing measure on the boundary of the realizability domain.

Definition 3.6 *Suppose $\mathbf{V} \in \mathcal{R}_{\mathbf{m}}^m$ is represented by a measure*

$$\gamma = \sum_i \alpha_i \delta(\mu - \mu_i)$$

composed only of Dirac measures. Then γ is called atomic representing measure for $\mathbf{V}$. Each Dirac measure $\delta(\mu - \mu_i)$ is called an atom.

A atomic representing measure composed of r atoms is called r-atomic representing measure.

Remark 3.2 *The proof of Lemma 3.3 and 3.4 provides the uniqueness of the representing measure, which is atomic, under the hypothesis of Lemmas 3.3 and*

3.4 which can be interpreted as requiring that $\mathbf{V} \in \partial \mathcal{R}^m_{\mathbf{m}}$ is on the boundary of the realizability domain.

With those lemmas, one obtains the following characterization of $\mathcal{R}^m_{\mathbf{m}}$ in 1D.

Theorem 3.1 (Truncated Hausdorff moment problem [1, 9]) ***Even case:***
Consider $\mathbf{V} \in \mathbb{R}^{2K}$. The vector $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}_{2K}}$ iff

$$R_{\mathbf{V}}(\mathbf{m}_K \otimes \mathbf{m}_K) \geq 0, \qquad R_{\mathbf{V}}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right) \geq 0. \tag{3.28}$$

Theorem 3.2 (Truncated Hausdorff moment problem [1, 9]) ***Odd case:***
Consider $\mathbf{V} \in \mathbb{R}^{2K+1}$. The vector $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}_{2K+1}}$ iff

$$R_{\mathbf{V}}\left((1-\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) \geq 0, \qquad R_{\mathbf{V}}\left((1+\mu)\mathbf{m}_K \otimes \mathbf{m}_K\right) \geq 0. \tag{3.29}$$

Proof Necessary condition: Using Propositions 3.2 and 3.3, if $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}}$ then (3.28) and (3.29) are satisfied.

Sufficient condition: The sufficiency part is proven by exhibiting one possible representing measure, *i.e.* a positive constant multiplied by the Lesbegues measure plus a sum of Dirac measures.

The proof is again given for the even case. The result for the odd case can be obtained by adapting the computations.

If one of the matrices (3.28) is singular, the existence of a representing measure is given by Lemma 3.3.

Suppose that the matrices (3.28) are non-singular, they are therefore strictly positive according to the hypothesis.

Define $\mathbf{V}_0$ the moments according to the Lebesgues measure

$$\mathbf{V}_0 = \int_{-1}^{+1} \mathbf{m}(\mu) d\mu,$$

and define the functions

$$\begin{aligned}
\mathbf{W}(x) &= \mathbf{V}_0 - x\mathbf{V}, \\
M_1(x) &= R_{\mathbf{W}(x)}(\mathbf{m}_K \otimes \mathbf{m}_K) \\
&= R_{\mathbf{V}_0}(\mathbf{m}_K \otimes \mathbf{m}_K) - xR_{\mathbf{V}}(\mathbf{m}_K \otimes \mathbf{m}_K), \\
M_2(x) &= R_{\mathbf{W}(x)}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right) \\
&= R_{\mathbf{V}_0}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right) - xR_{\mathbf{V}_0}\left((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1}\right).
\end{aligned}$$

Remark that M_1 and M_2 are matrices of linear functions of $x \in \mathbb{R}$, in particular their components are C^∞ functions of x. Furthermore, based on the hypothesis and on the definition of $\mathbf{V}_0$, they satisfy

$$\begin{aligned} \det(M_1(0)) > 0, &\qquad \lim_{x\to+\infty} \det(M_1(x)) < 0, \\ \det(M_2(0)) > 0, &\qquad \lim_{x\to+\infty} \det(M_2(x)) < 0. \end{aligned}$$

Using the intermediate value theorem, there exists positive scalars $x > 0$ such that $\det(M_1(x)) = 0$ and/or $\det(M_2(x)) = 0$. Write y the minimum of those values. At that point, one has

$$M_1(y) \geq 0, \qquad M_2(y) \geq 0,$$

and one of those matrices is singular. According to Lemma 3.3, there exists a unique representing measure γ for $W(y)$ which has the form

$$d\gamma(\mu) = \sum_i \alpha_i \delta(\mu - \mu_i),$$

where the scalars μ_i are the zeros of the characteristic polynomial of $M_1(y)$ or $M_2(y)$ (the one of minimal dimension or $M_1(y)$ if they have the same dimension) and the coefficients α_i are the unique (positive) scalars such that this measure has $W(y)$ for moments.

Finally one verifies that the following measure

$$d\gamma(\mu) + y d\mu \tag{3.30}$$

is positive and has $\mathbf{V}$ for moments. □

The idea of this atomic decomposition follows from the form of the unique representing measure on the boundary $\partial\mathcal{R}^m_{\mathbf{m}}$ and from the work of R. Curto and L. Fialkow ([9, 11, 12, 10]). This representing measure could be used to construct closure of 1D moment equations.

3.3.3 The realizability condition in 1D

The existence of a representing measure γ for a vector $\mathbf{V} \in int(\mathcal{R}^m_{\mathbf{m}})$ in the interior implies the existence of a positive $L^1([-1,1])^+$ representing function ψ according (3.6). This leads to

Corollary 3.1 ***Even case:***
Consider $\mathbf{V} \in \mathbb{R}^{2K}$. The vector $\mathbf{V} \in \mathcal{R}_{\mathbf{m}_{2K}}$ is realizable iff

$$R_{\mathbf{V}}(\mathbf{m}_N \otimes \mathbf{m}_N) > 0, \qquad R_{\mathbf{V}}\left((1-\mu^2)\mathbf{m}_{N-1} \otimes \mathbf{m}_{N-1}\right) > 0. \tag{3.31}$$

Corollary 3.2 ***Odd case:***

Consider $\mathbf{V} \in \mathbb{R}^{2K+1}$*. The vector* $\mathbf{V} \in \mathcal{R}_{\mathbf{m}_{2K+1}}$ *is realizable iff*

$$R_{\mathbf{V}}\left((1-\mu)\mathbf{m}_N \otimes \mathbf{m}_N\right) > 0, \qquad R_{\mathbf{V}}\left((1+\mu)\mathbf{m}_N \otimes \mathbf{m}_N\right) > 0. \tag{3.32}$$

The next chapter is devoted to the closure problem. Commonly one aims to define a closure ψ^{N+1} such that the vector $(\psi^0, ..., \psi^N, \psi^{N+1}) \in \mathcal{R}_{\mathbf{m}_{N+1}}$ is realizable. Corollaries 3.1 and 3.2 leads to the following requirements.

Corollary 3.3 **([16, 1])** ***Even case:***
Consider the vector $\mathbf{V} = (\psi^0, ..., \psi^{2K-1}, \psi^{2K}) \in \mathbb{R}^{2K}$ *and write*

$$\bar{\mathbf{V}} := (\psi^0, ..., \psi^{2K-1}, 0)$$

the vector $\mathbf{V}$ *with the last component replaced by 0.*
The vector $\mathbf{V} \in \mathcal{R}_{\mathbf{m}_{2K+2}}$ *is realizable iff*

1. *The truncated vector* $(\psi^0, ..., \psi^{2K-1}) \in \mathcal{R}_{\mathbf{m}_{2K-1}}$ *is realizable*

2. *The last component* ψ^{2K} *is bounded by*

$$-\frac{\det\left(R_{\bar{\mathbf{V}}}(\mathbf{m}_K \otimes \mathbf{m}_K)\right)}{\det\left(R_{\mathbf{V}}(\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1})\right)} \leq \psi^{2K}, \tag{3.33a}$$

$$\psi^{2K-2} + \frac{\det\left(R_{\bar{\mathbf{V}}}((1-\mu^2)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1})\right)}{\det\left(R_{\mathbf{V}}((1-\mu^2)\mathbf{m}_{K-2} \otimes \mathbf{m}_{K-2})\right)} \geq \psi^{2K}. \tag{3.33b}$$

Corollary 3.4 **([16, 1])** ***Odd case:***
Consider the vector $\mathbf{V} := (\psi^0, ..., \psi^{2K}, \psi^{2K+1}) \in \mathbb{R}^{2K+1}$ *and write*

$$\bar{\mathbf{V}} := (\psi^0, ..., \psi^{2K}, 0)$$

the vector $\mathbf{V}$ *with the last component replaced by 0.*
The vector $\mathbf{V} \in \mathcal{R}_{\mathbf{m}_{2K+1}}$ *is realizable iff*

1. *The truncated vector* $(\psi^0, ..., \psi^{2K}) \in \mathcal{R}_{\mathbf{m}_{2K}}$ *is realizable*

2. *The last component* ψ^{2K+1} *is bounded by*

$$-\psi^{2K} - \frac{\det\left(R_{\bar{\mathbf{V}}}((1+\mu)\,\mathbf{m}_K \otimes \mathbf{m}_K)\right)}{\det\left(R_{\mathbf{V}}((1+\mu)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1})\right)} \leq \psi^{2K+1}, \tag{3.34a}$$

$$\psi^{2K} + \frac{\det\left(R_{\bar{\mathbf{V}}}((1-\mu)\,\mathbf{m}_K \otimes \mathbf{m}_K)\right)}{\det\left(R_{\mathbf{V}}((1-\mu)\mathbf{m}_{K-1} \otimes \mathbf{m}_{K-1})\right)} \geq \psi^{2K+1}. \tag{3.34b}$$

The conditions (3.33a-3.33b) and (3.34a-3.34b) for the moment of order up to 3 have the following form ([16, 24]). By extension, one also observes the following condition for scalars to be a moment of order 0 of a positive function.

$$\begin{aligned}
&\textbf{N=0:} & 0 &< \psi^0, && (3.35a)\\
&\textbf{N=1:} & -\psi^0 &< \psi^1 < \psi^0, && (3.35b)\\
&\textbf{N=2:} & \frac{|\psi^1|^2}{\psi^0} &< \psi^2 < \psi^0, && (3.35c)\\
&\textbf{N=3:} & -\psi^2+\frac{(\psi^1+\psi^2)^2}{\psi^0+\psi^1} &< \psi^3 < \psi^2-\frac{(\psi^1-\psi^2)^2}{\psi^0-\psi^1}. && (3.35d)
\end{aligned}$$

Finally the realizability domain for the first order moments in 1D is characterized by

$$\begin{aligned}
\mathcal{R}_{(1,\mu)} &= \{(\psi^0,\psi^1) \in \mathbb{R}^2, \quad \text{s.t.} \quad |\psi^1| < \psi^0\}, && (3.36a)\\
\mathcal{R}_{(1,\mu,\mu^2)} &= \{(\psi^0,\psi^1,\psi^2) \in \mathbb{R}^3, \quad \text{s.t.} \quad |\psi^1| < \psi^0, \quad \frac{|\psi^1|^2}{\psi^0} < \psi^2 < \psi^0\}. && (3.36b)
\end{aligned}$$

The convex cones (3.36a) and (3.36b) are represented on Fig. 3.1 and 3.2. The first set is represented in the plan $(\psi^0,\psi^1) \in \mathbb{R}^2$, the second is represented in the plan $\left(\frac{\psi^1}{\psi^0}, \frac{\psi^2}{\psi^0}\right) \in \mathbb{R}^2$.

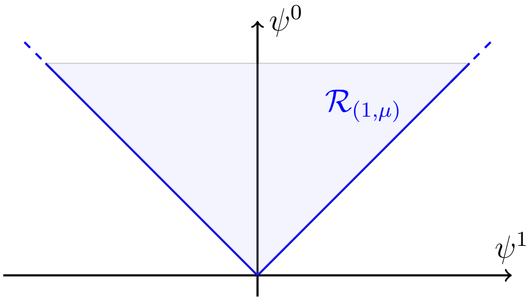

Figure 3.1: Realizability domain $\mathcal{R}_{(1,\mu)}$ for moments of order up to 1 in the space $(\psi^0,\psi^1) \in \mathbb{R}^2$.

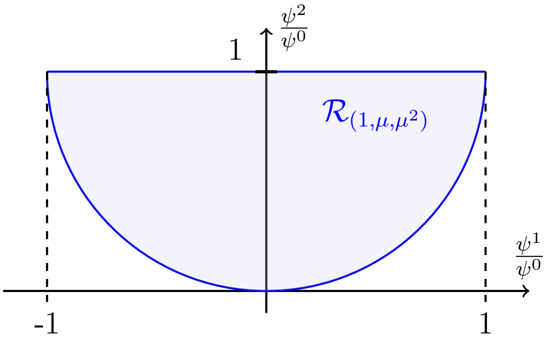

Figure 3.2: Realizability domain $\mathcal{R}_{(1,\mu,\mu^2)}$ for moments of order up to 2 in the space $(\frac{\psi^1}{\psi^0}, \frac{\psi^2}{\psi^0}) \in \mathbb{R}^2$.

3.4 Moment realizability in 3D

This section is devoted to introduce results about the realizability domain in 3D which are used in the next chapter. Here the truncated moment problem on the unit sphere S^2 is focused on.

3.4.1 Polynomials on S^2

Before studying the truncated moment problem, one needs to choose a set of polynomials $\mathbf{m}$ defined on S^2. Choosing S^2 as the set of integration (and as the set of definition of the polynomials) leads to the following major differences compared to the 1D case:

1. Choosing $\Omega \in S^2 \subset \mathbb{R}^3$ forces us to study multivariate polynomials $\mathbf{m}(\Omega)$. Such multivariate polynomials can not be factorized as a product of polynomials of degree one or two.
2. The set of integration S^2 is a compact algebraic variety defined by

$$S^2 = \left\{ \Omega \in \mathbb{R}^3, \quad \text{s.t.} \quad 1 - (\Omega_1^2 + \Omega_2^2 + \Omega_3^2) = 0 \right\} \subset \mathbb{R}^3.$$

Similarily to the 1D case, one typically aim to choose the vector of polynomials $\mathbf{m}$ such that it is composed of linearily independent polynomials and it generates all the polynomials of degree less or equal to N. Although, the meaning of "independent" and of "generating" is ambiguous for multivariate polynomials on S^2. For this purpose, the following set is defined.

Definition 3.7 *Define the polynomial*

$$p_0(X_1, X_2, X_3) := 1 - (X_1^2 + X_2^2 + X_3^2),$$

which is zero on S^2.

The set I^N is the set of polynomials of degree N generated by p_0

$$I^N := p_0 \mathbb{R}^{N-2}[X_1, X_2, X_3] = \left\{ p\ p_0, \quad p \in \mathbb{R}^{N-2}[X_1, X_2, X_3] \right\}. \tag{3.37}$$

First the following proposition provides an additional constraint on vectors to be realizable.

Proposition 3.5 *Consider a vector $\mathbf{V} \in \mathcal{R}_{\mathbf{m}}^m$, then*

$$R_{\mathbf{V}}(I^N) = \{0\}.$$

Proof Consider $\mathbf{V} \in \mathcal{R}_{\mathbf{m}}$, then $\mathbf{V}$ is composed of moments of a function $\psi \in L^1(S^2)^+$. All the polynomials of $p \in I^N$ are zero on S^2. Therefore the integral of their product

$$\int_{S^2} p(\Omega)\psi(\Omega) d\Omega = 0$$

is zero. By continuity, it also holds on the boundary of the realizability domain. □

This result is illustrated through the following example that will be used in the next chapter.

Example 3.4 *Consider a vector* $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}}$*, then*

$$R_{\mathbf{V}}\left(1-(\Omega_1^2+\Omega_2^2+\Omega_3^2)\right)=0. \tag{3.38}$$

Write

$$\psi^0 = R_{\mathbf{V}}(1), \qquad \psi^2 = R_{\mathbf{V}}(\Omega\otimes\Omega).$$

Equation (3.38) *provides the following relation between* ψ^0 *and* ψ^2

$$\psi^0-\left(\psi^2_{1,1}+\psi^2_{2,2}+\psi^2_{3,3})\right)=\psi^0-tr(\psi^2)=0.$$

The vector $\mathbf{m}$ is chosen to be linearly independent in the sense

$$Span(\mathbf{m})\cap I^N=\{0\}, \tag{3.39a}$$

and such that it generates $\mathbb{R}^N[X_1,X_2,X_3]$ in the following sense

$$Span(\mathbf{m})\cup I^N=\mathbb{R}^N[X_1,X_2,X_3]. \tag{3.39b}$$

i.e. $\mathbf{m}$ is a linearly independent and generating family modulo I^N.

The vector $\tilde{\mathbf{m}}_K$ is defined to be composed of all monomials of degree K is written

$$\tilde{\mathbf{m}}_K=\left(\Omega_1^K,\quad \Omega_1^{K-1}\Omega_2,\quad \Omega_1^{K-1}\Omega_3,\quad \Omega_1^{K-2}\Omega_2^2,\quad \dots\quad ,\Omega_3^K\right),$$

and the vector of all monomials of degree $K-1$ and K is written

$$\mathbf{m}_K=(\tilde{\mathbf{m}}_{K-1},\tilde{\mathbf{m}}_K).$$

The following choice of $\mathbf{m}$ is made

$$\mathbf{m}=\mathbf{m}_N.$$

Example 3.5 *For the first orders* $N=1$ *and* $N=2$*, one has*

$$\begin{aligned}\mathbf{m}_1(\Omega) &= (1,\quad \Omega_1,\ \Omega_2,\ \Omega_3),\\ \mathbf{m}_2(\Omega) &= \left(\Omega_1,\ \Omega_2,\ \Omega_3,\quad \Omega_1^2,\ \Omega_1\Omega_2,\ \Omega_2^2,\ \Omega_1\Omega_3,\ \Omega_2\Omega_3,\ \Omega_3^2\right).\end{aligned}$$

No polynomial of $Span(\mathbf{m})$ can be factorized by p_0, therefore this choice of $\mathbf{m}$ satisfy (3.39a). One shows that (3.39b) is satisfied by comparing the dimensions of $Span(\mathbf{m})$, I^N and $\mathbb{R}^N[X_1,X_2,X_3]$.

Notation 3.2 *After computations, the number of components of the vector* $\mathbf{m}_N$ *is* $(N+1)^2$*. It is afterward written*

$$c_N := Card(\mathbf{m}_N)=(N+1)^2.$$

For the problem of moments on S^2, positive polynomials are also used. Propositions 3.3 and 3.4 can be generalized into Putinar ([26]) or Schmüdgen ([28]) (see also Stengle [33]) Positivstellensätze in order to write a positive polynomial p on S^2 as a sum of squares

$$0 < p \quad \text{on } S^2 \quad \Leftrightarrow \quad p = \sum_i q_i^2. \tag{3.40}$$

Although for multivariate polynomials, the degree of such a sum of squares representation is not bounded by the degree of p, *i.e.* in (3.40) *a priori* $2deg(q_i) \leq deg(p)$ does not hold ([32, 25]).

Sum of squares are still positive on S^2, but there might be some positive polynomials of degree $2K$ which are not sum of squares of polynomials of degree K. Remark that Proposition 3.3 was not used in Subsection 3.3.2 and 3.3.3 but only motivated the use of moment matrices.

3.4.2 Realizability of first orders moments

Some explicit conditions were found for moments on the unit sphere of order up to one and to two. They recalled here because they will be used in the next chapter.

Realizability condition for moments of order up to one

The realizability condition for moments of order up to one on the unit sphere is characterized by the following proposition.

Proposition 3.6 ([16]) *Consider the vector* $\mathbf{V} \in \mathbb{R}^{c_1}$ *and write*

$$\psi^0 = R_{\mathbf{V}}(1), \qquad \psi^1 = R_{\mathbf{V}}(\Omega).$$

The vector $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}_1}$ *iff*

$$|\psi^1| \leq \psi^0. \tag{3.41}$$

Furthermore, in the equality case, there exists a unique representing measure, which has the form

$$\gamma(\Omega) = \psi^0 \delta\left(\Omega - \frac{\psi^1}{\psi^0}\right),$$

One can easily see that the condition (3.41) is necessary for $\mathbf{V}$ to be in $\mathcal{R}^m_{\mathbf{m}_1}$. This condition was also proven to be sufficient in [16] by exhibiting the existence of the following positive representing measure for $\mathbf{V}$

$$\gamma(\Omega) = \left(\psi^0 + |\psi^1|\right) \delta\left(\Omega - \frac{\psi^1}{|\psi^1|}\right) + \left(\psi^0 - |\psi^1|\right) \delta\left(\Omega + \frac{\psi^1}{|\psi^1|}\right).$$

Here γ is a positive measure as long as (3.41) is satisfied.

This result leads to the following characterization of the realizability domain $\mathcal{R}_{\mathbf{m}_1}$

$$\mathcal{R}_{\mathbf{m}_1} = \left\{ (\psi^0, \psi^1) \in \mathbb{R} \times \mathbb{R}^3, \quad \text{s.t.} \quad |\psi^1| < \psi^0 \right\}. \tag{3.42}$$

This condition can be interpreted at the physical level by the boundedness of the fluxes. Remark the similarilty with the 1D condition (3.35b).

Realizability condition for moments of order up to two

The realizability condition for moments of order up to two on the unit sphere is characterized by the following proposition.

Proposition 3.7 **([16])** *Consider the vector* $\mathbf{V} \in \mathbb{R}^{c_2}$ *and write*

$$\psi^1 = R_{\mathbf{V}}(\Omega), \qquad \psi^2 = R_{\mathbf{V}}(\Omega \otimes \Omega).$$

The vector $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}_2}$ *iff*

$$\psi^0 := tr(\psi^2) \quad \geq \quad 0, \tag{3.43a}$$
$$\psi^0\psi^2 - \psi^1 \otimes \psi^1 \quad \geq \quad 0. \tag{3.43b}$$

Furthermore, if one of the eigenvalues of (3.43b) *is zero, i.e.*

$$\exists W \in \mathbb{R}^3, \quad s.t. \quad W \left(\psi^0\psi^2 - \psi^1 \otimes \psi^1 \right) W = 0,$$

then the support of any representing measure γ *for* $\mathbf{V}$ *is in the set of zeros of the polynomial* $p := (\Omega - \psi^1)W$*, i.e.*

$$Supp(\gamma) \subset Z(p). \tag{3.44}$$

One verifies that the condition (3.43) is necessary for $\mathbf{V}$ to be in $\mathcal{R}^m_{\mathbf{m}_2}$. This condition was also proven to be sufficient in [16] by exhibiting the existence of one representing measure for $\mathbf{V}$ positive as long as $\mathbf{V}$ satisfies (3.43).

This result leads to the following characterization of the realizability domain $\mathcal{R}_{\mathbf{m}_2}$

$$\mathcal{R}_{\mathbf{m}_2} = \left\{ \mathbf{V} \in \mathbb{R}^{c_2}, \quad \text{s.t.} \quad \psi^0\psi^2 - \psi^1 \otimes \psi^1 > 0 \right\}. \tag{3.45}$$

Remark the similarilty with the 1D condition (3.35c).

3.4.3 Partial results for higher order moments

No practical characterization of the realizability condition for high order moments on the unit sphere S^2 is known. This subsection is devoted to (partially) complete the description of the realizability domain in multi-D and to provide few more results. Those results do not provide a practical characterization of the realizability property in the general case but they provide .

The necessary conditions for a vector $\mathbf{V}$ to be realizable proposed in (3.8) can be completed.

For instance, one can first remark that any realizable vector $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}_{2K}}$ verifies the following conditions.

Proposition 3.8 *If* $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}_{2K}}$ *then*

$$R_{\mathbf{V}}(I^{2K}) = \{0\}, \qquad R_{\mathbf{V}}(\mathbf{m}_K \otimes \mathbf{m}_K) \geq 0.$$

Suppose furthermore that there exists a polynomial $0 \leq p \in \mathbb{R}^{2K}[X_1, X_2, X_3]$ *non-negative over* S^2 *such that*

$$R_{\mathbf{V}}(p) = 0.$$

Then the support of γ *is included in the zero set of* p

$$Supp(\gamma) \subset Z(p).$$

Proof If $\mathbf{V} \in \mathcal{R}^m_{\mathbf{m}_{2K}}$ then it has a positive representing measure γ, *i.e.*

$$\mathbf{V} = \int_{S^2} \mathbf{m}_{2K}(\Omega) d\gamma(\Omega).$$

Based on the definition of the Riesz functional, for a non-negative polynomial $p \in \mathbb{R}^{2K}[X_1, X_2, X_3]$

$$R_{\mathbf{V}}(p) = \int_{S^2} p(\Omega) d\gamma(\Omega) = 0$$

implies that $Supp(\gamma) \subset Z(p)$.

The equality $R_{\mathbf{V}}(I^{2K}) = \{0\}$ is provided by Proposition 3.5. Any square of a polynomial is non-negative, and γ is a positive measure. This leads to the inequality. □

A 1D version of this result was also used in the proof of Lemmas 3.3 and 3.4.

One method to characterize the realizability for arbitrary high order moments over S^2 consists in exhibiting a representing measure having the form of a sum of Dirac measures. The following method aims to exhibit an atomic representing measure for a vector $\mathbf{V}^{2K} \in \mathbb{R}^{c_{2K}}$ composed of $2K+2$ atoms instead of $2K$ and, contrarily to the method used to exhibit the measure (3.30) in 1D, this representing measure is composed only of atoms and has no regular part $d\Omega$. For this purpose, the following notations are introduced.

Definition 3.8 *Consider* $\mathbf{V}^{2K} \in \mathbb{R}^{c_{2K}}$ *and write*

$$M^K := R_{\mathbf{V}^{2K}}(\mathbf{m}_K \otimes \mathbf{m}_K).$$

The vector $\mathbf{V}^{2K}$ *(and the associated matrix* M^K*) admits a flat extension* $\mathbf{V}^{2K+2} \in$

$\mathbb{R}^{c_{2K+2}}$ *(associated to* $M^{K+1} \in \mathbb{R}^{c_{K+1} \times c_{K+1}}$*) iff*

$$R_{\mathbf{V}^{2K+2}}(\mathbf{m}_{2K}) = \mathbf{V}^{2K}, \qquad \mathit{rank}\left(M^{K+1}\right) = \mathit{rank}\left(M^{K}\right).$$

Example 3.6 *Choose*

$$\mathbf{m}(\Omega) = \mathbf{m}_0(\Omega) = 1,$$

and a vector $\boldsymbol{\psi}$ *and its associated matrix* M *defined by*

$$\boldsymbol{\psi} = 1 \in \mathbb{R}^{c_0}, \qquad M = R_{\boldsymbol{\psi}}(\mathbf{m}_0 \otimes \mathbf{m}_0) = 1 \in \mathbb{R}^{c_0 \times c_0}.$$

The vector $\bar{\boldsymbol{\psi}}$ *and its associated matrix* $\bar{M}$ *defined by*

$$\begin{aligned} \bar{\boldsymbol{\psi}} &= \left(1,\ \frac{3}{5},\ \frac{4}{5}, 0,\ \frac{9}{25},\ \frac{12}{25},\ \frac{16}{25},\ 0,\ 0,\ 0\right), \\ \bar{M} &= R_{\bar{\boldsymbol{\psi}}}(\mathbf{m}_1 \otimes \mathbf{m}_1) = \begin{pmatrix} 1 & \frac{4}{5} & \frac{3}{5} & 0 \\ \frac{4}{5} & \frac{16}{25} & \frac{12}{25} & 0 \\ \frac{3}{5} & \frac{12}{25} & \frac{9}{25} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

are a flat extension of $\boldsymbol{\psi}$ *and* M*. Indeed one can verify that*

$$\mathit{rank}(M) = \mathit{rank}(\bar{M}) = 1, \qquad R_{\bar{\boldsymbol{\psi}}}(\mathbf{m}_0) = \boldsymbol{\psi} = (1).$$

In order to characterize the realizability of a vector $\mathbf{V}^{2K} \in \mathbb{R}^{c_{2K}}$, one may look for an extension $\mathbf{V}^{2K+2}$ the realizability of which can be proven, *e.g.* by exhibiting one representing measure. Typically one seeks an extension that is represented by a sum of Dirac measures.

Definition 3.9 *A* r*-atomic representing measure for* $\mathbf{V} \in \mathcal{R}^{m}_{\mathbf{m}}$ *is a representing measure* γ *for* $\mathbf{V}$ *of the form*

$$\gamma(\Omega) = \sum_{i=1}^{r} \alpha_i \delta(\Omega - \Omega_i),$$

with $(\alpha_i)_{\{i=1,\ldots,r\}} \in (\mathbb{R}^{*+})^r$ *and* $(\Omega_i)_{\{i=1,\ldots,r\}} \in (S^2)^r$*.*

The notion of flat extension is implicitly related to the existence of an atomic representing measure. The following theorem characterizes the part of the realizability domain, which is represented by r-atomic measures.

Theorem 3.3 ([12]) *Consider* $\mathbf{V}^{2K} \in \mathbb{R}^{c_{2K}}$ *and define*

$$r := \mathit{rank}\left(R_{\mathbf{V}^{2K}}(\mathbf{m}_K \otimes \mathbf{m}_K)\right).$$

There exists a r-atomic representing measure for $\mathbf{V}^{2K}$ *iff*

$$R_{\mathbf{V}^{2K}}(\mathbf{m}_K \otimes \mathbf{m}_K) \geq 0$$

and $\mathbf{V}^{2K}$ *admits a flat extension* $\mathbf{V}^{2K+2}$ *satisfying*

$$\begin{aligned} R_{\mathbf{V}^{2K+2}}(I^{2K+2}) &= \{0\}, & (3.46a)\\ R_{\mathbf{V}^{2K+2}}(\mathbf{m}_{K+1} \otimes \mathbf{m}_{K+1}) &\geq 0. & (3.46b)\end{aligned}$$

Furthermore, this flat extension $\mathbf{V}^{2K+2}$ *admits a unique representing measure, which is r-atomic.*

This theorem can be illustrated through Example 3.6. One can verifiy in this example that $\delta\left(\Omega.(\frac{3}{5},\frac{4}{5},0)-1\right)$ is a representing measure for $\boldsymbol{\psi}$ and the unique representing measure for $\bar{\boldsymbol{\psi}}$.

Remark 3.3 *This theorem provides a method to characterize the realizability property for even order moments. It can also be used for odd order moments in the following way.*

Consider $\mathbf{V}^{2K+1} \in \mathbb{R}^{c_{2K+1}}$ *and write its restriction to the* $2K$*-th order*

$$\mathbf{V}^{2K} := R_{\mathbf{V}^{2K+1}}(\mathbf{m}_{2K}).$$

There exists a r-atomic representing measure for $\mathbf{V}^{2K+1} \in \mathcal{R}^m_{\mathbf{m}_{2K+1}}$ *iff there exists a flat extension* $\mathbf{V}^{2K+2}$ *of* $\mathbf{V}^{2K}$ *satisfying* (3.46) *and such that*

$$R_{\mathbb{V}^{2K+2}}(\mathbf{m}_{2K+1}) = \mathbf{V}^{2K+1}.$$

3.5 Discussion

Due to the positivity of the solution ψ of kinetic equation of the form (2.1) (or (1.11) with the different collision operators (1.14), (1.34) or (1.36) for electron collisions), the solution $\boldsymbol{\psi}$ of the moment equations (2.2) (or (2.3) with the different collision operators (2.11), (2.12) or (2.13) for electron collisions) is expected to be in the realizability domain $\mathcal{R}_{\mathbf{m}}$.

In this chapter, link were made between the realizability domain and the set of positive measures over the domain of integration S. However no practical characterizations were found for arbitrarily large order moments on the unit sphere, and only partial results were provided, *i.e.* necessary but *a priori* not sufficient conditions for a vector to be realizable.

In practice, for every point of the realizability domain, one can reconstruct a representing measure on S. This is one common idea to construct the closure of moment equations, *e.g.* to compute the flux function F in (2.1). However, as described in this chapter, reconstructing such a measure (or a $L^1(S)^+$ function) is not an easy task. Especially, the method presented in this chapter is not general enough and can not be applied to construct representing measure for moments on

S^2 of order more than 2. For practical applications, the closure is often preferred to be written as an analytical smooth function from $\mathcal{R}^m_>$ to $\mathbb{R}^{3\times Card(m)}$ in 3D. In the next chapter, the characterization of the realizability domain in Theorems 3.1 and 3.2 in 1D and in Propositions (3.6) and (3.7) in 3D are exploited in order to construct analytical approximations of the entropy-based M_1 and M_2 closures.

Bibliography

[1] N. I. Akhiezer. *The classical moment problem.* Hafner Publ. Co., 1965.

[2] N. I. Akhiezer and M. G. Krein. *Some questions in the theory of moments.* Translations of mathematical monographs. American Mathematical Society, 1974.

[3] E. Artin. Über die Zerlegung definiter Funktionen in Quadrate. *Abh. Math. Sem. Hamburg*, 5(1):100–115, 1923.

[4] G. Blekherman and J.B. Lasserre. The truncated K-moment problem for closure of open sets. *arXiv:1108.0627*, 2011.

[5] J. Borwein and A. Lewis. Duality relationships for entropy-like minimization problems. *SIAM J. Control Optim.*, 29(2):325–338, 1991.

[6] J. Borwein and A. Lewis. Partially finite convex programming, Part I: Quasi relative interiors and duality theory. *Math. Program.*, 57:15–48, 1992.

[7] J. Borwein and A. Lewis. Partially finite convex programming, Part II. *Math. Program.*, 57:49–83, 1992.

[8] P. L. Chebyshev. Sur les fonctions qui différent le moins possible de zéro. *Notes Acad. Sci.*, XXII(1):189 – 215, 1873.

[9] R. Curto and L. A. Fialkow. Recusiveness, positivity, and truncated moment problems. *Houston J. Math.*, 17(4):603–634, 1991.

[10] R. Curto and L. A. Fialkow. The truncated complex K-moment problem. *T. Am. Math. Soc.*, 352(6):2825–2855, 2000.

[11] R. Curto and L. A. Fialkow. A duality prood to Tchakaloff's theorem. *J. Math. Anal. Appl.*, 269:519–536, 2002.

[12] R. Curto and L. A. Fialkow. Truncated K-moment problems in several variables. *arXiv preprint math/0507067*, 2005.

[13] D. W. Dubois. Note on Artin's solution of Hilbert's 17th problem. *Bull. Am. Math. Soc.*, 73(4):540 – 541, 1967.

[14] C. D. Hauck, C. D. Levermore, and A. L. Tits. Convex duality and entropy-based moment closures: Characterizing degenerate densities. *SIAM J. Control Optim.*, 2007.

[15] M. Junk. Maximum entropy for reduced moment problems. *Math. Mod. Meth. Appl. S.*, 10(1001–1028):2000, 1998.

[16] D. Kershaw. Flux limiting nature's own way. Technical report, Lawrence Livermore Laboratory, 1976.

[17] M. G. Krein and A. A. Nudelman. *The Markov moment problem and extremal problems : Ideas and problems of P. L. Cebysev and A. A. Markov and their further development.* American Mathematical Society, 1977.

[18] H. J. Landau. The classical moment problem: Hilbertian proofs. *J. Funct. Anal.*, 38(2):255 – 272, 1980.

[19] J. B. Lasserre. *Moments, positive polynomials and their applications.* Imperial College press optimization series, 2010.

[20] M. Laurent. Revisiting two theorems of Curto and Fialkow on moment matrices. *Proc. A.M.S.*, 133(10):2965–2976, 2005.

[21] A. A. Markov. Finding the smallest and the largest values of some function that deviates least from zero. *Soobshch. Kharkov Soc. Math.*, 1(1), 1884.

[22] A. A. Markov. On functions of least deviation from zero in a given interval. 1892.

[23] L. R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *J. Math. Phys.*, 25(8):2404–2417, 1984.

[24] P. Monreal. *Moment realizability and Kershaw closures in radiative transfer.* PhD thesis, Rheinisch-Westfälische Technische Hochschule, 2012.

[25] J. Nie and M. Schweighofer. On the complexity of Putinar's Positivstellensatz. *J. Complexity*, 23(1):135 – 150, 2007.

[26] M. Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana U. Math. J.*, 42(3):969–984, 1993.

[27] M. Riesz. Sur le problème des moments, troisième note. *Ark. Math. Astr. Fys.*, 17(16):1–52, 1923.

[28] K. Schmüdgen. The K-moment problem for compact semi-algebraic sets. *Math. Ann.*, 289(2):203–206, 1991.

[29] F. Schneider. *Moment models in radiation transport equations.* PhD thesis, Teschnische Universität Kaiserslautern, 2015.

[30] F. Schneider. Kershaw closures for linear transport equations in slab geometry I: Model derivation. *J. comput. phys.*, pages –, 2016.

[31] J. Schneider. Entropic approximation in kinetic theory. *ESAIM-Math. Model. Num.*, 38(3):541–561, 2004.

[32] M. Schweighofer. On the complexity of Schmüdgen's Positivstellensatz. *J. Complexity*, 20(4):529 – 543, 2004.

[33] G. Stengle. A Nullstellensatz and a Positivstellensatz in semialgebraic geometry. *Math. Ann.*, 207(2):87–97, 1974.

[34] T.-J. Stieltjes. Recherches sur les fractions continues. *Anns. Fac. Sci. Toulouse : Mathématiques*, 8(4):J1–J122, 1894.

Chapter 4

Moment closure

4.1 Introduction

In this chapter, both vectorial and tensorial notations for the moments are used.

As a continuity, the moment problem consists in studying the existence of a representing measure for a vector $\mathbf{V}$, the closure problem consists in chosing one of those representing measures. Indeed, in the last chapter, it was shown that the representing measures (or representing $L^1(S)^+$ functions) for a vector $\mathbf{V}$ was not unique except for particular vectors $\mathbf{V}$ on the boundary of the realizability domain $\partial\mathcal{R}_{\mathbf{m}}$.

The choice of the closure has a significant influence on the properties of the resulting moment model. Among the desired properties one expects from a moment model and that the choice of the closure influence, one can specify hyperbolicity, realizability preservation, entropy dissipation and the accuracy of modelling the physical phenomena presented in Chapter 1, especially the ability of modelling accurately beams of particles is non-trivial. Each of those properties are to be kept in mind when constructing the closure of a moment system of equations.

This chapter is organized as follow. In the next section, some preliminar notations are given and the closure problem is presented. Then the major properties expected from the closure are presented. In the following sections, several choices of closures are shown with their characteristics. The last section is devoted to concluding remarks.

4.2 The closure problem

For convenience, and to exhibit several properties of the closure, the system of moments of the equation (2.1) is studied. This equation is recalled here.

$$\textbf{in 1D:} \qquad \partial_t\psi + \mu\ \partial_x\psi = C(\psi), \tag{4.1a}$$

$$\textbf{in 3D:} \qquad \partial_t\psi + \Omega\nabla_x\psi = C(\psi), \tag{4.1b}$$

where $C(\psi)$ is a collision operator which satisfies certain properties that will be specified in the next section in order to illustrate certain requirement on the closure.

The system of moments extracted from this equation reads (2.2). For convenience, this system is recalled here.

$$\textbf{in 1D:} \qquad \begin{cases} \partial_t \boldsymbol{\psi} \ + \partial_x \mathbf{F} & = \ \mathbf{C}, \\ \partial_t \psi^i \ + \partial_x \psi^{i+1} & = \ C^i, \end{cases} \tag{4.2a}$$

$$\textbf{in 3D:} \qquad \begin{cases} \partial_t \boldsymbol{\psi} \ + \nabla_x \mathbf{F} & = \ \mathbf{C}, \\ \partial_t \psi^i \ + \nabla_x \psi^{i+1} & = \ C^i. \end{cases} \tag{4.2b}$$

The system (4.2) requires a closure (both in 1D or 3D), because it has more unknowns than equations. In practice, the closure is commonly defined by reconstructing an ansatz ψ_R from the first moments $(\psi^0, ..., \psi^N)$, *i.e.* such that

$$\textbf{in 1D:} \qquad \begin{cases} \left\langle (1, \mu, ..., \mu^N)\psi_R \right\rangle & = \ (\psi^0, ..., \psi^N), \\ \langle \mathbf{m}(\mu)\psi_R \rangle & = \ \boldsymbol{\psi}, \end{cases} \tag{4.3a}$$

$$\textbf{in 3D:} \qquad \begin{cases} \left\langle (1, \Omega, ..., \Omega^{\otimes N})\psi_R \right\rangle & = \ (\psi^0, ..., \psi^N), \\ \langle \mathbf{m}(\Omega)\psi_R \rangle & = \ \boldsymbol{\psi}, \end{cases} \tag{4.3b}$$

and then computing the other terms using this ansatz, *i.e.* the higher order moment and the moments of the collision operator

$$\textbf{in 1D:} \qquad \begin{cases} \begin{cases} \psi^{N+1}(\psi^0, ..., \psi^N) & \approx \ \left\langle \mu^{N+1}\psi_R \right\rangle, \\ C^i(\psi^0, ..., \psi^N) & \approx \ \left\langle \mu^i C(\psi_R) \right\rangle, \end{cases} \\ \begin{cases} F(\boldsymbol{\psi}) & \approx \ \langle \mu \mathbf{m} \psi_R \rangle, \\ \mathbf{C}(\boldsymbol{\psi}) & \approx \ \langle \mathbf{m} C(\psi_R) \rangle, \end{cases} \end{cases} \tag{4.4a}$$

$$\textbf{in 3D:} \qquad \begin{cases} \begin{cases} \psi^{N+1}(\psi^0, ..., \psi^N) & \approx \ \left\langle \Omega^{\otimes N+1}\psi_R \right\rangle, \\ C^i(\psi^0, ..., \psi^N) & \approx \ \left\langle \Omega^{\otimes i} C(\psi_R) \right\rangle, \end{cases} \\ \begin{cases} F(\boldsymbol{\psi}) & \approx \ \langle \Omega \otimes \mathbf{m} \psi_R \rangle, \\ \mathbf{C}(\boldsymbol{\psi}) & \approx \ \langle \mathbf{m} C(\psi_R) \rangle, \end{cases} \end{cases} \tag{4.4b}$$

This method corresponds to approximate the kinetic fluence ψ by an ansatz ψ_R.

The first problem when defining a closure for angular moment models is the construction of such an ansatz ψ_R.

in 1D:

$$\textit{find} \quad \psi_R \in L^1([-1,1]), \quad \textit{s.t.} \quad \left\langle (1, \mu, ..., \mu^N)\psi_R \right\rangle \ = \ (\psi^0, ..., \psi^N),$$

in 3D:

$$\textit{find} \quad \psi_R \in L^1(S^2), \quad \textit{s.t.} \quad \left\langle (1, \Omega, ..., \Omega^{\otimes N})\psi_R \right\rangle \ = \ (\psi^0, ..., \psi^N).$$

The second problem is the computation of the integrals (4.4) which may be non-analytical, depending on the obtained ansatz ψ_R.

The next section describes the main desired properties that one typically expects to obtain from the moment model and that are direct consequences of the choice of the closure. The following Sections 4.4, 4.5, 4.6 and 4.7 describe possible choices of ψ_R with consideration to those desired properties.

4.3 Properties of the closure

This section describes how the properties of hyperbolicity, positivity and entropy dissipation, described in Chapter 2 Section 2.1 at the kinetic level, can be interpreted after moment extraction and how the choice of the closure affects those properties.

4.3.1 Hyperbolicity

A partial differential equation of the form

$$\partial_t \boldsymbol{\psi} + \partial_x \mathbf{F}(\boldsymbol{\psi}) = 0, \tag{4.5}$$

is called hyperbolic if the Jacobian of the flux

$$J := \partial_{\boldsymbol{\psi}} \mathbf{F}(\boldsymbol{\psi}) \tag{4.6}$$

is diagonalizable with real eigenvalues, for all possible values of $\boldsymbol{\psi}$. Those possible values may be restrained to all realizable vectors $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$ depending on the chosen closure, if this property needs to be satisfied by the solution of moment system. This is the case when choosing *e.g.* the M_N closure (see Section 4.4 below), but not when chosing *e.g.* the P_N closure (see Subsection 4.7.1 below).

The left-hand side of the kinetic equation (4.1) and its associated moment system (4.2) are of the form (4.5). The kinetic equation is hyperbolic (see Section 2.1). The flux term of the moment system (4.2) is defined by the closure relation. One typically aims to define a closure such that the moment system is hyperbolic, *i.e.* find a flux function F such that

$$J := \partial_{\boldsymbol{\psi}} F(\boldsymbol{\psi}) \quad \text{is diagonalizable with real eigenvalues.}$$

This property may not be easy to check *a posteriori*, because the Jacobian J is a function of the unknown $\boldsymbol{\psi}$ and computing its eigenvalue for all values of $\boldsymbol{\psi}$ may not be possible. The following theorem proposes a characterization of the hyperbolicity when the closure has a certain form.

Theorem 4.1 **([17])** *Suppose that the unknown $\boldsymbol{\psi}$ is defined as a function of a vector $\boldsymbol{\lambda} \in \mathbb{R}^{Card(\mathbf{m})}$ into $\mathbb{R}^{Card(\mathbf{m})}$, i.e. such that (4.5) has the form*

$$\partial_t \boldsymbol{\psi}(\boldsymbol{\lambda}(t,x)) + \partial_x \mathbf{F}(\boldsymbol{\psi}(\boldsymbol{\lambda}(t,x))) = 0.$$

If the matrices

$$\begin{array}{ll} \partial_{\boldsymbol{\lambda}} \boldsymbol{\psi} \quad \text{and} \quad \partial_{\boldsymbol{\lambda}} F(\boldsymbol{\psi}) & \text{are symmetric,} \\ \qquad \partial_{\boldsymbol{\lambda}} \boldsymbol{\psi} & \text{is positive definite,} \end{array}$$

then the system (4.2) is hyperbolic. Such a system is called symmetric hyperbolic.

In 3D, the flux function has three directions, *i.e.* the value of the flux function $F(\psi) = (\mathbf{F_1}(\boldsymbol{\psi}), \mathbf{F_2}(\psi), \mathbf{F_3}(\psi))$ is a matrix in $\mathbb{R}^{3\times Card(\mathbf{m})}$. One generally requires the hyperbolicity property (4.6) to be fulfilled by the Jacobians

$$\mathbf{F_n}(\boldsymbol{\psi}) = n_1\mathbf{F_1}(\boldsymbol{\psi}) + n_2\mathbf{F_2}(\psi) + n_3\mathbf{F_3}(\psi) \tag{4.7}$$

according to all directions $n = (n_1, n_2, n_3) \in S^2$. Theorem 4.1 also holds in 3D, it was originally written in [17] in N dimension.

Corollary 4.1 **([17])** *Suppose that the unknown $\boldsymbol{\psi}$ is defined as a function of a vector $\boldsymbol{\lambda} \in \mathbb{R}^{Card(\mathbf{m})}$ into $\mathbb{R}^{Card(\mathbf{m})}$, i.e. such that (4.5) has the form*

$$\partial_t\boldsymbol{\psi}(\boldsymbol{\lambda}(t,x)) + \nabla_x.F(\boldsymbol{\psi}(\boldsymbol{\lambda}(t,x))) = 0.$$

If the matrices

$$\begin{array}{llll} \partial_{\boldsymbol{\lambda}}\boldsymbol{\psi} & \textit{and for} i = 1,2,3, & \partial_{\boldsymbol{\lambda}}F_i(\boldsymbol{\psi}) & \textit{are symmetric,} \\ & \partial_{\boldsymbol{\lambda}}\boldsymbol{\psi} & & \textit{is positive definite,} \end{array}$$

then the system (4.2) is hyperbolic.

One observes that if the matrix $\partial_{\boldsymbol{\lambda}}\mathbf{F_i}(\boldsymbol{\psi})$ is symmetric for all i then the matrix $\sum\limits_i n_i\partial_{\boldsymbol{\lambda}}F_i(\boldsymbol{\psi})$ is also symmetric and the results follow from the 1D result in Theorem 4.1.

4.3.2 Positivity

Typically the solution ψ of a kinetic equation of the form (4.1) is positive (see *e.g.* [15] for linear collision operator, or [9, 10]).

The positivity of a function ψ is translated at the moment level by the realizability of the unknown $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$, that was studied in the previous chapter. Therefore one typically expects the solution of the moment system (4.2) to be realizable.

As described in the introduction, the closure is tyically chosen by computing the moment of order $N+1$ of an ansatz ψ_R reconstructed from the first moments $\boldsymbol{\psi}$. If this ansatz is constructed as a positive $L^1(S)^+$ function of $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$, then the solution $\boldsymbol{\psi}$ is directly related to some underlying kinetic fluence, *i.e.* the constructed ansatz ψ_R. Therefore, by constructing a realizable closure, one relates the solution of the moment system to a positive kinetic fluence.

4.3.3 Entropy dissipation

The kinetic equation (4.1) dissipates an convex entropy η if

$$\langle\eta'(\psi)C(\psi)\rangle \leq 0 \tag{4.8}$$

for all possible $\psi \in L^1(S)$ such that $C(\psi) \in L^1(S)$. The entropy η may potentially be defined only on a part of $L^1(S)$, *e.g.* the Boltzmann entropy below is defined on $L^1(S)^+$.

Similarily as for the positivity property, when the closure of the moment system (4.2) is defined from a reconstructed ansatz ψ_R, one may expect the underling kinetic system to dissipate the entropy η, *i.e.* one may expect

$$\langle \eta'(\psi_R) C(\psi_R) \rangle \leq 0.$$

Based on the derivation of the linear transport equation of photons and electrons (1.11) from a (qudratic) Boltzmann operator (see *e.g.* [9]), one may expect the Boltzmann entropy

$$\eta(\psi) = \mathcal{H}(\psi) = \psi \log \psi - \psi \tag{4.9}$$

for the photons and the electron to be dissipated by this system of equation, *i.e.* the function $\mathcal{H}(\psi_\gamma) + \mathcal{H}(\psi_e)$. The entropy-based closure described in the next section is based on Boltzmann entropy.

In the next three sections, some closures are proposed. For each of these closures, the hyperbolicity, the realizability and the entropy-dissipation properties are considered.

4.4 The entropy-based (M_N) closure

The entropy-based closure, also called M_N closure (the "M" refers to G. Minerbo [30, 31]), is based on a entropy minimization procedure. It is obtained by reconstructing an ansatz $\psi_R = \psi_{M_N}$. This ansatz is chosen such that it minimizes an entropy function η under positivity and moment constraints.

Suppose $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$. The ansatz ψ_{M_N} is chosen such that it is non-negative and satisfies the moment constraints (4.3). Therefore, it is chosen in the set

$$\mathcal{F}(\boldsymbol{\psi}) := \left\{ \psi \in L^1(S)^+, \quad \text{s.t.} \quad \langle \mathbf{m}\psi \rangle = \boldsymbol{\psi} \right\}. \tag{4.10}$$

Among the possible candidates in $\mathcal{F}(\boldsymbol{\psi})$, the M_N ansatz is the one that minimizes the entropy η

$$\psi_{M_N} := \underset{\psi \in \mathcal{F}(\boldsymbol{\psi})}{\text{argmin}} \quad \eta(\psi). \tag{4.11}$$

The following theorem provides the existence and uniqueness of such an ansatz and its form.

Theorem 4.2 (**[29, 4, 5, 6, 37, 22, 20]**) *Consider a realizable vector $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$, and a convex entropy function η the deriveative of which $\eta' > 0$ is strictly positive.*

Then the M_N ansatz defined in (4.11) exists and is unique. Furthermore it has the form

$$\psi_{M_N} = \eta^{*\prime}(\boldsymbol{\lambda}^T \mathbf{m}). \tag{4.12}$$

> *where the superscript * refers to the Legendre dual (especially, it satisfies $\eta'^{-1} = \eta^{*\prime}$), and $\boldsymbol{\lambda} \in \mathbb{R}^{Card(\mathbf{m})}$ are Lagrange multipliers for the problem (4.11).*

A first verion of this result was proposed in [29] in 1D. Then it was proven in the serie [4, 5, 6] for the more general case where S is non-necessarily compact and non necessarily in 1D, that the optimization problem (4.11) a unique solution (under condition that can be simplified into the condition $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$ for the present problem). It was also proven that if the derivatice of the entropy is not necessarily positive then the problem (4.11) still has a unique solution solution which takes the form

$$\psi_{M_N} = (\eta + \mathbb{1}_{\mathbb{R}^+})^{*\prime}(\boldsymbol{\lambda}^T \mathbf{m}). \tag{4.13}$$

This result was afterward completed and applied to moment closure problems in [37, 22, 20, 26].

In the rest of the manuscript, η is always chosen to be Boltzmann entropy. However, other choices of entropy are also possible, especially more physically relevant ones such as bosons and fermions entropy (see *e.g.* [25]).

> **Example 4.1** *When chosing η to be the Boltzmann entropy function*
>
> $$\eta(\psi) = \mathcal{H}(\psi) := \psi \log(\psi) - \psi,$$
>
> *this leads to writing the M_N ansatz*
>
> $$\mathcal{H}'^{-1} = \mathcal{H}^{*\prime} = \exp \qquad \Rightarrow \qquad \psi_{M_N} = \exp(\boldsymbol{\lambda}^T \mathbf{m}). \tag{4.14}$$

This type of closure is commonly chosen because it offers several mathematical and physical properties.

Advantages

On the physical level, the entropy minimization procedure corresponds to chosing the most physically probable ansatz having the moments $\boldsymbol{\psi}$ ([21]).

On the mathematical level, based on its construction, the M_N model (*i.e.* the moment model with the M_N closure) is hyperbolic, realizable and entropy dissipative.

Hyperbolicity: With the M_N closure, one has

$$\begin{array}{rclcl} \boldsymbol{\psi} & = & \left\langle \mathbf{m} \exp(\boldsymbol{\lambda}^T \mathbf{m}) \right\rangle, & \partial_\lambda \boldsymbol{\psi} & \text{is symmetric positive definite} \\ \mathbf{F}(\boldsymbol{\psi}) & = & \left\langle \mu \mathbf{m} \exp(\boldsymbol{\lambda}^T \mathbf{m}) \right\rangle, & \partial_\psi \mathbf{F}(\boldsymbol{\psi}) & \text{is symmetric.} \end{array}$$

Therefore one can apply Theorem 4.1 and the M_N model is symmetric hyperbolic.

Realizability: According to (4.14), the M_N ansatz has the form $\exp(\boldsymbol{\lambda}^T \mathbf{m})$ which is obviously positive.

Entropy dissipation: Similarily, one can replace ψ by the ansatz ψ_{M_N} in (4.8). If (4.8) holds for all possible ψ it especially holds for ψ_{M_N}. And therefore the underlying kinetic model dissipates an entropy.

Drawbacks

Non-linearity: The original kinetic equations studied in this manuscript are all linear equations of ψ. Closing the moment equations using the entropy-based method leads to non-linear equations. Indeed the flux term $F(\boldsymbol{\psi})$ is not a linear function of $\boldsymbol{\psi}$.

Additionally to the difficulties introduced by non linear terms, such a non-linearity is also responsible for a non-physical effect appearing when studying multiple beam crossing each others. Indeed, consider ψ is the solution of the (linear) kinetic equation (1.11) with a boundary condition chosen to be a sum of incoming beams $\psi|_{\Gamma^-} = \sum\limits_i \psi_i^b$ then $\psi = \sum\limits_i \psi_i$ is the sum of the solutions ψ_i of (1.11) with boundary condition ψ_i^b. This property does not hold for non-linear equations such as the M_N equations.

Computational costs: The M_N closure is commonly computed by solving numerically (see *e.g.* [19, 2, 1]) the following optimization problem

$$\boldsymbol{\lambda}_{M_N} = \underset{\boldsymbol{\lambda}\in\mathbb{R}^{Card(\mathbf{m})}}{\operatorname{argmin}} \left(\left\langle \mathbf{m}\eta^{*\prime}(\boldsymbol{\lambda}^T\mathbf{m}) \right\rangle - \boldsymbol{\psi}^T\boldsymbol{\lambda} \right), \tag{4.15}$$

and then computing numerically the integral

$$F(\boldsymbol{\psi}) = \left\langle \eta^{*\prime}(\boldsymbol{\lambda}_{M_N}^T\mathbf{m}) \right\rangle. \tag{4.16}$$

Typically, the optimization problem (4.15) is solved with an iterative method which requires the computation of integrals at each step. Typically a quadrature formula is used to compute such integrals, and to compute the integral (4.16). Both the iterative method and the quadrature rule introduce computational costs and numerical errors.

Alternative methods reducing these numerical costs and errors can be found in [2, 1]. The next section describes approximated closures for the closures of the first two order angular moment models.

4.5 The M_1 closure

This section is devoted to the computation of an approximation of the M_1 closure. This approximation requires low computational costs and retains some characteristics of the exact closure.

For convenience the following notation is introduced.

Notation 4.1 *The i-th order moment normalized by the zeroth order moment is denoted*

$$N^i := \frac{\psi^i}{\psi^0}. \tag{4.17}$$

4.5.1 The approximated M_1 closure

The M_1 model is the first order entropy-based moment model. This model is commonly used in various fields of physics because it is accurate for a large range of physical phenomena and it is simple of implementation.

In 1D:

For convenience in this paragraph, we only focus on the moments according to $\Omega_1 = \mu$ and therefore the subscript are dropped from the moments, *i.e.* $N^1 \equiv N^1_1$ and $N^2 \equiv N^2_{1,1}$.

First, we focus on the M_1 model in 1D. The vector of polynomials $\mathbf{m}$ is

$$\mathbf{m}(\mu) = (1, \mu).$$

The 1D M_1 ansatz reads

$$\psi_{M_1}(\Omega) = \exp(\lambda_0 + \lambda_1 \mu), \tag{4.18}$$

where $(\lambda_0, \lambda_1) \in \mathbb{R}^2$ are the unique coefficients such that ψ_{M_1} has the moments

$$\langle \psi_{M_1}(\mu) \rangle = \psi^0, \quad \langle \mu \psi_{M_1}(\mu) \rangle = \psi^1.$$

The M_1 closure is typically computed through the Eddington factor $\chi_2 = N^2$, because χ_2 is a scalar function of only one scalar N^1 (see *e.g.* [25, 16]) which can easily be approximated. Indeed χ_2 is a function of the scalar λ_1

$$\chi_2 \quad = \quad \frac{2 - 2\lambda_1 \mathrm{coth}(\lambda_1) + \lambda_1^2}{\lambda_1^2},$$

and computations show that $N^1 \in]-1, 1[$ is in bijection with $\lambda_1 \in \mathbb{R}$

$$N^1 \quad = \quad \frac{\lambda_1 \mathrm{coth}(\lambda_1) - 1}{\lambda_1}.$$

One observes from these computations that χ_2 is an even function of λ_1 and N^1 is an odd function (bijection) of λ_1, therefore χ_2 is an even function of N^1.

There is no analytical formula for $\chi_2(N^1)$, except for particular choices of entropy η (see *e.g.* [25, 16]). It can be computed *e.g.* by solving numerically (4.15) and computing (4.16). In order to reduce computational costs compared to this methods, one can compute an (analytical) approximation of the Eddington factor.

We aim to construct an approximation of the Eddington factor that provides a closure $\psi^2 = \psi^0 \chi_2$ realizable and hyperbolic.

Property 4.1 ***Realizability:*** *Using* (3.35c), *the realizability condition on* ψ^2 *equivals to the following condition on* χ_2

$$|N^1|^2 < \chi_2(N^1) < 1, \qquad \forall N^1 \in]-1, 1[. \tag{4.19}$$

For this purpose, $\chi_2(N^1)$ *is rewritten as a convex combination of the bounds*

$|N^1|^2$ and 1

$$\chi_2(N^1) \quad = \quad \theta_1(N^1) \quad |N^1|^2 \quad + \quad (1-\theta_1(N^1)) \quad 1, \tag{4.20}$$

therefore we need only to approximate the function θ_1 from $[0,1[$ to $]0,1[$.
Hyperbolicity: *The Jacobian of the flux (in 1D) reads*

$$J = \partial_\psi F(\psi) = \begin{pmatrix} 0 & 1 \\ \chi_2(N^1) - N^1\chi_2'(N^1) & \chi_2'(N^1) \end{pmatrix}. \tag{4.21}$$

Computations show that J is diagonalizable with real eigenvalues iff

$$\left(\chi_2'(N^1)\right)^2 + 4\left[\chi_2(N^1) - N^1\chi_2'(N^1)\right] \geq 0, \qquad \forall N^1 \in]-1,1[. \tag{4.22}$$

For the approximation of χ_2 to provide an hyperbolic and realizable closure, it needs to interpolate the exact value and the exact derivative of χ_2 on the boundary of the realizability domain (for M_1 in 1D this corresponds to $N^1 = \pm 1$). Indeed having a wrong value of $\chi_2(1)$ or of $\chi_2'(1)$ leads to violate (4.19) or (4.22) or both.

Furthermore, we force the approximation to interpolate the value of $\chi_2(0)$ and $\chi_2'(0)$ when $N^1 = 0$. This corresponds to the case where ψ_{M_1} is isotropic.

In order to compute the values $\chi_2(0)$, $\chi_2(1)$, $\chi_2'(0)$ and $\chi_2'(1)$, the Eddington factor and its derivative can be written

$$\chi_2(N^1) = \frac{\langle \mu^2 \exp(\lambda_1(N^1)\mu)\rangle}{\langle \exp(\lambda_1(N^1)\mu)\rangle}, \qquad \chi_2'(N^1) = \frac{d\lambda_1}{dN^1}\frac{d}{d\lambda_1}\frac{\langle \mu^2 \exp(\lambda_1(N^1)\mu)\rangle}{\langle \exp(\lambda_1(N^1)\mu)\rangle},$$

where $\frac{d\lambda_1}{dN^1} = \left(\frac{dN^1}{d\lambda_1}\right)^{-1}$. Then one remarks that

$$N^1(\lambda_1 = 0) = 0, \qquad \underset{\lambda_1 \to +\infty}{lim} N^1 = 1,$$

and since N^1 is a bijection of λ_1, this implies that

$$\lambda_1(N^1 = 0) = 0, \qquad \underset{N^1 \to 1}{lim} \lambda_1 = +\infty,$$

then computing χ_2 and its derivative at those values using the symbolic computation software MAPLE ([27]) leads to

$$\chi_2(1) = 1, \quad \chi_2(0) = \frac{1}{3}, \quad \chi_2'(1) = 2, \quad \chi_2'(0) = 0. \tag{4.23}$$

In the end, we propose the following approximation of θ_1

$$\begin{aligned} \theta_1(N^1) \approx |N^1|^2 \quad &+ \quad \frac{2}{3}(1-|N^1|^2) \\ &+ \quad |N^1|^2(1-|N^1|^2)\left(c_0 + c_1|N^1|^2 + c_2|N^1|^4\right), \end{aligned} \tag{4.24}$$

where the coefficients $c_0 = -0.0954823981432433$, $c_1 = 0.229069986304953$ and $c_2 = -0.0344846229504588$ are computed with MAPLE ([27]) such that the approximated χ_2 given in (4.20) fit the exact χ_2 for 10^3 values of N^1 equally distributed in $[0,1]$, *i.e.* it minimizes the discrete L^2 error compared to the values obtained by solving (4.15) and computing numerically (4.16).

One verifies *a posteriori* that this approximation of θ_1 is in $]0,1[$, *i.e.* it provides a realizable closure, and that the condition (4.22) is satisfied, *i.e.* it provides a hyperbolic closure.

In 3D:

For the M_1 model in 3D, the vector $\mathbf{m}$ reads

$$\mathbf{m}(\Omega) = (1, \Omega_1, \Omega_2, \Omega_3).$$

Define the vector $\boldsymbol{\lambda}_1 = (\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^3$. The M_1 ansatz reads

$$\psi_{M_1}(\Omega) = \exp(\lambda_0 + \boldsymbol{\lambda}_1 \Omega), \tag{4.25}$$

where $\lambda_0 \in \mathbb{R}$ and $\boldsymbol{\lambda}_1 \in \mathbb{R}^3$ are the unique coefficients such that ψ_{M_1} has the moments

$$\langle \psi_{M_1}(\Omega) \rangle = \psi^0, \qquad \langle \Omega \psi_{M_1}(\Omega) \rangle = \psi^1.$$

Proposition 4.1 **([25])** *Define the vector*

$$n = \frac{\psi^1}{|\psi^1|}.$$

The M_1 closure has the form

$$\psi^2 = \psi^0 \left(\frac{1-\chi_2}{2} Id + \frac{3\chi_2 - 1}{2} n \otimes n \right). \tag{4.26}$$

Proof One observes that the ansatz ψ_{M_1} is invariant by rotation around the axis $\boldsymbol{\lambda}_1$. Define a rotation matrix $R \in \mathbb{R}^{3\times 3}$ along the axis $\boldsymbol{\lambda}_1$, *i.e.* such that

$$\boldsymbol{\lambda}_1 = R\boldsymbol{\lambda}_1.$$

Using the change of variable $\Omega" = R\Omega$ in the computation of ψ^0 read

$$\begin{aligned} \psi^0 = \langle \psi_{M_1}(\Omega) \rangle &= \langle \exp(\lambda_0 + \boldsymbol{\lambda}_1 \Omega) \rangle \\ &= \langle \exp(\lambda_0 + R\boldsymbol{\lambda}_1 \Omega) \rangle \\ &= \langle \exp(\lambda_0 + \boldsymbol{\lambda}_1 \Omega") \rangle = \psi^0 \end{aligned}$$

and for ψ^1

$$\begin{aligned} \psi^1 = \langle \Omega \psi_{M_1}(\Omega) \rangle &= \langle \Omega \exp(\lambda_0 + \boldsymbol{\lambda}_1 \Omega) \rangle \\ &= \langle \Omega \exp(\lambda_0 + R\boldsymbol{\lambda}_1 \Omega) \rangle \\ &= R^T \langle \Omega" \exp(\lambda_0 + \boldsymbol{\lambda}_1 \Omega") \rangle = R^T \psi^1. \end{aligned}$$

Since ψ is invariant by rotation around the axis $\boldsymbol{\lambda}_1$, its moment are too, which means that $\psi^1 = \alpha\boldsymbol{\lambda}_1$ for some $\alpha \in \mathbb{R}$. Similarily computing the moment of order 2 leads to write that

$$\psi^2 = R^T \psi^2 R,$$

i.e. ψ^2 is also invariant around the axis $\boldsymbol{\lambda}_1$. Therefore ψ^2 can only have the form

$$\psi^2 = \alpha Id + \beta\psi^1 \otimes \psi^1, \qquad \text{for some scalars} \alpha, \beta.$$

Finally, remarking that

$$\psi^0 = \langle \psi_{M_1} \rangle = \left\langle |\Omega|^2 \psi_{M_1} \right\rangle = \langle tr(\Omega \otimes \Omega)\psi_{M_1} \rangle = tr(\psi^2)$$

leads to the write ψ^2 under the form (4.26). □

Using the approximation of χ_2 proposed in the previous paragraph leads to a hyperbolic and realizable closure approximating the M_1 closure.

4.5.2 The advantages of the M_2 model

Before describing the approximated M_2 closure, such a work is motivated here.

Despite improving the accuracy of the model, due to the better flexibility of the M_2 ansatz compared to the M_1 ansatz ψ_{M_1}, the M_2 model is also able to model a larger range of physical phenomena than the M_1 model.

This improvement can be illustrated by two problems emerging in the field of radiotherapy. First the physics of the angular diffusion is better modeled with the M_2 model than with the M_1 model. Second the M_1 model is not able to model multiple beams of particles crossing each others.

Angular diffusion modelling

Suppose an elastic linear Boltzmann operator $G_{el} - P_{el}$ characterized by a differential cross section σ_{el}. Decomposing σ_{el} into polynomials reads

$$\sigma_{el}(\epsilon, \mu) = \sum_{i=0}^{\infty} \sigma_{el}^i(\epsilon)\mu^i.$$

Then extracting the N first moments of the collisional operator is equivalent to truncating this expansion at degree N. Clearly the collisions are better modelled as N increases. This phenomenon is illustrated through the test cases in Part III Chapter 5 Subsections 5.3.1 and 5.3.3.

Multiple beams

Consider two perfect beams of opposite direction ±e_1 crossing each other. This is modelled by a sum of two Dirac peaks

$$\psi = \delta(\Omega.e_1 - 1) + \delta(\Omega.(-e_1) - 1).$$

Extracting the moments of such a measure reads

$$\begin{aligned} \psi^0 = &\quad \langle \psi \rangle &= 2, \\ \psi^1 = &\quad \langle \Omega\psi \rangle &= 0_{\mathbb{R}^3}, \\ \psi^2 = &\quad \langle \Omega \otimes \Omega\psi \rangle &= 2e_1 \otimes e_1. \end{aligned}$$

Remark that the first two moments (ψ^0, ψ^1) are identical to the first two moments of an isotropic distribution $\psi = 1/2\pi$. Since only (ψ^0, ψ^1) are considered in the M_1 model, then the M_1 model is unable to distinguish two beams from a isotropic distribution.

In the M_2 model, the moment ψ^2 is also considered and therefore the M_2 is able to distinguish two beams from an isotropic distribution. This is illustrated through the test case of Part III Chapter 5 Subsection 5.3.2.

This problem also appears in multi-D when the two beams are not along the same direction. Consider a distribution of two beams of respective directions e_1 and e_2

$$\psi = \delta(\Omega.e_1 - 1) + \delta(\Omega.e_2 - 1).$$

The moments of such a measure are

$$\begin{aligned} \psi^0 = &\quad \langle \psi \rangle &= 2, \\ \psi^1 = &\quad \langle \Omega\psi \rangle &= e_1 + e_2, \\ \psi^2 = &\quad \langle \Omega \otimes \Omega\psi \rangle &= e_1 \otimes e_1 + e_2 \otimes e_2. \end{aligned}$$

Using the rotational invariance ψ_{M_1} (see proof of Proposition 4.1), the M_1 ansatz computed from the moment (ψ^0, ψ^1) of such a distribution has the form

$$\psi_{M_1}(\Omega) = \exp(\lambda_0 + \alpha(e_1 + e_2)\Omega),$$

and is therefore a single imperfect beam of direction $(e_1 + e_2)$. This is illustrated through the test case of Part III Chapter 5 Subsection 5.3.4.

As for the previous case, the M_2 model is able to distinguish those two beams.

This problem emerges from the form of the M_N ansatz ψ_{M_N} defined in (4.11). One simply observes that the M_N model for a given $N \in \mathbb{N}$ is able to distinguish multiple beams at the points $\Omega = \Omega_i^b$ if there exists $\boldsymbol{\lambda} \in \mathbb{R}^{Card(\mathbf{m})}$ such that the M_N ansatz $\psi_{M_N} = \exp(\boldsymbol{\lambda}^T \mathbf{m})$ presents a local maxima in each of the points Ω_i^b.

> **Remark 4.1** *Typically, the multiple beam problem is solved by exploiting the linearity of the underlying kinetic equation* (2.1) *(or* (1.11)*). Consider N boundary conditions ψ_{b_i} modeling at most one entering beam and initial condition $\psi_i(\epsilon = 0) = 0$. Then the sum $\sum\limits_i \psi_i$ of the solutions ψ_i to* (2.1) *with those initial-boundary conditions is the solution to the kinetic equation* (2.1) *with the bound-*

ary condition $\psi|_{\partial Z} = \sum_i \psi_{b_i}$ and initial condition $\psi(\epsilon = 0) = 0$. Each solution ψ_i can be approximated by solving the M_1 system (2.2) (or (2.3)), and one may obtain ψ by summing those contributions ψ_i.

4.6 The approximated M_2 closure

The M_2 ansatz in 3D does not present such a rotational invariance as the M_1 ansatz. The method proposed in the last section does not lead to such a simplification of the problem.

In this section, in 3D, the vector $\mathbf{m}$ is chosen to be

$$\mathbf{m}(\Omega) = \Big(\Omega_1, \quad \Omega_2, \quad \Omega_3, \quad \Omega_1^2, \quad \Omega_2^2, \quad \Omega_3^2, \quad \Omega_1\Omega_2, \quad \Omega_1\Omega_3, \quad \Omega_2\Omega_3\Big)^T,$$

which generates all polynomials on S^2 of degree 2, and the M_2 ansatz is denoted

$$\psi_{M_2}(\Omega) = \exp(\boldsymbol{\lambda}\mathbf{m}(\Omega)), \tag{4.27}$$

for $\boldsymbol{\lambda} \in \mathbb{R}^9$.

4.6.1 An hierarchy of sets

The strategy to construct an approximation of the M_2 closure consists in exploiting the hierarchical character of the M_N models, *i.e.* the fact that the M_{N-1} is a subcase of the M_N model. The hierarchical character of the M_N models can be illustrated through the following example.

Example 4.2 *The M_2 ansatz is mor flexible than the M_1 ansatz. Indeed one observes that the M_1 ansatz ψ_{M_1} is a particular value of M_2 ansatz*

$$\begin{aligned}\psi_{M_1}(\Omega) &= \exp(\lambda_0 + \lambda_1\Omega_1 + \lambda_2\Omega_2 + \lambda_3\Omega_3)\\ &= \exp(\lambda_1\Omega_1 + \lambda_2\Omega_2 + \lambda_3\Omega_3 + \lambda_0(\Omega_1^2 + \Omega_2^2 + \Omega_3^2)),\end{aligned}$$

i.e. the M_1 ansatz (4.25) is the M_2 ansatz (4.27) where the coefficients $\boldsymbol{\lambda}$ have the form

$$\boldsymbol{\lambda} = (\lambda_1, \quad \lambda_2, \quad \lambda_3, \quad \lambda_0, \quad \lambda_0, \quad \lambda_0, \quad 0, \quad 0, \quad 0)\,.$$

The following hierarchy of subdomains of $\mathbb{R}^9$ is considered

$$\mathcal{L}_0 := \mathbb{R}^9, \tag{4.28a}$$

$$\mathcal{L}_1 := \Big\{(\lambda_1, 0, 0, \lambda_4, \lambda_6, \lambda_9, 0, 0, 0), \quad \text{s.t.} \quad (\lambda_1, \lambda_4, \lambda_6, \lambda_9) \in \mathbb{R}^4\Big\} \subset \mathcal{L}_0, \tag{4.28b}$$

$$\mathcal{L}_2 := \Big\{(\lambda_1, 0, 0, \lambda_4, \lambda_6, \lambda_6, 0, 0, 0), \quad \text{s.t.} \quad (\lambda_1, \lambda_4, \lambda_6) \in \mathbb{R}^3\Big\} \subset \mathcal{L}_1, \tag{4.28c}$$

$$\mathcal{L}_3 := \Big\{(\lambda_1, 0, 0, \lambda_4, \lambda_4, \lambda_4, 0, 0, 0), \quad \text{s.t.} \quad (\lambda_1, \lambda_4) \in \mathbb{R}^2\Big\} \subset \mathcal{L}_2. \tag{4.28d}$$

Choosing $\boldsymbol{\lambda}$ in those sets in (4.27) leads to ansätze ψ_{M_2} in the sets

$$\mathcal{E}_0 := \left\{ \exp(\boldsymbol{\lambda}^T \mathbf{m}(\Omega)), \quad \boldsymbol{\lambda} \in \mathcal{L}_0 \right\} \subset L^1(S^2)^+, \tag{4.29a}$$

$$\mathcal{E}_1 := \left\{ \exp(\lambda_1 \Omega_1 + \lambda_4 \Omega_1^2 + \lambda_6 \Omega_2^2 + \lambda_9 \Omega_3^2), \quad (\lambda_1, \lambda_4, \lambda_6, \lambda_9) \in \mathbb{R}^4 \right\} \subset \mathcal{E}_0, \tag{4.29b}$$

$$\mathcal{E}_2 := \left\{ \exp(\lambda_6 + \lambda_1 \Omega_1 + (\lambda_4 - \lambda_6) \Omega_1^2), \quad (\lambda_1, \lambda_4, \lambda_6) \in \mathbb{R}^3 \right\} \subset \mathcal{E}_1, \tag{4.29c}$$

$$\mathcal{E}_3 := \left\{ \exp(\lambda_4 + \lambda_1 \Omega_1), \quad (\lambda_1, \lambda_4) \in \mathbb{R}^2 \right\} \subset \mathcal{E}_2. \tag{4.29d}$$

Remark 4.2 *One observes the following properties.*

- *The functions of the form* (4.29d) *are M_1 ansätze* (4.25).
- *The functions of the form* (4.29c) *are 1D ansätze, as they depend only on one variable Ω_1.*
- *Computations show that the moments ψ^1 and ψ^2 of functions of the form* (4.29b) *are the moments such that ψ^1 is an eigenvector of ψ^2.*

The hierarchy of the sets of exponential functions (4.29) is illustrated on Fig. 4.1, 4.2, 4.3, 4.4, 4.5 and 4.6. On those plots, the unit sphere is colored by the value of a possible function of the sets (4.29) (where blue corresponds to the lowest value and red the highest). As $\mathcal{E}_3$ and $\mathcal{E}_2$ are sets of 1D distributions, those distributions can be represented along the preferred axis, *i.e.* as a function of Ω_1.

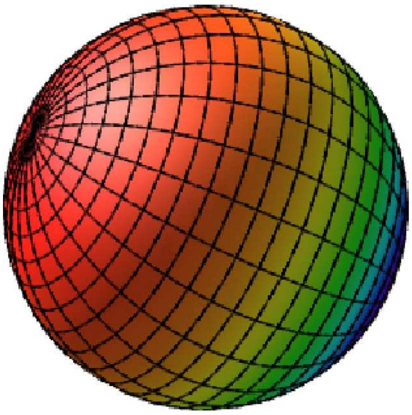

Figure 4.1: Unit sphere colored by one possible function in $\mathcal{E}_3$.

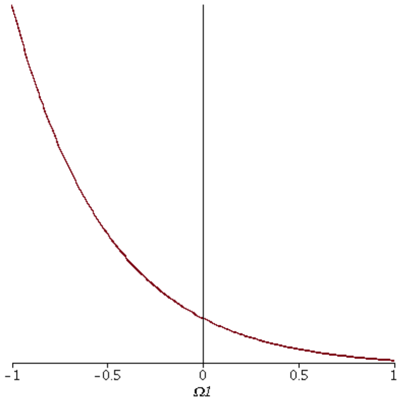

Figure 4.2: Function in $\mathcal{E}_3$ as a function of Ω_1.

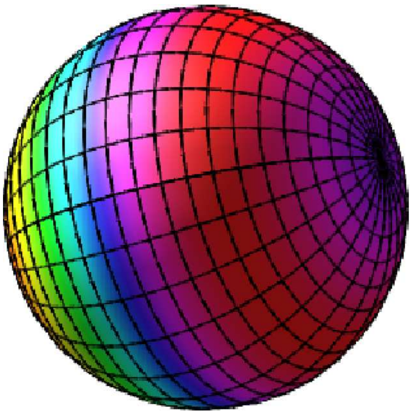

Figure 4.3: Unit sphere colored by one possible function in $\mathcal{E}_2 \backslash \mathcal{E}_3$.

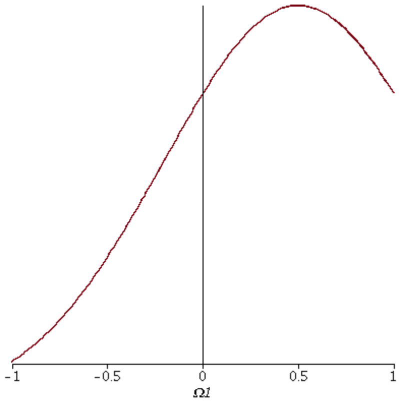

Figure 4.4: Function in $\mathcal{E}_3 \backslash \mathcal{E}_2$ as a function of Ω_1.

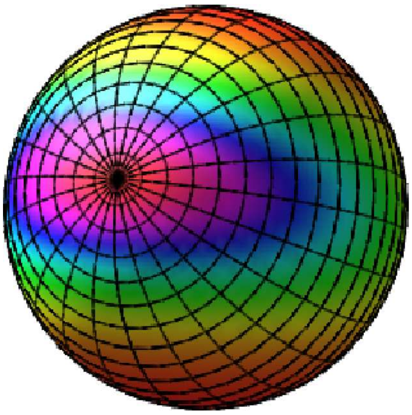

Figure 4.5: Unit sphere colored by one possible function in $\mathcal{E}_1 \backslash \mathcal{E}_2$.

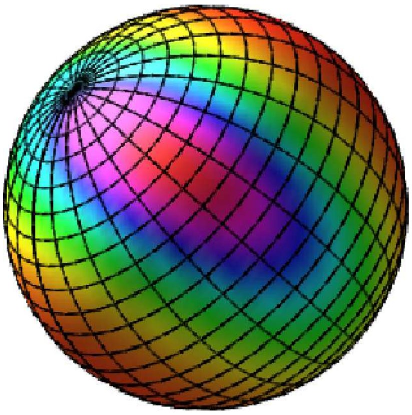

Figure 4.6: Unit sphere colored by one possible function in $\mathcal{E}_0 \backslash \mathcal{E}_1$.

Extracting the moments of order up to 2 of these functions leads to define the following hierarchy of subdomains of $\mathcal{R}_{\mathbf{m}}$

$$\mathcal{R}_0 := \{\langle \mathbf{m}\psi \rangle, \quad \psi \in \mathcal{E}_0\}, \tag{4.30a}$$

$$\mathcal{R}_1 := \{\langle \mathbf{m}\psi \rangle, \quad \psi \in \mathcal{E}_1\} \subset \mathcal{R}_0, \tag{4.30b}$$

$$\mathcal{R}_2 := \{\langle \mathbf{m}\psi \rangle, \quad \psi \in \mathcal{E}_2\} \subset \mathcal{R}_1, \tag{4.30c}$$

$$\mathcal{R}_3 := \{\langle \mathbf{m}\psi \rangle, \quad \psi \in \mathcal{E}_3\} \subset \mathcal{R}_2. \tag{4.30d}$$

Remark 4.3 *According to Theorem 4.2 ([4, 5, 6, 22, 37, 20]), each realizable vector* $\mathbf{V} \in \mathcal{R}_{\mathbf{m}}$ *are the moments of an exponential function of the form* (4.29a)*,* *i.e.*

$$\mathcal{R}_0 = \mathcal{R}_{\mathbf{m}}.$$

Furthermore the set $\mathcal{L}_0$ (resp. $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$) of Lagrange multipliers is in bijection with the sets of realizable moments $\mathcal{R}_0$ (resp. $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$).

4.6.2 Transformation of the realizability domain

As for the M_1 case (Section 4.5), one can use a rotation and a normalization on the moments and write

$$\psi^3 = \psi^0 Rot_3(R, N^3), \tag{4.31}$$

where $Rot_3(R, N^3)$ is the rotation of the tensor N^3 of order 3 according to the rotation R

$$Rot_3(R, N^3) = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} R_{i,i'}R_{j,j'}R_{k,k'}N^3_{i',j',k'}.$$

Furthermore, by computing the moments ψ^0, ψ^1, ψ^2 and ψ^3, and using again the change of variable $\Omega' = R\Omega$ (see Subsection 4.5.1), one observes that the tensor $Rot_3(R, N^3)$ in (4.31) depends only on RN^1 and RN^2R^T.

Using (3.43) and (4.17), the moment N^2 can be rewritten

$$N^2 = N^1 \otimes N^1 + (1 - |N^1|^2) R Diag(\gamma_1, \gamma_2, 1 - \gamma_1 - \gamma_2) R^T, \tag{4.32}$$

where R is a rotation matrix and $\gamma_1 \in]0, 1[$ and $\gamma_2 \in]0, 1 - \gamma_1[$.

For the purpose of the present approximation of the closure, the moments are rotated according to a rotation R that diagonalizes the matrix

$$N^2 - N^1 \otimes N^1.$$

If N^1 is an eigenvector of N^2, then we additionally require that R sends N^1 onto $|N^1|e_1$. For convenience, the sets of normalized realizable vectors rotated by R are written

$$\tilde{\mathcal{R}}_0 := \{\mathbf{V} \in \mathcal{R}_0, \quad \text{s.t.} \quad \psi^0 = 1, \quad N^2 - N^1 \otimes N^1 \quad \text{is diagonal}\}, \tag{4.33a}$$

$$\tilde{\mathcal{R}}_1 := \{\mathbf{V} \in \mathcal{R}_1, \quad \text{s.t.} \quad \psi^0 = 1, \quad N^1 = |N^1|e_1, \quad N^2 - N^1 \otimes N^1 \quad \text{is diagonal}\} \subset \tilde{\mathcal{R}}_0, \tag{4.33b}$$

$$\tilde{\mathcal{R}}_2 := \{\mathbf{V} \in \mathcal{R}_2, \quad \text{s.t.} \quad \psi^0 = 1, \quad N^1 = |N^1|e_1, \quad N^2 - N^1 \otimes N^1 \quad \text{is diagonal}\} \subset \tilde{\mathcal{R}}_1, \tag{4.33c}$$

$$\tilde{\mathcal{R}}_3 := \{\mathbf{V} \in \mathcal{R}_3, \quad \text{s.t.} \quad \psi^0 = 1, \quad N^1 = |N^1|e_1, \quad N^2 - N^1 \otimes N^1 \quad \text{is diagonal}\} \subset \tilde{\mathcal{R}}_2. \tag{4.33d}$$

Remark 4.4 *Consider* $\mathbf{V} \in \mathcal{R}_0$ *such that* $\psi^1 = R_{\mathbf{V}}(\Omega)$ *is an eigenvector of* $\psi^2 = R_{\mathbf{V}}(\Omega \otimes \Omega)$. *Rotating* ψ^1 *and* ψ^2 *with* R *reads*

$$R\psi^1 = |\psi^1| e_1, \quad R\psi^2 R^T \quad \textit{is diagonal.} \tag{4.34}$$

Using again Theorem 4.2 with $\mathbf{m}(\Omega) = (\Omega_1, \Omega_1^2, \Omega_2^2, \Omega_3^2)$ *leads to write that if* (ψ^1, ψ^2) *satisfy* (4.34) *then there exists an exponential function* $\psi \in \mathcal{R}_1$ *such that*

$$R\psi^1 = \int_{S^2} \Omega \psi(\Omega) d\Omega, \qquad R\psi^2 R^T = \int_{S^2} \Omega \otimes \Omega \psi(\Omega) d\Omega.$$

Computing the moments of order one and two of the functions (4.29) normalized by ψ^0 and rotated by R leads to

$$\psi \in \mathcal{E}_0, \;\Rightarrow\; \left\{ \begin{array}{l} N^1 \in B(0_{\mathbb{R}^3}, 1), \\ N^2 = \left[N^1 \otimes N^1 + (1 - |N^1|^2)\, Diag\,(\gamma_1, \gamma_2, 1 - \gamma_1 - \gamma_2)\right], \end{array} \right. \tag{4.35a}$$

$$\psi \in \mathcal{E}_1, \;\Rightarrow\; \left\{ \begin{array}{l} N^1 = N_1^1 e_1, \\ N^2 = \left[|N_1^1|^2 e_1 \otimes e_1 + (1 - |N_1^1|^2)\, Diag\,(\gamma_1, \gamma_2, 1 - \gamma_1 - \gamma_2)\right], \end{array} \right. \tag{4.35b}$$

$$\psi \in \mathcal{E}_2, \;\Rightarrow\; \left\{ \begin{array}{l} N^1 = N_1^1 e_1, \\ N^2 = \left[|N_1^1|^2 e_1 \otimes e_1 + (1 - |N_1^1|^2)\, Diag\left(\gamma_1, \dfrac{1-\gamma_1}{2}, \dfrac{1-\gamma_1}{2}\right)\right], \end{array} \right. \tag{4.35c}$$

$$\psi \in \mathcal{E}_3, \;\Rightarrow\; \left\{ \begin{array}{l} N^1 = N_1^1 e_1, \\ N^2 = \left[\dfrac{3\chi_2(N_1^1) - 1}{2} e_1 \otimes e_1 + \dfrac{1 - \chi_2(N_1^1)}{2} Id\right], \end{array} \right. \tag{4.35d}$$

where χ_2 is the Eddington factor (see Subsection 4.5.1 or *e.g.* [25]). For convenience, the following parametrizations of $\tilde{\mathcal{R}}_0$, $\tilde{\mathcal{R}}_1$, $\tilde{\mathcal{R}}_2$ and $\tilde{\mathcal{R}}_3$

$$\begin{array}{lcl} \mathcal{P}_0 & := & B(0_{\mathbb{R}^3}, 1) \times]0, 1[\times]0, 1 - \gamma_1[, \\ \mathcal{P}_1 & := & \left\{ (N^1, \gamma_1, \gamma_2) \in \mathcal{P}_0 \quad \text{s.t.} \quad N^1 = N_1^1 e_1 \right\}, \\ \mathcal{P}_2 & := & \left\{ (N^1, \gamma_1, \gamma_2) \in \mathcal{P}_1 \quad \text{s.t.} \quad \gamma_2 = \dfrac{1-\gamma_1}{2} \right\}, \\ \mathcal{P}_3 & := & \left\{ (N^1, \gamma_1, \gamma_2) \in \mathcal{P}_2 \quad \text{s.t.} \quad \gamma_1 = \dfrac{\chi_2(|N_1^1|) - |N_1^1|^2}{1 - |N_1^1|^2} \right\}, \end{array}$$

For a vector $\mathbf{V}$ to be in the set $\tilde{\mathcal{R}}_0$, respectively $\tilde{\mathcal{R}}_1$, $\tilde{\mathcal{R}}_2$, $\tilde{\mathcal{R}}_3$, then the normalized moments (N^1, N^2) satisfy (4.35a) where the parameters $(N^1, \gamma_1, \gamma_2)$ are in $\mathcal{P}_0$, respectively $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$.

The hierarchy of sets $\mathcal{P}_3 \subset \mathcal{P}_2 \subset \mathcal{P}_1$ is represented on Fig. 4.7 in the space $(N_1^1, \gamma_1, \gamma_2) \in \mathbb{R}^3$. On Fig. 4.7, the red curve represents $\mathcal{P}_3$ in the space $(N_1^1, \gamma_1, \gamma_2) \in \mathbb{R}^3$. This set is included into the green plane representing $\mathcal{P}_2$ which is itself included in the blue volume representing $\mathcal{P}_1$.

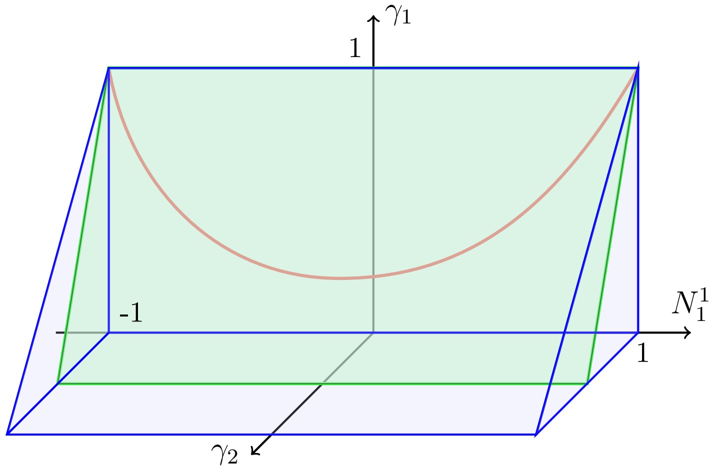

Figure 4.7: Representation of $\mathcal{P}_3$ (red line), $\mathcal{P}_2$ (green plane) and $\mathcal{P}_1$ (blue volume) in the space $(N_1^1, \gamma_1, \gamma_2) \in \mathbb{R}^3$.

Similarly, computing the third order moment of the functions (4.29) normalized by ψ^0 reads

$$\psi \in \mathcal{E}_1, \qquad \Rightarrow \qquad N^3 = \kappa_2 1_{1,1,1} + \kappa_3 T_{1,2,2} + (N_1^1 - \kappa_2 - \kappa_3) T_{1,3,3}, \tag{4.36a}$$

$$\psi \in \mathcal{E}_2, \qquad \Rightarrow \qquad N^3 = \kappa_1 1_{1,1,1} + \frac{N_1^1 - \kappa_1}{2}(T_{1,2,2} + T_{1,3,3}), \tag{4.36b}$$

$$\psi \in \mathcal{E}_3, \qquad \Rightarrow \qquad N^3 = \chi_3 1_{1,1,1} + \frac{N_1^1 - \chi_3}{2}(T_{1,2,2} + T_{1,3,3}), \tag{4.36c}$$

$$T_{i,j,j} \quad = \quad 1_{i,j,j} + 1_{j,i,j} + 1_{j,j,i}, \qquad 1_{i,j,k} = e_i \otimes e_j \otimes e_k,$$

where χ_3, κ_1, κ_2 and κ_3 are scalar coefficients depending on $(N^1, \gamma_1, \gamma_2)$ in $\mathcal{P}^3$ for χ_3, in $\mathcal{P}^2$ for κ_1, and in $\mathcal{P}^1$ for κ_2 and κ_3, *i.e.* χ_3 is a function of $|N^1| \in [0,1[$, κ_1 of $|N^1| \in [0,1[$ and $\gamma_1 \in]0,1[$ and κ_2 and κ_3 of $|N^1| \in [0,1[$, of $\gamma_1 \in]0,1[$ and of $\gamma_2 \in]0,1-\gamma_1[$.

Strategy of the approximation: The idea is to construct an approximation of N^3 for $\mathbf{V} \in \tilde{\mathcal{R}}_3$, using the same techniques as in Subsection 4.5.1 and to extend progressively this approximation to $\tilde{\mathcal{R}}_2$, to $\tilde{\mathcal{R}}_1$ and finally to $\tilde{\mathcal{R}}_0$. This corresponds to approximating progressively χ_3 as a function of N_1^1 in (4.36c), then κ_1 as a function of (N_1^1, γ_1) in (4.36b), then κ_2 and κ_3 as a function of $(N_1^1, \gamma_1, \gamma_2)$ in (4.36a) and finally N^3 as a function of $(N^1, \gamma_1, \gamma_2) \in \mathcal{P}_0$.

4.6.3 In $\tilde{\mathcal{R}}_3$, *i.e.* for M_1 ansätze

For convenience in this subsection and in the next one (in 1D), we only focus on the moments according to $\Omega_1 = \mu$ and therefore the subscript are dropped from the moments, *i.e.* $N^1 \equiv N_1^1$, $N^2 \equiv N_{1,1}^2$ and $N^3 \equiv N_{1,1,1}^3$.

Reproducing the computations of Subsection 4.5.1 for the moment of order 3 reads (4.36c). Therefore, after rotation and normalization, one only needs to study the following 1D problem.

For the M_1 ansatz, the vector of polynomials $\mathbf{m}$ is

$$\mathbf{m}(\mu) = (1, \mu).$$

For moments in $\tilde{\mathcal{R}}_3$, the associated ansatz $\psi_{\tilde{\mathcal{R}}_3}$, *i.e.* a M_1 ansatz, reads

$$\psi_{\tilde{\mathcal{R}}_3}(\Omega) = \exp(\lambda_0 + \lambda_1 \mu), \tag{4.37}$$

where $(\lambda_0, \lambda_1) \in \mathbb{R}^2$ are the unique coefficients such that ψ_{M_1} satisfies

$$\left\langle \mu \psi_{\tilde{\mathcal{R}}_3}(\mu) \right\rangle = \psi^1, \qquad \left\langle \mu^2 \psi_{\tilde{\mathcal{R}}_3}(\mu) \right\rangle = \psi^2.$$

Property 4.2 ***Realizability:*** *Using (3.35c), the realizability condition on ψ^3 equivals to the following condition on N^3*

$$b_-(N^1, N^2) < N^3 < b_+(N^1, N^2), \tag{4.38}$$

$$b_-(N^1, N^2) := -N^2 + \frac{(N^1 + N^2)^2}{(1 + N^1)}, \qquad b_+(N^1, N^2) := N^2 - \frac{(N^1 - N^2)^2}{(1 - N^1)}.$$

For this purpose, $\chi_3 = N^3$ is rewritten as a convex combination of b_- and b_+

$$\begin{aligned} \chi_3(N^1) \quad &= \quad b_-(N^1, \chi_2(N^1)) \quad \theta_2(N^1) \\ &\quad + \quad b_+(N^1, \chi_2(N^1)) \quad \left(1 - \theta_2(N^1)\right). \end{aligned} \tag{4.39}$$

therefore one only needs to approximate the function θ_3 from $(N^1, \gamma_1) \in [0, 1[\times]0, 1[$ to $]0, 1[$.

Using the change of variable $\Omega = -\Omega'$ in (4.36c), one observes that N^3 is an odd function of λ_1. Similarly N^1 is odd and N^2 is even in λ_1. This leads to write that χ_3 is an odd function of N^1.

The coefficient θ_2 is approximated such that χ_3 is odd and interpolates the values of χ_3 and its derivative χ_3' for $N^1 = \pm 1$ and for $N^1 = 0$. Those values read (using the same method as for (4.23))

$$\chi_3(1) = 1, \quad \chi_3(0) = 0, \quad \chi_3'(1) = 3, \quad \chi_3'(0) = \tfrac{1}{2}.$$

In the end, the following approximation of θ_2 is proposed

$$\theta_2(N^1) \quad \approx \quad \frac{1}{2} + N^1 \left(-\frac{1}{2} + (1 - |N^1|^2)(d_0 + d_1 |N^1|^2 + d_2 |N^1|^4) \right),$$

where the coefficients $d_0 = 0.386143553495150$, $d_1 = 0.488034553677475$ and $d_2 = -0.681343955348390$ are computed with MAPLE ([27]) such that the approximated χ_3 given in (4.40) approximates the exact χ_3 for 10^3 values of x equally distributed in $[0, 1]$, *i.e.* it minimizes the discrete L^2 error compared to the values obtained by solving numerically (4.15) and computing (4.16).

One verifies *a posteriori* that this approximation of θ_2 is in $]0, 1[$, *i.e.* it provides a realizable closure.

4.6.4 In $\tilde{\mathcal{R}}_2$, *i.e.* for 1D ansätze

Now, we focus on the M_2 model in 1D. The vector of polynomials $\mathbf{m}$ is

$$\mathbf{m}(\mu) = (1, \mu, \mu^2).$$

For moments in $\tilde{\mathcal{R}}_2$, the associated ansatz $\psi_{\tilde{\mathcal{R}}_2}$, *i.e.* a 1D M_2 ansatz, reads

$$\psi_{\tilde{\mathcal{R}}_2}(\Omega) = \exp(\lambda_0 + \lambda_1\mu + \lambda_2\mu^2), \tag{4.40}$$

where $(\lambda_0, \lambda_1, \lambda_2) \in \mathbb{R}^3$ are the unique coefficients such that $\psi_{\tilde{\mathcal{R}}_2}$ satisfies

$$\left\langle \psi_{\tilde{\mathcal{R}}_2}(\mu) \right\rangle = \psi^0, \qquad \left\langle \mu\psi_{\tilde{\mathcal{R}}_2}(\mu) \right\rangle = \psi^1, \qquad \left\langle \mu^2\psi_{\tilde{\mathcal{R}}_2}(\mu) \right\rangle = \psi^2.$$

Property 4.3 ***Realizability:*** *According to (4.38), the closure $N^3(N^1,\gamma_1)$ is realizable if it can be written as a convex combination of b_- and b_+*

$$\begin{aligned} N^3(N^1,\gamma_1) \quad &= \quad b_-\left(N^1, N^2(N^1,\gamma_1)\right) \quad \theta_3(N^1,\gamma) \\ &\quad + \quad b_+\left(N^1, N^2(N^1,\gamma_1)\right) \quad \left(1-\theta_3(N^1,\gamma_1)\right), \end{aligned} \tag{4.41}$$

where the functions b_- and b_+ are defined in (4.38), and $N^2(N^1,\gamma_1)$ is given by (4.35c).
Hyperbolicity: *The Jacobian of the flux (in 1D) reads*

$$\partial_{\psi}\mathbf{F}(\boldsymbol{\psi}) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a & b & c \end{pmatrix},$$
$$a = N^3(N^1,N^2) \; - \; N^1\partial_{N^1}N^3(N^1,N^2) \; - \; N^2\partial_{N^2}N^3(N^1,N^2),$$
$$b = \partial_{N^1}N^3(N^1,N^2) \qquad c = \partial_{N^2}N^3(N^1,N^2).$$

The Jacobian $\partial_{\psi}\mathbf{F}(\boldsymbol{\psi})$ has real eigenvalues iff the discriminant Δ of the characteristic polynomial (here of degree 3) is non-negative

$$\Delta \quad = \quad 18abc - 4ac^3 + b^2c^2 - 4b^2 - 27a^2 \geq 0. \tag{4.42}$$

Using again the change of variable $\Omega = -\Omega'$ in (4.36b) leads to write that N^3 is an odd function of N^1.

The coefficient θ_3 is approximated such that N^3 is an odd function of N^1 and such that it interpolates the following values of N^3

$$N^3\left(N^1,\gamma_1=0\right) = \left(N^1\right)^3, \quad N^3\left(N^1, \gamma_1 = \frac{\chi_2(|N_1^1|) - |N_1^1|^2}{1-|N_1^1|^2}\right) = \chi_3(N^1)$$
$$N^3\left(N^1,\gamma_1=1\right) = N^1,$$

i.e. it retreives the exact N^3 on the boundary of the realizbility domain ($\gamma_1 = 0$ or $\gamma_1 = 1$) where the exact values of N^3 are computed from (4.15) and it retreives the previous approximation (4.36c) in the M_1 case $(N^1,\gamma_1,\gamma_2) \in \mathcal{P}_3$.

Notation 4.2 *The polynomial of degree two interpolating the values A, B and C at the points a, b and c is denoted E.*
The polynomial of degree three which is zero in a, b and c is denoted Z.

$$\begin{aligned} E\left((A,a),(B,b),(C,c)\right)(x) &:= A\,\frac{x-b}{a-b}\frac{x-c}{a-c} + B\,\frac{x-a}{b-a}\frac{x-c}{b-c} \\ &\quad + C\,\frac{x-a}{c-a}\frac{x-b}{c-b}, \\ Z\left(a,b,c\right)(x) &:= (x-a)(x-b)(x-c). \end{aligned}$$

In the end, the following approximation of θ_3 is proposed

$$\begin{aligned} a_1 &:= 0, & A_1 &:= \frac{b_+ - \kappa_1}{b_+ - b_-}(N_1^1, a_1), \\ a_2 &:= \frac{\chi_2(|N_1^1|) - (N_1^1)^2}{1-(N_1^1)^2}, & A_2 &:= \frac{b_+ - \kappa_1}{b_+ - b_-}(N_1^1, a_2), \\ a_3 &:= 1, & A_3 &:= \frac{b_+ - \kappa_1}{b_+ - b_-}(N_1^1, a_3), \end{aligned} \tag{4.43}$$

$$\begin{aligned} \theta_3(N_1^1,\gamma_1) \approx{}& E\left((A_1,a_1),(A_2,a_2),(A_3,a_3)\right)(\gamma_1) \\ &+ Z\left(a_1,a_2,a_3\right)(\gamma_1)Q_1(N_1^1,\gamma_1), \end{aligned} \tag{4.44}$$

where Q_1 is a polynomial of N_1^1 and γ_1 of degree sixteen.

The coefficients of the polynomial Q_1 are computed to minimize the discrete L^2 distance between the approximation and the κ_1 computed by solving (4.15) for 10^4 values of $(N_1^1, \gamma_1) \in [0,1] \times [0,1]$. Those values are chosen from 100 values of N_1^1 equally distributed in $[0,1]$ and 100 of γ_1 equally distributed in $[0,1]$.

In this process, the minimization problem (4.15) is solved numerically using the minimization routine HUMSL of MINPACK ([33]) which calls the quadrature routine DQAGS of QUADPACK ([34]). Both those routines are iterative, and the error tolerance for both routines was fixed at 10^{-9}.

In the next subsections, an approximation of the closure is proposed in $\tilde{\mathcal{R}}_1$ and in $\tilde{\mathcal{R}}_0$ based on the approximation (4.41,4.44) and the error produced in the approximation (4.41,4.44) propagates and will be amplified in the next step of the approximation. Therefore a high accuracy is required, which explains why the polynomial Q_1 was chosen to be of such a high degree.

The discrete L^∞ error compared to the numerical solution of (4.15) for those 10^4 values of $(N_1^1, \gamma_1) \in [0,1] \times [0,1]$ is 8.43×10^{-3}.

One verifies *a posteriori* that this approximation of N^3 satisfies (4.42) and that $\theta_3 \in [0,1]$ for all $(N^1, \gamma_1) \in [-1,1] \times [0,1]$. Therefore this approximation of the 1D M_2 closure is hyperbolic and realizable.

4.6.5 In $\tilde{\mathcal{R}}_1$

In $\tilde{\mathcal{R}}_1$, the vector of polynomials $\mathbf{m}$ is

$$\mathbf{m}(\Omega) = (\Omega_1, \quad \Omega_1^1, \quad \Omega_2^2, \quad \Omega_3^2).$$

For moments in $\tilde{\mathcal{R}}_1$, the associated ansatz $\psi_{\tilde{\mathcal{R}}_1} \in \mathcal{E}_1$ reads

$$\psi_{\tilde{\mathcal{R}}_1}(\Omega) = \exp(\lambda_1\Omega_1 + \lambda_2\Omega_1^1 + \lambda_3\Omega_2^2 + \lambda_4\Omega_3^2), \tag{4.45}$$

where $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \in \mathbb{R}^4$ are the unique coefficients such that $\psi_{\tilde{\mathcal{R}}_1}$ satisfies

$$\begin{aligned} \left\langle \Omega_1\psi_{\tilde{\mathcal{R}}_1}(\Omega)\right\rangle &= \psi_1^1, & \left\langle \Omega_1^2\psi_{\tilde{\mathcal{R}}_1}(\Omega)\right\rangle &= \psi_{1,1}^2, \\ \left\langle \Omega_2^2\psi_{\tilde{\mathcal{R}}_1}(\Omega)\right\rangle &= \psi_{2,2}^2 & \left\langle \Omega_3^2\psi_{\tilde{\mathcal{R}}_1}(\Omega)\right\rangle &= \psi_{3,3}^2. \end{aligned}$$

Similarily as in the previous subsections, an approximation of the scalars κ_2 and κ_3, defined in (4.36a), interpolating the exact values of N^3 on the boundary of the realizability domain and interpolating the previous approximation when $\gamma_2 = (1-\gamma_1)/2$. Here the boundary of the realizability domain corresponds to fixing $\gamma_2 = 0$ and $\gamma_2 = 1-\gamma_1$. Using Proposition (3.7), those values reads

$$\begin{array}{cccccc} & & b_1 := 0, & b_2 := \dfrac{1-\gamma_1}{2}, & b_3 := 1-\gamma_1, & \\ B_1 & := & \kappa_2(N_1^1, \gamma_1, b_1) & = & (N_1^1)^3 + N_1^1(1-|N_1^1|^2)\gamma_1, \\ B_2 & := & \kappa_2(N_1^1, \gamma_1, b_2) & = & \kappa_1(N_1^1, \gamma_1), \\ B_3 & := & \kappa_2(N_1^1, \gamma_1, b_3) & = & (N_1^1)^3 + N_1^1(1-|N_1^1|^2)\gamma_1, \\ C_1 & := & \kappa_3(N_1^1, \gamma_1, b_1) & = & 0, \\ C_2 & := & \kappa_3(N_1^1, \gamma_1, b_2) & = & \frac{1}{2}(N_1^1 - \kappa_1(N_1^1, \gamma_1)), \\ C_3 & := & \kappa_3(N_1^1, \gamma_1, b_3) & = & (N_1^1)^3 + N_1^1(1-|N_1^1|^2)(1-\gamma_1). \end{array}$$

Similarly as in the previous subsection, the following approximations of κ_2 and κ_3 are proposed

$$\begin{aligned} \kappa_2(N^1, \gamma_1, \gamma_2) \quad &\approx \quad E\left((B_1, b_1),\ (B_2, b_2),\ (B_3, b_3)\right)(\gamma_2) \\ &\quad +Z(b_1, b_2, b_3)(\gamma_2)\ Q_2(N_1^1, \gamma_1, \gamma_2), \qquad (4.46a) \\ \kappa_3(N^1, \gamma_1, \gamma_2) \quad &\approx \quad E\left((C_1, b_1),\ (C_2, b_2),\ (C_3, b_3)\right)(\gamma_2) \\ &\quad +Z(b_1, b_2, b_3)(\gamma_2)\ Q_3(N_1^1, \gamma_1, \gamma_2) \qquad (4.46b) \end{aligned}$$

where Q_2 and Q_3 are polynomials of N^1, γ_1 and γ_2 of degree eight. As in the previous subsection, the coefficients of Q_2 and Q_3 are computed so that they minimize the discrete L^2 distance between the approximations and the coefficients κ_2 and κ_3 computed by solving (4.15) for 8×10^3 values of $(N_1^1, \gamma_1, \gamma_2) \in \mathcal{P}_1$. Those values are chosen from 20 values of N_1^1 equally distributed in $[0,1]$, 20 values of γ_1 in $[0,1]$ and 20 of γ_2 in $[0, 1-\gamma_1]$.

In this process, the minimization problem (4.15) is solved numerically using the minimization routine HUMSL of MINPACK ([33]) which calls the cubature routine DCUHRE from [3]. Both those routines are iterative, and the error tolerances for both routines was fixed at 10^{-9}.

The discrete L^∞ error compared to the numerical solution of (4.15) for those 8×10^3 values of $(N^1, \gamma_1, \gamma_2) \in \mathcal{P}_1$ is $2.09.10^{-2}$.

Studying the hyperbolicity and the realizability properties is no simpler for moments in $\tilde{\mathcal{R}}_1$ than in $\tilde{\mathcal{R}}_0$. Therefore, those properties are studied in the more general case of moments in $\tilde{\mathcal{R}}_0$ in the next subsection.

4.6.6 In the whole realizability domain $\tilde{\mathcal{R}}_0$

The approximation (4.36a-4.46) of the previous subsection provides a closure when N^1 is along one of the Cartesian axes. One can also compute the closure when $|N^1| = 1$ through the equality case of Proposition 3.6. In that case, all the moments of higher order read

$$N^i = (N^1)^{\otimes i}. \tag{4.47}$$

Now we aim to approximate N^3 at any point $P = (N^1, \gamma_1, \gamma_2) \in \mathcal{P}_0$. Define the following points (see Fig. 4.8 and 4.9)

$$P_1 = (N_1^1 e_1, \gamma_1, \gamma_2), \qquad P_2 = (N_2^1 e_2, \gamma_1, \gamma_2), \qquad P_3 = (N_3^1 e_3, \gamma_1, \gamma_2).$$

Those points correspond to the projections of P onto each Cartesian axis at fixed γ_1 and γ_2. At those points, N^1 is an eigenvalue of N^2 and therefore the previous approximation can be used, *i.e.* use the appropriate rotation to turn ψ^1 onto the axis e_1. Now define the following lines and their intersection points with the unit sphere (see Fig. 4.8 and 4.9)

$$\begin{aligned}
L_1 = (P_1, P), \qquad L_2 &= (P_2, P), \qquad L_3 = (P_3, P)\\
P_4 = L_1 \cap S^2 &= \left(N^1 + (N^1 - N_1^1 e_1)\sqrt{\frac{1 - |N_1^1|^2}{|N_2^1|^2 + |N_3^1|^2}}, \gamma_1, \gamma_2 \right),\\
P_5 = L_2 \cap S^2 &= \left(N^1 + (N^1 - N_2^1 e_2)\sqrt{\frac{1 - |N_2^1|^2}{|N_1^1|^2 + |N_3^1|^2}}, \gamma_1, \gamma_2 \right),\\
P_6 = L_3 \cap S^2 &= \left(N^1 + (N^1 - N_3^1 e_3)\sqrt{\frac{1 - |N_3^1|^2}{|N_1^1|^2 + |N_2^1|^2}}, \gamma_1, \gamma_2 \right).
\end{aligned}$$

According to (4.47), the closure N^3 at the points P_4, P_5 and P_6 read

$$\begin{aligned}
N^3(P_4) &= \left(N^1 + (N^1 - N_1^1 e_1)\sqrt{\frac{1 - |N_1^1|^2}{|N_2^1|^2 + |N_3^1|^2}} \right)^{\otimes 3},\\
N^3(P_5) &= \left(N^1 + (N^1 - N_2^1 e_2)\sqrt{\frac{1 - |N_2^1|^2}{|N_1^1|^2 + |N_3^1|^2}} \right)^{\otimes 3},\\
N^3(P_6) &= \left(N^1 + (N^1 - N_3^1 e_3)\sqrt{\frac{1 - |N_3^1|^2}{|N_1^1|^2 + |N_2^1|^2}} \right)^{\otimes 3}.
\end{aligned}$$

For $P \in \mathcal{P}_0$, the different components of the closure N^3 are approximated by different convex combinations. In particular, $N^3_{i,j,j}(P)$ is approximated by a convex combination of $N^3_{i,j,j}(P_i)$ and $N^3_{i,j,j}(P_{3+i})$.

Similarly, the value of $N^3_{1,2,3}$ is known at the points

$$\begin{aligned}
P_0 &= (0_{\mathbb{R}^3}, \gamma_1, \gamma_2), \qquad P_7 = (N^1/|N^1|, \gamma_1, \gamma_2),\\
N^3(P_0) &= 0_{\mathbb{R}^{3\times3\times3}}, \qquad N^3(P_7) = \left(\frac{N^1}{|N^1|} \right)^{\otimes 3}.
\end{aligned}$$

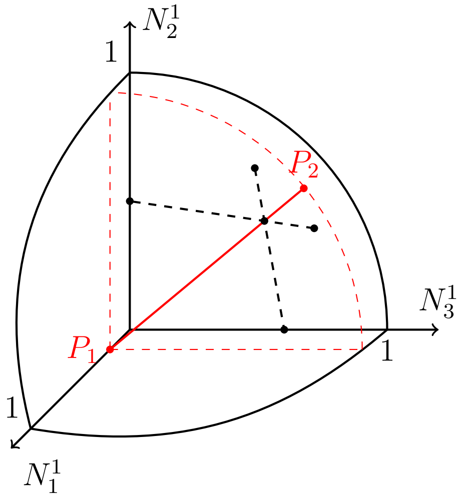

Figure 4.8: Configuration at fixed γ_1 and γ_2.

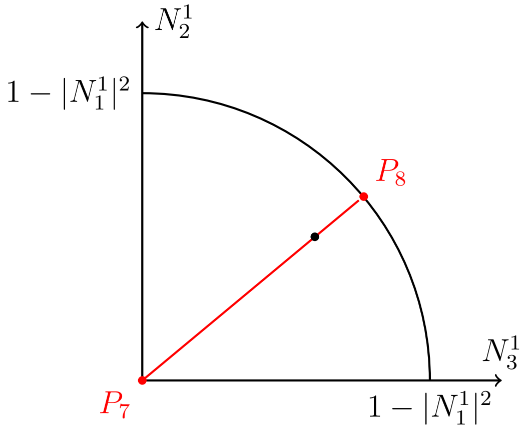

Figure 4.9: Configuration at fixed γ_1, γ_2 and N_3^1.

Therefore $N^3_{1,2,3}$ can be approximated by a convex combination of $N^3_{1,2,3}(P_0)$ and $N^3_{1,2,3}(P_7)$. In the end, for a point $P \in \mathcal{P}_0$ those approximations read

$$N^3_{1,1,1}(P) \approx (1-\alpha_1)N^3_{1,1,1}(P_1) + \alpha_1 N^3_{1,1,1}(P_4), \tag{4.48a}$$
$$N^3_{1,2,2}(P) \approx (1-\alpha_1)N^3_{1,2,2}(P_1) + \alpha_1 N^3_{1,2,2}(P_4), \tag{4.48b}$$
$$N^3_{1,3,3}(P) \approx (1-\alpha_1)N^3_{1,3,3}(P_1) + \alpha_1 N^3_{1,3,3}(P_4), \tag{4.48c}$$

$$N^3_{1,1,2}(P) \approx (1-\alpha_2)N^3_{1,1,2}(P_2) + \alpha_2 N^3_{1,1,2}(P_5), \tag{4.48d}$$
$$N^3_{2,2,2}(P) \approx (1-\alpha_2)N^3_{2,2,2}(P_2) + \alpha_2 N^3_{2,2,2}(P_5), \tag{4.48e}$$
$$N^3_{2,3,3}(P) \approx (1-\alpha_2)N^3_{2,3,3}(P_2) + \alpha_2 N^3_{2,3,3}(P_5), \tag{4.48f}$$

$$N^3_{1,1,3}(P) \approx (1-\alpha_3)N^3_{1,1,3}(P_3) + \alpha_3 N^3_{1,1,3}(P_6), \tag{4.48g}$$
$$N^3_{2,2,3}(P) \approx (1-\alpha_3)N^3_{2,2,3}(P_3) + \alpha_3 N^3_{2,2,3}(P_6), \tag{4.48h}$$
$$N^3_{3,3,3}(P) \approx (1-\alpha_3)N^3_{3,3,3}(P_3) + \alpha_3 N^3_{3,3,3}(P_6), \tag{4.48i}$$

$$N^3_{1,2,3}(P) \approx (1-\alpha_4)N^3_{1,2,3}(P_0) + \alpha_4 N^3_{1,2,3}(P_7), \tag{4.48j}$$

$$\begin{aligned} \alpha_1 &= \frac{|N_2^1|^2 + |N_3^1|^2}{1-|N_1^1|^2}, & \alpha_2 &= \frac{|N_1^1|^2 + |N_3^1|^2}{1-|N_2^1|^2}, \\ \alpha_3 &= \frac{|N_1^1|^2 + |N_2^1|^2}{1-|N_3^1|^2}, & \alpha_4 &= |N^1|. \end{aligned}$$

Remark that those linear combinations are only one choice, and several others were possible. However this choice was found to give the best approximation of the M_2 closure.

Precision and numerical cost

First the accuracy and the computational costs of the approximation (4.48) in $\mathcal{P}_0$ are compared to the ones of the minimization procedure (4.15).

The approximation is compared to the solution of (4.15) computed numerically for 3.2×10^6 values of $(N^1, \gamma_1, \gamma_2) \in \mathcal{P}_0$. Those values are obtained from 20 values of N^1_1 equally distributed in $[0, 1]$, 20 of N^1_2 in $[0, \sqrt{1 - |N^1_1|^2}]$, 20 of N^1_3 in $[0, \sqrt{1 - |N^1_1|^2 - |N^1_2|^2}]$, 20 of γ_1 in $[0, 1]$ and 20 of γ_2 in $[0, 1 - \gamma_1]$.

In this process, the minimization problem (4.15) is solved numerically using the minimization routine HUMSL of MINPACK ([33]) which calls the cubature routine DCUHRE from [3] (see *e.g.* [19, 2, 1] for more robust methods). Both those routines are iterative, and the error tolerance for both routines was fixed at 10^{-9}.

The discrete L^∞ error compared to the numerical solutions of the minimization problem (4.15) for 3.2×10^6 values of $P \in \tilde{\mathcal{P}}_0$ is of 3.12×10^{-2}.

In order to compare the computational costs between evaluating the value of the approximation and the numerical methods for solving (4.15), the error tolerance in this numerical method is fixed at 3×10^{-2}, *i.e.* such that both methods have the same accuracy.

The computational time required for computing N^3 for 3.2×10^6 values of $P \in \tilde{\mathcal{P}}_0$ with both methods are summarized in Table 4.1.

	Minimization solver	Approximation
Computation times	1654 sec = 27 min 34 sec	0.434 sec
L^∞ error	$\leq 3 \times 10^{-2}$	3.12×10^{-2}

Table 4.1: Computation times and discrete L^∞ error with the approximation and the numerical method for solving (4.15) for 3.2×10^6 values of $P \in \mathcal{P}_0$.

Hyperbolicity

For moments in $\tilde{\mathcal{R}}_1$ or in $\tilde{\mathcal{R}}_0$, the Jacobians J of the flux (defined in (4.6)) are 10×10 matrices. No analytical condition were found to verify that such matrices were diagonalizable with real eigenvalues for all $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$.

Instead, one can verify that those matrices are diagonalizable with real eigenvalues for a finite number of points in $P \in \mathcal{P}_0$. This does not prove that the approximation is hyperbolic, but this provides a good indication that the approximation is not non-hyperbolic.

For this purpose, consider 10^5 values of $P \in \mathcal{P}_0$ obtained from 10 values of N^1_1 in $[0, 1]$, 10 of N^1_2 in $[0, \sqrt{1 - |N^1_1|^2}]$, 10 of N^1_2 in $[0, \sqrt{1 - |N^1_1|^2 - |N^1_2|^2}]$, 10 of γ_1 in $[0, 1]$ and 10 of γ_2 in $[0, 1 - \gamma_1]$, each equally distributed in their respective intervals.

The Jacobians were found to be diagonalizable with real eigenvalues at each of those 10^5 points $P \in \mathcal{P}_0$.

Remark that improper numerical methods for solving the minimization problem (4.15) also introduces numerical errors which may result in losing the hyperbolicity property.

Realizability

In multi-D, no sufficient conditions for a vector to be realizable were found. As the closure proposed here is exact on the boundary of the realizability domain and approximates a realizable closure, one may expect the approximation to be realizable.

4.7 Other closures

The M_N closure was mainly studied in this thesis because it retains the major characteristics of the underlying kinetic model. Although other closures for angular moment models are available.

4.7.1 The polynomial (P_N) closure

The P_N (the "P" stands for polynomial) closure is often used in computational physics because it is very simple of implementation and it offers several desirable properties.

As for the M_N closure, the P_N closure is obtained by reconstructing an ansatz ψ_{P_N} which is simply chosen to be a polynomial of degree N

$$\psi_{P_N} = \mathbf{m}\boldsymbol{\lambda}, \tag{4.49}$$

where $\boldsymbol{\lambda}$ is the unique vector such that

$$\boldsymbol{\psi} = \langle \mathbf{m}\,(\mathbf{m}\boldsymbol{\lambda})\rangle = \langle \mathbf{m} \otimes \mathbf{m}\rangle\, \boldsymbol{\lambda},$$

which can simply be computed by inverting the matrix $\langle \mathbf{m} \otimes \mathbf{m}\rangle$. And the closure is again computed from the ansatz ψ_{P_N} through

$$F(\boldsymbol{\psi}) = \langle \Omega \otimes \mathbf{m} \otimes \mathbf{m}\rangle\, \boldsymbol{\lambda}.$$

Advantages

Linearity: Based on their definitions, both the P_N ansatz ψ_{P_N} and the P_N closure are linear functions of $\boldsymbol{\psi}$.
The P_N model can also be interpreted as a spectral method ([8], see also [7]), *i.e.* as a truncated polynomial expansion of ψ.
Hyperbolicity: The Jacobians of the fluxes do not depend on $\boldsymbol{\psi}$ and one verifies *a priori* that they are diagonalizable with real eigenvalues.

Entropic: The P_N closure can be interpreted as an entropy-based method (see *e.g.* [18]). Indeed, the P_N ansatz is the function minimizing the quadratic entropy

$$\eta(f) = f^2$$

over the set of $L^1(S)$ functions (here non-restricted to positive ones $L^1(S)^+$) having $\boldsymbol{\psi}$ for moments

$$\left\{\psi \in L^1(S), \quad \text{s.t.} \quad \langle \mathbf{m}\psi \rangle = \boldsymbol{\psi}\right\}. \tag{4.50}$$

(S)

Drawbacks

Non-realizable: The P_N ansatz ψ_{P_N} defined in (4.49) is simply a polynomial of degree N. There are no constraints on the coefficients $\boldsymbol{\lambda}$ to enforce the positivity of ψ_{P_N}. Therefore, for a realizable vector $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$, the vector $(\boldsymbol{\psi}, F(\boldsymbol{\psi}))$ is not always a vector of moments when F is chosen to be the P_N closure. This phenomenum may lead to instabilities ([18, 28]) or to modelling issues. This problem can also be adressed by choosing a positive ansatz $\psi \in L^1(S)^+$ in (4.50) (see [18]), although this requires to use similar numerical methods to compute such an ansatz as to compute a M_N ansatz.

Beam modelling issue: When considering a beam in the direction e_i, *i.e.* a Dirac measure, the P_N ansatz fails to approximate such a measure. Indeed extracting the moments $\boldsymbol{\psi}$ of such a Dirac measure and reconstructing the P_N ansatz ψ_{P_N} from those moments, one observes that the P_N ansatz differs from the original Dirac measure.

Example 4.3 *Consider a 1D measure of the form*

$$\psi = \delta\left(\mu - 1\right).$$

Extracting the first two order moments reads

$$\psi^0 = 1, \qquad \psi^1 = 1.$$

Computing the P_N ansatz from those reads

$$\psi_{P_N}(\mu) = \frac{1 + 3\mu}{2}.$$

Finally the closure differs from the only realizable closure in that case

$$\int_{-1}^{+1} \mu^2 \psi_{P_N}(\mu) d\mu = \frac{1}{3} \quad \neq \quad 1 = \int_{-1}^{+1} \mu^2 \delta(\mu - 1).$$

This problem does not appear with the M_N closure. This is simply due to the fact that the sum of N Dirac measures are in the closure of the set of possible M_N ansatz (a Dirac is the limit of an exponential function, see Example 3.1), while they are not in the set of possible P_N ansatz

$$\psi \in \overline{\{\exp(\boldsymbol{\lambda}\mathbf{m}), \quad \boldsymbol{\lambda} \in \mathbb{R}^{Card(\mathbf{m})}\}}, \qquad \psi \notin \overline{\mathbb{R}^{Card(\mathbf{m})}[X_1, X_2, X_3]}.$$

In external radiotherapy, modeling beams of particles is of major importance. In order to obtain accurate results when considering beams with a P_N model, one needs to use a high order model (N large) which results in raising the numerical costs as the size of $\boldsymbol{\psi}$ raises as $(N+1)^2$.

4.7.2 An atom-based (K_N) closure

The Kershaw closure, or K_N closure (the "K" stands for D. Kershaw [23]), is based on an atomic decomposition ([23, 32, 35, 36, 11, 24]).

In 1D, the idea proposed in [32, 35, 36] consists in exploiting the knowledge of the unique representing measure for a vector $\mathbf{V} \in \partial\mathcal{R}_{\mathbf{m}}^m$ on the boundary of the realizability domain (see the proof of Theorem 3.1). From this unique representing measure, one deduces the existence of a unique realizable closure on this boundary, which can be computed numerically.

When the vector $\mathbf{V} \in \mathcal{R}_{\mathbf{m}}$ is in the interior of the realizability domain, it can be written as a convex combination of vectors $\mathbf{V}_1$ and $\mathbf{V}_2$ on the boundary $\partial\mathcal{R}_{\mathbf{m}}^m$ and of an equilibrium point $\mathbf{V}_0 = \langle \mathbf{m} \rangle$ (see [32, 35, 36] for details)

$$\mathbf{V} = \alpha_0\mathbf{V}_0 + \alpha_1\mathbf{V}_1 + \alpha_2\mathbf{V}_2. \tag{4.51}$$

Remark that a similar method was used to construct (3.30). At each of these points a representing measure is known

$$\mathbf{V}_0 = \int_{-1}^{+1} \mathbf{m}(\mu)d\mu, \quad \mathbf{V}_1 = \int_{-1}^{+1} \mathbf{m}(\mu)d\gamma_1(\mu), \quad \mathbf{V}_2 = \int_{-1}^{+1} \mathbf{m}(\mu)d\gamma_2(\mu),$$

where γ_1 and γ_2 are sums of Dirac measures.

Finally the K_N closure consists in writing the flux $\mathbf{F}(\mathbf{V})$ from this representing measure

$$\mathbf{F}(\mathbf{V}) = \alpha_0 \int_{-1}^{+1} \mu\mathbf{m}(\mu)d\mu + \alpha_1 \int_{-1}^{+1} \mu\mathbf{m}d\gamma_1(\mu) + \alpha_2 \int_{-1}^{+1} \mu\mathbf{m}(\mu)d\gamma_2(\mu).$$

Remark that such an atom-based closure is non unique because different sets of coefficients α_0, α_1, α_2 and of vectors $\mathbf{V}_1$ and $\mathbf{V}_2$ satisfy (4.51). The ones proposed and studied in [32, 36] are possible atom-based closure that were shown to have certain properties. Although other choices of atom-based representing measures can be proposed. The one proposed in (3.30) was shown to have the minimum number of atoms in [11, 13, 14, 12].

Advantages

By construction this closure is **realizable**.

The 1D K_1 and K_2 closures were found to be **hyperbolic** ([32, 36, 35]).

Drawbacks

The main issue is the current lack of knowledge on the realizability property for moments over S^2. This closure is therefore restricted to 1D problems.

Bibliography

[1] G. W. Alldredge, C. D. Hauck, D. P. O'Leary, and A. L. Tits. Adaptive change of basis in entropy-based moment closures for linear kinetic equations. *J. Comput. Phys.*, 74(4), february 2014.

[2] G. W. Alldredge, C. D. Hauck, and A. L. Tits. High-order entropy-based closures for linear transport in slab geometry II: A computational study of the optimization problem. *SIAM J. Sci. Comput.*, 34(4):361–391, 2012.

[3] J. Berntsen, T. O. Espelid, and A. Genz. Algorithm 698: DCUHRE: An adaptive multidemensional integration routine for a vector of integrals. *ACM T. Math. Software*, 17(4):452–456, 1991. `http://netlib.org/toms/`.

[4] J. Borwein and A. Lewis. Duality relationships for entropy-like minimization problems. *SIAM J. Control Optim.*, 29(2):325–338, 1991.

[5] J. Borwein and A. Lewis. Partially finite convex programming, part I: Quasi relative interiors and duality theory. *Math. Program.*, 57:15–48, 1992.

[6] J. Borwein and A. Lewis. Partially finite convex programming: Part II. *Math. Program.*, 57:49–83, 1992.

[7] Y. Bourgault, D. Broizat, and P.-E. Jabin. Convergence rate for the method of moments with linear closure relations. *Kin. Rel. Mod.*, 8(1):1–27, 2015.

[8] C. Canuto, M. Y. Hussaini, A. Quarteroni, and Th. A. Zang. *Spectral Methods: Fundamental in single domains.* Springer, 2006.

[9] C. Cercignani. *The Boltzmann equation and its applications.* Springer, 1988.

[10] C. Cercignani. *The relativistic Boltzmann equation: Theory and applications.* Birkäuser, 2002.

[11] R. Curto and L. A. Fialkow. Recusiveness, positivity, and truncated moment problems. *Houston J. Math.*, 17(4):603–634, 1991.

[12] R. Curto and L. A. Fialkow. The truncated complex K-moment problem. *T. Am. Math. Soc.*, 352(6):2825–2855, 2000.

[13] R. Curto and L. A. Fialkow. A duality prood to Tchakaloff's theorem. *J. Math. Anal. Appl.*, 269:519–536, 2002.

[14] R. Curto and L. A. Fialkow. Truncated K-moment problems in several variables. *arXiv preprint math/0507067*, 2005.

[15] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology: Volume 6, Evolution problems II.* Springer, 2000.

[16] B. Dubroca and J.-L. Feugeas. Hiérarchie des modèles aux moments pour le transfert radiatif. *C. R. Acad. Sci. Paris*, 329:915–920, 1999.

[17] K. O. Friedrichs and P. D. Lax. Systems of conservation equations with a convex extension. *Proc. Nat. Acad. Sci.*, 68(8):1686–1688, 1971.

[18] C. Hauck and R. McClarren. Positive P_N closures. *SIAM J. Sci. Comput.*, 32(5):2603–2626, 2010.

[19] C. D. Hauck. High-order entropy-based closures for linear transport in slab geometry. *Commun. Math. Sci*, 9(1):187–205, 2011.

[20] C. D. Hauck, C. D. Levermore, and A. L. Tits. Convex duality and entropy-based moment closures: Characterizing degenerate densities. *SIAM J. Control Optim.*, 47(4):1977–2015, 2008.

[21] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630, 1957.

[22] M. Junk. Maximum entropy for reduced moment problems. *Math. Mod. Meth. Appl. S.*, 10(1001–1028):2000, 1998.

[23] D. Kershaw. Flux limiting nature's own way. Technical report, Lawrence Livermore Laboratory, 1976.

[24] J. B. Lasserre. *Moments, positive polynomials and their applications.* Imperial College press optimization series, 2010.

[25] C. D. Levermore. Relating Eddington factors to flux limiters. *J. Quant. Spectros. Radiat. Transfer*, 31:149–160, 1984.

[26] C. D. Levermore. Moment closure hierarchies for kinetic theories. *J. Stat. Phys.*, 83(5–6):1021–1065, 1996.

[27] Maple$^{\text{TM}}$. Technical report, Maplesoft, a division of Waterloo Maple Inc., 2016.

[28] R. G. McClarren and C. D. Hauck. Robust and accurate filtered spherical harmonics expansion for radiative transfer. *J. Comput. Phys.*, pages 5597–5614, october 2010.

[29] L. R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *J. Math. Phys.*, 25(8):2404–2417, 1984.

[30] G. N. Minerbo. Maximum entropy Eddington factors. *J. Quant. Spectros. Radiat. Transfer*, 20:541–545, 1978.

[31] G. N. Minerbo. Maximum entropy reconstruction from cone-beam projection data. *Comput. Biol. Med.*, 9(1):29–37, 1979.

[32] P. Monreal. *Moment realizability and Kershaw closures in radiative transfer.* PhD thesis, Rheinisch-Westfälische Technische Hochschule, 2012.

[33] J. J. Moré, B. S. Garbow, and K. E. Hillstrom. *User Guide for MINPACK-1*, 1980. http://www.netlib.org/minpack/.

[34] R. Piessens, E. De Doncker-Kapenga, and C. W. Überhuber. *QUADPACK: A subroutine package for automatic integration*, springer edition, 1983. http://www.netlib.org/quadpack/.

[35] F. Schneider. *Moment models in radiation transport equations.* PhD thesis, Teschnische Universität Kaiserslautern, 2015.

[36] F. Schneider. Kershaw closures for linear transport equations in slab geometry I: Model derivation. *J. Comput. Phys.*, pages –, 2016.

[37] J. Schneider. Entropic approximation in kinetic theory. *ESAIM-Math. Model. Num.*, 38(3):541–561, 2004.

Part III

Numerical methods

Chapter 5

Numerical schemes

5.1 Introduction

This chapter is devoted to constructing numerical schemes adapted to the M_1 and M_2 model constructed in the previous chapters. In addition to the common stability and consistency requirements, numerical schemes for M_N equations need to preserve the realizability property. Indeed if this property is lost during the computations, the M_N closure exists not and the computations break.

The numerical schemes constructed in this chapter are based on methods commonly used for hyperbolic systems of conservation laws, and some of them can only be used when considering hyperbolic operators. However, not all the moment equations studied possesses such operators, *e.g.* the moment equations for photons transport (2.3) do not. Therefore, numerical schemes are constructed for the following three types of equations.

- For the sake of simplicity, the issues emerging when constructing numerical schemes for moment equations are first described and solved on the following toy problem

$$\textbf{in 1D:} \qquad \partial_\epsilon \boldsymbol{\psi} - \frac{1}{\rho}\partial_x \mathbf{F}(\boldsymbol{\psi}) = 0, \tag{5.1a}$$

$$\textbf{in 3D:} \qquad \partial_\epsilon \boldsymbol{\psi} - \frac{1}{\rho}\nabla_x F(\boldsymbol{\psi}) = 0, \tag{5.1b}$$

  and afterward apply to the particle transport equations.

- As an extension, the transport of electrons alone (without photons) is considered, when the collision operator is chosen to be either a LBCSD operator or a LBFP operator. Such a transport equation contains an hyperbolic operator (see Chapter 2). For simplicity, the gain term of Møller's secondary electrons is also removed. Such an equation is rewritten

$$\textbf{in 1D:} \qquad \partial_\epsilon (S\boldsymbol{\psi}) - \frac{1}{\rho}\partial_x \mathbf{F}(\boldsymbol{\psi}) + M\boldsymbol{\psi} = 0, \tag{5.2a}$$

$$\textbf{in 3D:} \qquad \partial_\epsilon (S\boldsymbol{\psi}) - \frac{1}{\rho}\nabla_x F(\boldsymbol{\psi}) + M\boldsymbol{\psi} = 0, \tag{5.2b}$$

where $\boldsymbol{\psi} \equiv \boldsymbol{\psi}_e$ are the moments of the fluence of electrons and the matrix M is defined by

$$\begin{aligned} M(\epsilon) &= s_{Mott}(\epsilon) - \sigma_{T,Mott}(\epsilon)Id && (5.3a)\\ &\qquad\qquad\text{when considering a CSD operator,} \\ M(\epsilon) &= T(\epsilon)M_{FP} \qquad \text{when considering a FP operator,} && (5.3b)\end{aligned}$$

see Appendices 2.A.1 and 2.A.2 for the definitions of the matrices s and M_{FP}.

- Finally, the equations of transport of electrons and photons together are considered. Such a system contains no hyperbolic operator, they are rewritten

$$\begin{aligned} \textbf{in 1D:} \qquad & \partial_x \mathbf{F}(\boldsymbol{\psi}) - \rho\mathbf{Q}(\boldsymbol{\psi}) = 0, && (5.4a)\\ \textbf{in 3D:} \qquad & \nabla_x F(\boldsymbol{\psi}) - \rho\mathbf{Q}(\boldsymbol{\psi}) = 0, && (5.4b)\end{aligned}$$

where $\boldsymbol{\psi} \equiv (\boldsymbol{\psi}_\gamma, \boldsymbol{\psi}_e)$ are the moments of the fluences of photons and electrons and $\mathbf{Q}$ is the associated collision operator. In the present work, the electron collisions are chosen to be modeled by a LBCSD operator, so $\mathbf{Q}$ is chosen to be

$$\begin{aligned} \mathbf{Q}(\boldsymbol{\psi}) &= \Big(\mathbf{G}_{\mathbf{C},\gamma}(\boldsymbol{\psi}_\gamma) - \mathbf{P}_{\mathbf{C}}(\boldsymbol{\psi}_\gamma), && (5.5)\\ &\qquad \partial_\epsilon(S\boldsymbol{\psi}_e) + (\mathbf{G}_{M,2} + M)(\boldsymbol{\psi}_e) + \mathbf{G}_{\mathbf{C},\mathbf{e}}(\boldsymbol{\psi}_\gamma)\Big).\end{aligned}$$

Remark 5.1 *At the numerical level, the transport of electrons can not be modeled by the original linear Boltzmann model. Indeed, in Subsections 1.4.3 and 1.4.4, the differential cross sections for electronic collisions were shown to be very peaked along the line $\epsilon' = \epsilon$. For such an equation to be accurately discretized, one would need the energy grid to be fine enough to capture such a peak. For application to the present problem, using such a fine grid was found to be numerically too costly.*

Remark that even if the elastic cross section, i.e. σ_{Mott} is peaked in angle along the line $\Omega'.\Omega = 1$, this peak appears not after integration when considering moment models. Therefore the CSD approximation can be used at the numerical level with moment models, but not with the kinetic model.

In all this chapter, the energy variable ϵ is considered as a "numerical time".

As desribed in Remark 1.6, the equations of transport of electrons and photons need to be solved backward in energy, *i.e.* from a maximum energy $\epsilon^0 = \epsilon^{\max}$ to a minimum one $\epsilon^{n_{\max}} = \epsilon^{\min}$. This direction is unusual and for the sake of simplicity and in order to use the "classical" notations and retreive some results from the litterature, the toy problem (5.1) is solved forward in energy, while the transport equations (5.2) and (5.4) are solved backward in energy.

The superscript n to the cell $[\epsilon^{n+\frac{1}{2}}, \epsilon^{n-\frac{1}{2}}]$ and the superscript $n+\frac{1}{2}$ to the interface $\epsilon^{n+\frac{1}{2}}$ between ϵ^{n+1} and ϵ^n. Remark that $\epsilon^n > \epsilon^{n+1}$ and $\epsilon^{n-\frac{1}{2}} > \epsilon^{n+\frac{1}{2}}$ when the energy is considered in the backward direction.

Finite volume and finte differences schemes are used, and the spatial grid is always chosen to be Cartesian. In 1D, the subscript l refers to the cell $[x_{l-\frac{1}{2}}, x_{l+\frac{1}{2}}]$ and the subscript $l+\frac{1}{2}$ to the interface $x_{l+\frac{1}{2}}$.

5.2 Approximate Riemann solver

Several numerical schemes adapted to angular moment equations such as (5.1) were proposed in the litterature (see *e.g.* [10, 24, 17, 7]).

Such state-of-the-art methods are typically stable under a Courant-Friedrichs-Lewy (CFL) condition depending on density ρ. Such stability conditions typically become very restrictive when the medium contains low density region ([8, 22]), *e.g.* in the field of medical physics when irradiating a lung ($\rho_{lung} \approx 0.3$). This problem is first presented through the following standard technique for the 1D equation (5.1a) and addressed in Sections 5.4, 5.5 and 5.7.

The HLL approximate Riemann solver (named after A. Harten, P. D. Lax and B. Van Leer [15]) is commonly used for M_N equations because it is known to preserve realizability from one energy step to another (see *e.g.* [7]) and is easy to construct even when considering non-linear fluxes (which is the case with M_N equations). This scheme is obtained by approximating the solution of the Riemann problem at each interface $x_{l+\frac{1}{2}}$.

5.2.1 The Riemann problem in 1D

Consider that the relative density $\rho(x) = \rho$ is homogeneous in a 1D medium and $\boldsymbol{\psi}(x, \epsilon^n)$ is constant on each side of the interface $x_{l+\frac{1}{2}}$ and denote

$$\begin{cases} \boldsymbol{\psi}(\epsilon^n, x) &= \boldsymbol{\psi}_L, \\ \mathbf{F}(\boldsymbol{\psi})(\epsilon^n, x) &= \mathbf{F}_L, \end{cases} \qquad \forall x \in]-\infty, x_{l+\frac{1}{2}}], \tag{5.6a}$$

$$\begin{cases} \boldsymbol{\psi}(\epsilon^n, x) &= \boldsymbol{\psi}_R, \\ \mathbf{F}(\boldsymbol{\psi})(\epsilon^n, x) &= \mathbf{F}_R, \end{cases} \qquad \forall x \in [x_{l+\frac{1}{2}}, +\infty[. \tag{5.6b}$$

The solution of the Riemann problem (5.1a) with the initial condition (5.6) is composed of waves of different velocity which are the eigenvalues of $\partial_{\boldsymbol{\psi}}\mathbf{F}/\rho$. In the case of the M_N system of equations, those velocities are bounded using the following lemma.

Lemma 5.1 *The eigenvalues of the Jacobian $\partial_{\boldsymbol{\psi}}\mathbf{F}(\boldsymbol{\psi})$ are bounded by 1 for all realizable moments $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$, i.e.*

$$\forall \boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}, \quad Sp\left(\partial_{\boldsymbol{\psi}}\mathbf{F}(\boldsymbol{\psi})\right) \subset]-1, 1[.$$

Proof Using the form of the 1D M_N closure (4.37), define the matrices

$$\begin{aligned} A &:= \partial_{\boldsymbol{\psi}} \mathbf{F}(\boldsymbol{\psi}), \\ B &:= \partial_{\boldsymbol{\lambda}} \mathbf{F}(\boldsymbol{\psi}) = \int_{-1}^{+1} \mu \mathbf{m}(\mu) \otimes \mathbf{m}(\mu) \exp(\boldsymbol{\lambda}^T \mathbf{m}(\mu)) d\mu, \\ C &:= \partial_{\boldsymbol{\lambda}} \boldsymbol{\psi} = \int_{-1}^{+1} \mathbf{m}(\mu) \otimes \mathbf{m}(\mu) \exp(\boldsymbol{\lambda}^T \mathbf{m}(\mu)) d\mu. \end{aligned}$$

The Jacobian $\partial_{\boldsymbol{\psi}} \mathbf{F}(\boldsymbol{\psi})$ can be rewritten

$$\partial_{\boldsymbol{\psi}} \mathbf{F}(\boldsymbol{\psi}) = A = BC^{-1}.$$

One needs to check that the eigenvalues α_i associated to the eigenvectors V_i are of norm inferior to 1

$$AV_i = BC^{-1}V_i = \alpha_i V_i \quad \Rightarrow \quad BW_i = \alpha_i CW_i$$

where $W_i = C^{-1}V_i$. Then one has

$$\alpha_i = \frac{BW_i.W_i}{CW_i.W_i}.$$

Remark that the matrix C is strictly positive as long as $\boldsymbol{\psi}$ is realizable, so here $CW_i.W_i > 0$. Using a Poincaré inequality, one finds

$$\begin{aligned} BW_i.W_i &= \int_{-1}^{+1} \mu (\mathbf{m}(\mu)^T W_i)^2 \exp(\boldsymbol{\lambda}^T \mathbf{m}(\mu)) d\mu \\ &\leq \int_{-1}^{+1} (\mathbf{m}(\mu)^T W_i)^2 \exp(\boldsymbol{\lambda}^T \mathbf{m}(\mu)) d\mu = CW_i.W_i, \end{aligned}$$

which leads to the result. □

This result also holds in multi-D.

Define the cone $\mathcal{C}^n_{l+\frac{1}{2}}$ (see Fig. 5.1) by

$$\mathcal{C}^n_{l+\frac{1}{2}} := \left\{ (x, \epsilon) \in \mathbb{R} \times \mathbb{R}^+, \quad \text{s.t.} \quad \rho |x_{l+\frac{1}{2}} - x| \leq |\epsilon^n - \epsilon| \right\}.$$

Using Lemma 5.1, the velocities of the waves propagating from the interface $x_{l+\frac{1}{2}}$ are inferior in norm to $1/\rho$. This implies that the solution $\boldsymbol{\psi}$ of the Riemann problem is constant out of the cone $\mathcal{C}^n_{l+\frac{1}{2}}$, *i.e.* $\boldsymbol{\psi}(x, \epsilon) = \boldsymbol{\psi}^n_l$ on the left of the cone (when $(x - x_{l+\frac{1}{2}})\rho < -(\epsilon - \epsilon^n)$) and to $\boldsymbol{\psi}(x, \epsilon) = \boldsymbol{\psi}^n_{l+1}$ on the right of the cone (when $(x - x_{l+\frac{1}{2}})\rho > (\epsilon - \epsilon^n)$).

In order to construct the approximate Riemann solver, the value of $\boldsymbol{\psi}(x, \epsilon)$ inside the cone $\mathcal{C}^n_{l+\frac{1}{2}}$ is approximated by its average value written $\boldsymbol{\psi}^*_{l+\frac{1}{2}}$ which is given by

$$\boldsymbol{\psi}^*_{l+\frac{1}{2}} = \frac{1}{\Delta x} \int_{x_{l+\frac{1}{2}} - \frac{\Delta \epsilon}{\rho}}^{x_{l+\frac{1}{2}} + \frac{\Delta \epsilon}{\rho}} \boldsymbol{\psi}(\epsilon^n + \Delta \epsilon, x) dx.$$

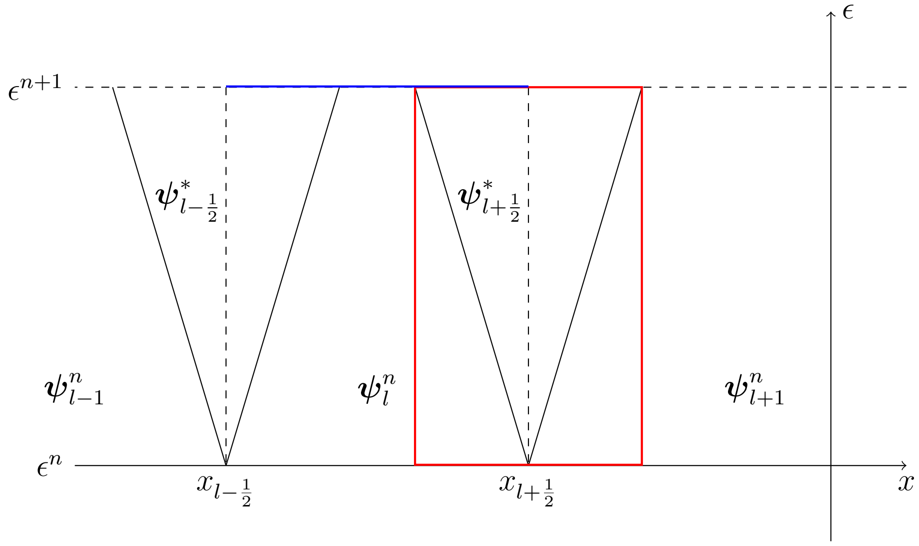
ϵ
ϵ^{n+1}
$\boldsymbol{\psi}^*_{l-\frac{1}{2}}$
$\boldsymbol{\psi}^*_{l+\frac{1}{2}}$
$\boldsymbol{\psi}^n_{l-1}$
$\boldsymbol{\psi}^n_{l}$
$\boldsymbol{\psi}^n_{l+1}$
ϵ^n
$x_{l-\frac{1}{2}}$
$x_{l+\frac{1}{2}}$
x

Figure 5.1: Configuration for the approximate Riemann solver.

This average value can be computed by integrating (5.1a) over $[x_{l+\frac{1}{2}} - \frac{\Delta\epsilon}{\rho}, x_{l+\frac{1}{2}} + \frac{\Delta\epsilon}{\rho}] \times [\epsilon^n + \Delta\epsilon, \epsilon^n]$ (in red on Fig. 5.1). It reads

$$\begin{aligned} 0 &= \int_{x_{l+\frac{1}{2}}-\frac{\Delta\epsilon}{\rho}}^{x_{l+\frac{1}{2}}+\frac{\Delta\epsilon}{\rho}} \int_{\epsilon^n}^{\epsilon^n+\Delta\epsilon} \left[\partial_\epsilon \boldsymbol{\psi} - \frac{1}{\rho}\partial_x \mathbf{F}(\boldsymbol{\psi})\right](\epsilon, x) d\epsilon dx \\ &= \frac{2\Delta\epsilon}{\rho}\left[\boldsymbol{\psi}^*_{l+\frac{1}{2}} - \frac{\boldsymbol{\psi}^n_l + \boldsymbol{\psi}^n_{l+1}}{2}\right] - \frac{\Delta\epsilon}{\rho}\left(\mathbf{F}(\boldsymbol{\psi}^n_{l+1}) - \mathbf{F}(\boldsymbol{\psi}^n_l)\right). \end{aligned}$$

This leads to write

$$\boldsymbol{\psi}^*_{l+\frac{1}{2}} = \frac{1}{2}\left[\boldsymbol{\psi}^n_{l+1} + \boldsymbol{\psi}^n_l + \left(\mathbf{F}(\boldsymbol{\psi}^n_{l+1}) - \mathbf{F}(\boldsymbol{\psi}^n_l)\right)\right]. \tag{5.7}$$

One can prove that this intermediate state is realizable using the following proposition.

Proposition 5.1 *Suppose* $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$ *and the flux* $\mathbf{F}(\boldsymbol{\psi})$ *is defined from a realizable closure, i.e. such that*

$$\exists \psi > 0 \quad s.t. \quad \boldsymbol{\psi} = \langle \mathbf{m}\psi \rangle, \quad \mathbf{F}(\boldsymbol{\psi}) = \langle \mu\mathbf{m}\psi \rangle,$$

then

$$\boldsymbol{\psi} \pm \mathbf{F}(\boldsymbol{\psi}) \in \mathcal{R}_{\mathbf{m}}.$$

Proof Define $\mathbf{V} \in \mathbb{R}^{N+1}$ such that

$$\boldsymbol{\psi} = R_{\mathbf{V}}(\mathbf{m}), \qquad \mathbf{F}(\boldsymbol{\psi}) = R_{\mathbf{V}}(\mu\mathbf{m}).$$

If $\mathbf{F}$ is computed from a realizable closure such as M_N closure, then $\mathbf{V} \in \mathcal{R}_{\mathbf{m}_{N+1}}$ is realizable, *i.e.*

$$\exists \psi \in L^1([-1,1])^+, \quad \text{s.t.} \quad \mathbf{V} = \langle \mathbf{m}_{N+1}\psi \rangle .$$

Then one has

$$\boldsymbol{\psi} \pm \mathbf{F}(\boldsymbol{\psi}) = \langle (1 \pm \mu)\mathbf{m}\psi \rangle ,$$

with $(1+\mu)\psi \in L^1([-1,1])^+$ is positive. Therefore $\boldsymbol{\psi} \pm \mathbf{F}(\boldsymbol{\psi})$ are the moments of a positive $L^1(S)^+$ function. □

Using this result, if $\boldsymbol{\psi}_l^n \in \mathcal{R}_{\mathbf{m}}$, $\boldsymbol{\psi}_{l+1}^n \in \mathcal{R}_{\mathbf{m}}$ and the fluxes $\mathbf{F}$ are defined from a M_N closure, then the intermediate state reads

$$\boldsymbol{\psi}^*_{l+\frac{1}{2}} = \left(\frac{\boldsymbol{\psi}_l^n - \mathbf{F}(\boldsymbol{\psi}_l^n)}{2} \right) + \left(\frac{\boldsymbol{\psi}_{l+1}^n + \mathbf{F}(\boldsymbol{\psi}_{l+1}^n)}{2} \right) \in \mathcal{R}_{\mathbf{m}}.$$

This is a positive combination of realizable vector, it is therefore realizable according to Property 3.1.

5.2.2 The fast characteristic problem

With those computations, one can approximate the value of $\boldsymbol{\psi}_l^{n+1}$ at new energy step.

Consider that $\boldsymbol{\psi}_l^n$ is constant in each cell $x \in [x_{l-\frac{1}{2}}, x_{l+\frac{1}{2}}]$. Approximate the solution of the Riemann problems at each interface $x_{l+\frac{1}{2}}$ with (5.7). The value of $\boldsymbol{\psi}_l^{n+1}$ is computed by the following integral over $[x_{l-\frac{1}{2}}, x_{l+\frac{1}{2}}]$ (in blue on Fig. 5.1)

$$\boldsymbol{\psi}_l^{n+1} = \frac{1}{\Delta x} \int_{x_{l-\frac{1}{2}}}^{x_{l+\frac{1}{2}}} \boldsymbol{\psi}(x, \epsilon^{n+1}) dx,$$

where the value of $\boldsymbol{\psi}(x,\epsilon^{n+1})$ is approximated by $\boldsymbol{\psi}^*_{l+\frac{1}{2}}$ in each cone $C^n_{l+\frac{1}{2}}$. This leads to

$$\begin{aligned} \boldsymbol{\psi}_l^{n+1} &= \frac{1}{\Delta x} \left[\frac{\Delta\epsilon^n}{\rho} (\boldsymbol{\psi}^*_{l-\frac{1}{2}} + \boldsymbol{\psi}^*_{l+\frac{1}{2}}) + \left(\Delta x - \frac{2\Delta\epsilon^n}{\rho} \right) \boldsymbol{\psi}_l^n \right] \\ &= \frac{1}{\Delta x} \left[\frac{\Delta\epsilon^n}{\rho} \frac{\boldsymbol{\psi}_{l-1}^n - \mathbf{F}(\boldsymbol{\psi}_{l-1}^n)}{2} + \frac{\Delta\epsilon^n}{\rho} \frac{\boldsymbol{\psi}_{l+1}^n + \mathbf{F}(\boldsymbol{\psi}_{l+1}^n)}{2} \right. \\ &\qquad \left. + \left(\Delta x - \frac{\Delta\epsilon^n}{\rho} \right) \boldsymbol{\psi}_l^n \right], \end{aligned} \tag{5.8}$$

which can be written under the finite volume form

$$\boldsymbol{\psi}_l^{n+1} = \boldsymbol{\psi}_l^n + \frac{\Delta\epsilon^n}{\rho\Delta x} \left(\mathbf{F}^n_{l+\frac{1}{2}} - \mathbf{F}^n_{l-\frac{1}{2}} \right), \tag{5.9a}$$

$$\mathbf{F}^n_{l+\frac{1}{2}} = \frac{1}{2} \left[\mathbf{F}(\boldsymbol{\psi}_{l+1}^n) + \mathbf{F}(\boldsymbol{\psi}_l^n) + (\boldsymbol{\psi}_{l+1}^n - \boldsymbol{\psi}_l^n) \right] . \tag{5.9b}$$

Remark that such an approximation is only possible if the cones $C^n_{l+\frac{1}{2}}$ emerging from all interfaces $x_{l+\frac{1}{2}}$ do not intersect each other, this means under the following condition

$$\Delta\epsilon^n \quad \leq \quad \rho\Delta x. \tag{5.10}$$

One observes through (5.8) that $\boldsymbol{\psi}^{n+1}_l$ is a positive combination of realizable vectors under the condition (5.10). Therefore this scheme preserves the realizability property from one energy step to another (see Property 3.1).

Remark 5.2 (Position of the problem) *This stability condition can be very restrictive when the density ρ has a very low value. In radiotherapy, very heterogeneous media are considered. Those media may contain very low density media, e.g. the relative density of the lungs compared to is typically around $\rho_{lung} \approx 0.3$ and the relative density of air is 10^{-3}. Such low density regions are less collisional, i.e. the particles collide less in it. However the numerical scheme requires smaller energy steps $\Delta\epsilon^n$. In the next subsections, alternatives to this approach non-constrained by stability conditions are described.*

5.2.3 Application to electrons transport equations

The approximate Riemann solver approach is only considered with an hyperbolic operator, therefore only the transport of electrons (5.2) is considered and the collisions are modeled by a CSD or a FP operator.

This method is now used for (5.2) which is considered backward in energy, *i.e.* $\epsilon^{n+1} < \epsilon^n$. For this purpose, a splitting method is used. One first solves the equiation without the term $M\boldsymbol{\psi}$ and add its influence *a posteriori*. In multi-D, the equation is solved explicitely in each direction of space and the influence of the term $M\boldsymbol{\psi}$ is added *a posteriori*.

The inhomogeneous density ρ is approximated by a constant $\rho_{l+\frac{1}{2}}$ in each dual cell $[x_l, x_{l+1}]$. This choice is motivated by two reasons: it simplifies the computation of the solution of the Riemann problem, and it leads to a conservative form of the numerical scheme (see (5.11) below).

Applying the previous computations to the hyperbolic operator of (5.2a), *i.e.* to

$$\rho\partial_\epsilon(S\boldsymbol{\psi}) - \partial_x\mathbf{F}(\boldsymbol{\psi}) = 0$$

leads to write

$$\begin{aligned} S^{n+1}\boldsymbol{\psi}^{n+\frac{1}{2}}_l &= S^n\boldsymbol{\psi}^n_l - \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{\mathbf{F}^n_{l+\frac{1}{2}}}{\rho_{l+\frac{1}{2}}} - \frac{\mathbf{F}^n_{l-\frac{1}{2}}}{\rho_{l-\frac{1}{2}}}\right), \\ \mathbf{F}^n_{l+\frac{1}{2}} &= \frac{1}{2}\left[\mathbf{F}(\boldsymbol{\psi}^n_{l+1}) + \mathbf{F}(\boldsymbol{\psi}^n_l) - (\boldsymbol{\psi}^n_{l+1} - \boldsymbol{\psi}^n_l)\right], \end{aligned} \tag{5.11}$$

where $S^n = S(\epsilon^n)$. The condition (5.10) turns into

$$\Delta\epsilon^n \quad \leq S^n\left(\frac{1}{2\Delta x}\max_l\left(\frac{1}{\rho_{l+\frac{1}{2}}} + \frac{1}{\rho_{l-\frac{1}{2}}}\right)\right)^{-1}. \tag{5.12}$$

Now one needs to add the influence of the operator $\rho M\boldsymbol{\psi}$ which is chosen to be discretized implicitly. In the end, the scheme reads

$$(S^{n+1}Id - \Delta\epsilon^n M^{n+1})\boldsymbol{\psi}_l^{n+1} \quad = \quad S^n\boldsymbol{\psi}_l^n - \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{\mathbf{F}^n_{l+\frac{1}{2}}}{\rho_{l+\frac{1}{2}}} - \frac{\mathbf{F}^n_{l-\frac{1}{2}}}{\rho_{l-\frac{1}{2}}}\right),$$

where the matrix $M^n = M(\epsilon^n)$, and it can be rewritten

$$\begin{aligned}\left(S^{n+1}Id - \Delta\epsilon^n M^{n+1}\right)\boldsymbol{\psi}_l^{n+1} \quad &= \quad \frac{1}{\Delta x}\left[\left(\Delta x - \frac{\Delta\epsilon^n}{S^n}\left(\frac{1}{\rho_{l-\frac{1}{2}}} + \frac{1}{\rho_{l+\frac{1}{2}}}\right)\right)S^n\boldsymbol{\psi}_l^n\right.\\ &+ \quad \left.\Delta\epsilon^n\left(\frac{\boldsymbol{\psi}^n_{l-1} + \mathbf{F}(\boldsymbol{\psi}^n_{l-1})}{2\rho_{l-\frac{1}{2}}} + \frac{\boldsymbol{\psi}^n_{l+1} + \mathbf{F}(\boldsymbol{\psi}^n_{l+1})}{2\rho_{l+\frac{1}{2}}}\right)\right].\end{aligned} \tag{5.13}$$

Proposition 5.2 *If the condition* (5.12) *is satisfied, then the scheme* (5.13) *preserves the realizability property from one energy step to another in the following sense.*
If

$$\forall l, \quad \forall n' < n+1, \quad \boldsymbol{\psi}_l^{n'} \in \mathcal{R}_{\mathbf{m}},$$

then

$$\forall l, \qquad \boldsymbol{\psi}_l^{n+1} \in \mathcal{R}_{\mathbf{m}}.$$

Proof

With a CSD operator: Consider that $M = (s^{n+1}_{Mott} - \sigma^{n+1}_{T,Mott})$ corresponds to the CSD operator. The scheme (5.13) can be rewritten

$$(S^{n+1}Id \quad + \quad \Delta\epsilon^n(\sigma^{n+1}_{T,Mott}Id - s^{n+1}_{Mott}))\boldsymbol{\psi}_l^{n+1} = \mathbf{R}_l^n, \tag{5.14a}$$

$$\begin{aligned}\mathbf{R}_l^n \quad = \quad &\frac{1}{\Delta x}\left[\Delta\epsilon^n\left(\frac{\boldsymbol{\psi}^n_{l-1} + \mathbf{F}(\boldsymbol{\psi}^n_{l-1})}{2\rho_{l-\frac{1}{2}}} + \frac{\boldsymbol{\psi}^n_{l+1} + \mathbf{F}(\boldsymbol{\psi}^n_{l+1})}{2\rho_{l+\frac{1}{2}}}\right)\right.\\ &\left. + \left(\Delta x - \frac{\Delta\epsilon^n}{S^n}\left(\frac{1}{\rho_{l-\frac{1}{2}}} + \frac{1}{\rho_{l+\frac{1}{2}}}\right)\right)S^n\boldsymbol{\psi}_l^n\right].\end{aligned} \tag{5.14b}$$

Using Proposition 5.1, one observes that the right-hand side $\mathbf{R}_l^n$ is a positive combination of realizable vectors, and therefore, using Property 3.1, it is realizable $\mathbf{R}_l^n \in \mathcal{R}_{\mathbf{m}}$.

Define the function

$$J(\boldsymbol{\psi}) = \frac{\mathbf{R}_l^n + \Delta\epsilon^n s^{n+1}_{Mott}\boldsymbol{\psi}}{S^{n+1} + \Delta\epsilon^n\sigma^{n+1}_{T,Mott}}.$$

The function J is a contraction because it is Lipschitz continuous with a Lipschitz constant lower than 1. Indeed, one has

$$J(\boldsymbol{\psi}_1) - J(\boldsymbol{\psi}_2) = \frac{\Delta\epsilon^n s_{Mott}^{n+1}(\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2)}{S^{n+1} + \Delta\epsilon^n \sigma_{T,Mott}^{n+1}},$$

and since $\sigma_{Mott}^i \leq \sigma_{Mott}^0 = \sigma_{T,Mott}$, the eigenvalues of the matrix

$$\frac{\Delta\epsilon^n s_{Mott}^{n+1}}{S^{n+1} + \Delta\epsilon^n \sigma_{T,Mott}^{n+1}}$$

are of norm inferior to 1. Therefore, Banach fixed point theorem provides the existence and uniqueness of a fixed point of J and the sequence

$$\boldsymbol{\psi}^{(k+1)} = J(\boldsymbol{\psi}^{(k)})$$

converges to it.

Since $\mathbf{R}_l^n \in \mathcal{R}_{\mathbf{m}}$, write $\mathbf{R}_l^n = \langle \mathbf{m} f_R \rangle$ with $f_R \in L^1([-1,1])^+$, and initialize

$$\boldsymbol{\psi}_l^{n+1,(0)} = \mathbf{R}_l^n \in \mathcal{R}_{\mathbf{m}}.$$

Now, by induction, suppose $\boldsymbol{\psi}_l^{n+1,(k)} = \left\langle \mathbf{m} f^{(k)} \right\rangle$ with $f^{(k)} \in L^1([-1,1])^+$, one has

$$\begin{aligned} \boldsymbol{\psi}_l^{n+1,(k+1)} &= J(\boldsymbol{\psi}_l^{n+1,(k)}) = \frac{\mathbf{R}_l^n + \Delta\epsilon^n s_{Mott}^{n+1} \boldsymbol{\psi}_l^{n+1,(k)}}{S^n + \Delta\epsilon^n \sigma_{T,Mott}^{n+1}} \\ &= \left\langle \mathbf{m} \frac{f_R(\mu) + \Delta\epsilon^n \displaystyle\int_{-1}^{+1} \sigma_{Mott}(\epsilon^{n+1}, \mu', \mu) f^{(k)}(\mu') d\mu'}{S^{n+1} + \Delta\epsilon^n \sigma_{T,Mott}(\epsilon^{n+1})} \right\rangle. \end{aligned}$$

Since f_R, $f^{(k)}$, σ_{Mott}^{n+1} and $\sigma_{T,Mott}^{n+1}$ are positive, $\boldsymbol{\psi}_l^{n+1,(k+1)}$ is the moment vector of a positive function.

Thus, at the limit, the fixed point $\boldsymbol{\psi}_l^{n+1}$ to (5.14a) is in the closure of the realizability domain $\mathcal{R}_{\mathbf{m}}^m$. Moreover, by definition of this fixed point, $\boldsymbol{\psi}_l^{n+1} = J(\boldsymbol{\psi}_l^{n+1})$ is a positive combination of a vector $s^{n+1}\boldsymbol{\psi}_l^{n+1}$ in the closure $\mathcal{R}_{\mathbf{m}}^m$ of $\mathcal{R}_{\mathbf{m}}$ and of a vector $\mathbf{R}_l^n$ in the interior $\mathcal{R}_{\mathbf{m}}$. Therefore such a combination $\boldsymbol{\psi}_l^{n+1} \in \mathcal{R}_{\mathbf{m}}$ is in the interior of the realizability domain.

With a FP operator: Consider now that $M = T(\epsilon) M_{FP}$ corresponds to the FP operator. The scheme (5.13) can be rewritten

$$
\begin{aligned}
(S^{n+1}Id \quad &- \quad \Delta\epsilon^n T^{n+1} M_{FP})\boldsymbol{\psi}_l^{n+1} = \mathbf{R}_l^n, && (5.15a)\\
\mathbf{R}_l^n \quad &= \quad \frac{1}{\Delta x}\left[\Delta\epsilon^n \left(\frac{\boldsymbol{\psi}_{l-1}^n + \mathbf{F}(\boldsymbol{\psi}_{l-1}^n)}{2\rho_{l-\frac{1}{2}}} + \frac{\boldsymbol{\psi}_{l+1}^n + \mathbf{F}(\boldsymbol{\psi}_{l+1}^n)}{2\rho_{l+\frac{1}{2}}}\right)\right.\\
&\quad \left. + \left(\Delta x - \frac{\Delta\epsilon^n}{S^n}\left(\frac{1}{\rho_{l-\frac{1}{2}}} + \frac{1}{\rho_{l+\frac{1}{2}}}\right)\right) S^n \boldsymbol{\psi}_l^n\right], && (5.15b)
\end{aligned}
$$

where the right-hand side $\mathbf{R}_l^n \in \mathcal{R}_{\mathbf{m}}$ is realizable as it is a sum of realizable vectors.

The equation (5.15a) is the moments of the following kinetic equation

$$
\begin{aligned}
0 \quad &= \quad S^{n+1}\boldsymbol{\psi}_l^{n+1} - \Delta\epsilon^n T^{n+1} M_{FP}\boldsymbol{\psi}_l^{n+1} - \mathbf{R}_l^n\\
&= \quad \left\langle \mathbf{m}\left(S^{n+1}\psi - \Delta\epsilon^n T^{n+1}\partial_\mu\left((1-\mu^2)\partial_\mu\psi\right) - f_R\right)\right\rangle.
\end{aligned}
$$

Consider the problem

$$
\left\{\begin{array}{rl}
S^{n+1}\psi \quad - \quad \Delta\epsilon^n T^{n+1}\partial_\mu\left((1-\mu^2)\partial_\mu\psi\right) \quad - \quad f_R & = 0\\
\partial_\mu\psi(\mu = \pm 1) & = 0
\end{array}\right. . \qquad (5.16)
$$

In the spirit of [12, 18, 19], define the function

$$
H(x) = \left\{\begin{array}{ll} x^2 & \text{if } x \leq 0\\ 0 & \text{otherwise}\end{array}\right. .
$$

Multiplying (5.16) by $H'(\psi)$ and integrating it over $\mu \in [-1,1]$, and using an integration by part leads to

$$
\left\langle S^{n+1}\psi H'(\psi)\right\rangle \; + \; \Delta\epsilon^n T^{n+1}\left\langle(1-\mu^2)(\partial_\mu\psi)^2 H''(\psi)\right\rangle \; - \; \langle f_R H'(\psi)\rangle \quad = \quad 0.
$$

Based on the definition of H, one verifies that each of those three terms is non-negative. Since their sum is zero, they are all zero. Especially, from the first term, one deduces the positivity of ψ.

Therefore, this positive function ψ which moments are $\boldsymbol{\psi}_l^{n+1}$. □

5.2.4 Application in multi-D

When considering the multi-D problem (5.1b), a splitting method is again used. One writes a Riemann solver at each interface (in each direction) and approximates its solution. Here, $X = (x, y)$ and the index m refers to the second space variable

y. In the end, in 2D the schemes reads

$$\begin{aligned}(S^{n+1}Id - \Delta\epsilon^n M^{n+1})\boldsymbol{\psi}^{n+1}_{l,m} \quad &= \quad S^n\boldsymbol{\psi}^n_{l,m} - \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{\mathbf{F}^n_{l+\frac{1}{2},m}}{\rho_{l+\frac{1}{2},m}} - \frac{\mathbf{F}^n_{l-\frac{1}{2},m}}{\rho_{l-\frac{1}{2},m}}\right)\\ &\quad -\frac{\Delta\epsilon^n}{\Delta y}\left(\frac{\mathbf{F}^n_{l,m+\frac{1}{2}}}{\rho_{l,m+\frac{1}{2}}} - \frac{\mathbf{F}^n_{l,m-\frac{1}{2}}}{\rho_{l,m-\frac{1}{2}}}\right), \qquad (5.17a)\end{aligned}$$

where the fluxes read

$$\mathbf{F}^n_{l+\frac{1}{2},m} \quad = \quad \frac{1}{2}\left[F(\boldsymbol{\psi}^n_{l+1,m})e_1 + F(\boldsymbol{\psi}^n_{l,m})e_1 - (\boldsymbol{\psi}^n_{l+1,m} - \boldsymbol{\psi}^n_{l,m})\right], \qquad (5.17b)$$

$$\mathbf{F}^n_{l,m+\frac{1}{2}} \quad = \quad \frac{1}{2}\left[F(\boldsymbol{\psi}^n_{l,m+1})e_2 + F(\boldsymbol{\psi}^n_{l,m})e_2 - (\boldsymbol{\psi}^n_{l,m+1} - \boldsymbol{\psi}^n_{l,m})\right], \qquad (5.17c)$$

and the matrix M is identical to the one of the previous subsection. This scheme is stable and preserves the realizability property under the condition

$$\Delta\epsilon^n \quad \leq \quad S^n \min_{l,m}\left[\frac{1}{2\Delta x}\left(\frac{1}{\rho_{l+\frac{1}{2},m}} + \frac{1}{\rho_{l-\frac{1}{2},m}}\right) + \frac{1}{2\Delta y}\left(\frac{1}{\rho_{l,m+\frac{1}{2}}} + \frac{1}{\rho_{l,m-\frac{1}{2}}}\right)\right]^{-1}. \qquad (5.18)$$

This method is efficient, although the problem emerging with the condition (5.12), described in Remark 5.2, is too restrictive for our medical applications. In Sections 5.4, 5.5 and 5.7, alternatives are presented. In the next section, numerical test cases are presented.

5.3 Tests on the method of moments

These first test cases are very basic test cases emerging in the field of radiation dose computation. The studied medium, commonly called phantom in the field of radiation therapy, is composed of homogeneous water, *i.e.* the relative density is fixed $\rho(x) = 1$, and one or two beams of electron are injected on the boundary of the medium. This problem is studied first in 1D, then in 2D.

Through such basic test cases, we aim to study the accuracy of the method of moments, and especially compare the M_1 and M_2 models. Only the numerical schemes (5.13) and (5.17) presented in the last section are used.

5.3.1 Single beam in 1D

In 1D, considering only the transport of electrons leads to considering the kinetic equation

$$\partial_\epsilon(S\psi) - \frac{\mu}{\rho}\partial_x\psi + T\partial_\mu\left((1-\mu^2)\partial_\mu\psi\right) = 0. \qquad (5.19)$$

Discretizing directly the 1D kinetic equation (5.19) was affordable. For this purpose, the dose results obtained with the different numerical schemes for moment equations were compared to the kinetic results obtained with the following scheme.

Using a first order explicit Euler energy discretization, an upwind scheme for the x-derivative and a midpoint quadrature rule leads to write the following scheme for the 1D kinetic equation (5.19)

$$\begin{aligned}
S^{n+1}\psi^{n+1}_{l,p} &= S^n\psi^n_{l,p} - \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{F^n_{l+\frac{1}{2},p}}{\rho_{l+\frac{1}{2}}} - \frac{F^n_{l-\frac{1}{2},p}}{\rho_{l-\frac{1}{2}}}\right) + \frac{\Delta\epsilon^n T^n}{\Delta\mu}\left(G^n_{l,p+\frac{1}{2}} - G^n_{l,p-\frac{1}{2}}\right), && \text{(5.20a)}\\
F^n_{l+\frac{1}{2},p} &= \mu^+_p\psi^n_{l,p} + \mu^-_p\psi^n_{l+1,p}, && \text{(5.20b)}\\
G^n_{l,p+\frac{1}{2}} &= \left(1 - \left(\frac{\mu_p + \mu_{p+1}}{2}\right)^2\right)\frac{(\psi^n_{l,p+1} - \psi^n_{l,p})}{\Delta\mu}, && \text{(5.20c)}
\end{aligned}$$

where $\mu^\pm = (\mu \pm |\mu|)/2$. Here the subscript p refers to the cosangle μ. This scheme is stable under the CFL condition

$$\Delta\epsilon^n \leq S^n\left(\frac{1}{\min\rho\Delta x} + \frac{2T^n}{\Delta\mu^2}\right)^{-1}. \tag{5.21}$$

The M_1 and M_2 equations extracted from (5.19) have the form (5.2) and are discretized using the numerical scheme (5.13) where the energy step $\Delta\epsilon^n$ is fixed so that

$$\Delta\epsilon^n_{M_N} = 0.95 S^n\Delta x, \tag{5.22}$$

this corresponds to the CFL condition (5.12) , while the energy step for the kinetic model is fixed by

$$\Delta\epsilon^n_{kinetic} = 0.95 S^n\left(\frac{1}{\Delta x} + \frac{2T^n}{\Delta\mu^2}\right)^{-1}. \tag{5.23}$$

The 1D medium is 6 cm long meshed with 600 cells in position, and for the kinetic model 128 cells were used for the μ variable.

A beam of $\epsilon_0 = 10$ MeV electrons is injected on the boundary of the medium, this is imposed by a boundary condition

$$\begin{aligned}
\psi^n_{0,p} &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon^n - \epsilon_0\right)^2\right)\exp\left(-\alpha_\mu\left(\mu_p - 1\right)^2\right), && \mu_p > 0, \quad \text{(5.24a)}\\
\psi^n_{l_{\max},p} &= 0, && \mu_p < 0, \quad \text{(5.24b)}
\end{aligned}$$

with $\alpha_\epsilon = 100$ and $\alpha_\mu = 1000$ for the kinetic model. For the moment models, the following condition is imposed on the boundary

$$\begin{aligned}
\boldsymbol{\psi}^n_0 &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon^n - \epsilon_0\right)^2\right)\left\langle \mathbf{m}(\mu)\exp\left(-\alpha_\mu\left(\mu - 1\right)^2\right)\right\rangle, && \text{(5.25a)}\\
\boldsymbol{\psi}^n_{l_{\max}} &= 0_{\mathbb{R}^{Card(\mathbf{m})}}. && \text{(5.25b)}
\end{aligned}$$

The initial energy is fixed at $\epsilon_{\max} = 1.2\epsilon_0$ and the final one at $\epsilon_{\min} = 10^{-3}$ MeV.

In order to compare, the dose results with the different methods are normalized by the quantity of electrons N_e injected in the medium defined for the kinetic model by

$$N_e = \sum_{p=1}^{p_{\max}} \sum_{n=1}^{n_{\max}} \left[\psi^n_{0,p} \mathbf{1}_{[0,1]}(\mu_p) + \psi^n_{l_{\max}+1,p} \mathbf{1}_{[-1,0]}(\mu_p) \right] \Delta\epsilon^n \Delta\mu,$$

and for the moment models by

$$N_e = \sum_{n=1}^{n_{\max}} \left[\psi_0^{0,n} + \psi_{l_{\max}+1}^{0,n} \right] \Delta\epsilon^n,$$

where $n_{\max}$ is the number of energy steps and $p_{\max} = 128$ is the number of angle cells.

The doses obtained with the kinetic, M_1 and M_2 models are plotted on Fig. 5.3, and the computational times are gathered in Table 5.1.

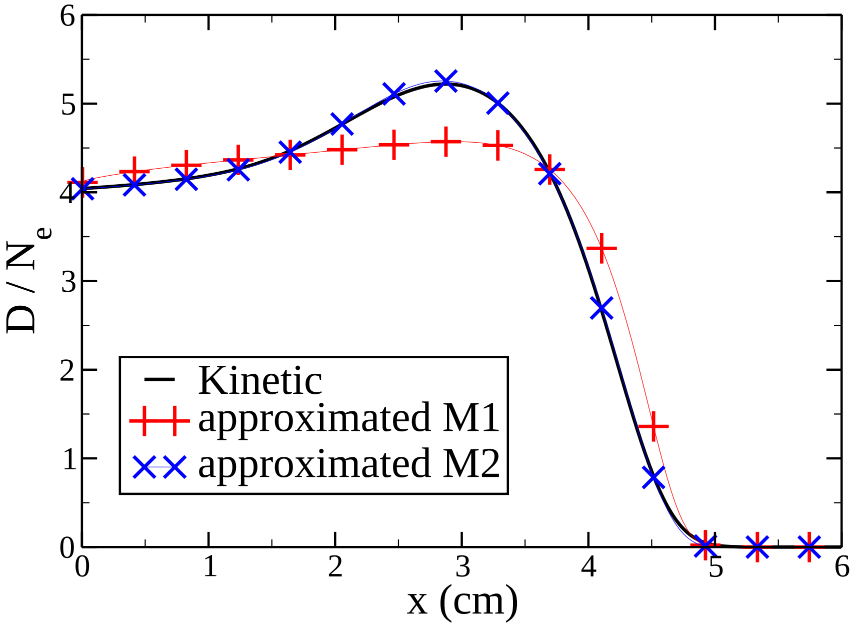

Figure 5.2: Doses obtained with the kinetic and approximated M_1 and M_2 models.

Furthermore, in 1D, it was also affordable to solve numerically the optimization problem (4.15) in order to construct the M_N closure. This optimization problem was solved numerically using the minimization routine HUMSL of MINPACK ([20]) which is calling the quadrature routine DQAGS of QUADPACK ([23]). Those routines are based on iterative methods the maximum residual of which is fixed at

Models	Kinetic	approximated M_1	approximated M_2
Computation times	32.17 sec	0.01885 sec	0.043063 sec
Number of energy steps	51294	634	634

Table 5.1: Computational times to obtain the dose results in the case of a 1D beam of 10 MeV electrons in water with the different models.

10^{-10}. This method was used because of its simplicity of implementation. For more evolved method adapted to this problem, the reader is referred *e.g.* to [16, 2, 1].

The results obtained with the approximated M_1 and M_2 closures are compared to those obtained with the closure computed by solving numerically the optimization problem (4.15) on Fig. 5.3 and the computational times are gathered in Table 5.2. Those computations were performed on one single processor. This first case shows

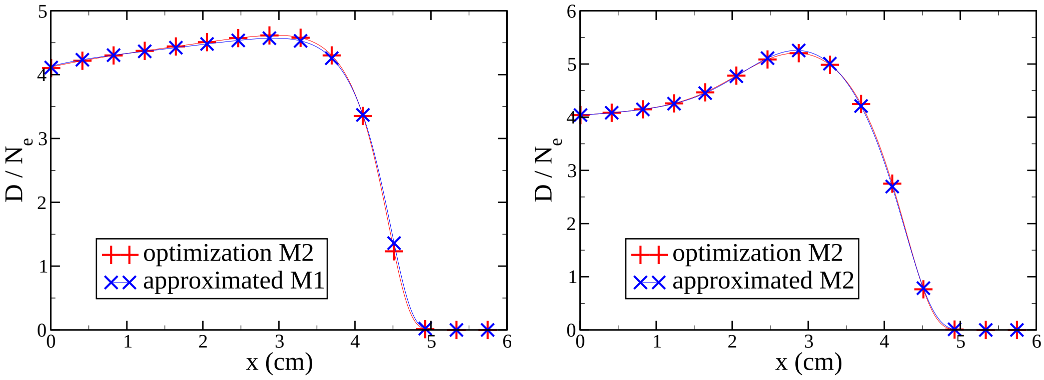

Figure 5.3: Doses obtained with the M_1 and M_2 models with the closures obtained by approximation or numerical optimization.

Closure	approximated M_1	optimization M_1
Computation times	0.01885 sec	4.107181 sec

Closure	approximated M_2	optimization M_2
Computation times	0.043063 sec	8.347469 sec

Table 5.2: Computational times to obtains the dose results in the case of a 1D beam of 10 MeV electrons in water with the different models.

that the M_N models have the right behaviour compared to the kinetic reference. Each dose obtained with the method of moments slowly raises at the entry, reaches its maximum value and drops to zero. The reference kinetic dose curve has a similar shape.

Although the dose obtained with the M_1 model is imprecise. The derivative of the dose at the entry is too high. The doses obtained with the M_2 and the kinetic models have a lower derivative at the entry and a higher maximum and they decrease faster after the maximum.

Such discrepancies are due to the approximation made when extracting the moments of the kinetic equation as described in Subsection 4.5.2. The M_2 model provides more accurate results than the M_1 model, the doses obtained with this model follows precisely the one of the kinetic model.

When comparing the dose profiles obtained with the approximated M_1 or M_2 closures to the ones obtained with a closure computed by a numerical optimization procedure, one observes very little discrepancies. Therefore, the approximations of the closures are accurate and they are assumed to be accurate enough dor the present applications.

Table 5.2 shows that the computations with the approximated M_N closures require a significantly lower computational times than the computations with a closure computed from a numerical minimization procedure (around 200 times faster). Those computations are themself faster than the kinetic computations (between 5 to 10 times faster).

5.3.2 Double beam in 1D

The non-linear effects emerging with the M_N models when stufying multiple beams crossing each other, described in Subsection 4.5.2, is studied through this test case. A beam of 10 MeV electrons is injected at both ends of a 8 cm long homogeneous water phantom. This is modeled by the following boundary conditions for the kinetic model

$$\begin{aligned}
\psi_{0,p}^n &= 10^{10} \exp\left(-\alpha_\epsilon \left(\epsilon^n - \epsilon_0\right)^2\right) \exp\left(-\alpha_\mu \left(\mu_p - 1\right)^2\right), & \mu_p > 0,\\
\psi_{l_{\max},p}^n &= 10^{10} \exp\left(-\alpha_\epsilon \left(\epsilon^n - \epsilon_0\right)^2\right) \exp\left(-\alpha_\mu \left(\mu_p + 1\right)^2\right), & \mu_p < 0,
\end{aligned}$$

with $\epsilon_0 = 10$ MeV, $\alpha_\epsilon = 100$ and $\alpha_\mu = 1000$ and for the moment models

$$\begin{aligned}
\boldsymbol{\psi}_0^n &= 10^{10} \exp\left(-\alpha_\epsilon \left(\epsilon^n - \epsilon_0\right)^2\right) \left\langle \mathbf{m}(\mu) \exp\left(-\alpha_\mu \left(\mu - 1\right)^2\right)\right\rangle,\\
\boldsymbol{\psi}_{l_{\max}}^n &= 10^{10} \exp\left(-\alpha_\epsilon \left(\epsilon^n - \epsilon_0\right)^2\right) \left\langle \mathbf{m}(\mu) \exp\left(-\alpha_\mu \left(\mu + 1\right)^2\right)\right\rangle.
\end{aligned}$$

The initial energy is fixed at $\epsilon_{\max} = 1.2\epsilon_0$ and the final one at $\epsilon_{\min} = 10^{-3}$ MeV. The domain is meshed with 800 cells in position and the energy step sizes $\Delta\epsilon^n$ are fixed by the same conditions (5.22) and (5.23) as in the previous test case. The 8 cm long water phantom is meshed with 800 cells, and for the kinetic model 128 cells are used for the μ variable in $[-1, 1]$.

The doses obtained with the kinetic, M_1 and M_2 models are plotted on Fig. 5.4, and the computational times are gathered in Table 5.3.

When using a non-linear moment model such as the M_N models on this problem, an artificial bump is observed in the middle of the medium, *i.e.* a peak of dose for

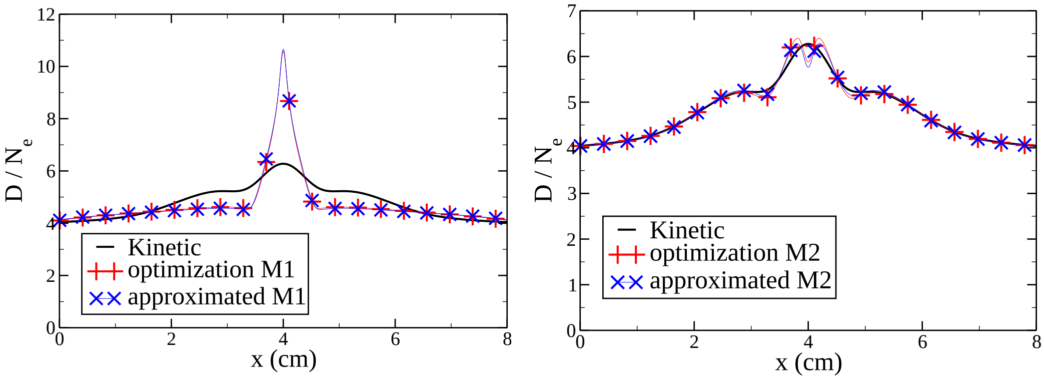

Figure 5.4: Doses obtained with the M_1 and M_2 models with the closures obtained by approximation or numerical optimization.

Models	Kinetic	approximated M_1	optimization M_1
Computation times	44.997 sec	0.0278 sec	6.064254 sec
Number of energy steps	51350	634	634

Models	approximated M_2	optimization M_2
Computation times	0.051825 sec	15.39329 sec
Number of energy steps	634	634

Table 5.3: Computational times to obtains the dose results in the case of two 1D beams of 10 MeV electrons in water with the different models.

the M_1 model and a drop for the M_2 model. This artificial phenomenum has smaller effect when the number of moments N raises. This problem was also studied *e.g.* in [16].

Although, as described in Subsection 4.5.2, the kinetic equation are linear while the M_N models are not. This effect is due to the non-linearity of the moment closure. By computing the dose created by each beam seperately and summing them, one can get rid of this artificial effect of the M_N model (see Remark 4.1). This method is afterward referred to as the double M_N models. The doses obtained using the M_N and double M_N models are compared on Fig. 5.5.

As in the previous test case, the double M_2 model is very accurate while the dose obtained with the double M_1 model is overdiffused. One observes that the artificial bump only appears in a small region in the center of the medium, *i.e.* in a 2 cm long interval for the M_1 model and a 1 cm long interval for the M_2 model. In the rest of the meidum, the M_N and double M_N models gives approximately the same dose.

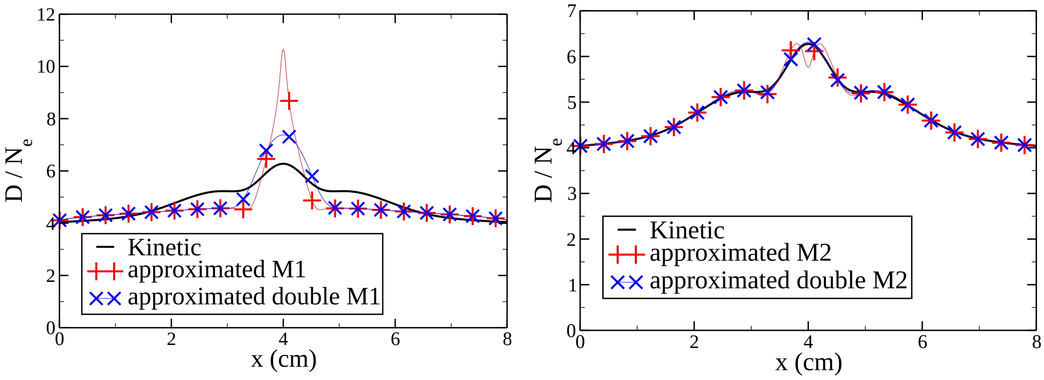

Figure 5.5: Dose obtained with the M_1, double M_1, M_2 and double M_2 models with the approximated closures.

5.3.3 Single beam in 2D

In 2D, discretizing directly the kinetic electrons transport equation, *i.e.* using a discrete ordinate method, was found too costly for our applications. Instead, the doses obtained with the moment models are compared to a reference given by the Monte Carlo code PENELOPE ([14, 5, 6]). The physics of the collisions considered in this code is more complete than the one presented in this manuscript, although in the energy range considered here, the collisions considered in Chapter 1 are predominant.

In order to compare the dose results with the Monte Carlo results, the doses are normalized by their maximum values. Such a normalized dose is commonly called percentage depth dose (PDD).

The transport of electrons in this code is based on the the CSD operator (1.30), therefore the moment equations solved here are (5.2b) where the matrix M is given by (5.3a) and it corresponds to the CSD operator.

Similarily, computing the closure by solving numerically the optimization problem (4.15) was possible in 1D and for a finite number of points, but it was found too costly for application in the numerical solver (5.13), so only the approximated closures are used.

The medium is a square of dimension 6 cm × 6 cm composed of water. The beam of 10 MeV electrons is injected on the boundary of the medium

$$\begin{aligned}
\psi(X,\epsilon,\Omega) &= 10^{10} \exp\left(-\alpha_\epsilon (\epsilon-\epsilon_0)^2\right) \exp\left(-\alpha_\mu (\Omega_1-1)^2\right) \mathbf{1}_B(X),\\
B &= \Big\{ X=(x,y), \qquad x=0 \text{ cm}, \qquad y\in[2.5 \text{ cm}, 3.5 \text{ cm}] \Big\}.
\end{aligned}$$

for all $(X,\Omega)\in\Gamma^-$, with $\alpha_\epsilon = 200$ and $\alpha_\mu = 1000$ for the kinetic model. For the moment models, the following conditions are imposed on the boundary

$$\begin{aligned}\boldsymbol{\psi}^n_{0,m} &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon^n-\epsilon_0\right)^2\right)\left\langle \mathbf{m}(\mu)\exp\left(-\alpha_\mu\left(\mu-1\right)^2\right)\right\rangle \mathbf{1}_B(X_{l,m}),\\ \boldsymbol{\psi}^n_{l,0} &= \boldsymbol{\psi}^n_{l_{\max},m} = \boldsymbol{\psi}^n_{l,m_{\max}} = 0_{\mathbb{R}^{Card(\mathbf{m})}},\end{aligned}$$

where number of cells in the first and second spatial direction are $l_{\max} = 600$ and $m_{\max} = 600$ and the energy step $\Delta\epsilon^n$ is fixed based on the CFL condition (5.18)

$$\Delta\epsilon^n = 0.95 S^n\left(\frac{1}{\Delta x}+\frac{1}{\Delta y}\right)^{-1}.$$

The initial energy is fixed at $\epsilon_{\max} = 1.2\epsilon_0$ and the final one at $\epsilon_{\min} = 10^{-3}$ MeV.

The doses obtained with the Monte Carlo solver and the M_1 and M_2 solver (5.17) are plotted on Fig. 5.6, and the computational times are gathered in Table 5.4. The solver for the moment models was parallelized and the computations were performed on 4 processors.

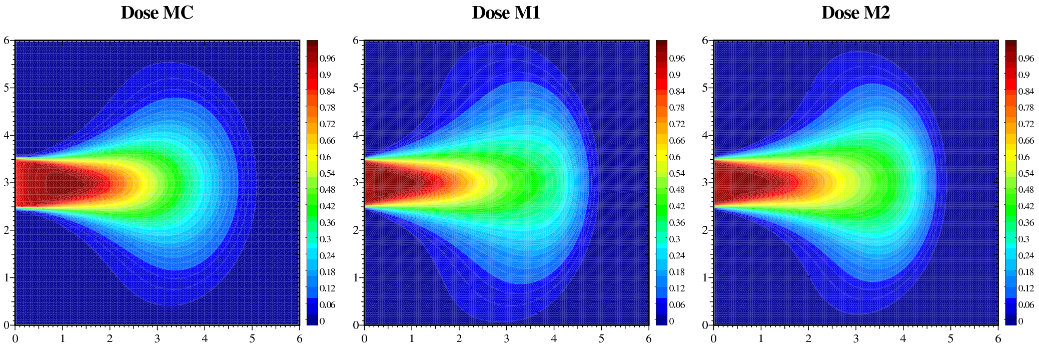

Figure 5.6: Doses obtained with the Monte Carlo solver and the approximated M_1 and M_2 models.

Models	Monte Carlo	approximated M_1	approximated M_2
Computation times	$\approx$ 10 hours	135.223 sec	510.794 sec

Table 5.4: Computational times required to obtain the dose results in the case of a 2D beam of 10 MeV electrons in water with the different solvers.

One observes a similar behavior as for the 1D case in Subsection 5.3.1. The M_1 results is slightly overdiffusive while the M_2 results are more accurate. Although the diffusive effect in multi-D is lower than in 1D.

5.3.4 Double beam in 2D

As in Subsection 5.3.4, the multi-beam instability is studied through this test case in 2D. Two orthogonal beams of 10 MeV electrons are injected on the boundary of a water phantom. This is modelled by the following boundary conditions for the kinetic model

$$\begin{aligned}
\psi(X,\epsilon,\Omega) &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon-\epsilon_0\right)^2\right)\Big[\exp\left(-\alpha_\mu\left(\Omega_1-1\right)^2\right)\mathbf{1}_{B_1}(X) \\
&\qquad\qquad + \exp\left(-\alpha_\mu\left(\Omega_2-1\right)^2\right)\mathbf{1}_{B_2}(X)\Big], \\
B_1 &= \Big\{X=(x,y), \qquad x=0 \text{ cm}, \qquad y\in[0.75 \text{ cm}, 1.25 \text{ cm}]\Big\}, \\
B_2 &= \Big\{X=(x,y), \qquad x\in[0.75 \text{ cm}, 1.25 \text{ cm}], \qquad y=0 \text{ cm}\Big\}.
\end{aligned}$$

for all $(X,\Omega)\in\Gamma^-$, with $\epsilon_0 = 10$ MeV, $\alpha_\epsilon = 200$ and $\alpha_\mu = 1000$ and for the moment models

$$\begin{aligned}
\boldsymbol{\psi}^n_{0,m} &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon^n-\epsilon_0\right)^2\right)\left\langle \mathbf{m}(\Omega)\exp\left(-\alpha_\mu\left(\Omega_1-1\right)^2\right)\right\rangle \mathbf{1}_{B_1}(X_{l,m}), \\
\boldsymbol{\psi}^n_{l,0} &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon^n-\epsilon_0\right)^2\right)\left\langle \mathbf{m}(\Omega)\exp\left(-\alpha_\mu\left(\Omega_2-1\right)^2\right)\right\rangle \mathbf{1}_{B_2}(X_{l,m}), \\
\boldsymbol{\psi}^n_{l_{\max},m} &= \boldsymbol{\psi}^n_{l,m_{\max}} = 0_{\mathbb{R}^{Card(\mathbf{m})}}.
\end{aligned}$$

The spatial domain and the energy spectrum are identical to the ones of the last test case in Subsection 5.3.3.

The dose obtained with the kinetic, M_1 and M_2 models are plotted on Fig. 5.7, and the computational times are identical to the ones of the previous test case (see Table 5.4).

As described in Subsection 4.5.2, when using the M_1 model, the two beams merge into one of direction e_1+e_2. This artificial effect appears not when considering the two beams seperately, *i.e.* when using the double M_1 model.

As in the 1D double beam case, the dose obtained with the M_2 model presents a small bump where the beams cross each other. Even if the M_2 model is non-linear, it is able to differentiate the two beams.

The comparisons with the reference Monte Carlo results show that the dose obtained with the double M_1 model is more diffused than the one with the M_2 model. The dose obtained with M_2 model is very close to the one obtained with the double M_2 model except in the small region where the beams cross each others.

5.4 Relaxation method

In order to construct inconditionnally stable schemes preserving the realizability property, a relaxation method is used. Commonly, such a method leads to construct numerical schemes having the same structures as the ones based on an approximate Riemann solver approach. However the relaxation method offers more flexibility which can be exploited to construct inconditionnally stable schemes.

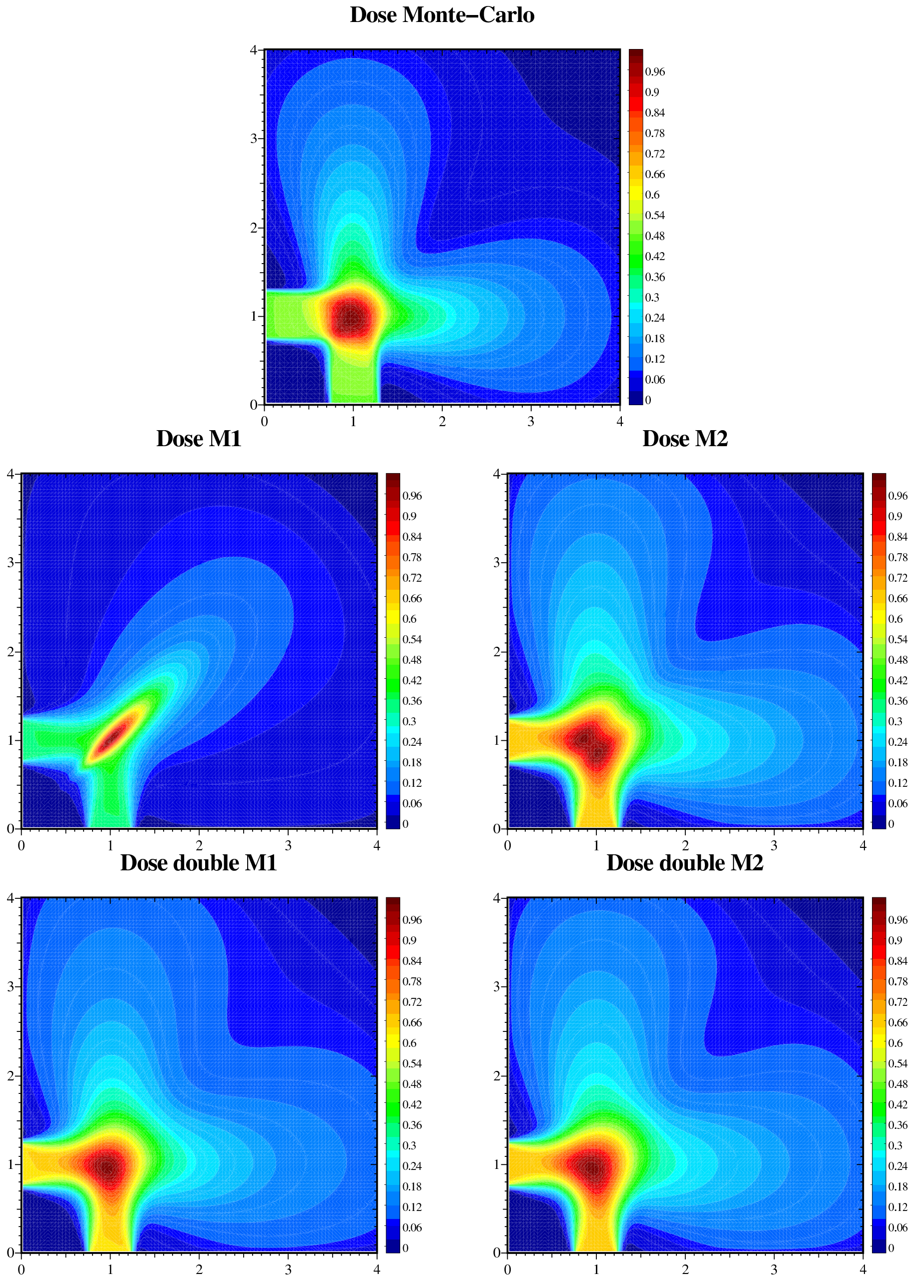

Figure 5.7: Doses obtained with the Monte Carlo solver and the approximated M_1, M_2, double M_1 and double M_2 models.

5.4.1 Relaxed equations in 1D

The relaxation method described here is based on the work of D. Aregba-Driollet and R. Natalini ([21, 3, 11, 4]). This method was originally developed for hyperbolic equations and afterward completed for parabolic ones.

The relaxation method is first described for 1D problems. Consider the 1D equation (5.1a). Chose J relaxation directions $(\lambda_j)_{j=1,...,J} \in \mathbb{R}^J$. To each relaxation direction, associate a Lipschitz continuous function $\mathbf{M}_j(\boldsymbol{\psi})$, such that

$$\sum_{j=1}^{J} \mathbf{M}_j(\boldsymbol{\psi}) = \boldsymbol{\psi}, \qquad \sum_{j=1}^{J} \lambda_j \mathbf{M}_j(\boldsymbol{\psi}) = \mathbf{F}(\boldsymbol{\psi}). \tag{5.26}$$

Such functions $\mathbf{M}_j$ are afterward referred to as Maxwellians (as it will represent an equilibrium). At this point, they are also assumed to be realizable $\mathbf{M}_j(\boldsymbol{\psi}) \in \mathcal{R}_{\mathbf{m}}$ as long as $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$.

The following system of equations is a relaxation system for (5.1a)

$$\partial_\epsilon \mathbf{f}_j^\tau - \frac{\lambda_j}{\rho} \partial_x \mathbf{f}_j^\tau = \frac{1}{\tau} \left(\mathbf{M}_j \left(\sum_{j=1}^{J} \mathbf{f}_j^\tau \right) - \mathbf{f}_j^\tau \right), \qquad \forall j = 1, ..., J, \tag{5.27}$$

where $\tau > 0$ is a relaxation parameter.

In the limit $\tau \to 0$, the solution of (5.27) corresponds to the solution of the original problem (5.1a). Indeed, formally, multiplying (5.27) by τ and having τ tends to zero leads to

$$\mathbf{f}_j^0 = \mathbf{M}_j \left(\sum_{j=1}^{J} \mathbf{f}_j^0 \right),$$

then replacing $\mathbf{f}_j^0$ by $\mathbf{M}_j(\sum_{j=1}^{J} \mathbf{f}_j^0)$ in (5.27) and summing over j reads exactly (5.1a).

This result was proven in [21] in the case of scalar hyperbolic equations under the requirement ([21, 9, 4]) that all the eigenvalues of $\partial_{\boldsymbol{\psi}} \mathbf{F}(\boldsymbol{\psi})/\rho$ are bounded by the extremal relaxation speeds, *i.e.*

$$\text{Spectrum} \left(\partial_{\boldsymbol{\psi}} \frac{\mathbf{F}(\boldsymbol{\psi})}{\rho} \right) \subset \left[\min_j \frac{\lambda_j}{\rho}, \max_j \frac{\lambda_j}{\rho} \right]. \tag{5.28}$$

For the present applications, the following choice was made

$$J = 2, \qquad \lambda_1 = 1 \text{ and } \lambda_2 = -1.$$

This leads to the following definition of the Maxwellians (which are uniquely determined by (5.31) with this choice of relaxation parameters)

$$\mathbf{M}_1(\boldsymbol{\psi}) = \frac{\boldsymbol{\psi} + \mathbf{F}(\boldsymbol{\psi})}{2}, \qquad \mathbf{M}_2(\boldsymbol{\psi}) = \frac{\boldsymbol{\psi} - \mathbf{F}(\boldsymbol{\psi})}{2}. \tag{5.29}$$

Using Lemma 5.1, one verifies that these relaxation parameters satisfy (5.28) and using Proposition 5.1, those Maxwellians are realizable when $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$ is realizable.

The idea to construct numerical schemes is to use the following splitting method.

1. *At the entry of each energy cell ϵ^n, the solution $\mathbf{f}_j$ is initialized at the value of its associated Maxwellian*

$$\mathbf{f}_1(\epsilon^n) := \mathbf{M}_1\left(\boldsymbol{\psi}(\epsilon^n)\right) \qquad \mathbf{f}_2(\epsilon^n) := \mathbf{M}_2\left(\boldsymbol{\psi}(\epsilon^n)\right).$$

2. *The homogeneous relaxed equations are solved, i.e. the part without the non-linear relaxation term*

$$\partial_\epsilon \mathbf{f}_1 - \frac{1}{\rho}\partial_x \mathbf{f}_1 = 0, \qquad \partial_\epsilon \mathbf{f}_2 + \frac{1}{\rho}\partial_x \mathbf{f}_2 = 0, \tag{5.30}$$

One obtains the intermediate values ${\mathbf{f}_1}^{n+\frac{1}{2}}$ and ${\mathbf{f}_2}^{n+\frac{1}{2}}$ from those computations.

3. *The influence of the relaxation term is added. In the limit $\tau \to 0$, this corresponds to projecting*

$$\boldsymbol{\psi}(\epsilon^{n+1}) = {\mathbf{f}_1}^{n+\frac{1}{2}} + {\mathbf{f}_2}^{n+\frac{1}{2}}. \tag{5.31}$$

Remark that the homogeneous equation (5.30) is linear which offers much flexibility in the choice of a numerical schemes to solve such an equation. In the next sections, some numerical schemes for solving (5.30) are presented.

Remark 5.3 *Using the upwind scheme on* (5.30) *and rewriting it in terms of $\boldsymbol{\psi}$, one verifies that the resulting scheme is exactly the approximate Riemann solver* (5.9) *(see also [10, 24, 17]).*

Applying the relaxation method to the particle transport equations corresponds to replacing (5.30) by

$$\partial_\epsilon(S\mathbf{f}_1) - \frac{1}{\rho}\partial_x \mathbf{f}_1 + M\mathbf{f}_1 = 0, \tag{5.32a}$$

$$\partial_\epsilon(S\mathbf{f}_2) + \frac{1}{\rho}\partial_x \mathbf{f}_2 + M\mathbf{f}_2 = 0, \tag{5.32b}$$

when considering the transport of electrons alone, or by

$$\partial_x \mathbf{f}_1 - \rho\mathbf{Q}(\mathbf{f}_1) = 0, \qquad -\partial_x \mathbf{f}_2 - \rho\mathbf{Q}(\mathbf{f}_2) = 0, \tag{5.33}$$

when considering the transport of photons and electrons together.

5.4.2 Extension to multi-D

As in 1D, choose J directions of relaxation $\lambda_j \in \mathbb{R}^3$ (for 3D problems or $\lambda_j \in \mathbb{R}^2$ in 2D), which are vectors instead of scalars in the multi-D case.

The requirement (5.28) on the λ_j turns into

$$\forall n \in S^2, \qquad \text{Spectrum}\left(\partial_{\boldsymbol{\psi}} \frac{\mathbf{F_n}(\boldsymbol{\psi})}{\rho}\right) \subset \left[\min_j \frac{\lambda_j.n}{\rho}, \max_j \frac{\lambda_j.n}{\rho}\right], \tag{5.34}$$

where $\mathbf{F_n}$ is defined in (4.7).

The 1D method did not use any 1D argument. One can rewrite the previous method with vectors λ_j instead of scalars. Then the method for solving the moment systems can be rewritten.

1. *At each energy step, initialize* $\mathbf{f}_j^n := \mathbf{M}_j^n$*, where the Maxwellians* $\mathbf{M}_j^n$ *satisfy*

$$\sum_{j=1}^{J} \mathbf{M}_j^n = \boldsymbol{\psi}^n \in \mathbb{R}^4, \qquad \sum_{j=1}^{J} \lambda_j \otimes \mathbf{M_j}^n = F(\boldsymbol{\psi}^n) \in \mathbb{R}^{3\times 4}, \tag{5.35}$$

where $\otimes$ *denotes tensorial product.*

2. *Then, compute* $\mathbf{f}_j^{n+\frac{1}{2}}$ *for each* $1 \leq j \leq J$ *by solving the homogeneous relaxed equations, i.e.*

$$\partial_\epsilon \mathbf{f}_j - \frac{\lambda_j}{\rho}.\nabla_x \mathbf{f}_j = 0, \quad 1 \leq j \leq J. \tag{5.36}$$

when considering (5.1b) *or*

$$\partial_\epsilon (S\mathbf{f}_j) - \frac{\lambda_j}{\rho}.\nabla_x \mathbf{f}_j - M\mathbf{f}_j = 0, \qquad \forall j = 1, ..., J, \tag{5.37}$$

when considering the electron transport (5.2) *or*

$$\lambda_j.\nabla_x \mathbf{f}_j - \rho \mathbf{Q}(\mathbf{f}_j) = 0, \qquad \forall j = 1, ..., J, \tag{5.38}$$

when considering the coupled photon and electron transport (5.4)

3. *Finally, update* $\boldsymbol{\psi}^{n+1} := \sum_{j=1}^{J} \mathbf{f}_j^{n+\frac{1}{2}}$.

For the sake of simplicity, only two dimensional problem are considered but the method can easily be extended to higher dimensional problems. This method was tested with the following sets of relaxation parameters:

- Cartesian relaxation

$$\text{Relaxation directions} \quad \lambda_1 = (2,0), \quad \lambda_2 = (-2,0), \tag{5.39a}$$

$$\lambda_3 = (0,2), \quad \lambda_4 = (0,-2),$$

$$\text{Associated Maxwellians} \quad \mathbf{M_i} = \frac{1}{4}\left(\boldsymbol{\psi} + \frac{\lambda_i}{|\lambda_i|}F(\boldsymbol{\psi})\right), \tag{5.39b}$$

- Diagonal relaxation

$$\text{Relaxation directions}\qquad \lambda_1 = \frac{1}{\sqrt{2}}(2,2), \quad \lambda_2 = \frac{1}{\sqrt{2}}(-2,2), \tag{5.40a}$$
$$\lambda_3 = \frac{1}{\sqrt{2}}(-2,-2), \quad \lambda_4 = \frac{1}{\sqrt{2}}(2,-2),$$
$$\text{Associated Maxwellians}\qquad \mathbf{M_i} = \frac{1}{4}\left(\boldsymbol{\psi} + \frac{\lambda_i}{|\lambda_i|}F(\boldsymbol{\psi})\right), \tag{5.40b}$$

- Star relaxation

$$\text{Relaxation directions}\qquad \lambda_1 = (4,0), \quad \lambda_2 = (0,4), \tag{5.41a}$$
$$\lambda_3 = (-4,0), \quad \lambda_4 = (0,-4),$$
$$\lambda_5 = \frac{1}{\sqrt{2}}(4,4), \quad \lambda_6 = \frac{1}{\sqrt{2}}(-4,4),$$
$$\lambda_7 = \frac{1}{\sqrt{2}}(-4,-4), \quad \lambda_8 = \frac{1}{\sqrt{2}}(4,-4),$$
$$\text{Associated Maxwellians}\qquad \mathbf{M_i} = \frac{1}{8}\left(\boldsymbol{\psi} + \frac{\lambda_i}{|\lambda_i|}F(\boldsymbol{\psi})\right). \tag{5.41b}$$

Remark that those relaxation directions λ_j satisfy the condition (5.34) and their associated Maxwellians $\mathbf{M_j}$ are realizable as long $\boldsymbol{\psi} \in \mathcal{R}_{\mathbf{m}}$ according to Property 3.1.

5.5 An explicit Finite Difference (FD) scheme for linear equations

This section is devoted to constructing unconditionally stable numerical schemes to solve the (linear) homogeneous relaxed equations (5.30), (5.32), (5.33), (5.36), (5.37) and (5.38).

The first method, presented in Subsection 5.5.1, is based on the method of characteristics. Similarily as the approximate Riemann solver, this method can only be applied when the considered equations contains an hyperbolic operator. Therefore they are only applied to the transport of electrons when considering whether a CSD (2.9) or a FP (2.10) collision term. Another alternative, *i.e.* an implicit solver, is presented in Section 5.7, and it can be applied when considering more general collision operators.

5.5.1 A FD scheme for the 1D toy equations

For the sake of simplicity, the numerical scheme is presented on the scalar equation

$$\partial_\epsilon \psi - \frac{1}{\rho(x)}\partial_x \psi = 0, \tag{5.42}$$

which corresponds to chosing $\lambda_j = 1$ in (5.30). The other cases, *i.e.* $\lambda = -1$ or $\lambda \in \mathbb{R}$, can be treated similarily. Remark also that the equations composing the system (5.30) are independent, therefore one can construct a numerical scheme in the scalar case (5.42) and apply this method to each equation of the system (5.30) separately.

The density ρ is considered constant equal $\rho_{l+\frac{1}{2}}$ in each dual cell $[x_l, x_{l+1}]$.

The method of characteristic

Using the method of characteristics, ψ is constant along the characteristic curves (see the configuration on Fig. 5.8)

$$\frac{d}{d\epsilon}\psi\left(\epsilon, y(\epsilon, e^0, x)\right) = 0, \tag{5.43}$$

$$\frac{d}{d\epsilon}y(\epsilon, e^0, x) = -\frac{1}{\rho(x)}, \qquad y(e^0, e^0, x) = x. \tag{5.44}$$

The characteristic curve the foot of which is x_l in e^0 reads

$$y(\epsilon, e^0, x_l) = x_l - \frac{\epsilon - e^0}{\rho_{l-\frac{1}{2}}}, \tag{5.45}$$

if the characteristic curve y reaches not the point x_{l+1} in the interval $]\epsilon, e^0]$, *i.e.* if

$$|\epsilon - e^0| \leq \rho_{l-\frac{1}{2}}|y(\epsilon, e^0, x_l) - x_l|. \tag{5.46}$$

The relation (5.45) can easily be inverted

$$y(\epsilon, e^0, x) = z \qquad \Rightarrow \qquad x(\epsilon, e^0, z) = z + \frac{\epsilon - e^0}{\rho_{l+\frac{1}{2}}},$$

A convex combination

With those computations, one obtains

$$\psi_l^{n+1} = \psi\left(\epsilon^{n+1}, x_l\right) = \psi\left(\epsilon^n, x_l + \frac{\epsilon^{n+1} - \epsilon^n}{\rho_{l+\frac{1}{2}}}\right).$$

Approximating $\psi\left(\epsilon^n, x\right)$ by a piecewise affine function having ψ_l^n for value at each node x_l leads to the well-known upwind scheme

$$\psi_l^{n+1} = \left(1 - \frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x}\right)\psi_l^n + \frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x}\psi_{l+1}^n,$$

and the condition (5.46) corresponds to the Courant-Friedrichs-Lewy (CFL) condition

$$\forall l, \qquad \frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x} \leq 1.$$

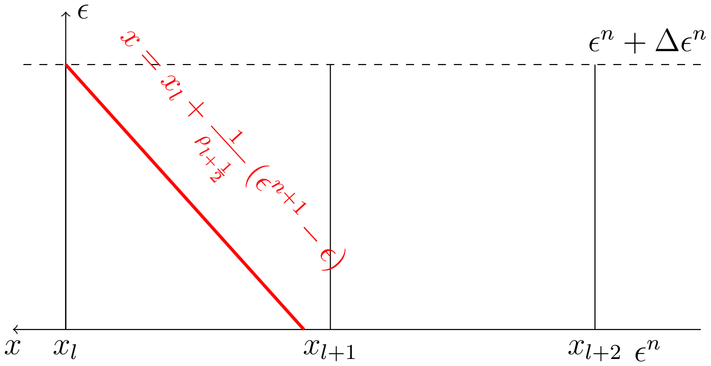
ϵ
$\epsilon^n + \Delta\epsilon^n$
$x = x_l + \frac{1}{\rho_{l+\frac{1}{2}}}(\epsilon^{n+1} - \epsilon)$
x
x_l
x_{l+1}
x_{l+2}
ϵ^n

Figure 5.8: Characteristic curves when $\Delta\epsilon^n \leq \rho_{l+\frac{1}{2}}\Delta x$.

An unconditionally stable scheme

Using the same method, one can construct a numerical scheme non constrained by the condition (5.46). Consider that the characteristic curve (5.45) can cross more than one cell (see configuration on Fig. 5.9). Equation (5.45) is modified into

$$y(\epsilon, e^0, x_l) = x_l - \sum_{k=0}^{K_l^- -1}\left(\frac{e^{k+1} - e^k}{\rho_{l-k-\frac{1}{2}}}\right) - \frac{\epsilon - e^{K_l^-}}{\rho_{l-K_l^- -\frac{1}{2}}} \tag{5.47}$$

where the scalars e^k and the integer K_l^- are such that

$$\forall k = 0, ..., K_l^- - 1, \qquad (e^{k+1} - e^k) = \rho_{l-k-\frac{1}{2}}\Delta x,$$
$$\sum_{k=0}^{K_l^- -1} \rho_{l-k-\frac{1}{2}}\Delta x \leq \epsilon - e^0 \leq \sum_{k=0}^{K_l^-} \rho_{l-k-\frac{1}{2}}\Delta x.$$

Inverting (5.47) reads

$$y(\epsilon, e^0, x) = z \quad \Rightarrow \quad x(\epsilon, e^0, z) = z + \sum_{j=0}^{K_l^+ -1}\left(\frac{e^{j+1} - e^j}{\rho_{l+j+\frac{1}{2}}}\right) + \frac{\epsilon - e^{K_l^+}}{\rho_{l+K_l^+ +\frac{1}{2}}} \tag{5.48}$$

where the scalars e^k and the integer K_l^+ are such that

$$\forall j = 0, ..., K_l^+ - 1, \qquad (e^{j+1} - e^j) = \rho_{l+j+\frac{1}{2}}\Delta x,$$
$$\sum_{j=0}^{K_l^+ -1} \rho_{l+j+\frac{1}{2}}\Delta x \leq \epsilon - e^0 \leq \sum_{j=0}^{K_l^+} \rho_{l+j+\frac{1}{2}}\Delta x.$$

One obtains

$$\begin{aligned}\psi_l^{n+1} = \psi\left(\epsilon^{n+1}, x_l\right) &= \psi\left(\epsilon^n,\ x_l + \sum_{j=0}^{K_l^+-1}\left(\frac{e^{j+1}-e^j}{\rho_{l+j+\frac{1}{2}}}\right) + \frac{\epsilon^{n+1}-e^{K_l^+}}{\rho_{l+K_l^++\frac{1}{2}}}\right)\\ &= \psi\left(\epsilon^n,\ x_{l+K_l^+-1} + \frac{\epsilon^{n+1}-e^{K_l^+}}{\rho_{l+K_l^++\frac{1}{2}}}\right).\end{aligned}$$

Approximating $\psi\left(\epsilon^n, x\right)$ by a piecewise affine function having ψ_l^n for value at each node x_l leads to

$$\psi_l^{n+1} = (1-\alpha_l)\psi^n_{l+K_l^+} + \alpha_l\psi^n_{l+K_l^++1}, \qquad \alpha_l = \frac{(\epsilon^{n+1}-e^{K_l^+})}{\rho_{l+K_l^++\frac{1}{2}}}. \tag{5.49}$$

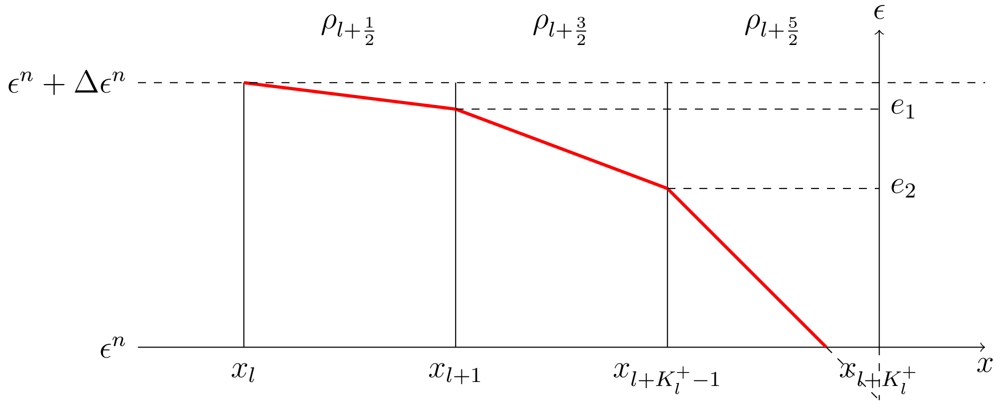

Figure 5.9: Configuration for the unconditionally Finite Difference scheme with $K_l^+ = 3$

Property 5.1

- *If the characteristic curves do not cross more than one cell, this scheme is equivalent to the original upwind scheme. This corresponds to the case where the common CFL condition (5.46) is satisfied.*
- *One verifies that the consistency error is of order 1 in Δx and in $\Delta\epsilon^n$.*
- *In the end, the scheme (5.49) expresses ψ_l^{n+1} as a convex combination of $\psi_{l'}^n$ for different l'. Therefore it is monotone, Total Variation (TV) stable and it preserves the realizability (according to Property 3.1).*

Remark 5.4 *There are no stability restrictions on this scheme, so it is more stable than the common upwind scheme. However, the precision of this scheme when extending the stencil, i.e. when $K_l^+ > 1$, is lower than the one of the upwind scheme the common CFL restriction.*

5.5.2 Extension to multi-D

The 2D linear transport equation read

$$\partial_\epsilon \psi - \frac{\lambda}{\rho}.\nabla_x \psi = 0.$$

Again, one can construct a FD scheme by following the characteristic curves (in 2D). Given a cell $C_{l+\frac{1}{2},m+\frac{1}{2}} = [x_l, x_{l+1}] \times [y_m, y_{m+1}]$ which center is $X_{l,m} = (x_l, y_m)$, one can find the origin $X_c = (x_c, y_c)$ of the characteristic which passes through $X_{l,m}$ at energy $\epsilon^n + \Delta\epsilon^n$ (see configuration on Fig. 5.10). From this, one defines a Finite Difference scheme by approximating the value of $\psi(\epsilon^n, X_c)$ using the values $\psi(\epsilon^n, X_{l',m'})$ at the nearest cell centers $X_{l',m'}$ around X_c. This reads

$$\begin{aligned} \psi_{l,m}^{n+1} &= \psi(\epsilon^{n+1}, X_{l,m}) = \psi(X_c, t^n) \\ &\approx \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{|x_c - x_{l'+i}|}{|x_{l'+1} - x_{l'}|} \frac{|y_c - y_{m'+j}|}{|y_{m'+1} - y_{m'}|} \psi_{l'+i,m'+j}^n, \end{aligned} \tag{5.50}$$

if $X_c \in C_{l'+\frac{1}{2},m'+\frac{1}{2}}$.

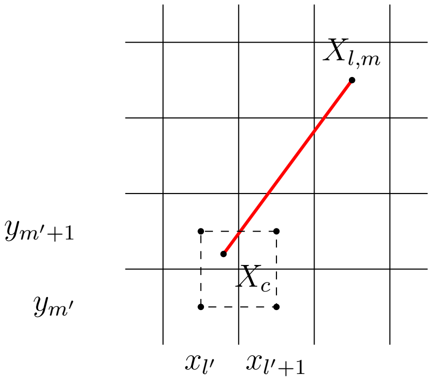

Figure 5.10: A characteristic curve in two dimensions. Configuration for the 2D Finite Difference scheme.

5.5.3 Application to electrons transport

For the relaxed electron transport equations (5.32), one can use a splitting method and solve first the homogeneous equation and add the remaining terms implicitly afterward. This reads

$$(S^{n+1}Id - \Delta\epsilon^n M^{n+1})\mathbf{f_1}_l^{n+1} = S^n\left[(1-\alpha_l^-)\mathbf{f_1}_{l-K_l^-}^n + \alpha_l^-\mathbf{f_1}_{l-K_l^--1}^n\right], \tag{5.51a}$$

$$(S^{n+1}Id - \Delta\epsilon^n M^{n+1})\mathbf{f_2}_l^{n+1} = S^n\left[(1-\alpha_l^+)\mathbf{f_2}_{l+K_l^+}^n + \alpha_l^+\mathbf{f_2}_{l+K_l^++1}^n\right], \tag{5.51b}$$

$$\alpha_l^- = \frac{(\epsilon^n - \epsilon^-)}{\rho_{l-K_l^--\frac{1}{2}}}, \qquad \alpha_l^+ = \frac{(\epsilon^n - \epsilon^+)}{\rho_{l+K_l^++\frac{1}{2}}}, \tag{5.51c}$$

where the scalars ϵ^- and ϵ^+ and the integer K_l^- and K_l^+ are determined by

$$\epsilon^- = \epsilon^n - \sum_{k=0}^{K_l^--1}\rho_{l-k-\frac{1}{2}}\Delta x, \qquad \epsilon^+ = \epsilon^n - \sum_{k=0}^{K_l^+-1}\rho_{l+k+\frac{1}{2}}\Delta x, \tag{5.51d}$$

$$\sum_{k=0}^{K_l^--1}\rho_{l-k-\frac{1}{2}}\Delta x \le \Delta\epsilon^n \le \sum_{k=0}^{K_l^-}\rho_{l-k-\frac{1}{2}}\Delta x, \quad \sum_{k=0}^{K_l^+-1}\rho_{l+k+\frac{1}{2}}\Delta x \le \Delta\epsilon^n \le \sum_{k=0}^{K_l^+}\rho_{l+k+\frac{1}{2}}\Delta x. \tag{5.51e}$$

When applied to the 2D problem (5.2), this reads

$$(S^{n+1}Id - \Delta\epsilon^n M^{n+1})\mathbf{f_i}_{l,m}^{n+1} = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{|x_c - x_{l'+i}|}{|x_{l'+1} - x_{l'}|}\frac{|y_c - y_{m'+j}|}{|y_{m'+1} - y_{m'}|}S^n\psi_{l'+i,m'+j}^n \tag{5.52}$$

when the foot of the characteristic X_c is in the cell $C_{l'+\frac{1}{2},m'+\frac{1}{2}}$.

Proposition 5.3 *The schemes (5.51) and (5.52) preserve the realizability property from one energy step to another.*

Proof The proof is identical to the one of Proposition 5.2. □

5.6 Tests with fast characteristics

The aim of this section is to show the efficiency of the explicit solvers (5.51) in 1D and (5.52) in 2D when the medium contains low density regions. Therefore those numerical schemes are compared to the reference approximate Riemann solvers (5.13) and (5.17) for moment equations, which equivals to the schemes (5.51) and (5.52) with the restrictive condition (5.12). As the purpose of this section is only to test the accuracy of the numerical methods for fast characteristics, kinetic (in 1D) or Monte Carlo (in 2D) results are provided only as indication.

The accuracy of the approximation of the moment closures were tested in the last section. Here, in order to accelerate the computations, only the approximated closures are used.

5.6.1 In a 1D medium containing air

For this test case, the domain is chosen to be 12 cm long. The density of th medium is chosen to be

$$\begin{aligned}\rho(x) \quad &= \quad 10^{-3}\Big(\mathbf{1}_{[0\text{ cm},2\text{ cm}]\cup[4\text{ cm},6\text{ cm}]\cup[8\text{ cm},10\text{ cm}]}(x) \\ &\qquad +\mathbf{1}_{[2\text{ cm},4\text{ cm}]\cup[6\text{ cm},8\text{ cm}]\cup[10\text{ cm},12\text{ cm}]}(x)\Big),\end{aligned}$$

which corresponds to a medium composed of 2 cm wide slabs of alternatively air ($\rho_{air} = 10^{-3}$) or water ($\rho_{water} = 1$).

A beam identical to the one of Subsection 5.3 is injected on the boundary of the medium. This corresponds to fixing the boundary conditions (5.24) for the kinetic model or (5.25) for the moment models.

The kinetic equation (5.19) is solved with the numerical scheme (5.20). The results with the explicit solver (5.51) (with different energy step sizes) for moment equations are compared on this test case.

The medium is uniformly meshed with 1200 cells, and 128 cells in μ were used for the kinetic model. The energy step is fixed by the condition (5.21) for the kinetic model. For the moment models, the results with two different energy step sizes. The first energy step size is

$$\Delta\epsilon^n_{air} = 0.95\rho_{air}S^n\frac{\Delta x}{\max\limits_j|\lambda_j|}. \tag{5.53}$$

It corresponds to (5.12). In that case, the explicit numerical scheme (5.51) is equivalent to the approximate Riemann solver (5.13) as the characteristic curves do not cross more than one cell (see Property 5.1). The second energy step size is

$$\Delta\epsilon^n_{water} = 0.95\rho_{water}S^n\frac{\Delta x}{\max\limits_j|\lambda_j|}. \tag{5.54}$$

The explicit numerical schemes (5.51) can be used without any condition on the energy step size so fixing a fine or a coarse energy step will only affect the accuracy and the computational time of the method.

The doses obtained with the kinetic scheme (5.20) and the M_1 and M_2 models with the explicit (5.51) numerical scheme are plotted on Fig. 5.11 and the computational times are gathered in Table 5.5.

The results with larger energy steps $\Delta\epsilon^n = \Delta\epsilon^n_{water}$ show good agreements with those with fine ones $\Delta\epsilon^n = \Delta\epsilon^n_{air}$.

When using fine energy steps, one needs around 1000 times more steps than when using the coarse steps because

$$\frac{\Delta\epsilon^n_{water}}{\Delta\epsilon^n_{air}} = 10^3.$$

This leads to a large difference in the computational times, the explicit solver is around 900 times faster with coarse steps than with fine ones.

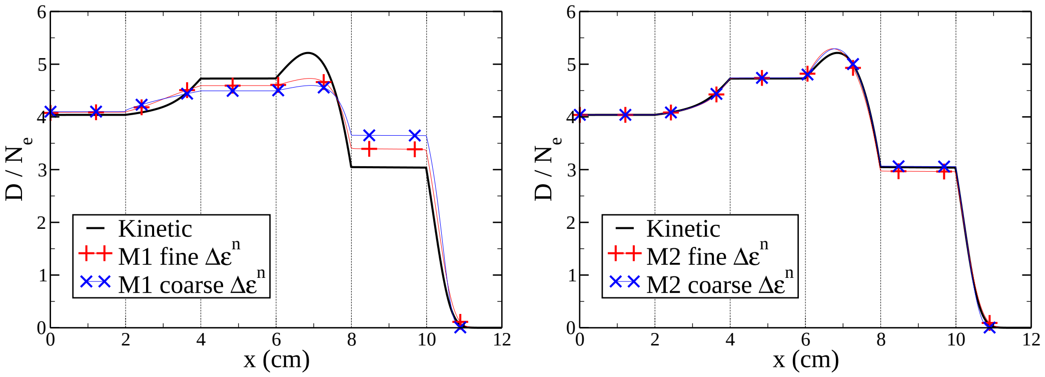

Figure 5.11: Doses obtained with the kinetic solver and the M_1 and M_2 solvers with fine and coarse energy steps $\Delta\epsilon^n$.

Model and energy step	Kinetic	M_1 with $\Delta\epsilon^n_{air}$	M_1 with $\Delta\epsilon^n_{water}$
Computation times	860.685 sec	19.041089 sec	0.073788 sec
Number of energy steps	683219	632468	634

Model and energy step	M_2 with $\Delta\epsilon^n_{air}$	M_2 with $\Delta\epsilon^n_{water}$
Computation times	64.149026 sec	0.136977 sec
Number of energy steps	632468	634

Table 5.5: Computational times with the kinetic solver (5.20), and the explicit solver (5.51) for the M_1 and M_2 models with a fine $\Delta\epsilon^n_{air}$ and a coarse $\Delta\epsilon^n_{water}$ energy step size.

5.6.2 In a 2D cut of a chest

The density map presented in the last test case contained extremely low density regions ($\rho_{air} = 10^{-3}$). The 2D density map in the present test case corresponds to a cut of a human chest. The aim of this test case is to show that the numerical schemes are efficient for more practical applications.

The domain is of size 22.3 cm × 29.5 cm, meshed with 223 × 295 cells.

A beam is imposed on the boundary. It is modeled by the following boundary condition

$$\begin{aligned} \psi(X,\epsilon,\Omega) &= 10^{10}\exp\left(-\alpha_\epsilon\,(\epsilon-\epsilon_0)^2\right)\exp\left(-\alpha_\mu\,(\Omega_1+1)^2\right)\mathbf{1}_B(X), \\ B &= \Big\{X=(x,y), \qquad x=0\text{ cm}, \qquad y\in[18\text{ cm},20\text{ cm}]\Big\}. \end{aligned}$$

for all $(X,\Omega)\in\Gamma^-$, with $\epsilon_0 = 10$ MeV, $\alpha_\epsilon = 200$ and $\alpha_\mu = 1000$ and for the moment models

$$\begin{aligned}\boldsymbol{\psi}^n_{l_{\max},m} &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon^n-\epsilon_0\right)^2\right)\left\langle \mathbf{m}(\Omega)\exp\left(-\alpha_\mu\left(\Omega_1+1\right)^2\right)\right\rangle \mathbf{1}_B(X_{l,m}),\\ \boldsymbol{\psi}^n_{l,0} &= \boldsymbol{\psi}^n_{0,m}=\boldsymbol{\psi}^n_{l,m_{\max}} = 0_{\mathbb{R}^{Card(\mathbf{m})}}.\end{aligned}$$

This case is meant to test the relaxation method and the numerical schemes for the relaxed equations. Especially the influence of the choice of the relaxation parameter, *i.e.* with the caretesian, diagonal or star relaxation directions, is tested on this problem.

The dose results obtained with coarse energy steps $\Delta\epsilon_{water}$ (5.54) using the explicit scheme are compared to the result with fine energy steps $\Delta\epsilon_{air}$ (5.53).

The isodose curves of 5% (red), 10% (orange), 25% (green), 50% (light blue), 70% (dark blue) and 80% (violet) of the maximum dose are plotted on Fig. 5.12 over the density map in grayscale. The computational times for this test case are gathered in Table 5.6.

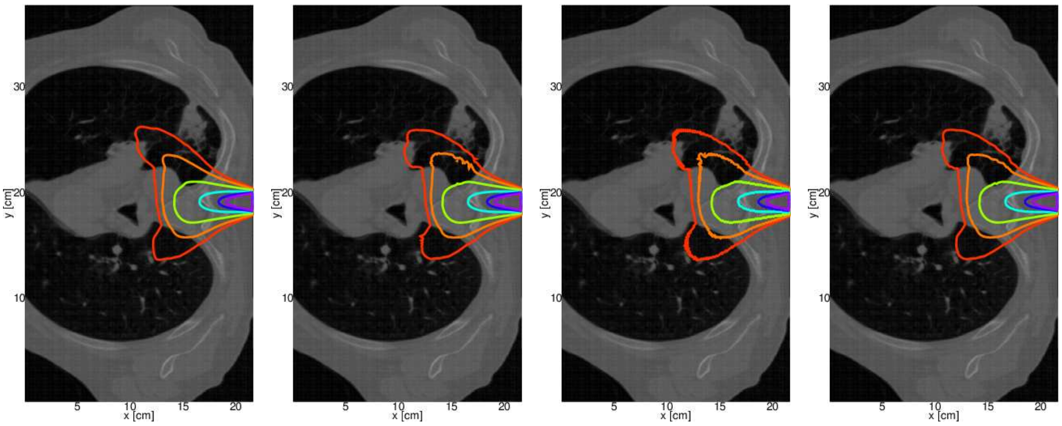

Figure 5.12: Isodose curves in a chest at 5% (red), 10% (yellow), 25% (green), 50% (cyan) and 80% (blue) of the maximum dose with a fine $\Delta\epsilon^n$ (left) and a coarse $\Delta\epsilon^n$ using cartesian (middle left), diagonal (middle right) and star (right) directions of relaxation.

The isocurves of absolute error compared to the explicit scheme with a fine energy step $\Delta\epsilon^n_{air}$ at 0.2% (yellow), 0.5% (light blue), 1% (red) are plotted on Fig. 5.13.

The dose obtained with the explicit scheme with coarse energy steps $\Delta\epsilon^n_{water}$ with cartesian directions of relaxation with the approximated M_1 and M_2 models are compared to a reference Monte Carlo results of Fig. 5.14, and the computational times are gathered in Table 5.14.

The shape of the dose obtained with the different relaxation parameters with a coarse $\Delta\epsilon^n_{water}$ are close to the one obtained with the cartesian relaxation parameters when using fine $\Delta\epsilon^n_{air}$. The absolute error is smaller than 5.3140% of the maximum dose when using the cartesian directions of relaxation, than 12.702% with the diagonal directions, and than 3.4072% with the star directions. The maximum errors are

Relaxation directions and $\Delta\epsilon^n$	Cartesian, $\Delta\epsilon^n_{air}$	Cartesian, $\Delta\epsilon^n_{water}$
Computation times	5939.6982 sec	20.8869 sec
Number of energy steps	247930	885

Relaxation directions and $\Delta\epsilon^n$	Diagonal, $\Delta\epsilon^n_{water}$	Star, $\Delta\epsilon^n_{water}$
Computation times	22.6130 sec	61.2220 sec
Number of energy steps	885	1768

Table 5.6: Computational times with the explicit solver (5.51) with the different relaxation directions and the different energy step sizes.

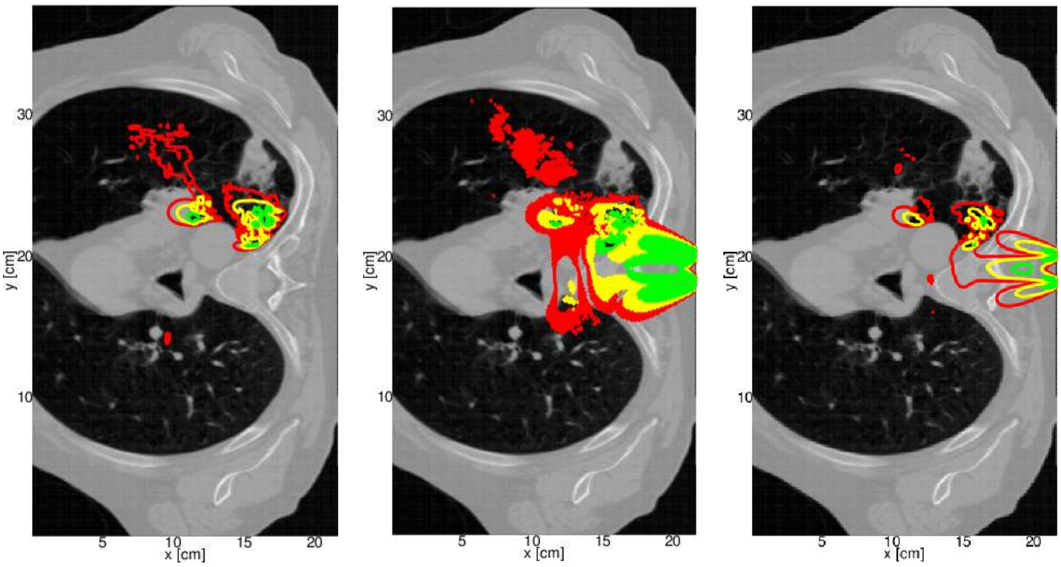

Figure 5.13: Isodose curves of absolute error compared to the dose obtained with explicit scheme with fine energy steps $\Delta\epsilon^n_{air}$ in a chest at 0.5% (red), 1% (yellow) and 2% (green) of the maximum dose with coarse energy steps $\Delta\epsilon^n$ with cartesian (left), diagonal (middle) and star (right) directions of relaxation.

Solver	Monte Carlo	M_1 model	M_2 model
Computation times	14 hours	20.8869 sec	74.3470 sec

Table 5.7: Computational times with the Monte Carlo solver, and the explicit solver (5.51) for the M_1 and M_2 models using the Cartesian directions of relaxation.

located in the middle of the medium at about 2 cm and 6 cm depth and on each side of the beam at the entry of the low density regions (lungs). When using the diagonal

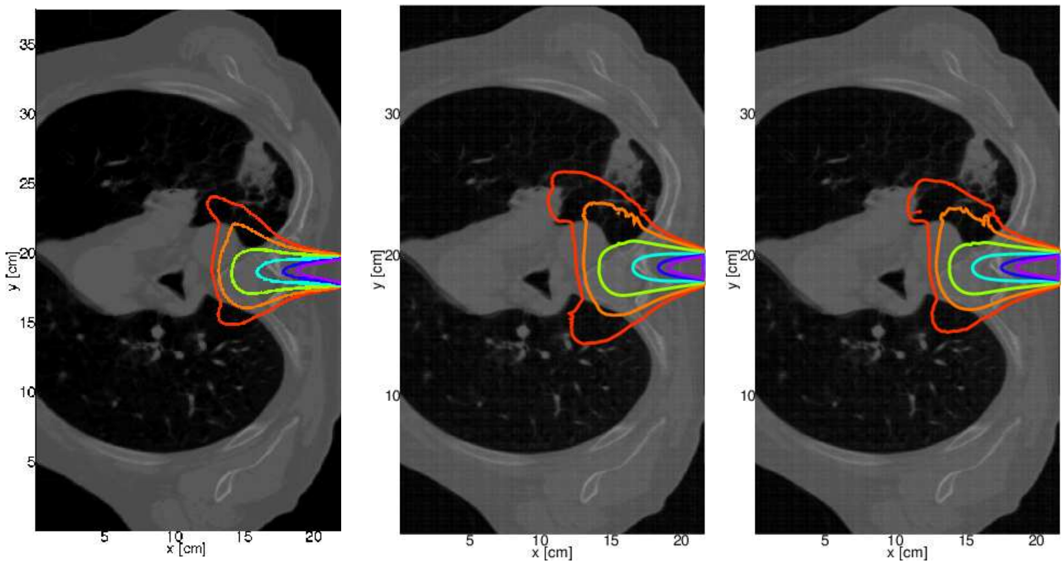

Figure 5.14: Isodose curves in a chest at 5% (red), 10% (yellow), 25% (green), 50% (cyan) and 80% (blue) of the maximum dose with a Monte Carlo solver (left), with the explicit scheme with coarse energy steps $\Delta\epsilon^n$ for the M_1 model (middle) and M_2 model (right) with cartesian directions of relaxations.

directions of relaxation, the information is transported in diagonal direction. Thus, when transporting particles along the x-axis, the scheme does not transport them from one cell to its neighboor. This results in some irregularities which can be seen in Fig. 5.13. The relaxed models are better when the directions of relaxation are collinear to the mesh directions (*i.e.* cartesian directions).

Using a the numerical schemes with coarse energy steps provides a significant speed-up factor compared to the reference method and it provides a relatively good accuracy compared to those reference results.

The Cartesian directions of relaxation provides more accurate results than the diagonal and the star directions of relaxation. Therefore, Cartesian directions of relaxations are chosen for the next 2D tests.

5.7 An implicit scheme for linear equations

The last approach consists in treating implicitly the non-linear flux term in the transport equations.

5.7.1 For the toy problem

On the homogeneous relaxed equations (5.30), such a scheme simply reads

$$\left(1+\frac{\Delta\epsilon^n}{\rho_{l-\frac{1}{2}}\Delta x}\right)(\mathbf{f_1}^{\tau})_l^{n+1} = (\mathbf{f_1}^{\tau})_l^{n} + \frac{\Delta\epsilon^n}{\rho_{l-\frac{1}{2}}\Delta x}(\mathbf{f}_1^{\tau})_{l-1}^{n+1} = 0,$$

$$\left(1+\frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x}\right)(\mathbf{f_2}^{\tau})_l^{n+1} = (\mathbf{f_2}^{\tau})_l^{n} + \frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x}(\mathbf{f}_2^{\tau})_{l+1}^{n+1} = 0,$$

which can be obtained *e.g.* by a Finite Difference approach. When applying this scheme to (5.30) in the relaxation method, this scheme leads to write the following scheme for $\boldsymbol{\psi}$

$$\boldsymbol{\psi}_l^{n+1} - \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{\mathbf{F}^{n+1}_{l+\frac{1}{2}}}{\rho_{l+\frac{1}{2}}} - \frac{\mathbf{F}^{n+1}_{l-\frac{1}{2}}}{\rho_{l-\frac{1}{2}}}\right) = \boldsymbol{\psi}_l^n, \tag{5.55a}$$

$$\mathbf{F}^{n+1}_{l+\frac{1}{2}} = \frac{1}{2}\left[\mathbf{F}(\boldsymbol{\psi}^{n+1}_{l+1}) + \mathbf{F}(\boldsymbol{\psi}^{n+1}_{l}) + (\boldsymbol{\psi}^{n+1}_{l+1} - \boldsymbol{\psi}^{n+1}_{l})\right]. \tag{5.55b}$$

which equivals to rewriting (5.9) with implicit fluxes.

5.7.2 For the transport equations

When applied to the transport equations of electrons (5.2), this scheme reads

$$\left(S^{n+1}Id - \Delta\epsilon^n M^{n+1}\right)\boldsymbol{\psi}_l^{n+1} + \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{\mathbf{F}^{n+1}_{l+\frac{1}{2}}}{\rho_{l+\frac{1}{2}}} - \frac{\mathbf{F}^{n+1}_{l-\frac{1}{2}}}{\rho_{l-\frac{1}{2}}}\right) = S^n\boldsymbol{\psi}_l^n, \tag{5.56a}$$

$$\mathbf{F}^{n+1}_{l+\frac{1}{2}} = \frac{1}{2}\left[\mathbf{F}(\boldsymbol{\psi}^{n+1}_{l+1}) + \mathbf{F}(\boldsymbol{\psi}^{n+1}_{l}) - (\boldsymbol{\psi}^{n+1}_{l+1} - \boldsymbol{\psi}^{n+1}_{l})\right], \tag{5.56b}$$

Such a discretization can also be applied for the transport of both photons and electrons (5.4), *i.e.* to the relaxed equation (5.33). This leads to write the numerical scheme

$$A^{n+1}_{\gamma\to\gamma}\boldsymbol{\psi_\gamma}_l^{n+1} + \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{\mathbf{F_\gamma}^{n+1}_{l+\frac{1}{2}}}{\rho_{l+\frac{1}{2}}} - \frac{\mathbf{F_\gamma}^{n+1}_{l-\frac{1}{2}}}{\rho_{l-\frac{1}{2}}}\right) = \sum_{n'=1}^{n} B^{n',n+1}_{C,\gamma}\boldsymbol{\psi_\gamma}_l^{n'} \tag{5.57a}$$

$$A^{n+1}_{e\to e}\boldsymbol{\psi_e}_l^{n+1} + \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{\mathbf{F_e}^{n+1}_{l+\frac{1}{2}}}{\rho_{l+\frac{1}{2}}} - \frac{\mathbf{F_e}^{n+1}_{l-\frac{1}{2}}}{\rho_{l-\frac{1}{2}}}\right) = \sum_{n'=1}^{n} B^{n',n+1}_{e\to e}\boldsymbol{\psi_e}_l^{n'} + B^{n',n+1}_{\gamma\to e}\boldsymbol{\psi_\gamma}_l^{n'}, \tag{5.57b}$$

$$\text{for } \alpha=\gamma,e \qquad \mathbf{F_\alpha}^{n+1}_{l+\frac{1}{2}} = \frac{1}{2}\left[\mathbf{F}(\boldsymbol{\psi_\alpha}^{n+1}_{l+1}) + \mathbf{F}(\boldsymbol{\psi_\alpha}^{n+1}_{l}) - (\boldsymbol{\psi_\alpha}^{n+1}_{l+1} - \boldsymbol{\psi_\alpha}^{n+1}_{l})\right], \tag{5.57c}$$

where the matrices $A^{n+1}_{\gamma\to\gamma}$, $A^{n+1}_{e\to e}$, $B^{n',n+1}_{\gamma\to\gamma}$, $B^{n',n+1}_{e\to e}$ and $B^{n',n+1}_{\gamma\to e}$ read

$$\begin{aligned}
A^{n+1}_{\gamma\to\gamma} &= \sigma^{n+1}_{T,C} Id - s^{n+1,n+1}_{C,\gamma}\Delta\epsilon^n, & B^{n',n+1}_{\gamma\to\gamma} &= s^{n',n+1}_{C,\gamma}\Delta\epsilon^{n'},\\
A^{n+1}_{e\to e} &= \frac{S^{n+1}}{\Delta\epsilon^n} Id - M^{n+1}, & B^{n',n+1}_{\gamma\to e} &= s^{n',n+1}_{C,e}\Delta\epsilon^{n'},\\
B^{n',n+1}_{e\to e} &= \frac{S^n Id}{\Delta\epsilon^n}\delta_{n',n},
\end{aligned}$$

In 2D, this numerical scheme reads

$$\begin{aligned}
A^{n+1}_{\gamma\to\gamma}{\boldsymbol{\psi}_{\gamma}}^{n+1}_{l,m} &+ \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{{\mathbf{F}_{\gamma}}^{n+1}_{l+\frac{1}{2},m}}{\rho_{l+\frac{1}{2},m}} - \frac{{\mathbf{F}_{\gamma}}^{n+1}_{l-\frac{1}{2},m}}{\rho_{l-\frac{1}{2},m}}\right)\\
&+ \frac{\Delta\epsilon^n}{\Delta y}\left(\frac{{\mathbf{F}_{\gamma}}^{n+1}_{l,m+\frac{1}{2}}}{\rho_{l,m+\frac{1}{2}}} - \frac{{\mathbf{F}_{\gamma}}^{n+1}_{l,m-\frac{1}{2}}}{\rho_{l,m-\frac{1}{2}}}\right) = \sum_{n'=1}^{n} B^{n',n+1}_{C,\gamma}{\boldsymbol{\psi}_{\gamma}}^{n'}_{l,m} \qquad (5.58a)\\
A^{n+1}_{e\to e}{\boldsymbol{\psi}_{\mathbf{e}}}^{n+1}_{l,m} &+ \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{{\mathbf{F}_{\mathbf{e}}}^{n+1}_{l+\frac{1}{2},m}}{\rho_{l+\frac{1}{2},m}} - \frac{{\mathbf{F}_{\mathbf{e}}}^{n+1}_{l-\frac{1}{2},m}}{\rho_{l-\frac{1}{2},m}}\right) \qquad (5.58b)\\
&+ \frac{\Delta\epsilon^n}{\Delta y}\left(\frac{{\mathbf{F}_{\mathbf{e}}}^{n+1}_{l,m+\frac{1}{2}}}{\rho_{l,m+\frac{1}{2}}} - \frac{{\mathbf{F}_{\mathbf{e}}}^{n+1}_{l,m-\frac{1}{2}}}{\rho_{l,m-\frac{1}{2}}}\right) = \sum_{n'=1}^{n} B^{n',n+1}_{e\to e}{\boldsymbol{\psi}_{\mathbf{e}}}^{n'}_{l,m} + B^{n',n+1}_{\gamma\to e}{\boldsymbol{\psi}_{\gamma}}^{n'}_{l,m},
\end{aligned}$$

for $\alpha = \gamma, e$, $${\mathbf{F}_{\alpha}}^{n+1}_{l+\frac{1}{2},m} = \frac{1}{2}\left[\mathbf{F}({\boldsymbol{\psi}_{\boldsymbol{\alpha}}}^{n+1}_{l+1,m}) + \mathbf{F}({\boldsymbol{\psi}_{\boldsymbol{\alpha}}}^{n+1}_{l,m}) - ({\boldsymbol{\psi}_{\boldsymbol{\alpha}}}^{n+1}_{l+1,m} - {\boldsymbol{\psi}_{\boldsymbol{\alpha}}}^{n+1}_{l,m})\right], \qquad (5.58c)$$

$${\mathbf{F}_{\alpha}}^{n+1}_{l,m+\frac{1}{2}} = \frac{1}{2}\left[\mathbf{F}({\boldsymbol{\psi}_{\boldsymbol{\alpha}}}^{n+1}_{l,m+1}) + \mathbf{F}({\boldsymbol{\psi}_{\boldsymbol{\alpha}}}^{n+1}_{l,m}) - ({\boldsymbol{\psi}_{\boldsymbol{\alpha}}}^{n+1}_{l,m+1} - {\boldsymbol{\psi}_{\boldsymbol{\alpha}}}^{n+1}_{l,m})\right]. \qquad (5.58d)$$

5.7.3 Computing $\boldsymbol{\psi}^{n+1}_l$

The numerical schemes (5.56), (5.56), (5.57) and (5.58) can be rewritten under the form

$$J(\boldsymbol{\psi}^{n+1})_l = \mathbf{R}^n_l, \qquad (5.59)$$

where J is a non-linear function of the unknown $\boldsymbol{\psi}^{n+1}$ and $\mathbf{R}^n_l$ contains all the terms $\boldsymbol{\psi}^{n'}$ for $n' < n+1$. In order to use the implicit schemes, the function J needs to be invertible and its inverse needs to be realizable.

In order to compute $\boldsymbol{\psi}^{n+1}$ and to prove the existence of a realizable solution to (5.59), an iterative method can be used. This method is inspired of [13].

Rewrite (5.59) under the form

$$-L(\boldsymbol{\psi}^{n+1}_{l-1}) + D(\boldsymbol{\psi}^{n+1}_l) - R(\boldsymbol{\psi}^{n+1}_{l+1}) = \mathbf{R}^n_l, \qquad (5.60)$$

where $L(\boldsymbol{\psi}_{l-1}^{n+1})$, respectively $D(\boldsymbol{\psi}_{l}^{n+1})$ and $R(\boldsymbol{\psi}_{l+1}^{n+1})$, contains all the terms depending of $\boldsymbol{\psi}_{l-1}^{n+1}$, respectively $\boldsymbol{\psi}_{l}^{n+1}$ and $\boldsymbol{\psi}_{l+1}^{n+1}$. When applied to (5.56), this reads

$$\begin{aligned} L(\boldsymbol{\psi}_{l-1}^{n+1}) &= \frac{\Delta\epsilon^n}{\rho_{l-\frac{1}{2}}\Delta x}\frac{\boldsymbol{\psi}_{l-1}^{n+1} - F\left(\boldsymbol{\psi}_{l-1}^{n+1}\right)}{2}, && (5.61a)\\ R(\boldsymbol{\psi}_{l+1}^{n+1}) &= \frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x}\frac{\boldsymbol{\psi}_{l+1}^{n+1} + F\left(\boldsymbol{\psi}_{l+1}^{n+1}\right)}{2}, && (5.61b)\\ D(\boldsymbol{\psi}_{l}^{n+1}) &= \left(1 + \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{1}{\rho_{l-\frac{1}{2}}} + \frac{1}{\rho_{l+\frac{1}{2}}}\right)\right)\boldsymbol{\psi}_{l}^{n+1}, && (5.61c)\\ \mathbf{R}_l^n &= \boldsymbol{\psi}_l^n, && (5.61d) \end{aligned}$$

when applied to (5.56), this reads

$$\begin{aligned} L(\boldsymbol{\psi}_{l-1}^{n+1}) &= \frac{\Delta\epsilon^n}{\rho_{l-\frac{1}{2}}\Delta x}\frac{\boldsymbol{\psi}_{l-1}^{n+1} - F\left(\boldsymbol{\psi}_{l-1}^{n+1}\right)}{2}, && (5.62a)\\ R(\boldsymbol{\psi}_{l+1}^{n+1}) &= \frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x}\frac{\boldsymbol{\psi}_{l+1}^{n+1} + F\left(\boldsymbol{\psi}_{l+1}^{n+1}\right)}{2}, && (5.62b)\\ D(\boldsymbol{\psi}_{l}^{n+1}) &= \left(S^{n+1} + \frac{\Delta\epsilon^n}{\Delta x}\left(\frac{1}{\rho_{l-\frac{1}{2}}} + \frac{1}{\rho_{l+\frac{1}{2}}}\right)Id - \Delta\epsilon^n M^{n+1}\right)\boldsymbol{\psi}_{l}^{n+1}, && (5.62c)\\ \mathbf{R}_l^n &= S^n\boldsymbol{\psi}_l^n, && (5.62d) \end{aligned}$$

and when applied to (5.57), this reads

$$L(\boldsymbol{\psi}_{l-1}^{n+1}) = \left(\frac{\Delta\epsilon^n}{\rho_{l-\frac{1}{2}}\Delta x}\frac{\boldsymbol{\psi_\gamma}_{l-1}^{n+1} - F\left(\boldsymbol{\psi_\gamma}_{l-1}^{n+1}\right)}{2}, \qquad \frac{\Delta\epsilon^n}{\rho_{l-\frac{1}{2}}\Delta x}\frac{\boldsymbol{\psi_e}_{l-1}^{n+1} - F\left(\boldsymbol{\psi_e}_{l-1}^{n+1}\right)}{2}\right), \tag{5.63a}$$

$$R(\boldsymbol{\psi}_{l+1}^{n+1}) = \left(\frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x}\frac{\boldsymbol{\psi_\gamma}_{l+1}^{n+1} + F\left(\boldsymbol{\psi_\gamma}_{l+1}^{n+1}\right)}{2}, \qquad \frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2}}\Delta x}\frac{\boldsymbol{\psi_e}_{l+1}^{n+1} + F\left(\boldsymbol{\psi_e}_{l+1}^{n+1}\right)}{2}\right), \tag{5.63b}$$

$$\begin{aligned} D(\boldsymbol{\psi}_{l}^{n+1}) = \Bigg(&\left(A_{\gamma\to\gamma}^{n+1} + \frac{\Delta\epsilon^n Id}{\Delta x}\left(\frac{1}{\rho_{l-\frac{1}{2}}} + \frac{1}{\rho_{l+\frac{1}{2}}}\right)\right)\boldsymbol{\psi_\gamma}_l^{n+1}, && (5.63c)\\ &\left(A_{e\to e}^{n+1} + \frac{\Delta\epsilon^n Id}{\Delta x}\left(\frac{1}{\rho_{l-\frac{1}{2}}} + \frac{1}{\rho_{l+\frac{1}{2}}}\right)\right)\boldsymbol{\psi_e}_l^{n+1} + A_{\gamma\to e}^{n+1}\boldsymbol{\psi_\gamma}_l^{n+1}\Bigg), \\ \mathbf{R}_l^n = \Bigg(&\sum_{n'=1}^{n} B_{\gamma\to\gamma}^{n'n+1}\boldsymbol{\psi_\gamma}^{n'}, \qquad \sum_{n'=1}^{n} B_{e\to e}^{n'n+1}\boldsymbol{\psi_e}^{n'} + B_{\gamma\to e}^{n'n+1}\boldsymbol{\psi_\gamma}^{n'}\Bigg). && (5.63d) \end{aligned}$$

Remark that in each case the diagonal term D is linear and can be inverted numerically. The following iterative algorithm can be used to compute the solution of (5.59).

Algorithm 5.1

Initialization: *Set*

$$\boldsymbol{\psi}_l^{n+1,(0)} = \boldsymbol{\psi}_l^n \qquad \forall l. \tag{5.64}$$

Iteration: *Solve iteratively*

$$\boldsymbol{\psi}_l^{n+1,(k+1)} = D^{-1}\left(\mathbf{R}_l^n + L(\boldsymbol{\psi}_{l-1}^{n+1,(k)}) + R(\boldsymbol{\psi}_{l+1}^{n+1,(k)})\right), \tag{5.65}$$

until the following residual error

$$r^{(k)} = \left\| D(\boldsymbol{\psi}_{l-1}^{n+1,(k)}) - \mathbf{R}_l^n - L(\boldsymbol{\psi}_{l-1}^{n+1,(k)}) - R(\boldsymbol{\psi}_{l+1}^{n+1,(k)}) \right\|_\infty \tag{5.66}$$

reaches a desired values maximum values $r_{\max}$ or until k reaches a maximum value $k_{\max}$.

Proposition 5.4 *Suppose $\boldsymbol{\psi}_l^{n'}$ is realizable for all l and all $n' < n+1$.*
Then there exists a unique set of realizable vectors $(\boldsymbol{\psi}^{n+1})_{\{l=1,\dots,l_{\max}\}}$ satisfying (5.59) for all l.
Moreover the algorithm 5.1 converges to this solution and $\boldsymbol{\psi}_l^{n+1,(k+1)} \in \mathcal{R}_{\mathbf{m}}$ is realizable for all l at each step (k).

Proof The Jacobian of J is a block matrix which is defined by

$$\left(\frac{\partial J(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)_{l,k} = \frac{\partial J(\boldsymbol{\psi})_l}{\partial \boldsymbol{\psi}_k} = D^{-1}\left[\frac{\partial L}{\partial \boldsymbol{\psi}_{l-1}}(\boldsymbol{\psi}_{l-1})\delta_{k,l-1} + \frac{\partial R}{\partial \boldsymbol{\psi}_{l+1}}(\boldsymbol{\psi}_{l+1})\delta_{k,l+1}\right].$$

Then using a Gershgörin theorem for block matrices ([25]) leads to write that all the eigenvalues of the Jacobian of J are bounded by

$$Sp\left(\frac{\partial J(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right) \subset [-r, +r],$$
$$r = |||D|||^{-1} \max_l \left(\frac{\partial L}{\partial \boldsymbol{\psi}_l}(\boldsymbol{\psi}_l) + \frac{\partial R}{\partial \boldsymbol{\psi}_l}(\boldsymbol{\psi}_l)\right).$$

The eigenvalues of the Jacobians $\frac{\partial L}{\partial \boldsymbol{\psi}_l}$ and $\frac{\partial R}{\partial \boldsymbol{\psi}_l}$ can be bounded using Lemma 5.1. Comparing those eigenvalues to $|||D|||^{-1}$, where D is given by (5.61), (5.62) or (5.63), leads to write that the spectral radius r is strictly inferior to 1. This implies that the function J is contractant, so it has a unique fixed point, and Algorithm 5.1 converges to this fixed point.

It remains to verify that if $\boldsymbol{\psi}_l^{n,(k)}$ is realizable for all l, then $\boldsymbol{\psi}_l^{n,(k+1)} = J(\boldsymbol{\psi}_l^{n,(k)})$ remains realizable.

By assumption $\boldsymbol{\psi}^n = \boldsymbol{\psi}^{n+1,(0)}$ is realizable. As $\mathbf{R}_l^n$ is a positive combination of realizable vectors, it is also realizable.

Then, by iteration, if $\boldsymbol{\psi}_l^{n+1,(k)}$ is realizable for all l, then $L(\boldsymbol{\psi}_l^{n+1,(k)})$ and

$R(\boldsymbol{\psi}_l^{n+1,(k)})$ are realizable (according to Proposition 5.1), and it remains to prove that D^{-1} preserves realizability.

In the case of the toy problem (5.55), $D = \alpha Id$ with $\alpha > 0$, therefore the realizability is obviously preserved. In the case of the electron transport (5.56), it was proven in Proposition 5.2. And in the case of coupled photons and electrons transport (5.57), it can be proven by reproducing the proof of Proposition 5.2. □

Remark 5.5 *The iterative method proposed in Algorithm 5.1 can be seen as a Jacobi method with non-linear extra-diagonal terms L and R. Similarily, a Gauss-Seidel method for this non-linear problem can be written. It consists in solving alternatively*

$$\boldsymbol{\psi}_l^{n+1,(k+1)} = D^{-1}\left(\mathbf{R}_l^n + L(\boldsymbol{\psi}_{l-1}^{n+1,(k)}) + R(\boldsymbol{\psi}_{l+1}^{n+1,(k+1)})\right), \tag{5.67a}$$

$$\boldsymbol{\psi}_l^{n+1,(k+1)} = D^{-1}\left(\mathbf{R}_l^n + L(\boldsymbol{\psi}_{l-1}^{n+1,(k+1)}) + R(\boldsymbol{\psi}_{l+1}^{n+1,(k)})\right), \tag{5.67b}$$

instead of (5.65) in Algorithm 5.1. This algorithm was tested experimentally and converged to the desired state. However, it was not proven to be convergent at the theoretical level.

5.8 Tests on the implicit solver

The aim of those test is to study the implicit scheme (5.57). Especially, two convergence rates are studied: the convergence of Algorithm (5.1) according to the number $k_{\max}$ of iterations to reach the resdual $r_{\max}$, and the convergence of the nummerical scheme according to the cells size $\Delta\epsilon^n$ and Δx.

Two test cases are proposed, a 1D electron beam and a 2D photon beam.

5.8.1 A 1D electron beam

This test case is identical to the one in Subsection 5.3.1, but only the implict solver (5.57) is used. However, as the implicit numerical scheme is more flexible than the explicit one, all the physics of the electronic collisions described by Equation 5.4 was included for this test case.

The objective of this subsection is only to study the convergence rates. Dose results obtained with Algorithm 5.1 are presented in the next subsection.

Convergence of Algorithm 5.1

The number of spatial cells is fixed at 600 and the energy step is fixed by

$$\Delta\epsilon^n = 5S^n\Delta x, \tag{5.68}$$

with an initial energy $\epsilon_{\max} = 1.2\epsilon_0 = 12$ MeV and a final energy at $\epsilon_{\min} = 10^{-3}$ MeV.

The iterative method of Algorithm 5.1 requires a criterium to stop. Either the number of iterations k_{max} is fixed or the residual (5.66) is reached.

As a first test, the number of iterations $k_{\max}$ is fixed at different values, *i.e.* 10, 30, 50 and 70, and the residual $r^{n,k_{\max}}$ is plotted as a function of the energy step n on Fig. 5.15. As a second test, the maximum residual $r_{\max}$ is fixed at different values, *i.e.* 0.1, 0.01, 0.001 and 0.0001, and the number of iterations k required to reach this residual is plotted as a function of the energy step n on Fig. 5.16.

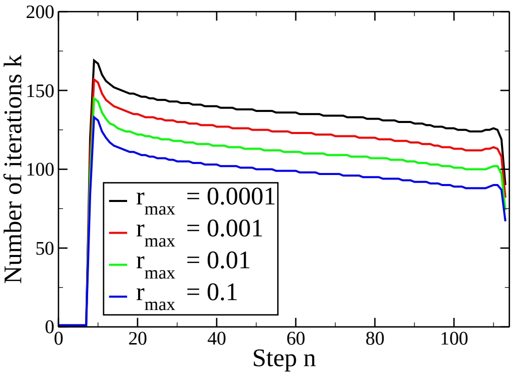

Figure 5.15: Number of iterations k as a function of the energy step n for a given maximum residual $r_{\max}$

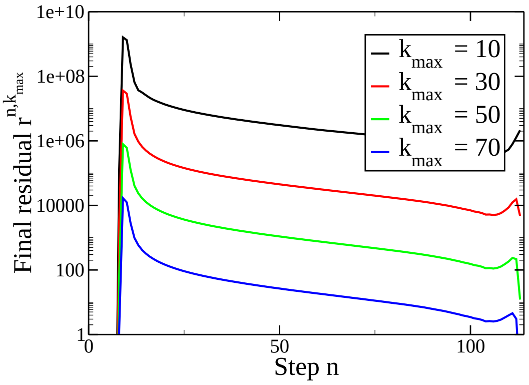

Figure 5.16: Final residual $r^{n,k_{\max}}$ as a function of the energy step n for a given number of iterations $k_{\max}$

During the first steps (up to $n = 8$), the values of $\boldsymbol{\psi}$ on the boundary (see (5.25)) is very low and the fluence ψ inside the medium is negligible. Therefore the values of k and of $r^{n,k_{\max}}$ are also very low.

In the first steps where the values of ψ is non-negligible, Algorithm 5.1 requires a large number of iterations to converge. At those energy the energy derivative $\partial_\epsilon \boldsymbol{\psi} \equiv (\boldsymbol{\psi}^n - \boldsymbol{\psi}^{n+1})/\Delta\epsilon^n$ is large and the intialization $\boldsymbol{\psi}^{n+1,(0)} = \boldsymbol{\psi}^n$ of Algorithm 5.1 is inaccurate which forces the algorithm to make more iterations to converge.

The convergence rate progressively raises, *i.e.* the final residual $r^{n,k_{\max}}$ or the number of itarations k reduce.

Finally, near the end of the simulation, the particles leave the system by reaching an energy lower than threshold $\epsilon_{\min}$. As the fluence ψ drops, the energy derivative too and the algorithm requires less iterations to converge. This explains the final drop in Fig. 5.15 and 5.16.

Convergence of the numerical scheme

Based on the construction of the implicit scheme (5.57), one may expect the truncation error to be of order 1 in Δx and in $\Delta\epsilon^n$.

The energy step size is fixed by the condition (5.68).

No analytical solution is known for equations (5.4). Therefore the solution obtain with the implicit scheme with different grid sizes are compared to a reference that was computed with this same scheme with a large number of cells, *i.e.* $l_{\max} = 9600$.

The number of spatial cells $l_{\max}$ is chosen to be 300, 600, 1200, 2400 and 4800 cells and for the reference solution 9600 cells.

The convergence rate of the numerical scheme is represented through the discrete L^2 error between the reference solution ψ computed with 9600 spatial cells and the less refined results. This error is plotted on Fig. 5.17 as a function of Δx. On this test case, one obtains an experimental convergence rate in Δx of 1.1046894.

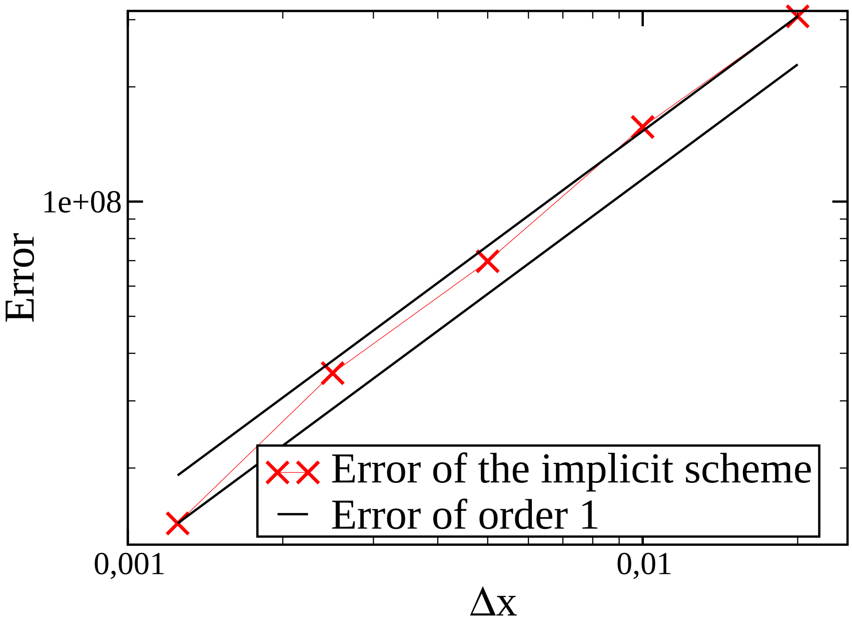

Figure 5.17: Discrete L^2 error compared to the most refined solution as a function of Δx.

5.8.2 Comparison with the explicit solver

This test case is identical to the 1D test case containing slabs of air of Subsection 5.6.1. The same parameters are used, *i.e.* 1200 spatial cells, the same beam (5.25), and the energy step $\Delta\epsilon^n$ is fixed either to $\Delta\epsilon^n_{water}$ (5.54) or to $\Delta\epsilon^n_{air}$ (5.53).

The aim of this case is only to compare the explicit (5.51) and implicit (5.56) numerical schemes. For this purpose, only the M_1 results are shown.

The maximum residual (5.66) for the implicit scheme (5.56) is fixed at $r_{\max} = 10^3$ and the maximum number of iterations is fixed $k_{\max} = 10000$. Experimentally this number of iterations was never reached, meaning that the maximum residual was always reached.

The doses obtained with the M_1 model with the explicit (5.51) and implicit (5.56) numerical schemes are plotted on Fig. 5.18 and the computational times are gathered in Table 5.8 and 5.9.

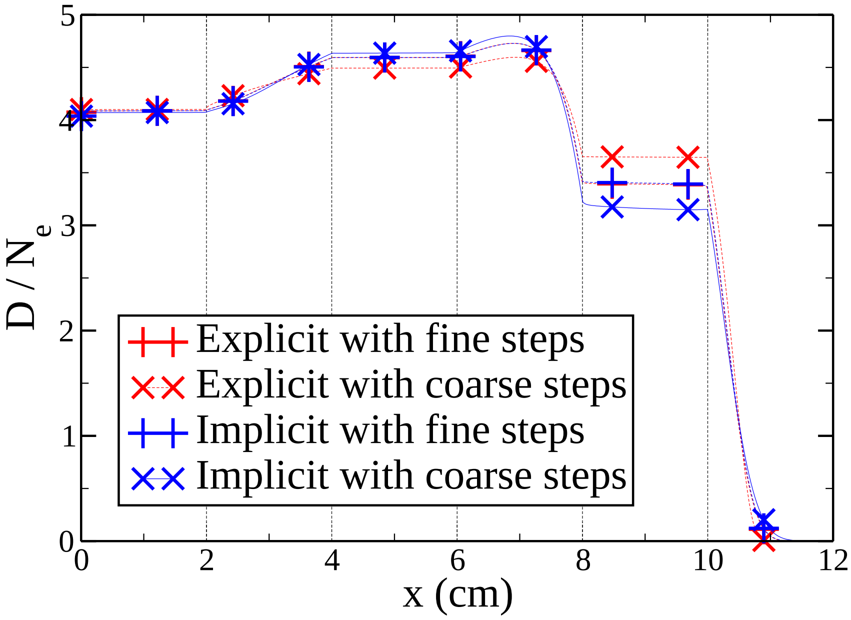

Figure 5.18: Doses obtained with the M_1 model with the explicit and implicit schemes with fine and coarse energy steps $\Delta\epsilon^n$.

Numerical scheme	Explicit with $\Delta\epsilon^n_{air}$	Implicit with $\Delta\epsilon^n_{air}$
Computation times	63.592346 sec	118.435718 sec
Number of energy steps	632468	632468

Table 5.8: Computational times with the explicit (5.51) and implicit (5.56) solver with a fine energy step size $\Delta\epsilon^n_{air}$.

Numerical schemes	Explicit with $\Delta\epsilon^n_{water}$	Implicit with $\Delta\epsilon^n_{water}$
Computation times	0.073788 sec	1.409635 sec
Number of energy steps	634	634

Table 5.9: Computational times with the explicit (5.51) and implicit (5.56) solver with a coarse energy step size $\Delta\epsilon^n_{water}$.

The difference between the dose results with the explicit and implicit schemes with a fine energy steps is very small. This shows that both schemes converge to the same results, which was expected.

As for the explicit scheme, the results with a larger energy step $\Delta\epsilon^n = \Delta\epsilon^n_{water}$ show good agreements with those with a fine $\Delta\epsilon^n = \Delta\epsilon^n_{air}$.

The implicit and explicit schemes with coarse steps present an error of the same order compared to the results with fine energy steps.

The implicit scheme requires an interne loop to solve the scheme. It is more flexible (it can be used on equations without hyperbolic operators), however it introduces additional computational costs and errrors that can be observed on this test case. Remark that with the maximum residual chosen for this test case, *i.e.* $r_{\max} = 10^3$, the error introduced by the iterative method is small enough such that the dose results are comparable to those of the explicit solver which does not introduce such an error.

5.8.3 A 2D photon beam

For this test case, photons are injected in a 2D homogeneous water phantom instead of electrons. The size of the medium is 2 cm × 10 cm, and a 0.5 cm large beam of 500 keV photons is injected on the left boundary. This is modeled by the boundary condition

$$\begin{aligned}\psi_\gamma(X,\epsilon,\Omega) &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon-\epsilon_0\right)^2\right)\exp\left(-\alpha_\mu\left(\Omega_1-1\right)^2\right)\mathbf{1}_B(X), \quad \forall(X,\Omega)\in\Gamma^-,\\ B &= \Big\{X=(x,y), \qquad x=0, \qquad y\in[1\text{ cm},2\text{ cm}]\Big\}.\end{aligned}$$

with $\epsilon_0 = 500$ keV, $\alpha_\epsilon = 20000$ and $\alpha_\mu = 10000$ and for the moment models

$$\begin{aligned}\boldsymbol{\psi}^n_{0,m} &= 10^{10}\exp\left(-\alpha_\epsilon\left(\epsilon^n-\epsilon_0\right)^2\right)\left\langle\mathbf{m}(\Omega)\exp\left(-\alpha_\mu\left(\Omega_1-1\right)^2\right)\right\rangle\mathbf{1}_B(X_{l,m}),\\ \boldsymbol{\psi}^n_{l,0} &= \boldsymbol{\psi}^n_{l_{\max},m} = \boldsymbol{\psi}^n_{l,m_{\max}} = 0_{\mathbb{R}^{Card(\mathbf{m})}}.\end{aligned}$$

First results

The dose obtained with the M_1 and M_2 model with the implicit scheme (5.58) are compared to a reference Monte Carlo results on Fig. 5.19 and the computational times are gathered on Table 5.10.

Solver	Monte Carlo	M_1 solver	M_2 solver
Computation times	14 hours	49.78699 sec	215.0480 sec

Table 5.10: Computational times with the implicit solver with the Cartesian direction of relaxation.

The doses along the axis $y = 1$ cm is plotted Fig. 5.20 and along the axes $x = 2$ cm and $x = 8$ cm on Fig. 5.21.

The total cross section for photons is much lower than the one for electrons. This implies that the photons collide much less than electrons and therefore travel longer distances without being deflected.

This can be observed on the Monte Carlo results on Fig. 5.19. However, the M_1 and M_2 results with the implicit scheme are almost identical and are very diffused in the direction transverse to the beam contrarily to the Monte Carlo results. This is due to the relaxation parameters chosen (5.39).

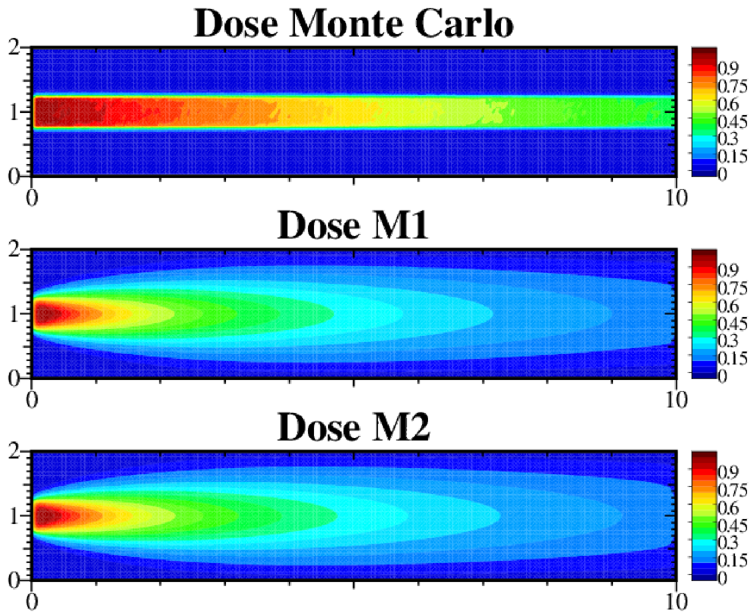

Figure 5.19: Doses obtained with the Monte Carlo solver (top) and the approximated M_1 (middle) and M_2 (below) models.

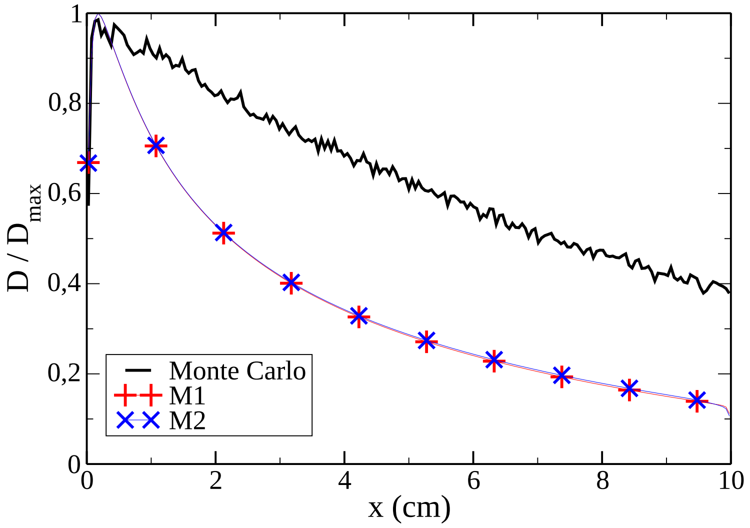

Figure 5.20: Dose obtained with the Monte Carlo solver and the approximated M_1 and M_2 models along the axis $y = 1$ cm.

Transverse diffusion

In practice, the relaxation method can be used under the stability condition (5.28) on the relaxation speeds. However, in 1D, the relaxation method is known to be

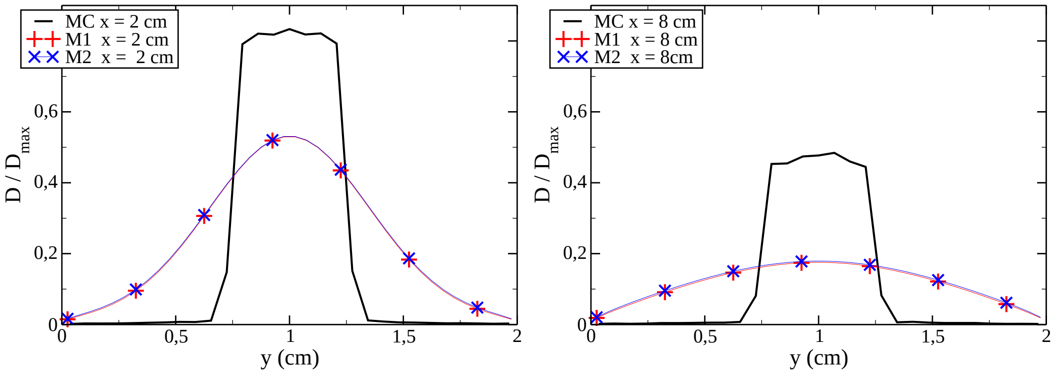

Figure 5.21: Dose obtained with the Monte Carlo solver and the approximated M_1 and M_2 models along the axes $x = 2$ cm and $x = 8$ cm.

overdiffusive when the relaxation speeds λ_j are too large compared to the eigenvalues of the Jacobian of the flux $\partial_{\boldsymbol{\psi}}\mathbf{F}(\boldsymbol{\psi})$.

Bounds of the eigenvalues of the Jacobian of the fluxes

The relaxation speeds were chosen to be of norm $|\lambda_j| = 1/\rho$ which was enough to satisfy (5.28) according to Lemma 5.1. However, when $\boldsymbol{\psi}$ is the moment vector of an imperfect beam in the direction e_1 modeled by

$$\boldsymbol{\psi} = \int_{S^2} \mathbf{m}(\Omega) \exp(-\alpha_\mu(\Omega_1 - 1)) d\Omega,$$

then the spectral radius of the Jacobian of the flux $\mathbf{F_2}$ transverse to the direction of the beam is zero. Indeed, define

$$\begin{aligned}
A &:= \partial_{\boldsymbol{\psi}}\mathbf{F_2}(\boldsymbol{\psi}),\\
B &:= \partial_{\boldsymbol{\lambda}}\mathbf{F_2}(\boldsymbol{\psi}) = \int_{S^2} \Omega_2 \mathbf{m}(\Omega) \otimes \mathbf{m}(\Omega) \exp(\boldsymbol{\lambda}^T \mathbf{m}(\Omega)) d\Omega\\
&= \int_{S^2} \Omega_2 \mathbf{m}(\Omega) \otimes \mathbf{m}(\Omega) \exp(-\alpha_2(\Omega_1 - 1)) d\Omega,\\
C &:= \partial_{\boldsymbol{\lambda}}\boldsymbol{\psi} = \int_{S^2} \mathbf{m}(\Omega) \otimes \mathbf{m}(\Omega) \exp(\boldsymbol{\lambda}^T \mathbf{m}(\Omega)) d\Omega\\
&= \int_{S^2} \mathbf{m}(\Omega) \otimes \mathbf{m}(\Omega) \exp(-\alpha_\mu(\Omega_1 - 1)) d\Omega.
\end{aligned}$$

Then the eigenvalues α_i (associated to the eigenvector V_i) of the Jacobian of the flux $\mathbf{F_2}$ can be bounded

$$\begin{aligned}
BW_i &= \alpha_i C W_i, \qquad W_i = C^{-1} V_i,\\
BW_i.W_i &= \int_{S^2} \Omega_2 (\mathbf{m}(\Omega)^T W_i)^2 \exp(-\alpha_\mu(\Omega_1 - 1)) d\Omega = 0,\\
CW_i.W_i &= \int_{S^2} (\mathbf{m}(\Omega)^T W_i)^2 \exp(-\alpha_\mu(\Omega_1 - 1)) d\Omega > 0.
\end{aligned}$$

Therefore all the eigenvalues α_i of the Jacobian of the transverse flux are zero.

In a more general way, those eigenvalues can be computed in the case of the M_1 model. Consider that ψ^1 is colinear to e_1 (otherwise just use a rotation to work in such a reference frame). Using the form (4.26) of the closure, the fluxes in the direction e_1 and in the transverse direction e_2 read

$$\begin{aligned}
\boldsymbol{\psi} &= \left(\psi^0,\ \psi^1\right), \\
\mathbf{F_n}(\boldsymbol{\psi}) &= \left(\psi^1.n,\ \frac{\psi^0}{2}\left[(1-\chi_2)n + (3\chi_2-1)\frac{(\psi^1.n)\psi^1}{|\psi^1|^2}\right]\right), \\
\partial_{\boldsymbol{\psi}}\left(\mathbf{F_n}(\boldsymbol{\psi})\right) &= \begin{pmatrix} 0 & V_1 \\ V_2 & M \end{pmatrix}, \\
V_1 &= \partial_{\psi^1}(\psi^1.n) = (n_1,\ n_2,\ n_3)\,, \\
V_2 &= \partial_{\psi^0}\left(\frac{\psi^0}{2}\left[(1-\chi_2)n + (3\chi_2-1)\frac{(\psi^1.n)\psi^1}{|\psi^1|^2}\right]\right), \\
&= \frac{1}{2}\left[(1-\chi_2)n + (3\chi_2-1)\frac{(\psi^1.n)\psi^1}{|\psi^1|^2}\right] - \frac{|\psi^1|}{2\psi^0}\chi_2'\left(-n + 3\frac{(\psi^1.n)\psi^1}{|\psi^1|^2}\right), \\
M &= \partial_{\psi^1}\left(\frac{\psi^0}{2}\left[(1-\chi_2)n + (3\chi_2-1)\frac{(\psi^1.n)\psi^1}{|\psi^1|^2}\right]\right) \\
&= \frac{\chi_2'}{2}\frac{\psi^1}{|\psi^1|} \otimes \left(-n + 3\frac{(\psi^1.n)\psi^1}{|\psi^1|^2}\right) \\
&\quad + \frac{\psi^0}{2}(3\chi_2-1)\left(\frac{n\otimes\psi^1}{|\psi^1|^2} + \frac{\psi^1.n}{|\psi^1|^2}Id - 2(\psi^1.n)\psi^1 \otimes \frac{\psi^1}{|\psi^1|^4}\right).
\end{aligned} \tag{5.69}$$

Chose a reference frame such that $\psi^1 = \psi^1_1 e_1$ with $\psi^1_1 \geq 0$. In this reference frame, the Jacobian of the flux F_1 along the direction e_1 reads

$$\partial_{\boldsymbol{\psi}}\left(\mathbf{F_1}(\boldsymbol{\psi})\right) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \chi_2 - N_1^1\chi_2' & \chi_2' & 0 & 0 \\ 0 & 0 & \frac{3\chi_2-1}{2N_1^1} & 0 \\ 0 & 0 & 0 & \frac{3\chi_2-1}{2N_1^1} \end{pmatrix},$$

and so the spectrum of the Jacobian of the flux is

$$\begin{aligned}
S_n(N_1^1) &= Sp\left(\partial_{\boldsymbol{\psi}}\left(\mathbf{F}(\boldsymbol{\psi})e_1\right)\right) \\
&= \Bigg(\frac{3\chi_2-1}{2N_1^1},\ \frac{3\chi_2-1}{2N_1^1}, \\
&\qquad \frac{\chi_2' + \sqrt{\chi_2'^2 + 4(\chi_2 - N_1^1\chi_2')}}{2},\ \frac{\chi_2' - \sqrt{\chi_2'^2 + 4(\chi_2 - N_1^1\chi_2')}}{2}\Bigg).
\end{aligned}$$

Now computing the Jacobian of the flux in a transverse direction, *e.g.* e_2 reads

$$\partial_{\boldsymbol{\psi}}\left(\mathbf{F_2}(\boldsymbol{\psi})\right) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{\chi_2'}{2} & 0 \\ \frac{1-\chi_2+N_1^1\chi_2'}{2} & \frac{3\chi_2-1}{2N_1^1} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and so the spectrum of the Jacobian of the flux is

$$\begin{aligned} S_t(N_1^1) &= Sp\left(\partial_{\boldsymbol{\psi}}\left(\mathbf{F_2}(\boldsymbol{\psi})\right)\right) \\ &= \left(0,\ 0,\ \sqrt{\frac{1-\chi_2+N_1^1\chi_2'-\frac{3\chi_2-1}{2N_1^1}\chi_2'}{2}},\ -\sqrt{\frac{1-\chi_2+N_1^1\chi_2'-\frac{3\chi_2-1}{2N_1^1}\chi_2'}{2}}\right). \end{aligned}$$

Now, in order to come back to the computations in any reference frame, one can simply use a rotation R such that $R\psi^1 = \psi_1^1 e_1$. One has

$$\begin{aligned} \partial_{\boldsymbol{\psi}}\left(\mathbf{F_n}(\boldsymbol{\psi})\right) &= \partial_{(\psi^0, R\psi_1^1 e_1)}\left(\mathbf{F_n}(\psi^0, R\psi_1^1 e_1)\right) \\ &= R_2 \partial_{\psi}\mathbf{F_{R^T n}}\left(\psi^0, |\psi^1| e_1\right) R_2^T, \\ R_2 &= \begin{pmatrix} 1 & 0_{\mathbb{R}^3} \\ 0_{\mathbb{R}^3} & R \end{pmatrix}. \end{aligned}$$

The spectrum of such a matrix can be bounded using the previous computations

$$\begin{aligned} Sp\left(\partial_{\boldsymbol{\psi}}\left(\mathbf{F_n}(\boldsymbol{\psi})\right)\right) &\subset [b_-, b_+], && (5.70a) \\ b_-(N^1, n) &= (1-\theta)\min S_t(|N^1|) + \theta \min S_n(|N^1|), && (5.70b) \\ b_+(N^1, n) &= (1-\theta)\max S_t(|N^1|) + \theta \max S_n(|N^1|), && (5.70c) \\ \theta &= \frac{N^1.n}{|N^1|}. \end{aligned}$$

Remark that those bounds b_- and b_+ are not the exact minimum and maximum eigenvalues of the Jacobians of the flux, which could be computed analytically through (5.69). Although computing numerically such eigenvalues using the analytical formulae introduces errors which can be non-negligible, while computing the bounds in (5.70) is easier. Furthermore, those bounds b_- and b_+ are sufficient for the present applications. Indeed, one verifies that they are zero when considering the flux of a perfect beam in the transverse direction of this beam.

The modified relaxation parameters

Now one can propose new relaxation parameters by replacing those bounds in the definition of the Cartesian parameter of relaxation, *e.g.*

$$\text{Relaxation directions} \quad \lambda_1 = (b_1, 0)\,, \quad \lambda_2 = (b_2, 0)\,, \tag{5.71a}$$

$$\lambda_3 = (0, b_3)\,, \quad \lambda_4 = (0, b_4)\,,$$

$$\text{Associated Maxwellians} \quad \mathbf{M_1} = \frac{|b_1|\boldsymbol{\psi}}{2(|b_1|+|b_2|)} + \frac{F\lambda_1}{|\lambda_1|(|b_1|+|b_2|)}, \tag{5.71b}$$

$$\mathbf{M_2} = \frac{|b_2|\boldsymbol{\psi}}{2(|b_1|+|b_2|)} + \frac{F\lambda_2}{|\lambda_2|(|b_1|+|b_2|)}, \tag{5.71c}$$

$$\mathbf{M_3} = \frac{|b_3|\boldsymbol{\psi}}{2(|b_3|+|b_4|)} + \frac{F\lambda_3}{|\lambda_3|(|b_3|+|b_4|)}, \tag{5.71d}$$

$$\mathbf{M_4} = \frac{|b_4|\boldsymbol{\psi}}{2(|b_3|+|b_4|)} + \frac{F\lambda_4}{|\lambda_4|(|b_3|+|b_4|)}. \tag{5.71e}$$

All the schemes based on the relaxation methods require that the Maxwellians $\mathbf{M_i} \in \mathcal{R}_\mathbf{m}$ are realizable as long as $\boldsymbol{\psi} \in \mathcal{R}_\mathbf{m}$. In practice, this implies that there is an additional requirement on the bounds b_1, b_2, b_3 and b_4.

For the M_1 model, those requirements can easily be computed using (3.41), it reads

$$\left(\frac{|b_i|}{2} + N^1.\frac{\lambda_i}{|\lambda_i|}\right)^2 > \left|\frac{|b_i|}{2}N^1 + N^2\frac{\lambda_i}{|\lambda_i|}\right|^2,$$

and therefore

$$|b_i| > \max\left(0, b_{\min}\left(N^1, \frac{\lambda_i}{|\lambda_i|}\right)\right),$$

$$b_{\min}(N^1, n) := \frac{-\beta - \sqrt{\beta^2 - \alpha\gamma}}{\alpha},$$

$$\alpha = \frac{1-|N^1|^2}{4}, \quad \beta = \frac{1}{2}\left(N^1.n - N^1.(N^2 n)\right), \quad \gamma = (N^1.n)^2 - |N^2 n|^2.$$

This leads to fix the bounds

$$\begin{aligned}
b_1(\boldsymbol{\psi}) &= \max\left(10^{-8}, \quad b_+(N^1, e_1), \quad b_{\min}(N^1, e_1)\right), \\
b_2(\boldsymbol{\psi}) &= \min\left(-10^{-8}, \quad b_-(N^1, e_1), \quad -b_{\min}(N^1, -e_1)\right), \\
b_3(\boldsymbol{\psi}) &= \max\left(10^{-8}, \quad b_+(N^1, e_2), \quad b_{\min}(N^1, e_2)\right), \\
b_4(\boldsymbol{\psi}) &= \min\left(-10^{-8}, \quad b_-(N^1, e_2), \quad -b_{\min}(N^1, -e_2)\right),
\end{aligned}$$

where the constants $\pm 10^{-8}$ are chosen arbitrarily low to avoid divisions by zero.

The new numerical scheme

Using the Maxwellians (5.71) leads to rewrite the fluxes

$$
\begin{aligned}
\mathbf{F}^{n+1}_{l+\frac{1}{2},m} &= \frac{1}{|b^{n+1}_{1,l+\frac{1}{2},m}| + |b^{n+1}_{2,l+\frac{1}{2},m}|}\Big[|b^{n+1}_{2,l+\frac{1}{2},m}|\mathbf{F}(\boldsymbol{\psi}^{n+1}_{l+1,m})e_1 + |b^{n+1}_{1,l+\frac{1}{2},m}|\mathbf{F}(\boldsymbol{\psi}^{n+1}_{l,m})e_1 \\
&\qquad + |b^{n+1}_{1,l+\frac{1}{2},m} b^{n+1}_{2,l+\frac{1}{2},m}|(\boldsymbol{\psi}^{n+1}_{l+1,m} - \boldsymbol{\psi}^{n+1}_{l,m})\Big], \\
\mathbf{F}^{n+1}_{l,m+\frac{1}{2}} &= \frac{1}{|b^{n+1}_{3,l,m+\frac{1}{2}}| + |b^{n+1}_{4,l,m+\frac{1}{2}}|}\Big[|b^{n+1}_{4,l,m+\frac{1}{2}}|\mathbf{F}(\boldsymbol{\psi}^{n+1}_{l,m+1})e_2 + |b^{n+1}_{3,l,m+\frac{1}{2}}|\mathbf{F}(\boldsymbol{\psi}^{n+1}_{l,m})e_2 \\
&\qquad + |b^{n+1}_{3,l,m+\frac{1}{2}} b^{n+1}_{4,l,m+\frac{1}{2}}|(\boldsymbol{\psi}^{n+1}_{l,m+1} - \boldsymbol{\psi}^{n+1}_{l,m})\Big], \\
b^{n+1}_{1,l+\frac{1}{2},m} &= \max\left(b_1(\boldsymbol{\psi}^{n+1}_{l,m}), b_1(\boldsymbol{\psi}^{n+1}_{l+1,m})\right), \quad b^{n+1}_{2,l+\frac{1}{2},m} = \min\left(b_2(\boldsymbol{\psi}^{n+1}_{l,m}), b_2(\boldsymbol{\psi}^{n+1}_{l+1,m})\right), \\
b^{n+1}_{3,l,m+\frac{1}{2}} &= \max\left(b_3(\boldsymbol{\psi}^{n+1}_{l,m}), b_3(\boldsymbol{\psi}^{n+1}_{l+1,m})\right), \quad b^{n+1}_{4,l,m+\frac{1}{2}} = \min\left(b_4(\boldsymbol{\psi}^{n+1}_{l,m}), b_4(\boldsymbol{\psi}^{n+1}_{l+1,m})\right),
\end{aligned}
$$

in the implicit numerical scheme (5.58) which turns into

$$
\begin{aligned}
\mathbf{R}^n_{l,m} &= -L_1(\boldsymbol{\psi}^{n+1}_{l-1,m}) - L_2(\boldsymbol{\psi}^{n+1}_{l,m-1}) + D(\boldsymbol{\psi}^{n+1}_{l,m}) && (5.72) \\
&\quad -R_1(\boldsymbol{\psi}^{n+1}_{l-1,m}) - R_2(\boldsymbol{\psi}^{n+1}_{l,m-1}), && (5.73)
\end{aligned}
$$

where the rest term reads

$$
\mathbf{R}^n_{l,m} = \left(\sum_{n'=1}^{n} B^{n'n+1}_{\gamma\to\gamma}\psi_{\gamma}{}^{n'}_{l,m}, \qquad \sum_{n'=1}^{n} B^{n'n+1}_{e\to e}\psi_{e}{}^{n'}_{l,m} + B^{n'n+1}_{\gamma\to e}\psi_{\gamma}{}^{n'}_{l,m}\right),
$$

the operators L_1, L_2, R_1 and R_2 read

$$
\begin{aligned}
L_1(\boldsymbol{\psi}^{n+1}_{l-1,m}) &= \left(c_1\psi_{\gamma}{}^{n+1}_{l-1,m} - a_1\mathbf{F_1}\left(\psi_{\gamma}{}^{n+1}_{l-1,m}\right), \quad c_1\psi_{e}{}^{n+1}_{l-1,m} - a_1\mathbf{F_1}\left(\psi_{e}{}^{n+1}_{l-1,m}\right)\right), \\
L_2(\boldsymbol{\psi}^{n+1}_{l,m-1}) &= \left(c_2\psi_{\gamma}{}^{n+1}_{l,m-1} - a_2\mathbf{F_2}\left(\psi_{\gamma}{}^{n+1}_{l,m-1}\right), \quad c_2\psi_{e}{}^{n+1}_{l,m-1} - a_2\mathbf{F_2}\left(\psi_{e}{}^{n+1}_{l,m-1}\right)\right), \\
R_1(\boldsymbol{\psi}^{n+1}_{l+1,m}) &= \left(c_3\psi_{\gamma}{}^{n+1}_{l+1,m} + a_3\mathbf{F_1}\left(\psi_{\gamma}{}^{n+1}_{l+1,m}\right), \quad c_3\psi_{e}{}^{n+1}_{l+1,m} + a_3\mathbf{F_1}\left(\psi_{e}{}^{n+1}_{l+1,m}\right)\right), \\
R_2(\boldsymbol{\psi}^{n+1}_{l,m+1}) &= \left(c_4\psi_{\gamma}{}^{n+1}_{l,m+1} + a_4\mathbf{F_2}\left(\psi_{\gamma}{}^{n+1}_{l,m+1}\right), \quad c_4\psi_{e}{}^{n+1}_{l,m+1} + a_4\mathbf{F_2}\left(\psi_{e}{}^{n+1}_{l,m+1}\right)\right).
\end{aligned}
$$

With this particular choice of bounds b_i, the operator D is not necessarily linear. Indeed it reads

$$
\begin{aligned}
D(\boldsymbol{\psi}^{n+1}_{l,m}) &= \Big(\left(A^{n+1}_{\gamma\to\gamma} + (c_1+c_2+c_3+c_4)Id\right)\psi_{\gamma}{}^{n+1}_{l,m} \\
&\qquad + (a_1 - a_3)\mathbf{F_2}(\psi_{\gamma}{}^{n+1}_{l,m}) + (a_2 - a_4)\mathbf{F_2}(\psi_{\gamma}{}^{n+1}_{l,m}), \\
&\quad \left(A^{n+1}_{e\to e} + (c_1+c_2+c_3+c_4)Id\right)\psi_{e}{}^{n+1}_{l,m} + A^{n+1}_{\gamma\to e}\psi_{\gamma}{}^{n+1}_{l,m} \\
&\qquad + (a_1 - a_3)\mathbf{F_1}(\psi_{e}{}^{n+1}_{l,m}) + (a_2 - a_4)\mathbf{F_2}(\psi_{e}{}^{n+1}_{l,m})\Big),
\end{aligned}
$$

which is non-linear when $a_1 \neq a_3$ or $a_2 \neq a_3$. In those computations, the scalars a_i and c_i read

$$\begin{aligned}
a_1 &= \frac{\Delta\epsilon^n}{\rho_{l-\frac{1}{2},m}\Delta x} \frac{|b^{n+1}_{1,l-\frac{1}{2},m}|}{|b^{n+1}_{1,l-\frac{1}{2},m}| + |b^{n+1}_{2,l-\frac{1}{2},m}|}, & c_1 &= a_1|b^{n+1}_{2,l-\frac{1}{2},m}|,\\
a_2 &= \frac{\Delta\epsilon^n}{\rho_{l,m-\frac{1}{2}}\Delta y} \frac{|b^{n+1}_{3,l,m-\frac{1}{2}}|}{|b^{n+1}_{3,l,m-\frac{1}{2}}| + |b^{n+1}_{4,l,m-\frac{1}{2}}|}, & c_2 &= a_2|b^{n+1}_{4,l,m-\frac{1}{2}}|,\\
a_3 &= \frac{\Delta\epsilon^n}{\rho_{l+\frac{1}{2},m}\Delta x} \frac{|b^{n+1}_{2,l+\frac{1}{2},m}|}{|b^{n+1}_{1,l+\frac{1}{2},m}| + |b^{n+1}_{2,l+\frac{1}{2},m}|}, & c_3 &= a_3|b^{n+1}_{1,l+\frac{1}{2},m}|,\\
a_4 &= \frac{\Delta\epsilon^n}{\rho_{l,m+\frac{1}{2}}\Delta y} \frac{|b^{n+1}_{4,l,m+\frac{1}{2}}|}{|b^{n+1}_{3,l,m+\frac{1}{2}}| + |b^{n+1}_{4,l,m+\frac{1}{2}}|}, & c_4 &= a_4|b^{n+1}_{3,l,m+\frac{1}{2}}|.
\end{aligned}$$

The new operator D is non-linear and is therefore not trivial to inverse. In order to use Algorithm 5.1 with those operators, D is decomposed into a linear and a non-linear part under the form

$$\begin{aligned}
D(\boldsymbol{\psi}) &= D_i(\boldsymbol{\psi}) - D_e(\boldsymbol{\psi}), \qquad (5.74)\\
D_i(\boldsymbol{\psi}) &= \Big(\Big(A^{n+1}_{\gamma\to\gamma} + (c_1+c_2+c_3+c_4+\alpha)Id\Big)\boldsymbol{\psi_\gamma},\\
&\qquad \Big(A^{n+1}_{e\to e} + (c_1+c_2+c_3+c_4+\alpha)Id\Big)\boldsymbol{\psi_e} + A^{n+1}_{\gamma\to e}\boldsymbol{\psi_\gamma}\Big),\\
D_e(\boldsymbol{\psi}) &= (\alpha\boldsymbol{\psi_\gamma} - (a_1-a_3)F(\boldsymbol{\psi_\gamma})e_1 - (a_2-a_4)F(\boldsymbol{\psi_\gamma})e_2,\\
&\qquad \alpha\boldsymbol{\psi_e} - (a_1-a_3)F(\boldsymbol{\psi_e})e_1 - (a_2-a_4)F(\boldsymbol{\psi_e})e_2),
\end{aligned}$$

where the coefficient α is chosen such that the operator D_e preserves the realizability. In practice, the following coefficient α is chosen

$$\alpha = |a_1 - a_3| + |a_2 - a_4|.$$

Finally, Algorithm 5.1 is rewritten by replacing (5.65) by

$$\begin{aligned}
\boldsymbol{\psi}^{n+1,(k+1)}_{l,m} &= D_i^{-1}\Big(R^n_{l,m} + L_1(\boldsymbol{\psi}^{n+1,(k)}_{l-1,m}) + L_2(\boldsymbol{\psi}^{n+1,(k)}_{l,m-1})\\
&\qquad +D_e(\boldsymbol{\psi}^{n+1,(k)}_{l,m}) + R_1(\boldsymbol{\psi}^{n+1,(k)}_{l-1,m}) + R_2(\boldsymbol{\psi}^{n+1,(k)}_{l,m-1})\Big).
\end{aligned}$$

One verifies that this modified algorithm is still convergent and preserves realizability.

Computing the eigenvalues of the Jacobian of the flux is more difficult, and this method was only tested for the M_1 model.

The dose results with the modified scheme

Using this modified algorithm on the 2D photon beam test case provides the dose result on Fig. 5.22 with the computational times in Table 5.11.

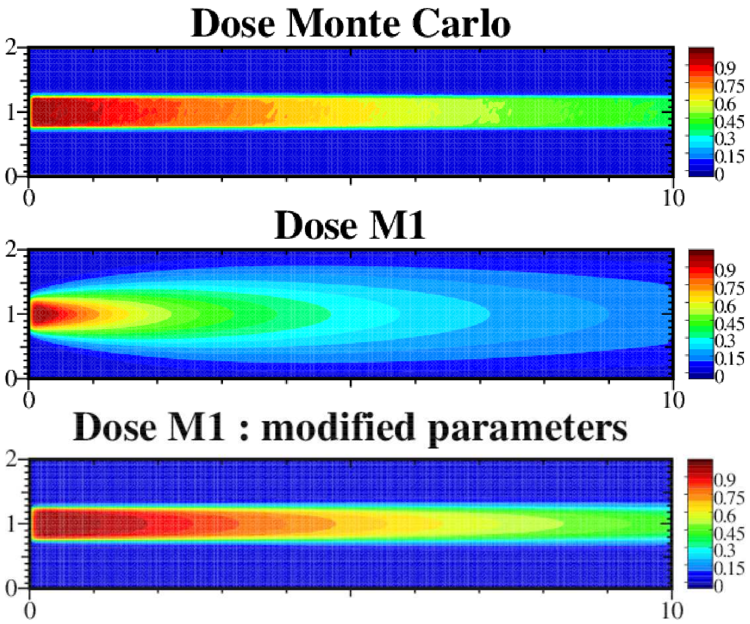

Figure 5.22: Doses obtained with the Monte Carlo solver (top) and the approximated M_1 model with the original Cartesian (middle) and modified (below) relaxation parameters.

Solver	Monte Carlo	M_1 solver	modified M_1 solver
Computation times	14 hours	49.78699 sec	204.1239 sec

Table 5.11: Computational times with the implicit solver with the Cartesian parameters of relaxation and the modified ones.

The dose along the axis $y = 1$ cm is plotted on Fig. 5.23 and along the axes $x = 2$ cm and $x = 8$ cm on Fig. 5.24.

The dose results with the modified relaxation parameters are much closer to the reference Monte Carlo results. The dose is not diffused in the tranverse direction which is the expected results. The plots along the transverse direction show that the diffusion phenomenum is accurately modeled when using the modified relaxation parameters.

Due to the noise in the Monte Carlo and the normalization by $\max D$, the M_1 dose curves with the modified relaxation parameters are slightly above the Monte Carlo reference.

Remark that the computational time is higher with the modified relaxation parameters. When dividing the operator D in two parts (5.74) in order to enforce the preservation of the realizability, a parameter αId was artificially added on both sides of (5.73). This method stabilizes the algorithm, but it reduces the convergence rate which explains the difference of computational times. The computational times

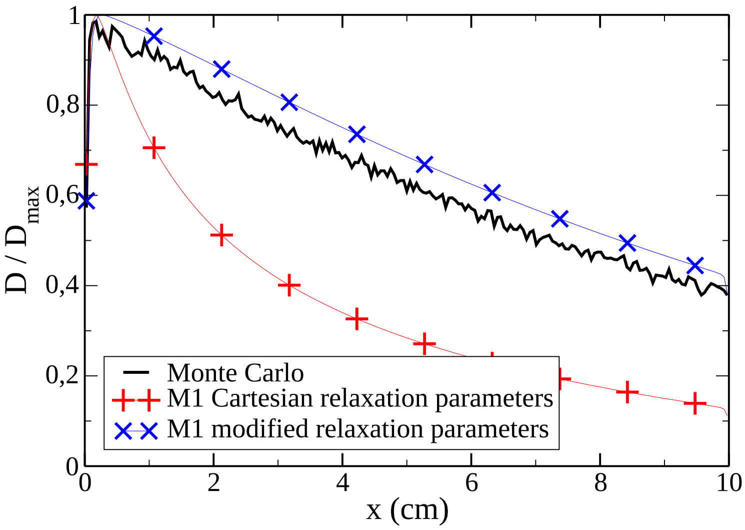

Figure 5.23: Doses obtained with the Monte Carlo solver and the approximated M_1 model with the Cartesian and the modified relaxation parameters along the axis $y = 1$ cm.

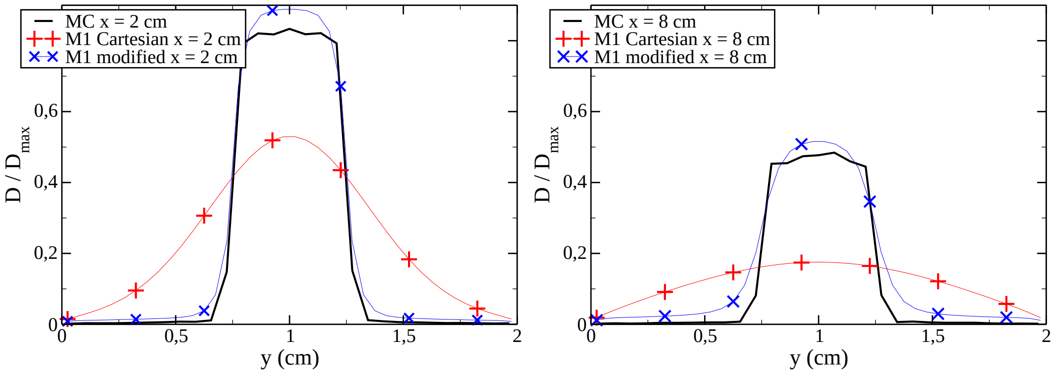

Figure 5.24: Doses obtained with the Monte Carlo solver and the approximated M_1 model with the Cartesian and the modified relaxation parameters along the axes $x = 2$ cm and $x = 8$ cm.

with this method is still much lower than with the Monte Carlo reference.

Bibliography

[1] G. W. Alldredge, C. D. Hauck, D. P. O'Leary, and A. L. Tits. Adaptive change of basis in entropy-based moment closures for linear kinetic equations. *J. Comput. Phys.*, 74(4), february 2014.

[2] G. W. Alldredge, C. D. Hauck, and A. L. Tits. High-order entropy-based closures for linear transport in slab geometry II: A computational study of the optimization problem. *SIAM J. Sci. Comput.*, 34(4):361–391, 2012.

[3] D. Aregba-Driollet and R. Natalini. Discrete kinetic schemes for multidimensional systems of conservation laws. *SIAM J. Numer. Anal.*, 6:1973–2004, 2000.

[4] D. Aregba-Driollet, R. Natalini, and S. Tang. Explicit diffusive kinetic schemes for nonlinear degenerate parabolic systems. *Math. Comp.*, 73:63–94, 2004.

[5] J. Barò, J. Sempau, J.M. Fernández-Varea, and F. Salvat. PENELOPE: An algorithm for Monte Carlo simulation of the penetration and energy loss of electrons in matter. *Nuclear instruments and methods*, 100:31–46, 1995.

[6] J. Barò, J. Sempau, J.M. Fernández-Varea, and F. Salvat. PENELOPE: An algorithm and computer code for Monte Carlo simulation of electron-photon shower. *Ciemat technical report*, pages 31–46, 1996.

[7] C. Berthon, P. Charrier, and B. Dubroca. An HLLC scheme to solve the M_1 model of radiative transfer in two space dimensions. *J. Sci. Comput.*, 31(3):347–389, 2007.

[8] C. Berthon, M. Frank, C. Sarazin, and R. Turpault. Numerical methods for balance laws with space dependent flux: Application to radiotherapy dose calculation. *Commun. Comput. Phys.*, 10(5), 2011.

[9] F. Bouchut. Construction of BGK models with a family of kinetic entropies for a given system of conservation law. *J. Stat. Phys.*, 95(1-2):113–170, 1999.

[10] F. Bouchut. *Nonlinear stability of Finite Volume methods for hyperbolic conservation laws and well-balanced schemes for sources.* Birkhäuser, 2004.

[11] F. Bouchut, F. R. Guarguaglini, and R. Natalini. Diffusive BGK approximations for nonlinear multidimensional parabolic equations. *Indiana Univ. Math. J.*, 49:723–749, 2000.

[12] S. Brull. Discrete coagulation-fragmentation system with transport and diffusion. *Anns. Fac. Sci. Toulouse : Mathématiques*, 17(3):439–460, 2008.

[13] B. Dubroca and M. Frank. An iterative method for transport equations in radiotherapy. *Progress in Industrial Mathematics at ECMI 2008*, pages 407–412, 2010.

[14] J. Sempau F. Salvat, J. M. Fernández-Varea. *PENELOPE-2011: A code system for Monte Carlo of electron and photon transport*, 2011.

[15] A. Harten, P. Lax, and B. Van Leer. On upstream differencing and Gudonov-type schemes for hyperbolic conservation laws. *SIAM Rev.*, 25(1):35–61, 1983.

[16] C. D. Hauck. High-order entropy-based closures for linear transport in slab geometry. *Commun. Math. Sci*, 9(1):187–205, 2011.

[17] R. J. LeVeque. *Finite Volume methods for hyperbolic problems*, volume 31. Cambridge university press, 2002.

[18] J. Mallet, S. Brull, and B. Dubroca. An entropic scheme for an angular moment model for the classical Fokker-Planck-Landau equation of electrons. *Commun. Comput. Phys.*, 15(2):422–450, 2014.

[19] J. Mallet, S. Brull, and B. Dubroca. General moment system for plasma physics based on minimum entropy principle. *Kin. rel. mod.*, 8(3):533–558, 2015.

[20] J. J. Moré, B. S. Garbow, and K. E. Hillstrom. *User guide for MINPACK-1*, 1980. http://www.netlib.org/minpack/.

[21] R. Natalini. A discrete kinetic approximation of entropy solutions to multidimensional scalar conservation laws. *J. Differ. Equations*, 148(2):292 – 317, 1998.

[22] T. Pichard, D. Aregba-Driollet, S. Brull, B. Dubroca, and M. Frank. Relaxation schemes for the M_1 model with space-dependent flux: Application to radiotherapy dose calculation. *Commun. Comput. Phys.*, 19:168–191, 2016.

[23] R. Piessens, E. De Doncker-Kapenga, and C. W. Überhuber. *QUADPACK: A subroutine package for automatic integration*, springer edition, 1983. http://www.netlib.org/quadpack/.

[24] E. Toro. *Riemann solvers and numerical methods for fluid dynamics.* Springer, 2009.

[25] A. van der Sluis. Gershgörin domains for partitioned matrices. *Linear Algebra Appl.*, 26:265 – 280, 1979.

Chapter 6

Dose optimization

6.1 Introduction

The final objective of the present work is to provide a method to optimize the source of particles such that the resulting dose is maximum in the tumor cells and minimum in the rest of the medium.

New emerging techniques in external radiotherapy such as the adaptative radiation therapy (ART, see *e.g.* [11]), the image guided radiation therapy (IGRT, see *e.g.* [12]) or thecommon intensity modulated radiation therapy (IMRT, see *e.g.* [4, 3, 13, 11]) require fast numerical optimization of pre-computed dose between two irradiations.

In particular, several techniques of IMRT present optimization problems which fit with the present framework. One technique of IMRT, called volumetic modulated arc therapy (VMAT), consists in having the source of particles turn around the patient. In this process, the intensity of the source is modulated using blades shading part of the beam. This purpose of this modulation is to obtain a uniform maximum dose deposited in the tumor and a dose below a certain threshold in the organs (see [5] for the standard applied in France on the maximum dose deposited per organ).

The main objective of such an optimization procedure is to find one possible source to irradiate the tumor without damaging the organs at risk (OAR). In practice, the optimal source is not always seeked because of time constraints but good potential dose distributions are sufficient.

In this chapter, basic techniques of optimization (based on [17, 10] and applied to dose optimization *e.g.* in [7, 9, 8, 1]) are presented and adapted to the present radiotherapy problem. Such optimization algorithms typically consist in solving iteratively the direct and the adjoint problem, the computational costs of which may be significant. In order to reduce such computational costs, the optimization procedure is typically based on a coarse spatial mesh. Then the dose with the optimized source resulting of the optimization algorithm is recomputed on a finer mesh. Finally, a physician accepts, or not, *a posteriori* the optimized dose to be delivered.

For the sake of simplicity in the present work, the spectrum of the beam, *i.e.* the energy distribution of the source, is assumed to be optimizable while the direction of the source is fixed and directed to one point in the medium.

The optimization method presented in this chapter is valid only when considering optimization problems under linear PDE constraints. The numerical approach described in the previous chapter is based on the M_N equations which are non-linear, even if the original kinetic equations were linear. In the present chapter, the numerical schemes developped in the previous chapter are used, but they are interpreted here only as a numerical approach, assumed to be accurate enough, for solving linear kinetic equations, and all the optimization procedure is written at the kinetic level, that is an optimize-then-discretize method.

6.2 Problem statement

The particle motion is governed by the kinetic equation (1.11) with a LBCSD collision operator (1.34) modelling the electron collisions. For writing purposes, this is rewritten under the form

$$\left\{\begin{array}{rcll} A\psi - Q\psi &=& 0 & \text{in the interior } [\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2, \\ \psi &=& \psi^b & \text{on the boundary } [\epsilon_{\min},\epsilon_{\max}]\times \Gamma^-, \\ \psi(\epsilon_{\max},x,\Omega) &=& 0 & \text{in } Z\times S^2, \end{array}\right. \tag{6.1a}$$

where $\psi = (\psi_e,\psi_\gamma)^T$ and $\psi^b = (\psi^b_e,\psi^b_\gamma)^T$ and the operators A and Q are given by

$$A\psi = \left(\Omega.\nabla_x\psi_\gamma, \quad \Omega.\nabla_x\psi_e\right)^T, \tag{6.1b}$$

$$Q\psi = \left(\rho(G_{\gamma\to\gamma}-P_\gamma)(\psi_\gamma), \quad \rho(Q_{LBCSD}(\psi_e)+G_{\gamma\to e}(\psi_\gamma))\right)^T, \tag{6.1c}$$

and the operator Q_{LBCSD}, $G_{\alpha\to\beta}$ and P_α are recalled here

$$Q_{LBCSD}(\psi_e) = \partial_\epsilon(S\psi_e) + (G_{e\to e}-P_e)(\psi_e), \tag{6.1d}$$

$$G_{\alpha\to\beta}(\psi_\alpha) = \int_\epsilon^{\epsilon_{\max}}\int_{S^2}\sigma_{\alpha\to\beta}(\epsilon',\epsilon,\Omega'\Omega)\psi_\alpha(\epsilon',x,\Omega')d\Omega'd\epsilon', \tag{6.1e}$$

$$P_\alpha(\psi_\alpha) = \sigma_{T,\alpha}\psi_\alpha. \tag{6.1f}$$

The aim of this chapter is to define a theoretical framework for optimization and to propose a numerical method to optimize a source ψ^b such that the resulting dose is as close as possible to an optimal dose $\bar{D}\in L^2(Z)$ *a priori* given. More biologically relevant objectives are available. For instance, the linear-quadratic model [15] describes the impact of radiations on cells through the fraction of cells surviving to the treatement. An meaningful objective consists in maximizing the fraction of tumor cells killed while minimizing the fraction of healthy tissue damaged, this optimization problem is studied in [1].

The objective functional J is the function to minimize. It is chosen to be the $L^2(Z)$ distance to the optimal dose $\bar{D}$ with an additional regularization term. It has the form

$$\begin{aligned} J(\psi,\psi^b) &= \int_Z \frac{c_D(x)}{2}\left(D(\psi)-\bar{D}\right)^2(x)dx \qquad (6.2)\\ &+ \int_{\epsilon_{\min}}^{\epsilon_{\max}}\int_{\Gamma^-}\frac{\Omega.n(x)}{2}\Big(c_\gamma^b(x,\epsilon,\Omega)(\psi_\gamma^b)^2(\epsilon,x,\Omega)^2\\ &\qquad +c_e^b(x,\epsilon,\Omega)(\psi_e^b)^2(\epsilon,x,\Omega)\Big)\,d\Omega dx d\epsilon, \end{aligned}$$

where the dose is written $D(\psi)$ as a function of ψ and it is defined in (1.64). The coefficients $c_D > 0$ and $c_\alpha^b > 0$ (for $\alpha = \gamma, e$) are positive scalars chosen according to the problem to solve and will be defined later.

The optimization method presented in this chapter is written in a general framework where several parameters are left free and optimizable. However, for practical applications, there are commonly more constraints on the sources. One may easily reproduce the computations and the techniques presented here with different constraints on the source ψ^b. In the numerical experiments of Section 6.5, the sources are chosen to be always directed to one point P inside the medium, and to be composed only of electrons (it is fixed at $\psi^b = (0,\psi_e^b)$) or only of photons (it is fixed at $\psi^b = (\psi_\gamma^b, 0)$). This can be interpreted either by a particular choice of the optimization parameter or by a straightforward adaptation of the method presented here.

6.3 Preliminaries

First, some notations are defined. Then the control-to-state operator sending a source ψ^b onto the solution ψ of (6.1) is studied, and a first existence result of a solution to the optimization problem is provided.

6.3.1 Notations

The following inner products defined in Chapter 1 are recalled

$$\begin{aligned} (\psi_\alpha,\lambda_\alpha)_i &= \int_{\epsilon_{\min}}^{\epsilon_{\max}}\int_Z\int_{S^2}\psi_\alpha(\epsilon,x,\Omega)\lambda_\alpha(\epsilon,x,\Omega)d\Omega dx d\epsilon,\\ (\psi_\alpha^b,\lambda_\alpha^b)_{b_-} &= \int_{\epsilon_{\min}}^{\epsilon_{\max}}\int_{\Gamma^-}|\Omega.n(x)|\psi_\alpha^b(\epsilon,x,\Omega)\lambda_\alpha^b(\epsilon,x,\Omega)d\Omega dx d\epsilon,\\ (\psi_\alpha^b,\lambda_\alpha^b)_{b_+} &= \int_{\epsilon_{\min}}^{\epsilon_{\max}}\int_{\Gamma^+}|\Omega.n(x)|\psi_\alpha^b(\epsilon,x,\Omega)\lambda_\alpha^b(\epsilon,x,\Omega)d\Omega dx d\epsilon, \end{aligned}$$

and by extension the following inner products and norms are defined

$$(\psi,\lambda)_I = (\psi_\gamma,\lambda_\gamma)_i + (\psi_e,\lambda_e)_i, \qquad (\psi^b,\lambda^b)_{B_-} = (\psi_\gamma^b,\lambda_\gamma^b)_{b_-} + (\psi_e^b,\lambda_e^b)_{b_-},$$
$$(\psi^b,\lambda^b)_{B_+} = (\psi_\gamma^b,\lambda_\gamma^b)_{b_+} + (\psi_e^b,\lambda_e^b)_{b_+},$$
$$\|\psi_\alpha\|_i = \sqrt{(\psi_\alpha,\psi_\alpha)_i}, \qquad \|\psi_\alpha^b\|_{b_-} = \sqrt{(\psi_\alpha^b,\psi_\alpha^b)_{b_-}}, \qquad \|\psi_\alpha^b\|_{b_+} = \sqrt{(\psi_\alpha^b,\psi_\alpha^b)_{b_+}},$$
$$\|\psi\|_I = \sqrt{(\psi,\psi)_I}, \qquad \|\psi^b\|_{B_-} = \sqrt{(\psi,\psi)_{B_-}}, \qquad \|\psi^b\|_{B_+} = \sqrt{(\psi,\psi)_{B_+}}.$$

The well-posedness of the problem (6.1) was shown in Theorem 1.3 under conditions on the the physical parameters S, $\sigma_{T,\gamma}$, $\sigma_{T,e}$, $\sigma_{\gamma\to\gamma}$, $\sigma_{\gamma\to e}$ and $\sigma_{e\to e}$.

Assumption 6.1 *In all this chapter, the density ρ, the stopping power S and the cross sections $\sigma_{\gamma\to\gamma}$, $\sigma_{\gamma\to e}$, $\sigma_{T,\gamma}$, $\sigma_{e\to e}$ and $\sigma_{T,e}$ are assumed to satisfy the requirements for the problem (6.1) to be well-posed.*

Furthermore, the solution ψ of (6.1) is assumed to be non-negative ($\psi_\gamma \geq 0$ and $\psi_e \geq 0$) as long as the source ψ^b is non-negative ($\psi_\gamma^b \geq 0$ and $\psi_e^b \geq 0$).

6.3.2 The control-to-state mapping

For writing purposes, the following operator is defined.

Notation 6.1 *The operator Ξ, afterward called control-to-state mapping, is the operator sending a source ψ^b onto the solution ψ of (6.1)*

$$\Xi := \begin{cases} (L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-))^2 & \to & (L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2))^2, \\ \psi^b & \mapsto & \psi. \end{cases} \tag{6.3}$$

Remark that in practice, the image set of Ξ, *i.e.* the set of solutions ψ of (6.1) is a subset of $(L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2))^2$ as each solution ψ of (6.1) satisfy

$$\psi \geq 0, \qquad \|(A-Q)(\psi)\|_I = 0, \qquad \|\psi\|_B < \infty \qquad \text{and} \qquad \psi(\epsilon_{\max}, x, \Omega) = 0.$$

The following properties of the control-to-state mapping Ξ are required in order to prove the existence of a unique optimal source.

Proposition 6.1 *The operator Ξ is continuous linear and bounded.*

Proof This was proven through Theorem 1.3 under the conditions (1.47), (1.57a) and (1.57b) on the stopping power S and the cross sections $\sigma_{T,\gamma}$, $\sigma_{T,e}$, $\sigma_{\gamma\to\gamma}$, $\sigma_{\gamma\to e}$ and $\sigma_{e\to e}$. See also [6, 7, 16]. □

On can rewrite the objective functional using this operator.

Notation 6.2 *The minimization of J is studied only under the constraints (6.1). Thus, the function to minimize is the reduced objective functional j*

$$j(\psi^b) := J\left(\Xi(\psi^b), \psi^b\right).$$

6.3.3 Existence and uniqueness of a minimizer

Standard techniques of the litterature (see *e.g.* [17, 10]) provide the existence of a unique optimal solution to the problem (6.1).

Define two positive functions $U_\gamma > 0$ and $U_e > 0$ over $[\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-$ and denote

$$(L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-))^2_{ad} = \Big\{ \psi^b \in (L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-))^2, \\ \text{s.t.} \quad \text{for} \quad \alpha = \gamma, e, \quad 0 \leq \psi^b_\alpha \leq U_\alpha \quad \text{on} \ \in [\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^- \Big\}.$$

The following existence result provides the existence of a solution to the optimization problem.

Theorem 6.1 *[17] The functional j has a unique minimizer in the admissible set $(L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-))^2_{ad}$.*

Proof The operator Ξ is continuous linear and bounded (according to Proposition 6.1) and the reduced objective functional j is convex and coercive (quadratic) over the non-empty closed convex subset $(L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-))^2_{ad}$ of the Hilbert space $(L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-))^2$. Then the result follows from Theorem 2.14 of [17]. □

6.4 Computing the optimal solution

A method to compute the optimal source $\psi^b \in (L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-))^2_{ad}$ minimizing the convex functional j under the PDE constraint (6.1) is seeked. Numerical method to optimize such a source typically requires the computation of the Fréchet derivative of the objective functional j. For this purpose, the Lagrange multiplier method is used. Define the Lagrangian

$$L(\psi, \psi^b, \lambda, \lambda^b) = J(\psi, \psi^b) + ((A - Q)(\psi), \lambda)_I + \left(\psi - \psi^b, \lambda^b\right)_{B_-}, \tag{6.4}$$

associated to the problem of minimizing J under the constraint (6.1).

In order to find the optimal solution $\bar{\psi}^b$, one typically computes the Fréchet derivative of L according to each of its variables ψ, ψ^b, λ and λ^b, obtains the derivative of j after those computations and then finds a descent direction.

6.4.1 The adjoint equation

In order to compute the Fréchet derivative of L, the adjoint of the operator A and Q are computed.

The adjoint of the advection operator

Computing the adjoint of the operator A reads

$$(A\psi, \lambda)_I = (\psi, A^T\lambda)_I - (\psi, \lambda)_{B_-} + (\psi, \lambda)_{B_+},$$

where the operator A^T is given by

$$A^T\lambda \quad = \quad \big(-\Omega.\nabla_x\lambda_\gamma, \quad -\Omega.\nabla_x\lambda_e \big)^T.$$

The adjoint of the collision operator

Computing the adjoint of Q reads

$$(Q\psi,\lambda)_I = (\psi, Q^T\lambda)_I - \int_Z\int_{S^2} S(\epsilon_{\min})\psi_e(\epsilon_{\min},x,\Omega)\lambda_e(\epsilon_{\min},x,\Omega)d\Omega dx,$$

where the operator Q^T is given by

$$Q^T\lambda \quad = \quad \left(\rho(G^T_{\gamma\to\gamma} - P^T_\gamma)(\lambda_\gamma) + \rho G^T_{\gamma\to e}(\lambda_e), \quad \rho Q^T_{LBCSD}(\lambda_e)\right)^T,$$

and the operators Q^T_{LBCSD}, $G^T_{\alpha\to\beta}$ and P^T_α read

$$\begin{aligned} Q^T_{LBCSD}(\lambda_e)(\epsilon,x,\Omega) &= -S(\epsilon)\partial_\epsilon\lambda_e(\epsilon,x,\Omega) + (G^T_{e\to e} - P^T_e)(\lambda_e)(\epsilon,x,\Omega), \\ G^T_{\alpha\to\beta}(\lambda_\beta)(\epsilon,x,\Omega) &= \int_{\epsilon_{\min}}^{\epsilon}\int_{S^2}\sigma(\epsilon,\epsilon',\Omega.\Omega')\lambda_\beta(\epsilon',x,\Omega')d\Omega'd\epsilon', \\ P^T_\alpha(\lambda_\alpha)(\epsilon,x,\Omega) &= P_\alpha(\lambda_\alpha)(\epsilon,x,\Omega). \end{aligned}$$

The adjoint equation

Define the following problem

$$\begin{cases} A^T\lambda - Q^T\lambda = q & \text{in the interior } [\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2, \\ \lambda = 0 & \text{on the boundary } [\epsilon_{\min},\epsilon_{\max}]\times\Gamma^+, \\ \lambda(\epsilon_{\min},x,\Omega) = 0 & \text{for } (x,\Omega)\in Z\times S^2. \end{cases} \tag{6.5}$$

Similarily to Theorem 1.3, one can prove the following.

Theorem 6.2 *Suppose that the source q satisfies*

$$q \in (L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2))^2,$$

and that Assumption 6.1 is valid.

Then the problem (6.5) has a unique solution λ satisfying

$$\begin{aligned} \lambda &\in (L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2))^2, \\ (A^T - Q^T)(\lambda) &\in (L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2))^2, \\ \lambda|_{[\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-} &\in (L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-))^2, \\ \lambda_e(\epsilon_{\max},.,.) &\in L^2(Z\times S^2). \end{aligned}$$

Proof This can be proven by adaptating the proof of Theorem 1.3. See also [6] and applications to radiotherapy problems in [7, 16]. □

For writing purposes, the following operators are defined.

Notation 6.3 *Similarily to the control-to-state mapping, define the operator E that sends a source $q \in (L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2))^2$ onto the solution λ of (6.5)*

$$E := \begin{cases} (L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2))^2 & \to \quad (L^2([\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2))^2, \\ q & \mapsto \quad \lambda, \end{cases} \tag{6.6}$$

and the operator Ξ^* adjoint to Ξ

$$\Xi^* := \left\{ \begin{array}{rcl} (L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2))^2 & \to & (L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-), \\ q & \mapsto & \lambda|_{[\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-})^2. \end{array} \right. \tag{6.7}$$

According to their definitions, the operators Ξ^* and E are related to each other through

$$\forall q \in (L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2))^2, \qquad \Xi^*(q) = E(q)|_{[\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-}.$$

See further study of the trace operator *e.g.* in [16].

6.4.2 Computing the derivative of j

The Lagrange multiplier method provides a method to compute the Fréchet derivative of j. In the following, d_X denotes the Fréchet derivative according to X.

Fréchet derivating the Lagrangian (6.4) according to λ, λ^b, ψ and ψ^b read

$$\begin{array}{rcl} d_\lambda L(\psi, \psi^b, \lambda, \lambda^b)(h) & = & ((A-Q)(\psi),\ h)_I\,, \\ d_{\lambda^b} L(\psi, \psi^b, \lambda, \lambda^b)(h) & = & \Big(\psi - \psi^b,\ h\Big)_{B_-}\,, \\ d_\psi L(\psi, \psi^b, \lambda, \lambda^b)(h) & = & \Big((A^T - Q^T)(\lambda) + c_D c_\epsilon (D(\psi) - \bar{D}),\ h\Big)_I - (\lambda - \lambda^b,\ h)_{B_-} \\ & & +(\lambda,\ h)_{B_+} - \displaystyle\int_Z \int_{S^2} S(\epsilon_{\min}) h(\epsilon_{\min}, x, \Omega) \lambda(\epsilon_{\min}, x, \Omega) d\Omega dx, \\ d_{\psi^b} L(\psi, \psi^b, \lambda, \lambda^b)(h) & = & (c^b \psi^b - \lambda^b,\ h)_{B_-}, \end{array}$$

where

$$c^b = \begin{pmatrix} c^b_\gamma & 0 \\ 0 & c^b_e \end{pmatrix}$$

and the coefficient c_ϵ is computed by derivating the dose (1.64) according to ψ, and using (1.31) leads to

$$\begin{array}{rcl} c_\epsilon(\epsilon) & = & \Big(\displaystyle\int_{\epsilon_B}^{\epsilon} \int_{S^2} (\epsilon - \epsilon') \sigma_{C,\gamma}(\epsilon, \epsilon', \Omega.\Omega') + \epsilon' \sigma_{C,e}(\epsilon, \epsilon', \Omega.\Omega') d\Omega' d\epsilon', \\ & & \qquad S(\epsilon) + \displaystyle\int_{\epsilon_B}^{\epsilon} \int_{S^2} \epsilon' \sigma_{M,2}(\epsilon, \epsilon', \Omega.\Omega') d\Omega' d\epsilon \Big). \end{array}$$

Fixing $d_\lambda L = 0$ and $d_{\lambda^b} L = 0$ means imposing the constraints (6.1), that is $\psi = \Xi(\psi^b)$.

The remaining terms leads to the computation of the derivative of j. Computing the derivative of L according to ψ and ψ^b read

$$\begin{array}{rcl} d_\psi L(\psi, \psi^b, \lambda, \lambda^b)(h) & = & d_\psi J(\Xi(\psi^b), \psi^b)(h), \\ d_{\psi^b} L(\psi, \psi^b, \lambda, \lambda^b)(h) & = & d_{\psi^b} J(\Xi(\psi^b), \psi^b)(h), \end{array}$$

and computing the derivative of j read

$$d_{\psi^b} j(\psi^b)(h) \quad = \quad d_{\psi^b} J(\Xi(\psi^b), \psi^b)(h) \ + \ d_\psi J(\Xi(\psi^b), \psi^b)\ d_{\psi^b} \Xi(\psi^b)(h). \tag{6.8}$$

In order to compute the derivative $d_{\psi^b} j(\psi^b)$, compute ψ, λ and λ^b such that

$$d_\lambda L(\psi,\psi^b,\lambda,\lambda^b)=0,\quad d_{\lambda^b} L(\psi,\psi^b,\lambda,\lambda^b)=0,\quad d_\psi L(\psi,\psi^b,\lambda,\lambda^b)=0, \tag{6.9}$$

then (6.8) can be rewritten

$$d_{\psi^b} j(\psi^b)(h) = d_{\psi^b} L(\psi,\psi^b,\lambda,\lambda^b)(h) = (c^b\psi^b-\lambda^b,\ h)_{B_-}. \tag{6.10}$$

The optimal source ψ^b is characterized by the following theorem.

Theorem 6.3 *The unique admissible source* $\bar{\psi}^b \in (L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-))^2_{ad}$ *minimizing* j *satisfies*

$$\forall \psi^b \in (L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-))^2_{ad}, \qquad d_{\psi^b} j(\bar{\psi}^b)(\bar{\psi}^b-\psi^b)\le 0. \tag{6.11}$$

Proof This theorem is Lemma 2.21 in [17] reformulated for the present radiotherapy problem. See also [2, 10]. □

The equations (6.9) and (6.11) form the so-called first order necessary optimality conditions. They can be rewritten under the following form

$$\left\{\begin{array}{rcll} (A-Q)(\bar\psi)(\epsilon,x,\Omega) &=& 0 & \text{for } (\epsilon,x,\Omega)\in[\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2,\\ \bar\psi(\epsilon,x,\Omega) &=& \bar\psi^b(\epsilon,x,\Omega) & \text{for } (\epsilon,x,\Omega)\in[\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-,\\ \bar\psi(\epsilon_{\max},x,\Omega) &=& 0 & \text{for } (x,\Omega)\in Z\times S^2, \end{array}\right. \tag{6.12a}$$

$$\left\{\begin{array}{rcll} (A^T-Q^T)(\bar\lambda)(\epsilon,x,\Omega) &=& c_D(x)c_\epsilon(\epsilon)\left(\bar D - D(\bar\psi)\right)(x) & \\ & & & \text{for } (\epsilon,x,\Omega)\in[\epsilon_{\min},\epsilon_{\max}]\times Z\times S^2,\\ \bar\lambda(\epsilon,x,\Omega) &=& 0 & \text{for } (\epsilon,x,\Omega)\in[\epsilon_{\min},\epsilon_{\max}]\times\Gamma^+,\\ \bar\lambda(\epsilon_{\min},x,\Omega) &=& 0 & \text{for } (x,\Omega)\in Z\times S^2, \end{array}\right. \tag{6.12b}$$

$$\bar\lambda^b = \bar\lambda|_{[\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-}, \tag{6.12c}$$

$$\forall\psi^b\in(L^2([\epsilon_{\min},\epsilon_{\max}]\times\Gamma^-))^2_{ad}, \qquad (c^b\bar\psi^b-\bar\lambda^b,\bar\psi^b-\psi^b)_{B_-}\le 0. \tag{6.12d}$$

6.4.3 A projected gradient method

The numerical approach is based on an iterative method. At each iteration, a descent direction is seeked.

Definition 6.1 *The direction* g *defined by*

$$g := -(c^b\psi^b-\lambda^b)$$

is called the steepest descent direction.

Using (6.10), one observes that this direction satisfies

$$d_{\psi^b} j(\psi^b)(g) = -\|g\|_I^2 \leq 0.$$

Therefore, using the definition of the derivative of j, one obtains

$$\exists \delta > 0 \quad \text{s.t.} \quad \frac{j(\psi^b + \delta g) - j(\psi^b)}{\delta} \leq 0,$$

and in particular the value of j at the point $\psi^b + \delta g$ is lower than at the point ψ^b.

For writing purposes, the following notation is defined.

Notation 6.4 *The projection of a function $\psi^b \in (L^2([\epsilon_{\min}, \epsilon_{\max}] \times \gamma^-))^2$ into the admissible set $(L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-))^2_{ad}$ is denoted*

$$\underset{ad}{proj}\ \psi^b(\epsilon, x, \Omega) = \left(\underset{[0, U_\gamma(\epsilon,x,\Omega)]}{proj}\ \psi^b_\gamma(\epsilon, x, \Omega), \quad \underset{[0, U_e(\epsilon,x,\Omega)]}{proj}\ \psi^b_e(\epsilon, x, \Omega) \right),$$

for all (ϵ, x, Ω) in $[\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-$, and where $\underset{[a,b]}{proj}\ c$ denotes the projection of c onto the interval $[a, b]$

$$\underset{[a,b]}{proj}\ c = \begin{cases} a & \text{if} \quad c < a, \\ c & \text{if} \quad c \in [a, b], \\ b & \text{if} \quad c > b. \end{cases}$$

The following algorithm was proven to converge to the desired optimal solution in [17, 10, 2].

Algorithm 6.1 (Projected gradient, [2])

Initialization: *Provide an initial source $(\psi^b)^{(0)}$.*

Iteration: *Iterate the following steps 1 to 6 until the following residual $r^{(k)}_{opt}$ reaches a desired value*

$$r^{(k)}_{opt} = \left\| c^b(\psi^b)^{(k)} - \lambda^{(k)} \right\|_{B_-} \leq r_{opt,\max},$$

or until k reaches a maximum value $k_{opt,\max}$.

1. *Compute the fluence*

$$\psi^{(k+1)} = \Xi\left((\psi^b)^{(k)} \right) \tag{6.13}$$

by solving (6.12a).

2. *Compute the dose*

$$D^{(k+1)} = D\left(\psi^{(k+1)} \right).$$

3. *Compute the ajoint*

$$\lambda^{(k+1)} = E\left(c_\epsilon c_D\left(\bar{D} - D^{(k+1)}\right)\right), \tag{6.14a}$$

   by solving (6.12b).

4. *Compute the steepest descent direction*

$$g^{(k+1)} = -\left(c^b(\psi^b)^{(k)} - \lambda^{(k+1)}\right).$$

5. *Find the optimal step size* $s^{(k+1)}$*, i.e. the scalar in* $]0,1]$ *such that*

$$\begin{aligned} s^{(k+1)} &= \underset{s\in]0,1]}{\mathit{argmin}} \; j^{(k+1)}(s), \\ j^{(k+1)}(s) &:= j\left(\underset{ad}{\mathit{proj}}\left((\psi^b)^{(k)} + sg^{(k+1)}\right)\right). \end{aligned}$$

6. *Compute the new source*

$$(\psi^b)^{(k+1)} = \underset{ad}{\mathit{proj}}(\psi^b)^{(k)} + s^{(k+1)}g^{(k+1)}.$$

In practice, Step 5 of algorithm 6.1 requires some more computations because computing the optimal step size is non trivial. For simplicity here, a bisection algorithm is used.

Algorithm 6.2 (Bisection)
Initialization: *Set*

$$s^{(k+1,0)} = 0, \qquad s^{(k+1,1)} = \frac{1}{2}.$$

Iteration: *Iterate*

$$s^{(k+1,n+1)} = \begin{cases} s^{(k+1,n)} + \frac{1}{2^{n+1}} & \text{if} \quad j^{(k+1)}\left(s^{(k+1,n-1)}\right) > j^{(k+1)}\left(s^{(k+1,n)}\right) \\ s^{(k+1,n)} - \frac{1}{2^{n+1}} & \text{otherwise} \end{cases}$$

until the number of iterations n *reaches a maximum value* $n_{bis,\max}$ *or until the following residual*

$$r_{bis}^{(n)} = \left| j^{(k+1)}\left(s^{(k+1,n+1)}\right) - j^{(k+1)}\left(s^{(k+1,n)}\right) \right|$$

reaches a threshold $r_{bis,\max}$.

6.4.4 Numerical approximations

The M_N models presented in Chapter 4 and the numerical schemes of Chapter 5 require several approximations to be applied to the present optimization problem.

Solving the direct and adjoint equations

At each iteration of Algorithm 6.1, the direct equation (6.1) is solved, *i.e.* $\Xi(\psi^b)$ is computed for some source $\psi^b \in (L^2([\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-))^2_{ad}$, and the adjoint equation (6.5) is solved, *i.e.* $E(q)$ is computed for some source $q \in (L^2([\epsilon_{\min}, \epsilon_{\max}] \times Z \times S^2))^2$.

The adjoint equation (6.5) has a similar form as the direct equation (6.1) and one easily adapts the schemes presented in Chapter 5 to the adjoint equation. However, one remarks that the source term $c_\epsilon c_D \left(\bar{D} - D^{(k+1)}\right)$ in the adjoint equation (6.5) is non-signed. The M_1 and M_2 models described in Chapter 4 are only valid under realizability condition which equivals to requiring the positivity of the solution to the kinetic equation. In order to use the M_N models on the adjoint equation, such equation is solved separately for the positive and negative part of the source in (6.5), *i.e.*

$$\lambda = \lambda^+ - \lambda^-, \tag{6.15a}$$

$$\lambda^+ = E\left(c_\epsilon c_D \max\left(0, \bar{D} - D^{(k+1)}\right)\right), \tag{6.15b}$$

$$\lambda^+ = E\left(c_\epsilon c_D \max\left(0, -(\bar{D} - D^{(k+1)})\right)\right). \tag{6.15c}$$

The scheme (5.58) is used to solve (6.1) and an adaptation of this scheme is also used to solve the adjoint equation under the form (6.15).

In practice, the M_1 or the M_2 equations are solved instead of the kinetic equations. The M_N equations are non-linear while the kinetic equations are linear and the optimization process described in the present chapter requires linear PDE constraints. Thus, the moment method and the discretization (5.58) are used here only as a numerical approximation of the operators Ξ and E. In practice, the present method using the M_N model can not converge to the desired solution. However, these approximations are assumed to be sufficiently accurate approximations of the kinetic solution for the desired accuracy of the optimal source ψ^b to be attainable.

The boundary conditions

In practice, it is difficult to relate directly the kinetic boundary conditions to the moment ones (see discussion in Subsection 2.5.2).

For simplicity, at the discrete level, the boundary conditions

$$\psi = \psi^b \quad \text{on} \quad [\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-, \qquad \lambda = 0 \quad \text{on} \quad [\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^+$$

are replaced by

$$\boldsymbol{\psi}^n_{l,m} = \int_{S^2} \mathbf{m}(\Omega) \tilde{\psi}^b(\epsilon^n, X_{l,m}, \Omega) d\Omega, \qquad \boldsymbol{\lambda}^n_{l,m} = 0_{\mathbb{R}^{Card(\mathbf{m})}}, \qquad \text{for } X_{l,m} \in \partial Z,$$

where

$$\tilde{\psi}^b(\epsilon, x, \Omega) = \begin{cases} \psi^b(\epsilon, x, \Omega) & \text{on } [\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^-, \\ 0 & \text{on } [\epsilon_{\min}, \epsilon_{\max}] \times \Gamma^+, \end{cases}$$

These approximations are assumed to be accurate enough for the present applications.

6.5 Numerical experiments

Two test cases are studied. The first one is a 1D test where a source of electrons alone (without photons) is optimized in a 1D medium composed of water ($\rho = 1$). In the second test, a source of photons alone (without electrons) is optimized in a 2D medium.

The coefficients c^b_α used in the regularization term of the objective functional (6.2) are fixed at $c^b_\alpha = 10^{-3}$ (for $\alpha = \gamma, e$) in the whole domain. In order to simulate more practical problems, three types of media are differentiated which lead to fix the coefficient c_D and the optimal dose $\bar{D}$:

- the tumor (T) regions : $c_D(x) = c_{D,T} = 20 \qquad \bar{D}(x) = 10^{10}$,
- the organs at risk (OAR) : $c_D(x) = c_{D,OAR} = 100 \qquad \bar{D}(x) = 0$,
- the healthy (H) tissues : $c_D(x) = c_{D,H} = 1 \qquad \bar{D}(x) = 0$.

The constraints U are chosen such that all the beams on the boundary of the medium target a chosen point P in the medium. In practice, this corresponds to choosing the following projection

$$\begin{aligned} \underset{ad}{\mathrm{proj}}(\psi^b)(x,\epsilon,\Omega) &= (\psi^b_\gamma, \psi^b_e)(x,\epsilon,\Omega)\delta\left(\Omega - d\right), \\ d = \frac{P-x}{\|P-x\|}, \end{aligned} \tag{6.16}$$

which can be interpreted at the moment level by imposing the following projection on the moment $(\psi^b_\alpha)^i$ of order i of ψ^b_α

$$\underset{\mathcal{R}^{U_\alpha}_{\mathbf{m}}}{\mathrm{proj}}(\psi^b_\alpha)^i = \begin{cases} ((\psi^b_\alpha)^1.d)d^{\otimes i} & \text{if} \qquad ((\psi^b_\alpha)^1.d) > 0, \\ 0 \quad \text{otherwise} \end{cases}$$

In Subsection 4.5.2, artificial effects caused by the M_N models when multiple beams cross each others were exhibited. This problem emerging with the non-linear M_N models was addressed in Subsections 4.5.2, 5.3.2 and 5.3.4 by solving (6.13) for each source of particles and then sum their influences together. The same method can be used here. Once the source $(\psi^b)^{(k)}$ is computed, one can solve (6.13) for each source seperately. In 1D, this corresponds to computing the doses obtained with the sources at each end separately, as in Subsection 5.3.2. In 2D, the boundary is splitted into the four sides of the medium, and the dose obtained with the source on each side of the medium is computed seperately. Remark that several sources emerging from the same side still cross each other in the same point, so this method does not solve the multi-beam problem, but this only reduces its effect. This method is afterward referred to as multi-M_N model. One may decompose the boundary differently depending on the expectations on the optimal sources.

6.5.1 1D electron source optimization

A 10 cm long 1D medium composed of uniform water ($\rho = 1$) is meshed with 80 cells. The energy bounds are fixed at $\epsilon_{\max} = 12$ MeV and $\epsilon_{\min} = 10^{-3}$ MeV, and the energy interval $[\epsilon_{\min}, \epsilon_{\max}]$ is meshed such that

$$\epsilon^0 = \epsilon_{\max}, \qquad \epsilon^{n+1} = \epsilon^n - S(\epsilon^n)\Delta x.$$

The target point P is located inside the medium, which means that the sources at both ends are always directed inside the medium.

The tumors are located in the slabs $[1 \text{ cm}, 2 \text{ cm}]$ and $[8 \text{ cm}, 9 \text{ cm}]$, an organ at risk is located in the middle of the medium $[4.5 \text{ cm}, 5.5 \text{ cm}]$ and the rest of the medium is composed of healthy tissues.

Only electrons are used in this test case, the source of photons is always fixed at 0. The initial boundary conditions for electrons are given by

$$\begin{aligned}
(\boldsymbol{\psi}^b)_0^n &= 10^{10} \exp\left(-\alpha_\epsilon \left(\epsilon^n - \epsilon_0\right)^2\right) \left\langle \mathbf{m}(\mu) \exp\left(-\alpha_\mu \left(\mu - 1\right)^2\right)\right\rangle, \\
(\boldsymbol{\psi}^b)_{l_{\max}}^n &= 10^{10} \exp\left(-\alpha_\epsilon \left(\epsilon^n - \epsilon_0\right)^2\right) \left\langle \mathbf{m}(\mu) \exp\left(-\alpha_\mu \left(\mu + 1\right)^2\right)\right\rangle.
\end{aligned}$$

with $\epsilon_0 = 10$ MeV, $\alpha_\mu = 1000$ and $\alpha_\epsilon = 200$. It models two opposite beams of 10 MeV crossing each other.

The maximum residual for the implicit solver (5.58) is fixed at $r_{\max} = 10^5$ and the maximum number of iterations is fixed at $k_{\max} = 500$. The residual for the bisection algorithm 6.2 is fixed at $r_{bis,\max} = 10^5$ and the maximum number of iterations at $n_{bis,\max} = 30$. Finally the residual for the optimization algorithm 6.1 is fixed at $r_{opt,\max} = 5.10^5$ and a maximum number of iterations $k_{opt,\max} = 20$.

The doses obtained in the end of this optimization process with the multi-M_1 and the multi-M_2 model are represented on Fig. 6.1. The computational times to obtain those results and the total number of times the equations (6.13) and (6.14) were solved numerically are gathered in Table 6.1. The spectra (energy distribution) of the optimized sources with the two methods are represented on Fig. 6.2.

Models	multi-M_1	multi-M_2
Computation times	23.280 sec	68.424 sec
Total number of equation solved	611	537

Table 6.1: Computational times to obtain the optimized doses with the approximated multi-M_1 and multi-M_2 models.

The optimized doses obtained with the multi-M_1 and the multi-M_2 models have very similar shapes. The highest values of the dose are located in the middle of the two tumors, while the minimum is in the middle of the medium, in the organ at risk. The value of the dose in the organ at risk is around 30% of the maximum dose. It is therefore non-zero but at the minimum value of the dose in the medium.

The spectra of the optimized sources with the two models also have very similar shapes. Contrarily to the different sources used in the test cases of Chapter 5,

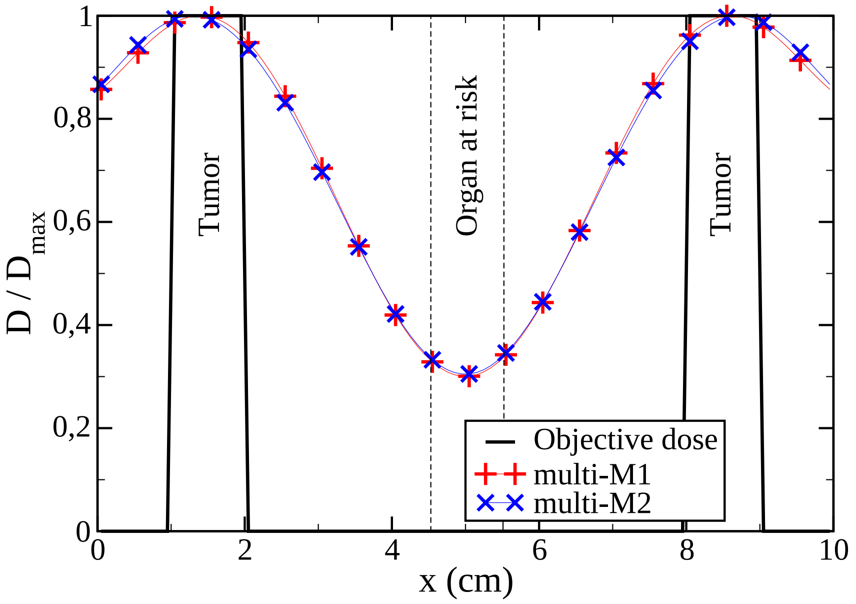

Figure 6.1: Optimized doses obtained with the approximated multi-M_1 and multi-M_2 models.

the optimized sources have very smooth spectra which are minimum in 0 MeV and maximum around 11 MeV. Therefore most of the electrons are injected with a high energy, but the quantity of electrons of lower energy injected and their impact on the dose are non-negligible.

In Subsection 5.3.2, the doses obtained with the M_1 and M_2 models were found to be different when the sources for the two models were identical and modelling a beam of 10 MeV electrons. Further experiments showed that such discrepancies vanish when considering sources of lower energy. This explains the similarities in the doses with the two models on Fig. 6.1 when considering the diffused optimized sources.

6.5.2 2D photon source optimization

The following test case is inspired of one proposed *e.g.* in [14, 1].

A 40 cm × 40 cm 2D medium composed of uniform water ($\rho = 1$) is meshed with 40 × 40 cells. The energy bounds are fixed at $\epsilon_{\text{max}} = 1.2$ MeV and $\epsilon_{\text{min}} = 10^{-3}$ MeV, and the energy interval $[\epsilon_{\text{min}}, \epsilon_{\text{max}}]$ is meshed such that

$$\epsilon^0 = \epsilon_{\text{max}}, \qquad \epsilon^{n+1} = \epsilon^n - 0.01 S(\epsilon^n)\Delta x.$$

The medium contains a C-shaped tumor, an organ at risk enveloped in the C-shaped tumor and the rest of the medium is composed of healthy tissues. The

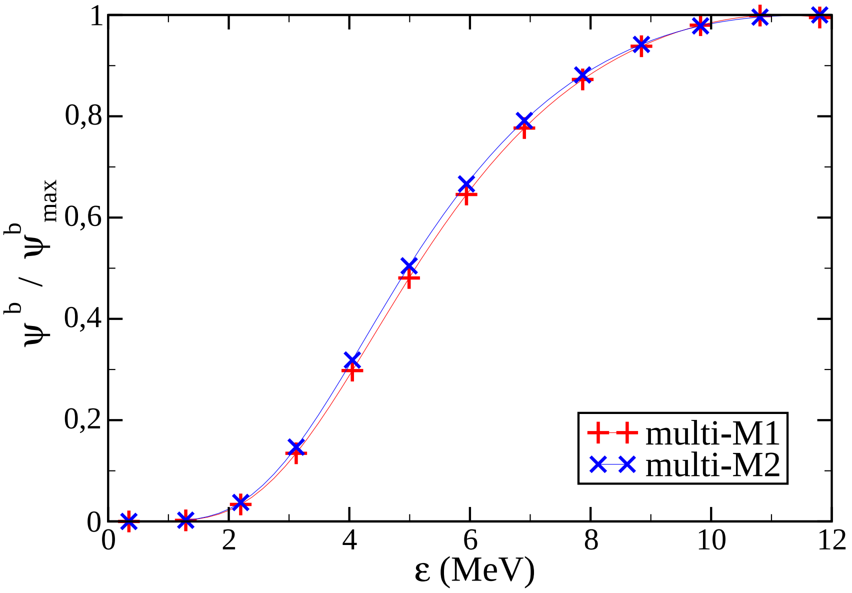

Figure 6.2: Spectra of the optimized sources obtained with the approximated multi-M_1 and multi-M_2 models.

medium is represented on Fig. 6.3, the light blue region corresponds to the tumor, the red region to the organ at risk and the dark blue region to the healthy tissues. The point P targeted by the source is located at $(10\text{ cm}, 20\text{ cm})$, *i.e.* in the center region of the tumor. It is indicated with a black plus on Fig. 6.3.

For this test case, the source of electrons ψ_e^b is fixed at zero. The sources of photons provided at the initial step of the optimization algorithm are chosen to be given by

$$\begin{aligned}
(\boldsymbol{\psi}_\gamma^b)_{l,m}^n &= 10^{10}\exp\left(-\alpha_\epsilon(\epsilon^n-\epsilon_0)^2\right)\left\langle \mathbf{m}\exp\left(-\alpha_\mu(1-\Omega.d(X_{l,m}))^2\right)\right\rangle \sum_{i=1}^{4}\mathbf{1}_{B_i}(X_{l,m}),\\
B_1 &= \left\{X\in\mathbb{R}^2,\quad \text{s.t.}\quad x\in[5\text{ cm},15\text{ cm}],\quad y=0\text{ cm}\right\},\\
B_2 &= \left\{X\in\mathbb{R}^2,\quad \text{s.t.}\quad x\in[5\text{ cm},15\text{ cm}],\quad y=40\text{ cm}\right\},\\
B_3 &= \left\{X\in\mathbb{R}^2,\quad \text{s.t.}\quad x=40\text{ cm},\quad y\in[0\text{ cm},5\text{ cm}]\right\},\\
B_4 &= \left\{X\in\mathbb{R}^2,\quad \text{s.t.}\quad x=40\text{ cm},\quad y\in[35\text{ cm},40\text{ cm}]\right\},
\end{aligned}$$

where $\epsilon_0 = 1$ MeV, $\alpha_\mu = 1000$, $\alpha_\epsilon = 200$ and the direction $d(X)$ is given in (6.16). This boundary condition corresponds to four beams targeting either the two branches of the C-shaped tumor or the center of the tumor.

The residual for the solution of the implicit solver (5.58) is fixed at $r_{\max} = 10^3$ and the maximum number of iterations is fixed at $k_{\max} = 1000$. The residual for

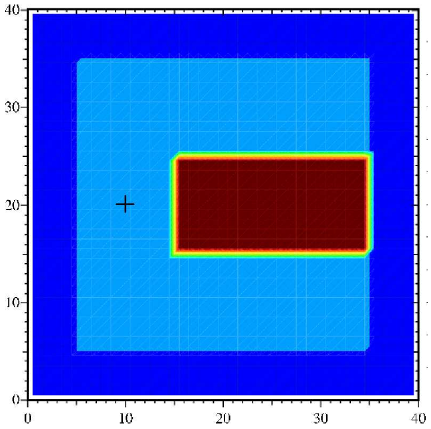

Figure 6.3: Representation of the 2D domain for the optimization test case: in light blue the tumors, in red the organ at risk and in dark blue the healthy tissue.

the bisection algorithm 6.2 is fixed at $r_{bis,\max} = 10^5$ and the maximum number of iteration at $k_{bis,\max} = 10$. Finally the residual for the optimization algorithm 6.1 is fixed at $r_{opt,\max} = 5.10^5$ and a maximum number of iteration $k_{opt,\max} = 10$.

The doses obtained in the end of this optimization process with the multi-M_1 and the multi-M_2 models are represented on Fig. 6.4. The computational times to obtain those results and the total number of times the equations (6.13) and (6.14) were solved numerically are gathered in Table 6.2.

Models	multi-M_1	multi-M_2
Computation times	2h 52min 9sec	10h 5min 13sec
Total number of equation solved	1009	235

Table 6.2: Computational times to obtain the optimized doses with the approximated multi-M_1 and multi-M_2 models.

This test case is extreme as the organ at risk is enveloped in the tumor region. Therefore, finding a position for source such that it irradiates not the organ at risk is complicated.

The optimized doses with the multi-M_1 and multi-M_2 models have similar shapes. As all the sources on the boundary target the point P, the dose around this point is very high. In the organ at risk, the dose is lower but non-zero. The maximum value of the dose in the organ at risk is located in the region the closest to the point P. It reaches 65% of the maximum dose.

The largest discrepancies in the dose between the two models are away from the maximum point P. The dose inside the organ at risk decreases faster away

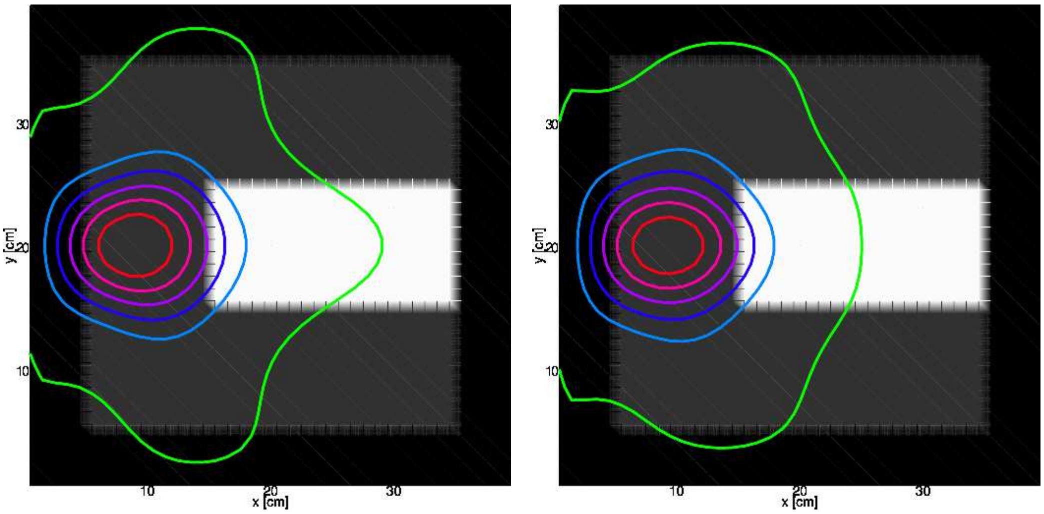

Figure 6.4: Isocurves of optimized dose obtained with the approximated multi-M_1 (left) and multi-M_2 (right) models at 25% (green), 50% (light blue), 60% (dark blue), 70% (purple), 80% (pink), 90% (red) of the maximum dose, over a representation of the domain in gray scale, in white the organ at risk, in gray the tumor and in black the healthy tissues.

from P when using the multi-M_2 model than when using the multi-M_1 model. At the contrary, in the two branches of the C-shaped tumor the dose is higher with multi-M_2 model than with the multi-M_1 model.

As in the 1D case, the source on the boundary is diffused in energy and, here in 2D, also in space. The M_1 and M_2 models were shown to provide similar dose results when considering a photon beam in 2D in Subsection 5.8.3. The (small) differences in Fig. 6.4 are therefore due to the non-linear effects emerging when multiple beams cross each other.

6.6 Discussion

The M_N approximations present several drawbacks that makes their use for optimization problems difficult. The major drawback of those models is their non-linearity which creates parasite effects polluting both the accuracy and the stability of Algorithm 6.1.

The use of the method of moments for optimization is although possible when using methods circumvented such issues at the numerical level. In the present case, splitting the computations in several parts was found to provide senseful results. Such methods make the computation possible, but their accuracy remain discussable as it does not remove entirely the artifical non-linear effects, especially for the M_1

model which is more sensitive to such effects.

The P_N closures presented in Subsection 4.7.1 are linear and they represent a viable alternative to M_N closures for applications to optimization problems (see *e.g.* [8] for application of the P_N models to dose optimization). However, in order to obtain an acceptable accuracy on the dose results with those models, the number of moments N needs to be larger than when using M_N models which considerably raises the computational costs.

Algorithm 6.1 is a basic optimization technique. The convergence rate, the stability or the computational cost of this algorithm can be improved. For instance, quasi-Newton methods (see *e.g* [10]) typically have a better convergencce rate than gradient methods.

The optimization algorithm presented in this chapter is not sufficient for practical applications with medical purposes. In practice, the dose is required to be more homogeneous in the tumor and there are more constraints on the dose in the organs at risk. For instance, for non small cell lung cancer, one requires that every part of the heart receives at most 35 Gy (the gray is the unit of the dose given by 1 Gy = 1 J.kg^{-1} = 6.24.10^{12} MeV.kg^{-1}) and the total dose deposited in the heart is below 40 Gy. This type of refinement of the present method can be adressed by modifying the objective functional j. However, such modifications lead to complications in the optimization process.

Furthermore, in practice, the point P targeted by the beams is also optimizable, and in certain cases with heterogeneous densities, this point may not be located inside a tumor.

The present (basic) optimization technique already provides senseful results but those techniques needs to be extended and improved to be applied to more practical problems.

Bibliography

[1] R. Barnard, M. Frank, and M. Herty. Optimal radiotherapy treatment planning using minimum entropy models. *Appl. Math. Comput.*, 219(5):2668 – 2679, 2012.

[2] P. H. Calamai and J. J. Moré. Projected gradient methods for linearly constrained problems. *Math. Program.*, 39(1):93–116, 1987.

[3] Memorial Sloan-Kettering Cancer Center. *A practical guide to intensity-modulated radiation therapy.* Medical Physics Publishing, 2003.

[4] International commission on radiation units and measurements. Prescribing, recording, and reporting photon-beam intensity-modulated radiationtherapy (IMRT). Technical Report 1, 2010.

[5] A. Costa and J.-P. Gerard, editors. *Guide des procédures de radiothérapie externe.* Société française de radiothérapie oncologie, 2007.

[6] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology: Volume 6, Evolution problems II.* Springer, 2000.

[7] M. Frank, M. Herty, and A. N. Sandjo. Optimal radiotherapy treatment planning governed by kinetic equations. *Math. Mod. Meth. Appl. S.*, 20(04):661–678, 2010.

[8] M. Frank, M. Herty, and M. Schäfer. Optimal treatment planning in radiotherapy based on Boltzmann transport calculations. *Math. Mod. Meth. Appl. S.*, 18(04):573–592, 2008.

[9] M. Herty and A. N. Sandjo. On optimal treatment planning in radiotherapy governed by transport equations. *Math. Mod. Meth. Appl. S.*, 21(02):345–359, 2011.

[10] C. T. Kelley. *Iterative methods for optimization.* SIAM, 1999.

[11] X. Allen Li. *Adaptative radiation therapy.* CRC press, 2011.

[12] T. R. Mackie, J. Kapatoes, K. Ruchala, G. Olivera W. Lu, C. Wu, L. Forrest, W. Tome, J. Welsh, R. Jeraj, P. Harari, P. Reckwerdt, B. Paliwal, M. Ritter, H. Keller, J. Fowler, and M. Mehta. Image guidance for precise conformal radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.*, 56(1):89–105, 2003.

[13] P. Mayles, A. Nahum, and J.C. Rosenwald, editors. *Handbook of radiotherapy physics: Theory and practice.* Taylor & Francis, 2007.

[14] D. M. Shepard, M. C. Ferris, G. H. Olivera, and T. Rockwell Mackie. Optimizing the delivery of radiation therapy to cancer patients. *SIAM Rev.*, 41(4):721–744, 1999.

[15] C. P. South, M. Partridge, and P. M. Evans. A theoretical framework for prescribing radiotherapy dose distributions using patient-specific biological information. *Med. Phys.*, 35(10):4599–4611, 2008.

[16] J. Tervo, P. Kokkonen, M. Frank, and M. Herty. On existence of L^2-solutions of coupled Boltzmann continuous slowing down transport equation system. *arXive*, 2016.

[17] F. Tröltzsch. *Optimal control of partial differential equations: Theory, methods and applications.* American Mathematical Society, 2005.

Conclusion and perspectives

Conclusion

The present thesis aims to propose numerical approaches for dose computation and optimization in the field of radiotherapy that are competitive in terms of accuracy and numerical costs compared to reference methods such as [6, 14].

The present work is based on linear transport models for photons and electrons. Studying the physics of the predominant interactions lead to constructing collision operators for a kinetic model. Then, the well-posedness of the resulting equations was studied.

Numerical methods obtained by discretizing directly such kinetic equations require considerable numerical costs. Those computational costs were reduced by using the method of moments. However, several difficulties emerge when using this method. The moment equations have more unknowns than equations and therefore require an additional closure equation. The entropy-based closures, so-called M_N closures, was preferred in this manuscript. One of the novelties presented in this manuscript was the construction of an approximation of the entropy-based closure for the second order moment equations, *i.e.* the M_2 closure, in three dimensions of space. This approximation is based on the study the domain of definition of the M_N closure, so-called realizability domain. The main characteristics of the M_N models, *i.e.* the realizability and the hyperbolicity were focused on during the construction of this approximation.

Standard numerical schemes applied to such moment equations present stiff terms which may considerably slow down the computations when the considered medium contains low density regions. Numerical methods for dose computation adapted to the entropy-based moment models were constructed, with a special focus on the preservation of the realizability property, and efficient even when considering low-density media.

Finally, a numerical approach was proposed to optimize a source of radiations, *i.e.* such that the dose is as clase as possible to a desired dose. Those numerical methods, both dose computation and optimization, were tested on practical test cases and showed good agreement with reference results.

Perspectives

The following points were introduced in this manuscript. A deeper understanding of these problems would lead to improving the numerical techniques of dose computation and optimization described in Part III.

Realizability and "Generalized moment problem"

The moment models with entropy-based closures are only valid under realizability constraints. This means that all the moments need to be weighted integrals of the same positive function. There exist practical characterizations of the realizability property in 1D (see Section 3.3.2), and such characterizations lead to constructing a new type of realizable closure for high order moment equations (see [8, 9, 11, 10] and Subsection 4.7.2). In the general case of arbitrary order moments on the unit sphere, there is no practical characterizations of the realizability property, and only few results are known. Practical realizability conditions exist for moments up to order two. Some criteria exist for moments of higher order, although they are technical and difficult to use and they are *a priori* not applicable in the whole realizability domain.

Perspectives:

- Define a practical criteria of realizability for moments on the unit sphere of higher order than two.
- Define a realizable and hyperbolic closure for high order angular moment models for multi-dimensional problems that is easy to compute numerically.

Boundary conditions for the moment models

The boundary conditions for kinetic equations are typically chosen by fixing the flux of particles entering inside the medium. Angular moments are obtained by integrating a kinetic equation over all directions of flight Ω. Prescribing the incoming flux on the boundary for the kinetic model is not enough to construct a flux for the moment equations. As the moment equations are hyperbolic, one typically choose boundary conditions for the moment equations based on the theory of hyperbolic equations (see *e.g.* [4, 1, 5]).However, those conditions are not directly related to the underlying kinetic boundary conditions. Furthermore, kinetic effects may appear near the boundary (see *e.g.* [3, 7, 2]) and they are not well modeled with standard moment boundary conditions in particular cases.

Perspective: Define practical boundary conditions for moment equations with an entropy-based closure based on underlying kinetic boundary conditions.

Dose optimization using the entropy-based moment models

Typically, dose optimization algorithms are obtained through a variational approach at the kinetic level. The moment approach is only seen as a numerical method to solve a kinetic equation. The existence and uniqueness of a minimizer were proven at a kinetic level. However, no such results are known at the moments level, except for the linear P_N models.
The method based on the non-linear M_1 and M_2 models showed good experimental results on simple test cases.

Perspectives:

- Improve the dose optimization algorithm based on moment models with entropy-based closure.
- Adapt this method to more practical problems.

Numerical methods for the moment equations

The numerical schemes presented in this manuscript are of order 1. Higher order numerical schemes can be developed based on the present approaches. Furthermore, the numerical scheme (5.57) applicable to the transport of photons and electrons together is based on an iterative algorithm. This method was solved directly using basic techniques which introduce both numerical errors and numerical costs.

Perspective:

- Complete the numerical approach to apply it to more realistic models.
- Construct higher order numerical schemes for the radiotherapy equations.
- Improve the numerical method to raise the convergence rate of Algorithm 5.1.

Biological effects

As a first approach, the effects of the radiations on the tissues is often assumed to be a function of the dose, *e.g.* the linear-quadratic model ([12]) states that the proportion of cells surviving to the radiations is of the form $\exp(-\alpha D - \beta D^2)$. This approach is simple to use and one needs not to compute the whole fluences $(\psi_\gamma,\ \psi_e)$, because only the dose is required. This model is empirical, however in practice, the biological effects of the radiations on the cells follow a more complicated model.

Furthermore, other effects neglected by this empirical approach are non-negligible in particular cases, *e.g.* the bystander effect ([13]) is non-local because it affects the cells neighboring those receiving the radiations.

In the present framework, a better understanding of those effects would lead to proposing a radiobiology-based objective functional J in the optimization algorithms of Chapter 6.

Perspective: Complete the present model to include the biological effects of the radiations on the cells.

Bibliography

[1] C. Bardos, A. Y. Leroux, and J. C. Nedelec. First order quasilinear equations with boundary conditions. *Commun. Part. Diff. Eq.*, 4(9):1017–1034, 1979.

[2] A. Bensoussan, J.-L. Lions, and G. C. Papanicolaou. Boundary layers and homogenization of transport processes. *Publ. RIMS, Kyoto Univ.*, 15:53–157, 1979.

[3] F. Coron, F. Golse, and C. Sulem. A classification of well-posed kinetic layer problems. *Commun. Pur. Appl. Math.*, 41(4):409–435, 1988.

[4] R. J. Diperna. Uniqueness of solutions to hyperbolic conservation laws. *Indiana Univ. Math. J.*, 28(1):137–188, 1979.

[5] B. Dubroca and G. Gallice. Resultats d'existence et d'unicite du problème mixte pour des systems hyperboliques de lois de conservation monodimensionnels. *Commun. Part. Diff. Eq.*, 15(1):59–80, 1990.

[6] J. Sempau F. Salvat, J. M. Fernández-Varea. *PENELOPE-2011: A code system for Monte Carlo simulation of electron and photon transport*, 2011.

[7] F. Golse, S. Jin, and C. D. Levermore. A domain decomposition analysis for a two-scale linear transport problem. *ESAIM-Math. Model. Num.*, 37(6):869–892, 2003.

[8] D. Kershaw. Flux limiting nature's own way. Technical report, Lawrence Livermore Laboratory, 1976.

[9] P. Monreal. *Moment realizability and Kershaw closures in radiative transfer.* PhD thesis, Rheinisch-Westfälische Technische Hochschule, 2012.

[10] F. Schneider. *Moment models in radiation transport equations.* PhD thesis, Teschnische Universität Kaiserslautern, 2015.

[11] F. Schneider. Kershaw closures for linear transport equations in slab geometry I: Model derivation. *J. comput. phys.*, pages –, 2016.

[12] C. P. South, M. Partridge, and P. M. Evans. A theoretical framework for prescribing radiotherapy dose distributions using patient-specific biological information. *Med. Phys.*, 35(10):4599–4611, 2008.

[13] C. A. Waldren. Classical radiation biology dogma, bystander effects and paradigm shifts. *Hum. Exp. Toxicol.*, 24(1):537–542, 2005.

[14] T.A. Wareing, J.M. McGhee, Y. Archambault, and S. Thompson. Acuros XB ® advanced dose calculation for the Eclipse $^{\mathrm{TM}}$ treatement planning system. *Clinical perspectives*, 2010.