Models and Numerical Methods for Time- and Energy-Dependent Particle Transport

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften
der RWTH Aachen University zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von
Diplom-Technomathematiker
Edgar Olbrant
aus Maikain (Kasachstan)

Berichter: Universitätsprofessor Dr. rer. nat. Martin Frank
Universitätsprofessor Cory D. Hauck, Ph.D.

Tag der mündlichen Prüfung: 13.04.2012

Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.

To my late grandfather Jakob Olbrant,
To my mother Ekaterina and my father Jakob,
To my sisters Renata and Inga,
To my nieces Angelina and Sophia and their father Waldemar,
To my caring partner Sophie.

ACKNOWLEDGEMENTS

Special thanks belongs to my advisor Martin Frank who assisted my scientific career for more than four years. His supervision has always been encouraging; his calm and clear way of explaining mathematical issues gave a valuable insight into the extensive field of transport theory and the Boltzmann equation. The scientific way of working which he has taught me molded my character and will remain a guidance for the future.

I also wish to express my gratitude to Cory D. Hauck who taught me a lot about moment models and DG numerics. Our conversations on Skype (which sometimes lasted for hours) and my short stay in Oak Ridge greatly pushed the projects forward.

I am greatly indebted to Edward Larsen for making my stay in Ann Arbor very comfortable. Our long discussions about neutron transport in his office led us eventually to one decisive afternoon: And all of a sudden, there it was – the key idea. That incident will always stay in my mind.

I am grateful to Benjamin Seibold for helpful conversations about the numerics and test cases in Chapter 2. His knowledge of MATLAB programming greatly accelerated the computations.

I thank Bruno Dubroca for his great hospitality during my stays in Bordeaux and for putting me in touch with physicists from CELIA. He also provided his 3D M1-code which was the source for 3D visualizations of the deposited radiation dose in the human body.

I also thank José M. Fernández-Varea and Josep Sempau for helping me with electron cross sections from Chapter 1 and their support on the Monte Carlo codes PENELOPE and penEasy.

I am grateful to Neal Carron for our email communications pointing me to the proton database from Chapter 1 and to the information about its difficult access and usage.

I thank Ryan G. McClarren for giving me the P_{99} transport solution which was used for comparisons in Chapter 4.

I acknowledge support from my fellows at MathCCES (RWTH) and the Technomathematics group (TU Kaiserslautern) who contributed to a relaxed working environment. I would particularly like to thank Raul Borsche for many friendly chats and his consultation on hyperbolic equations; Kai Krycki for giving advice on any problem I wrote on the blackboard; Richard Barnard for introducing me to the software ParaView; Pascal Richter for informal discussions as an important source for inspiration; and most of all Philipp Monreal for many helpful discussions on DG and the M1-model.

Working on this thesis, I spent most of my time at the two universities TU Kaiserslautern and RWTH Aachen. Although I suffered many setbacks during this time, it was the small sense of achievement which always served as motivation and piqued my curiosity to learn more about a certain problem. In the end, I believe that a mixture of my own interest, long discussions with my colleagues and the support from friends and family made this work possible. I am grateful for this experience.

Aachen, May 13, 2012

Edgar Olbrant

Abstract

Particles passing through a medium can be described by the Boltzmann transport equation. Therein, all physical interactions of particles with matter are given by cross sections. We compare different analytical models of cross sections for photons, electrons and protons to state-of-the-art databases. The large dimensionality of the transport equation and its integro-differential form make it analytically difficult and computationally costly to solve. In this work, we focus on the following approximative models to the linear Boltzmann equation: (i) the time-dependent simplified P_N (SP_N) equations, (ii) the M_1 model derived from entropy-based closures and (iii) a new perturbed M_1 model derived from a perturbative entropy closure. In particular, an asymptotic analysis for SP_N equations is presented and confirmed by numerical computations in 2D. Moreover, we design an explicit Runge-Kutta discontinuous Galerkin (RKDG) method to the M_1 model of radiative transfer in slab geometry and construct a scheme ensuring the realizability of the moment variables. Among other things, M_1 numerical results are compared with an analytical solution in a Riemann problem and the Marshak wave problem is considered. Additionally, we rigorously derive a new hierarchy of kinetic moment models in the context of grey photon transport in one spatial dimension. Numerical examples, such as the two beam instability or the analytical benchmark due to *Su and Olson* [173], are shown for the perturbed M_1 model and compared to the standard M_1 as well as transport solutions.

Contents

Preface **5**

1 Electron, Photon and Proton Scattering Processes **7**
- 1.1 Introduction 7
  - 1.1.1 Generalized Fokker-Planck Approximation 9
  - 1.1.2 Method of Moments 10
  - 1.1.3 Goals and Data Structure 11
- 1.2 Electron Interactions 12
  - 1.2.1 Physical Interactions 13
  - 1.2.2 Terminology 14
  - 1.2.3 ICRU Database 15
  - 1.2.4 Elastic Cross Sections 15
  - 1.2.5 Inelastic Cross Sections 20
  - 1.2.6 Comparison 21
  - 1.2.7 Stopping Power 23
  - 1.2.8 Dose Calculations 26
- 1.3 Photon Interactions 28
  - 1.3.1 Terminology 30
  - 1.3.2 Elastic (Coherent Rayleigh) Scattering 30
  - 1.3.3 Photoelectric Effect 32
  - 1.3.4 Compton (Incoherent) Scattering 33
  - 1.3.5 Comparison 36
- 1.4 High Energy Proton Interactions 39
  - 1.4.1 Terminology 39
  - 1.4.2 Elastic Cross Sections 40
  - 1.4.3 Stopping Power 50

2 Time-Dependent Simplified P_N Equations **53**
- 2.1 Introduction 53
- 2.2 Classic (Ad-Hoc) Derivation 55
- 2.3 Formal Asymptotic Derivation 62
  - 2.3.1 SP_1 Equations 66
  - 2.3.2 SP_2 Equations 67
  - 2.3.3 SP_3 Equations 68
- 2.4 Numerical Results in 2D 69
  - 2.4.1 Equivalence of SP_N and P_N equations 71
  - 2.4.2 A Moving Rod 72

2.4.3 Checkerboard 75
2.4.4 Line Source 81

3 A Realizability-Preserving DG Method: The M_1 Model 85
3.1 Introduction 85
3.2 The M_1 Model for Radiative Transfer 87
3.2.1 The Radiative Transfer Equation 87
3.2.2 The M_1 Closure 88
3.2.3 Realizability and Properties of the M_1 Model 89
3.3 DG Formulation 91
3.3.1 Spatial Discretization 91
3.3.2 Slope Limiter 94
3.3.3 Time Discretization: Explicit SSP Runge-Kutta Schemes 95
3.4 The Realizability-Preserving Limiter 96
3.5 Numerical Experiments 102
3.5.1 Accuracy Tests 103
3.5.2 Under-resolved Plane Source 106
3.5.3 Riemann Problem 106
3.5.4 Heated Wall 108
3.5.5 Two Opposing Beams 110
3.5.6 Coupling to the Material Energy Equation 110

4 Perturbed, Entropy-Based Closure 117
4.1 Introduction 117
4.2 Radiative Transfer and Moment Equations 118
4.2.1 The Radiative Transfer Equation 119
4.2.2 Moment Equations 119
4.3 Entropy-Based and Perturbed Entropy-Based Closures 120
4.3.1 Entropy-Based Closures 120
4.3.2 Perturbed Entropy-Based (PEB) Closures 121
4.3.3 The Perturbed M_1 (PM_1) model 124
4.3.4 Controlling the Perturbations 128
4.4 Numerical Simulation using Discontinuous Galerkin 129
4.4.1 Spatial Discretization 129
4.4.2 Time Discretization: Explicit SSP Runge-Kutta Schemes 133
4.4.3 Limiters 134
4.5 Numerical Results 137
4.5.1 Accuracy Test: Manufactured Solutions 138
4.5.2 Two Beam Instability 139
4.5.3 Source-Beam Problem 141
4.5.4 Gaussian Source 142
4.5.5 Including the Material Energy Equation 142

5 Discussion and Conclusions 147
5.1 Electron, Photon and Proton Scattering Processes 147
5.2 Time-Dependent Simplified P_N Equations 148
5.3 A Realizability-Preserving DG Method: The M_1 Model 150
5.4 Perturbed, Entropy-Based Closure 151

A Transport Coefficients: Different Polynomial Kernels 153

B Time-Dependent SP_N Equations 155
B.1 SP_N equations of Arbitrary Order 155
B.1.1 SP_1 Equations 157
B.1.2 SP_2 Equations 157
B.1.3 SP_3 Equations 157
B.1.4 SP_4 Equations 157
B.1.5 SP_5 Equations 157
B.2 Expanding Operators 158
B.3 Accuracy Tests: Manufactured Solutions 162
B.3.1 Diffusion Equation 162
B.3.2 SP_N Equations 162
B.3.3 P_N Equations 164

C Slope Limiting with Characteristic Variables 168

Preface

Particles passing through a medium can be described by the Boltzmann transport equation which, in a general form, can be written as

$$\frac{1}{v}\frac{\partial \psi}{\partial t}(\underline{x}, \underline{\Omega}, E, t) + \underline{\Omega} \cdot \nabla \psi(\underline{x}, \underline{\Omega}, E, t) = \mathcal{C}(\psi), \tag{1}$$

where ψ is related to the particle distribution in $\underline{x} \in \mathbb{R}^3$ at time t and is often referred to as angular flux. Particles with energy E move at velocity v in direction $\underline{\Omega} \in \mathbb{R}^3$. The first term in (1) is the time rate of change of the angular flux; the second term specifies the net leakage rate at which particles flow out of an incremental volume.

All interactions they experience with matter are specified by $\mathcal{C}(\psi)$. This collision operator contains integral terms of the solution ψ itself which depends on seven dimensions (three in space, two in angle, one in energy and time). The large dimensionality of the system and the integro-differential form of (1) make this equation difficult and computationally costly to solve. This is why much effort has been spent on developing approximative models to (1).

The accuracy of computed results strongly depends on realistic input data characterizing physical interactions of projectile-target collisions. In Chapter 1, we are primarily interested in irradiation of human tissue. The main goal is therefore to provide physical quantities which can be directly included in deterministic codes solving the transport equation or approximative models thereof. Partial results of this chapter have already been published in [137].

We compare different analytical models of cross sections for photons, electrons and protons to state-of-the-art databases discussed in literature. The accuracy of presented electron cross sections is demonstrated by comparing deterministic calculations of the radiation dose in water to stochastic Monte Carlo results.

The spherical harmonics (P_N) equations have been a standard approximation to the linear Boltzmann equation (1). A big drawback of the P_N equations in 3-D is their complicated form and the large number of equations growing as $(N+1)^2$. To lower the computational effort, the steady-state simplified P_N (SP_N) equations have been first derived in an ad-hoc way [63–65] and later proven to be higher-order corrections to the diffusion equation in certain physical systems [18, 102, 148, 175]. However, even the ad-hoc derivation relies on the special structure of the steady-state P_N equations. Thus, the same procedure cannot be extended in a straight-forward way to the time-dependent equations because of the additional time derivative.

In Chapter 2, we present an asymptotic analysis for the *time-dependent* simplified P_N equations up to $N = 3$. The final SP_N equations are hyperbolic and differ from those investigated in [56]. Additionally, SP_N equations of arbitrary order are derived in an ad-hoc way. In two space dimensions, numerical calculations for the P_N and

SP_N equations are performed. We simulate neutron distributions of a moving rod and present results for the checkerboard problem. Moreover, computations demonstrate that there are even cases outside the asymptotic limit where P_N and SP_N are equivalent in 2D. Comparisons between SP_N and diffusion solutions confirm significant improvements. To a large extent, this chapter relies on [139] which is in preparation.

The P_N equations can be assigned to the more general framework of moment methods used to derive approximate models. However, this approach always leads to an underdetermined system of equations which requires a closure. P_N equations are then obtained by simply setting the highest moment to zero.

A different goal is pursued by entropy-based closures where approximations to the highest moment are derived (instead of completely neglecting it). The M_1 model is the first member in this hierarchy of models which rely on the physical principle of maximum entropy. In Chapter 3, we study the M_1 model of radiative transfer in slab geometry which is a system of hyperbolic equations and additionally couple this system to the material energy equation. For this model to be well-posed, its moment variables must fulfill certain realizability conditions. For example, the evolution of the particle number (related to the radiation energy in the model) must remain non-negative.

We address the problem of realizability from a numerical point of view. Our main focus is on design and implementation of an explicit Runge-Kutta discontinuous Galerkin method which, in general, does not preserve such a property. However, we construct a realizability-preserving scheme which, under a more restrictive CFL condition, guarantees the realizability of the moment variables.

An analytical proof for our realizability-preserving scheme, which also includes a slope-limiting technique, is provided and confirmed by various numerical examples. Among other things, we present accuracy tests showing convergence up to fourth-order, compare our results with an analytical solution to a Riemann problem, and consider Marshak wave problems. The contents of Chapter 3 relies on [138] already submitted to the Journal of Computational Physics.

A different ansatz for moment closures is studied in Chapter 4. We rigorously derive a new hierarchy of kinetic moment models in the context of frequency integrated photon transport in slab geometry. These new models are perturbations of entropy-based models mentioned above; we therefore refer to them as *perturbative entropy-based* models. Our derivations of the perturbative model reveal final equations containing an additional convective and diffusive term which are added to the flux term of the standard entropy closure. The resulting system of equations is a convection-diffusion system. This is different to perturbations to standard P_N closures [156] which only gain a diffusive component.

The perturbed M_1 model, the first member in the moment hierarchy, is discretized by using a Runge-Kutta discontinuous Galerkin method. As for the standard M_1, we enforce the realizability property by a similar limiting approach. However, an additional control parameter needs to be introduced. It modifies the pressure term of the perturbed M_1 equations and ensures cell averages of the moments to remain realizable. Numerical simulations show improvements to the standard M_1 model in cases where particles move in opposing directions. The main parts of Chapter 4 are also supposed to be published in [55].

Chapter 1

Electron, Photon and Proton Scattering Processes

1.1 Introduction

The International Agency for Research on Cancer (IARC) provides estimates on incidence of and mortality from major types of cancers for all countries of the world [51]. Latest information are presented for the year 2008: Worldwide, there are about 12.6 million new cancer cases per year and about 7.6 million people died of cancer. The number of cancer cases between the sexes is almost equally distributed with 6.6 million male and 6 million female cases. However, the most often occuring cancer types differ between the sexes: Women develop breast and colorectal cancers most frequently whereas men often suffer from lung and prostate cancer. These statistics are average values and strongly depend on the corresponding region in the world. Due to the demographic effect, these numbers are expected to increase in future. By the year 2020, approximately 17 million cancer cases per year are predicted worldwide.

Above numbers indicate that the global cancer burden is fairly high. One possibility to effect the cure of cancer is radiotherapy. According to the International Atomic Energy Agency there are about 7500 radiotherapy centers, offering certain kinds of radiation therapy, and approximately 10000 linear particle accelerators worldwide [1].

Dose calculations are one decisive part in treatment planing for external beam radiation therapy. They are based on a detailed description of particle transport in tissue. Several mathematical methods and algorithms have been developed to address this type of problem [137]. They all have two major properties in common:

- In some way or other, they all look for solutions to the Boltzmann transport equation.
- The accuracy of their results *strongly* depends on realistic input data characterizing physical interactions of projectile-target collisions.

Consequently, this field of research ties together a variety of different topics from mathematics, medicine, biology and radiation physics. From the mathematical point of view, particles traversing a medium can be described by the Boltzmann transport equation which, in a general form, can be written as

$$\frac{1}{v}\frac{\partial \psi}{\partial t}(\underline{x}, \underline{\Omega}, E, t) + \underline{\Omega} \cdot \nabla \psi(\underline{x}, \underline{\Omega}, E, t) = \mathcal{C}(\psi), \tag{1.1}$$

where ψ is related to the particle distribution in $\underline{x} \in \mathbb{R}^3$ at time t and is often referred to as *angular flux.* Particles with energy E move at velocity v in direction $\underline{\Omega} \in \mathbb{R}^3$. The first term in (1.1) is the rate of change in ψ with respect to time; the second term specifies the net leakage rate at which particles flow out of an incremental volume.

All interactions they experience with matter are specified by $\mathcal{C}(\psi)$. This collision operator contains integral terms of the solution ψ itself which depends on seven dimensions (three in space, two in angle, one in energy and time). The large dimensionality of the system and the integro-differential form of (2.1) make this equation difficult and computationally costly to solve. This is why much effort has been spent on developing approximative models to (2.1).

If particle transport takes place in an isotropic and homogeneous medium in which interaction processes are Markovian and particles do not interact with themselves, their distribution can be described by the solution of the *linear* Boltzmann equation. For practical applications in radiotherapy we are faced with issues like distributions of the deposited energy in biological tissues or penetration depths of the beam. For many purposes it is therefore sufficient to know the steady-state solution. Discarding time-dependence and internal sources the collision operator can be written as:

$$\mathcal{C}(\psi) = \int_0^\infty \int_{S^2} \sigma_s(E', E, \underline{\Omega} \cdot \underline{\Omega}') \psi(\underline{x}, E', \underline{\Omega}') d\Omega' dE' - \Sigma_t(E) \psi(\underline{x}, E, \underline{\Omega}). \tag{1.2}$$

Above integral contains the *differential scattering cross section* $\sigma_s(E', E, \underline{\Omega} \cdot \underline{\Omega}')$ characterizing interaction mechanisms in which particles are deflected. The dot product $\underline{\Omega} \cdot \underline{\Omega}' = \cos(\theta_0) = \mu_0$ indicates that the scattering probability that a particle will scatter from direction of flight $\underline{\Omega}'$ to direction $\underline{\Omega}$ only depends on the scattering angle. Hence, all scattered directions $\underline{\Omega}$, which form a cone of a fixed scattering angle θ_0, are equally probable.

Integrating the scattering kernel $\sigma_s(E', E, \underline{\Omega} \cdot \underline{\Omega}')$ over all angles and energies, one gets the *total scattering* cross section

$$\Sigma_s(E) = 2\pi \int_0^\infty \int_{-1}^1 \sigma_s(E', E, \mu_0) d\mu_0 dE'. \tag{1.3}$$

$\Sigma_t(E)$ is called *total* cross section and determines the probability that a particle with energy E will undergo a collision. It is a sum of *absorption* and *total scattering* cross section:

$$\Sigma_t(E) = \Sigma_a(E) + \Sigma_s(E).$$

$\Sigma_a(E)$ describes the probability that particles are absorbed by the material.

Generally speaking, particle interactions can be described by two interaction processes:

- *Elastic scattering:* It is a non-radiative interaction between the projectile and the target particle in which the internal energy of the target is not changed.
- *Inelastic scattering:* Similarly, inelastic scattering summarizes all interactions of the projectile and an atom or molecule in which the internal energy of the target is changed by vibrational, rotational or electronical excitation of the target. The latter also includes ionization of the atom or molecule which plays an important role in radiotherapy.

Above classification is very rough and evidently, there are many more physical interactions depending on the particle type and energy range. At this point, this is, however, sufficient and we postpone detailed descriptions to the upcoming sections.

As already mentioned, the large phase space of the system in (1.2) makes direct numerical simulation computationally expensive. Thus, approximate models are needed to reduce the size of the system. However, up to now, there is no predominant method used for all types of particles. In fact, depending on specific particle properties and the regarding quantities you are interested in, one has to choose an appropriate approximation. We want to sketch two procedures briefly to clarify which physical quantities are supposed to be computed for the corresponding model.

1.1.1 Generalized Fokker-Planck Approximation

Some particles, e.g., electrons, show two crucial features when travelling through matter: First, the elastic differential cross section forms a sharp peak in the forward direction [86]. This can be expressed mathematically by introducing the positive *n-th scattering transport coefficient*

$$\xi_n(E) := 2\pi \int_0^\infty \int_{-1}^1 (1-\mu_0)^n \sigma_s(E', E, \mu_0) d\mu_0 dE', \quad \text{for all } n \geq 0. \tag{1.4}$$

A transport process with sufficiently forward-peaked scattering then implies that for increasing n, the coefficients ξ_n fall off sufficiently fast [107], i.e.,

$$\xi_{n+1}(E) \ll \xi_n(E), \quad \text{for all } n \geq 0. \tag{1.5}$$

Second, collision events mostly entail small energy losses. Therefore, one natural approximation is the expansion of the scattering kernel $\sigma_s(E', E, \mu_0)$ around $\mu_0 = 1$ and $E = E'$. This is how Pomraning [147] shows that the already known *Fokker-Planck* operator is the lowest-order asymptotic limit of the integral operator in $\mathcal{C}(\psi)$. In [107] this Fokker-Planck operator is derived as a first order angular approximation to $\mathcal{C}(\psi)$:

$$\mathcal{C}(\psi) = L_{FP}\psi + \mathcal{O}(\varepsilon), \quad \text{for } \varepsilon \approx 0, \quad \text{where} \quad L_{FP} := \frac{\xi_1}{2} L. \tag{1.6}$$

$\varepsilon \ll 1$ is a small scaling parameter characterizing small energy losses and small angular deflections in a system with characteristic size $\mathcal{O}(1)$.

Since the spherical Laplace-Beltrami operator

$$L = \left[\frac{\partial}{\partial \mu}(1-\mu^2)\frac{\partial}{\partial \mu} + \frac{1}{1-\mu^2}\frac{\partial^2}{\partial \phi^2}\right], \quad \text{with } \mu = \cos(\theta),$$

is differential in angle, the nonlocal *integral* Boltzmann operator $\mathcal{C}(\psi)$ is now approximated by a local *differential* operator. The crucial point is that an integro-differential equation is transformed into a partial differential equation. Although discretizations of differential equations often lead to large linear systems, their numerical effort turns out to be much lower. This is due to the local character of differential equations, which bring along much sparser matrices.

Pomraning's resulting Fokker-Planck equation for particle transport in an isotropic medium reads:

$$\begin{aligned}\Sigma_a(E)\psi(\underline{x},E,\underline{\Omega})+\underline{\Omega}\cdot\nabla\psi(\underline{x},E,\underline{\Omega}) = {} & \frac{\xi_1(E)}{2}L\psi(\underline{x},E,\underline{\Omega}) \\ & +\frac{\partial}{\partial E}(S(\underline{x},E)\psi(\underline{x},E,\underline{\Omega})).\end{aligned} \tag{1.7}$$

$S(\underline{x},E)$ is called *stopping power* defined by

$$S(\underline{x},E)=2\pi\int_0^\infty\int_{-1}^1 E'\sigma_s(E',E,\mu_0)\,d\mu_0 dE'. \tag{1.8}$$

The Fokker-Planck approximation is a frequently used method to describe transport processes in media where large-angle and large-energy loss scattering are negligible. Comparisons to real data, however, reveal that many scattering processes of interest contain a small but sufficient amount of large-angle scattering. To gain higher order asymptotic approximations to $\mathcal{C}(\psi)$, *generalized Fokker-Planck* equations were developed which incorporate large-angle scattering:

$$\begin{aligned}\Sigma_a(E)\psi(\underline{x},E,\underline{\Omega})+\underline{\Omega}\cdot\nabla\psi(\underline{x},E,\underline{\Omega}) = {} & L_{GFP_n}\psi(\underline{x},E,\underline{\Omega}) \\ & +\frac{\partial}{\partial E}(S(\underline{x},E)\psi(\underline{x},E,\underline{\Omega})),\end{aligned} \tag{1.9}$$

where, for $m\in\mathbb{N}$, the operators are defined by

$$L_{GFP_{2m}}:=\sum_{i=1}^m\alpha_i(E)L(I-\beta_i(E)L)^{-1}\quad\text{and}\quad L_{GFP_{2m+1}}:=L_{GFP_{2m}}+\alpha_{m+1}(E)L.$$

All coefficients $\alpha_i(E),\beta_i(E)>0$ are linear combinations of the scattering transport coefficients in (1.4). Indeed, numerical examples containing large angle-scattering confirm that above equations are more accurate than the conventional Fokker-Planck equation [137].

1.1.2 Method of Moments

In radiotherapy, the main quantity of interest is the absorbed dose. Caused by energy deposition in particle interactions, it can be computed by

$$D(\underline{x})=\frac{\tau}{\rho(\underline{x})}\int_0^\infty S(\underline{x},E)\Phi_0(\underline{x},E)dE, \tag{1.10}$$

where

$$\Phi_0(\underline{x},E)=\int_{S^2}\psi(\underline{x},E,\underline{\Omega})\,d\Omega. \tag{1.11}$$

τ is hereby the duration of the irradiation of the patient and ρ the mass density of the irradiated tissue so that $D(\underline{x})$ leads to SI unit J/kg or Gy.

Consequently, detailed information about $\psi(\underline{x},E,\underline{\Omega})$ is not necessary. Instead, it is sufficient to know the integral of ψ over all spatial directions. A common approach which directly incorporates this fact is the method of moments. A brief overview of moment models can be found in [52].

Derivation of any moment system begins with the choice of a vector-valued function $\mathbf{m} : S^2 \to \mathbb{R}^N$, whose N components are linearly independent functions of $\underline{\Omega}$. Evolution equations for the moments

$$\Phi_{n-1}(\underline{x}, t) := \int_{S^2} m_n \psi(\underline{x}, \underline{\Omega}, t) d\Omega,$$

where m_n is the n-th component of $\mathbf{m}$, can then be found by multiplying the corresponding equation by $\mathbf{m}$ and integrating over all angular directions. However, this system of N equations will always contain the $(N+1)$-st moment of the distribution. It is therefore inevitable to define a closure to make it solvable. Without going too much into detail here, this approach enables to derive a system of hyperbolic PDEs with N unknowns.

We apply the above approach to the Fokker-Planck equation (1.7):

$$\begin{aligned}\Sigma_a(E) \int_{S^2} \mathbf{m}\, \psi(\underline{x}, E, \underline{\Omega}) d\Omega + \int_{S^2} (\nabla \cdot \underline{\Omega})\, \mathbf{m}\, \psi(\underline{x}, \underline{\Omega}, E) d\Omega = \\ \frac{\xi_1(E)}{2} \int_{S^2} \mathbf{m}\, L\psi(\underline{x}, E, \underline{\Omega}) d\Omega + \frac{\partial}{\partial E} \left(S(\underline{x}, E) \int_{S^2} \mathbf{m}\, \psi(\underline{x}, E, \underline{\Omega}) d\Omega \right).\end{aligned} \tag{1.12}$$

A first member in a hierarchy of moment models, which relies on the physical principle of maximum entropy as a tool for deriving angular moment closures, is the M1 model [21, 47, 127]. Hereby, the components of $\mathbf{m}$ are monomials or, more precisely, $\mathbf{m} = [1, \underline{\Omega}]^T$. For the M1 model, (1.12) simplifies to:

$$\Sigma_a(E)\Phi_0(\underline{x}, E) + \nabla \cdot \Phi_1(\underline{x}, E) = \frac{\partial}{\partial E} \left(S(\underline{x}, E)\Phi_0(\underline{x}, E) \right) \tag{1.13}$$

$$\Sigma_a(E)\Phi_1(\underline{x}, E) + \nabla \cdot D_e\Phi_0(\underline{x}, E) + \xi_1(E)\Phi_1(\underline{x}, E) = \frac{\partial}{\partial E} \left(S(\underline{x}, E)\Phi_1(\underline{x}, E) \right) \tag{1.14}$$

where D_e is the Eddington tensor [46].

1.1.3 Goals and Data Structure

Primarily, we are interested in irradiation of human tissue by photons, electrons and protons. The main goal is to provide physical quantities which can directly be included in deterministic codes solving the transport equation or approximative models thereof. However, it is important to emphasize that our purpose is neither to study *all* possible interaction processes nor to provide detailed theoretical background. It is rather a selection of some *most relevant* interactions whose values are chosen *as accurate as possible*. Additionally, they are transformed in such a way that it should be possible to include them in deterministic codes without too much effort. Detailed overviews for interactions in photon and electron transport can be found in [154]

Altogether, our equations depicted above reveal the following physical quantities to be studied in detail:

- differential scattering cross section $\sigma_s(E', E, \underline{\Omega} \cdot \underline{\Omega}')$,

- absorption cross section $\Sigma_a(E)$,
- scattering transport coefficients $\xi_n(E)$ and
- stopping power $S(\underline{x}, E)$.

We compare different models of cross sections which are discussed in literature. Analytic formulae are used to compute certain quantities which are compared to results from several databases. The framework is kept as general as possible. Nevertheless, as human tissue mainly consists of water and data is not always available for many types of atoms/molecules, some physical constants or cross sections are restricted to water molecules. More precisely, three sections can be found in the following which are organized according to the special type of particles:

- **Section 1.2: Electrons.**
  Elastic scattering in electron-nucleus events and inelastic cross sections for electron-electron interactions in liquid water are extracted from the ICRU77 database [85] as well as computed analytically. Results of both methods are compared on an energy interval between $5 \cdot 10^{-5}$ MeV and 100 MeV. The inelastic Møller cross section is validated by the ICRU77 stopping power [85] whose energy range is between $5 \cdot 10^{-5}$ MeV and 10^3 MeV. To demonstrate the accuracy of the calculated cross sections generalized Fokker-Planck dose calculations in water are contrasted with Monte Carlo results.
- **Section 1.3: Photons.**
  We study elastic, photoelectric and Compton scattering which are the dominant contributions for our energy range of interest between 10^{-3} MeV and 25 MeV. We omit pair production because it only prevails at large energies. Cross sections from PENELOPE [155] are provided for energies between $5 \cdot 10^{-5}$ MeV and 10^3 MeV. Analytical expressions are also given for the same energy interval and compared to PENELOPE's total cross sections which always serve as a benchmark in these notes.
- **Section 1.4: Protons.**
  Four different approaches for the description of elastic scattering of protons at water molecules are presented. Although two analytic expressions are introduced, the main focus is on extracting and validating cross sections for oxygen and hydrogen from the ENDF/B-VII library [126]. Depending on the corresponding model, the energy range differs but the core interval is between 1 MeV and 150 MeV. We use the stopping power from the NIST database [83] for energies between 10^{-3} MeV and 10^5 MeV and additionally discuss briefly Bethe's formula.

1.2 Electron Interactions

Electron beams are nowadays a tool in cancer therapy for targets which are not deeply located in the body. Electron dose profiles first provide a high surface dose, increase to a maximum at a certain depth and drop off with a steep slope afterwards. Known as the bremsstrahlung tail they level off at a small dose. This characteristic shape (shown in Section 1.2.8) offers a good option for the treatment of superficial tumours (less than

5cm deep). Typical electron beams, provided by high energy linear accelerators, range from 4 MeV to 22 MeV (7.8-43.1 electron rest energies).

1.2.1 Physical Interactions

During irradiation of human tissue electrons interact with matter through several competing mechanisms:

1. *Elastic Scattering:* This is usually a non-radiative interaction between electrons and the atomic shell. Projectiles experience small deflections and lose little energy. High energy electrons can also penetrate through atomic shells and are afterwards scattered at the bare nucleus without any energy loss. With kinetic energies above 1 keV elastic scattering in water dominantly occurs in the forward direction [106].

2. *Soft Inelastic e^--e^- Scattering:* Electrons interact with other electrons of the outer atomic shell which usually leads to excitation or ionisation of the target particle. Here binding energies are only a few eV so that projectile electrons transfer little energy and are hardly deflected.

3. *Hard Inelastic e^--e^- Scattering:* These collisions are determined by large transfer energies to the target electron. What 'large' exactly means is specified in Monte Carlo (MC) codes by cutoff energies. In PENELOPE [155], for example, the default value of this simulation parameter is set to 1 % of the maximum energy of all particles. As a consequence, the target electrons are ejected with larger scattering angles and higher kinetic energies (delta rays). They act as an additional source in the transport equation.

4. *Bremsstrahlung:* Caused by the electrostatic field of atoms, electrons are accelerated and hence emit bremsstrahlung photons. However, for energies below 1 MeV this phenomenon can be neglected. Bremsstrahlung photons are not mainly emitted in the forward direction. The lower their kinetic energy the more isotropic their angle distribution becomes [96].

Evidently, there are more interaction processes like ejection of Auger electrons or characteristic X-ray photons. But they are very unlikely in the energy range considered.

Although inelastic collisions are decisive for the energy transfer, the radiation damage in the patient strongly depends on the spatial distribution of electrons in their passage through matter. They dominantly undergo multiple scattering events with small deviations. However, single backward scattering events also occur frequently which leads to tortuous trajectories of electrons. To a large extent, such trajectories are due to elastic collisions. We therefore focus on realistic simulation of elastic processes in our model. This is achieved by transport coefficients extracted from the ICRU77 database [85]. Inelastic transport coefficients are obtained in the same way.

Elastic and soft inelastic events lead to small energy loss. With kinetic energies above 1 keV, electrons are assumed to lose their energy continuously [62]. Because of this, the Continuous Slowing Down (CSD) approximation can be implemented to model energy loss of electrons. However, it implies that large energy loss fluctuations, caused by hard inelastic collisions, are neglected. For example, the Boltzmann-CSD approximation models electrons changing their direction of flight discretely [100]. The

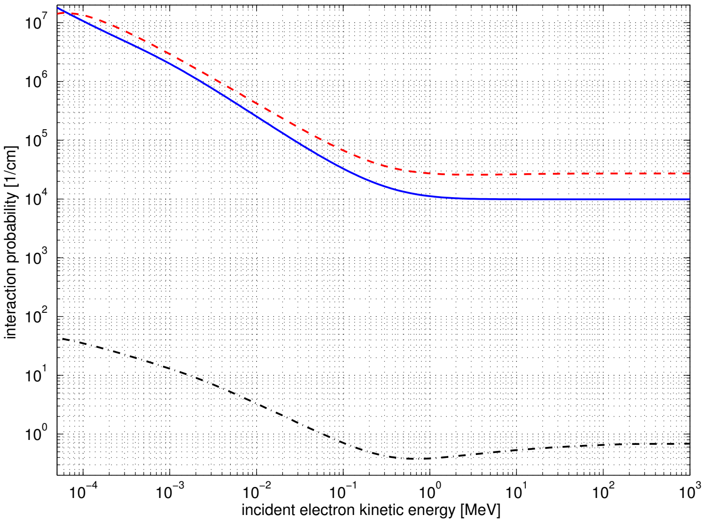


Figure 1.1: Interaction probability per unit path length in water [155]: inelastic (red dashed line), elastic (blue solid line) and bremsstrahlung (black dash-dot line).

advantage over the FP approximation from Section 1.1.1 is that it allows large-angle scattering.

To achieve a reliable prediction of the dose it is necessary to model all physical phenomena prevailing in the applied energy range. This can be determined by considering the interaction probability. Fig. 1.1 shows the probability of interaction per unit path length in water which is the inverse of the mean free path between two collisions. Compared to elastic and inelastic scattering, bremsstrahlung events are very improbable in the whole energy range. Their difference is a factor of roughly 10^5. Hence, bremsstrahlung transport coefficients might be discarded in modelling electron transport.

However, this does not mean that bremsstrahlung emission is completely negligible. Although there is a small number of bremsstrahlung collisions the energy loss per interaction increases rapidly for higher energies (Section 1.2.7). For an accurate description of energy deposition in tissue, bremsstrahlung effects must be an important part of our model at high energies.

1.2.2 Terminology

Some differential cross sections are scaled to electron rest energies, i.e., incoming particles of kinetic energy ε have an absolute kinetic energy of $E = 0.51099906 \cdot \varepsilon$ MeV or $E' = 0.51099906 \cdot \varepsilon'$ for outgoing particles, respectively. Nondimensional expressions often simplify physical formula. Therefore, we also introduce the scaled velocity $\beta = v/c$ of an incoming particle. Its square can be written as $\beta^2 = \varepsilon(\varepsilon + 2)/(\varepsilon + 1)^2$.

The scattering angle in the laboratory system is denoted by $\theta \in [0, 2\pi]$ or $\mu := \cos(\theta)$. Z is the atomic number of the irradiated medium and Z_{eff} is referred to as effective atomic number ($Z = 10$ and $Z_{\text{eff}} = 7.51$ for water). ε_B is known as the mean ionization energy of electrons which is equal to $\varepsilon_B = 1.4677 \cdot 10^{-4}$ in units of the electron rest energy for water molecules (PENELOPE table for liquid water [155]). Moreover ρ_e denotes the electron density in matter ($\rho_e = 3.3428847 \cdot 10^{23}\text{cm}^{-3}$ for water), $\rho_c = \rho_e/Z$

is the density of atomic cores and $r_e = 2.8179 \cdot 10^{-13}$cm the classic electron radius.

A useful quantity is also the molecular density defined as

$$\mathcal{N} := N_A \frac{\rho}{A}$$

where $N_A = 6.02214 \cdot 10^{23}\text{mol}^{-1}$ is the Avogadro number, A the molar mass and ρ the mass density of the medium. Its numerical value for water is $\mathcal{N}_{H_2O} = 3.3428847 \cdot 10^{22}\text{cm}^{-3}$. As cross sections are of a tiny order it is convenient to use the unit barn: 1 barn $= 10^{-24}$ cm^2.

1.2.3 ICRU Database

One important achievement in electron transport problems was the calculation of accurate scattering cross sections. The ICRU Report 77, provided by the International Commission on Radiation Units & Measurements, contains differential cross sections for elastic and inelastic scattering of electrons and positrons for different materials and energies between 50 eV and 100 MeV [85]. To obtain transport coefficients $\xi_{e,n}(E)$ we use these cross sections and proceed in the following way:

(a) The ELSEPA code system, distributed with the report, calculates elastic and inelastic angular differential cross sections for a fixed energy E,

$$\sigma^{\text{el,inel}}(E, \mu) = \int_0^\infty \sigma_s^{\text{el,inel}}(E', E, \mu) dE',$$

in tabulated form for discrete μ and $\sigma^{\text{el,inel}}(E, \mu)$. For a predetermined set of energy values between 50eV and 100MeV, data for $\sigma^{\text{el,inel}}(E, \mu)$ are extracted from these files.

(b) With this we calculate the n-th transport coefficient for a fixed energy E,

$$\xi_{e,n}^{\text{el,inel}}(E) = 2\pi\mathcal{N} \int_{-1}^{1} (1-\mu)^n \sigma^{\text{el,inel}}(E, \mu) d\mu, \quad \text{with} \quad \mu = \cos(\theta),$$

via numerical integration of the tabulated cross sections $\sigma^{\text{el,inel}}(E, \mu)$ by means of the trapezodial rule. Additionally, we multiply the result by the molecular density of the transmitted matter $\mathcal{N}$.

(c) Again, all computed results of $\xi_{e,n}^{\text{el,inel}}(E)$ are stored and used as a look-up table. To obtain the n-th transport coefficient at the desired energy E, this tabulated data is linearly interpolated.

Finally, the desired transport coefficient for our equations reads:

$$\xi_{e,n}(E) := 2\pi\mathcal{N} \int_{-1}^{1} (1-\mu)^n (\sigma^{\text{el}}(E, \mu) + \sigma^{\text{inel}}(E, \mu)) d\mu.$$

1.2.4 Elastic Cross Sections

Although inelastic collisions are decisive for the energy transfer, the radiation damage in the patient strongly depends on the spatial distribution of electrons in their passage through matter. Tortuous trajectories of electrons are mainly due to elastic collisions. That is why knowledge of realistic elastic scattering cross sections is important for predicting spatial dose distributions.

Analytic Formulas

Various types of approximation formula have been published in literature. In the end, it has been illustrated for water molecules that the integrated screened Rutherford formula fits the experimental data of the total elastic cross section quite well. However, this does not imply that the angular dependence is also described with the same accuracy. On the contrary, LaVerne and Pimblott emphasize that the screened Rutherford cross section predicts too little forward scattering [106]. At relativistic energies, the screened Rutherford cross section has to be modified by a factor which accounts for spin effects. This leads to the Mott formula which still depends on the choice of a certain screening parameter:

Mott Angular Differential Cross Section

$$\sigma^{e}_{\mathrm{Mott}}(\varepsilon,\mu) = \frac{Z^2 r_e^2 (1+\varepsilon)^2}{[\varepsilon(\varepsilon+2)]^2 (1+2\eta(\varepsilon)-\mu)^2}\left[1-\frac{\varepsilon(\varepsilon+2)}{2(1+\varepsilon)^2}(1-\mu)\right] \tag{1.15}$$

The screening parameter is denoted by $\eta(\varepsilon)$. We choose the following four expressions and compare them to tabulated data:

$$\text{Davisson/Evans [40]:} \quad \eta_{DE}(\varepsilon) = \frac{Z^{2/3}\pi^2}{137^2\varepsilon(\varepsilon+2)}, \tag{1.16}$$

$$\text{Wentzel [72]:} \quad \eta_{W}(\varepsilon) = \frac{1.7\cdot 10^{-5} Z^{2/3}}{\varepsilon(\varepsilon+2)}, \tag{1.17}$$

$$\text{Grosswendt/Waibel [69]:} \quad \eta_{GW}(\varepsilon) = (1.64 - 0.0825\ln(\varepsilon\cdot 0.511\cdot 10^6))\,\eta_W(\varepsilon), \tag{1.18}$$

$$\text{Molière [72]:} \quad \eta_{M}(\varepsilon) = \left[1.13 + 3.76\frac{Z^2(\varepsilon+1)^2}{137^2\varepsilon(\varepsilon+2)}\right]\eta_W(\varepsilon). \tag{1.19}$$

In collisions with atoms or molecules, elastic scattering is the dominant contribution to angular deflections of electrons. Although a certain amount of energy is transferred to the target, this effect is negligible due to the relatively small electron mass. Since the proton mass is roughly 3600 times larger, the recoil of the target can be discarded.

ICRU77 Database

A good agreement of calculated elastic cross sections with experimental observations can be achieved by accessing tabulated data. We generate tables for angular differential cross sections $\sigma^{\mathrm{el}}(E,\mu)$ and calculate the necessary quantities as described in Section 1.2.3.

We conclude from Figure 1.2 that for large energies the cross section is monotonically decreasing and has an *extremely* high peak at tiny scattering angles (i.e., it levels off in logarithmic scale). At $E = 100$ MeV the cross section decreases within 23 orders of magnitude. Hence, electrons are hardly deflected and travel in an almost straight line. However, for increasing energies, the situation changes: The maximum value is much smaller and the scattering behavior becomes less forward-peaked. Even more, electrons are scattered backwards at $E = 10^{-4}$ MeV.

Figure 1.3 illustrates integrated elastic cross sections $\xi^{\mathrm{el}}_{e,n}/\mathcal{N}$ in liquid water for different orders n. They are all monotonically decreasing. For $E \geq 10^{-3}$, $\xi^{\mathrm{el}}_{e,1}$ is

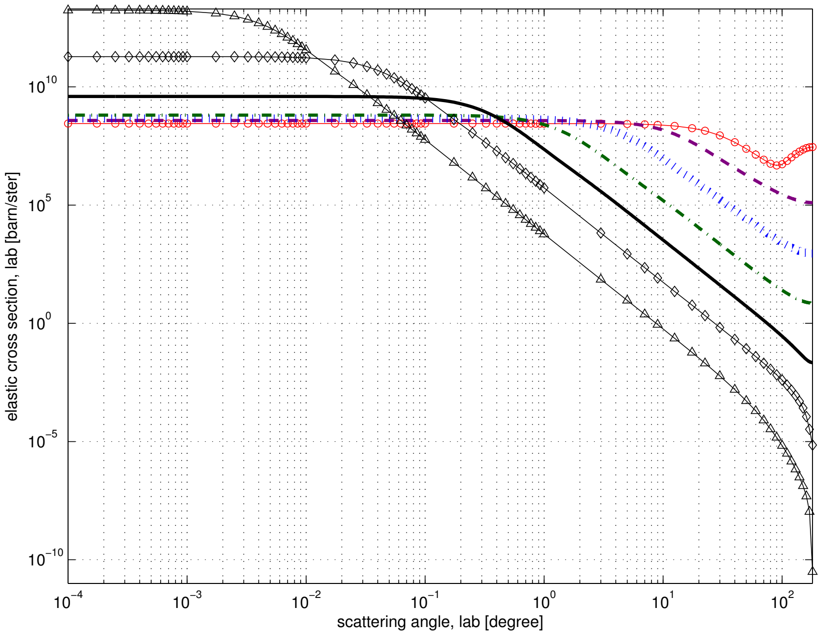


Figure 1.2: Elastic cross sections σ^{el} (extracted from ICRU77): 10^{-4} MeV (red solid circle line), 10^{-3} MeV (purple dotted line), 10^{-2} MeV (blue dashed line), 0.1 MeV (dark green dash-dot line), 1 MeV (black solid line), 10 MeV (black solid diamond line), 100 MeV (black solid triangle line).

always larger than $\xi^{\text{el}}_{e,2}$ but, as E decreases, their difference reduces more and more. For increasing $n \geq 2$ the deviation between two consecutive $\xi^{\text{el}}_{e,n}$ is so small that the assumption in the generalized FP-asymptotics (Section 1.1.1) is not fulfilled.

Comparison

We compare the first transport coefficient calculated by aforementioned analytic Mott formulas to those extracted from ICRU77 data:

$$\text{ICRU77:} \qquad \xi^{\text{el}}_{e,n}(\varepsilon) = 2\pi \mathcal{N}_{H_2O} \int_{-1}^{1} (1-\mu)^n \sigma^{\text{el}}(\varepsilon,\mu) d\mu, \tag{1.20a}$$

$$\text{Analytical:} \qquad \Sigma^{\text{Mott}}_{e,n}(\varepsilon) = 2\pi \frac{\rho_e}{Z} \int_{-1}^{1} (1-\mu)^n \sigma^{e}_{\text{Mott}}(\varepsilon,\mu) d\mu. \tag{1.20b}$$

Note a subtle difference in the notation: The cross section σ^{el} from ELSEPA is scaled by the molecular density $\mathcal{N}_{H_2O}$ whereas the analytical expression σ^{e}_{Mott} is multiplied by the core density $\rho_c = \rho_e/Z$.

The results are shown in Figure 1.4. All displayed functions reveal a common behavior: At high energies, they are all close to each other. Nevertheless, at low energies (between 50 eV and 0.1 MeV) deviations can be observed. The proposed formula of Molière first increases, achieves a maximum value and decreases afterwards. However, this is not the case for the remaining functions: They are all monotonically decreasing. The largest difference to ICRU77 data shows Molière's transport coefficient. Using the screening parameter proposed by Davisson/Evans in (1.16) also makes the first transport coefficient deviate in the order of 2 magnitudes at low energies. A good

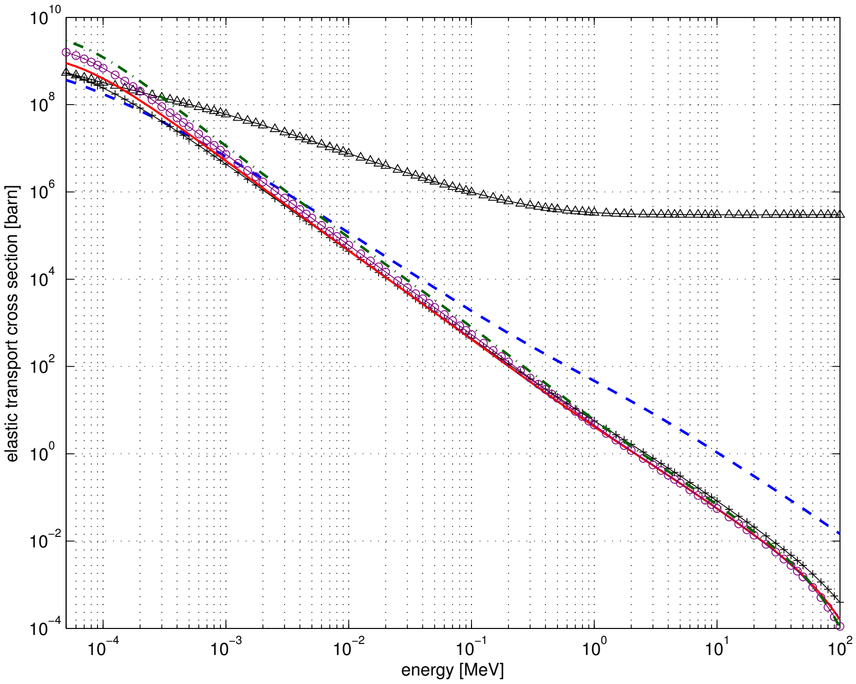


Figure 1.3: Integrated elastic cross sections $\xi_{e,n}^{\mathrm{el}}/\mathcal{N}$ (extracted from ICRU77): $n = 0$ (black solid triangle line), $n = 1$ (blue dashed line), $n = 2$ (black solid plus line), $n = 3$ (red solid line), $n = 4$ (purple solid circle line), $n = 5$ (dark green dash-dot line).

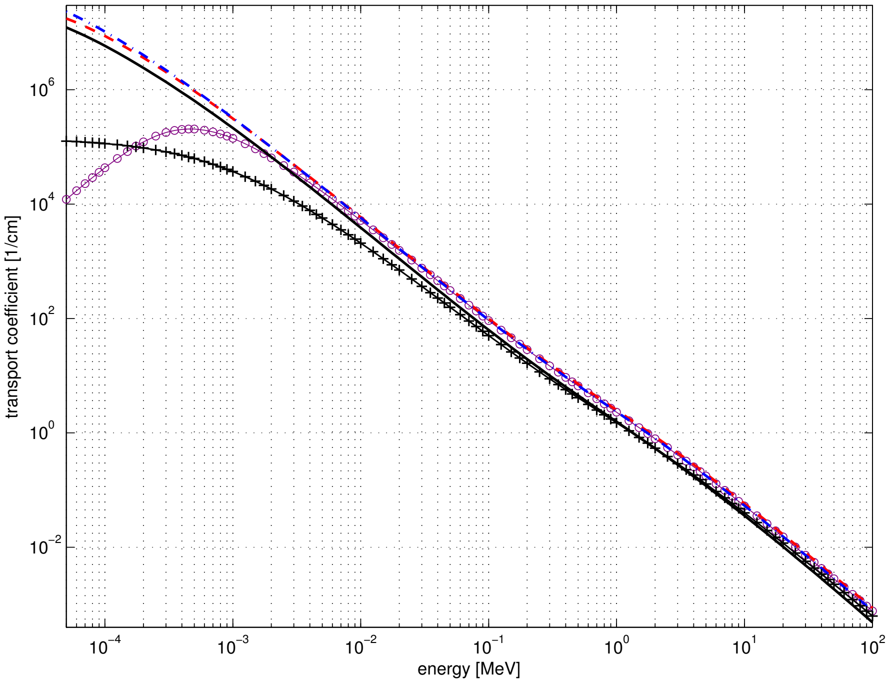
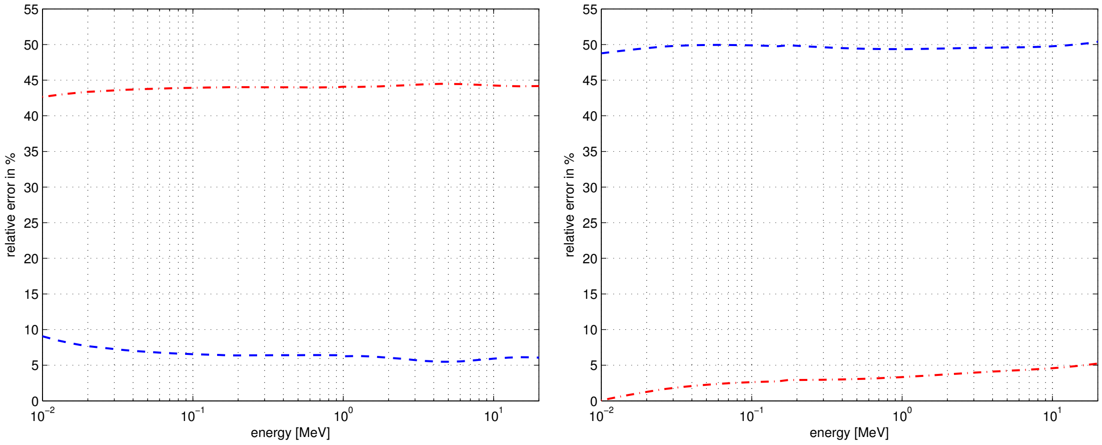
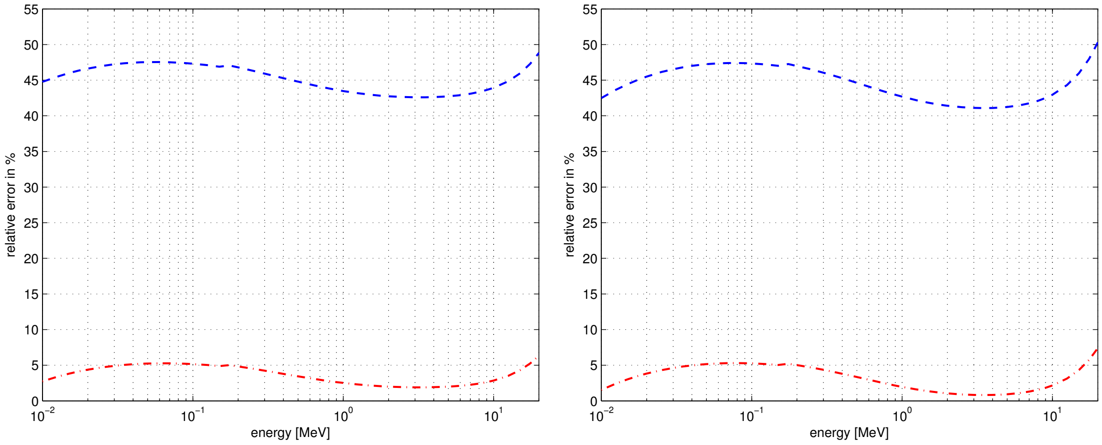


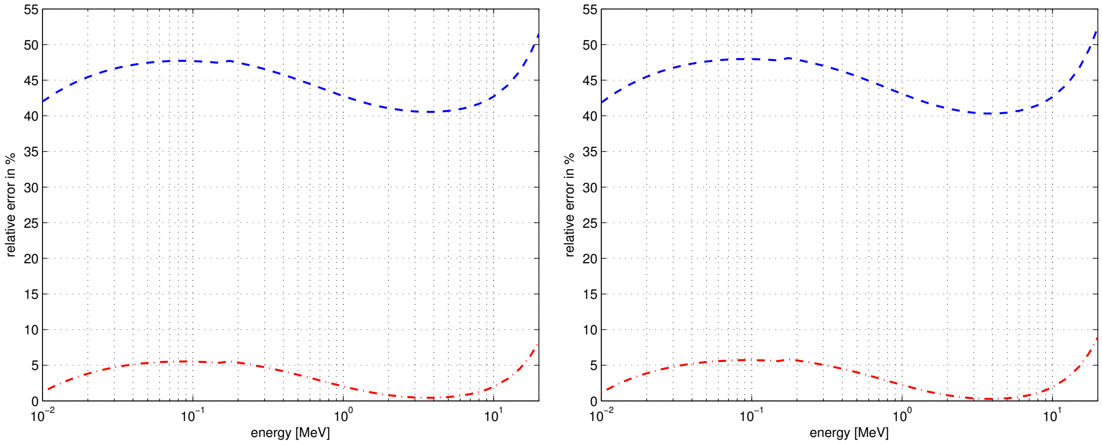
Figure 1.4: 1st elastic transport coefficient: Molière (purple solid circle line), Grosswendt/Waibel (red dashed line), Wentzel (blue dash-dot line), Davisson/Evans (black solid cross line), ICRU77 (black solid line)



(a) 0th elastic transport coefficient (b) 1st elastic transport coefficient



(c) 2nd elastic transport coefficient (d) 3rd elastic transport coefficient



(e) 4th elastic transport coefficient (f) 5th elastic transport coefficient

Figure 1.5: Relative errors for elastic transport coefficients in water: Wentzel-Mott formula from (1.17) (blue dashed line), fitted Wentzel-Mott formula with parameters from Table 1.1 (red dash-dot line).

	Z	ρ_e $[10^{23}\text{cm}^{-3}]$
physical values	10	3.34
Wentzel formula fitted in Z, ρ_e	15.82	1.51
Grosswendt/Waibel formula fitted in Z, ρ_e	8.62	2.42

Table 1.1: Fitting the model parameters: least squares fit of the first transport coefficient to ICRU77 data.

agreement with ICRU77 data is observed with the Mott formulas from (1.17) and (1.18). It is striking that there are hardly discrepancies between both functions and they are throughout close to the reference.

Although screening parameters by Wentzel and Grosswendt/Waibel give quite good results we increase their accuracy by performing a least squares fit for $\xi^{\text{el}}_{e,1}$. Note that this is purely formal and has no physical justification. It solely pursues the goal of implementing transport parameters which are as realistic as possible. The first transport coefficient is chosen because our most frequently used approximation for dose calculations is the M1-model so far.

Using ρ_e and Z as model parameters we obtain fitting parameters given in Table 1.1. Similar to the behavior in Figure 1.4, it turns out that no significant differences between the fitted Wentzel and Grosswendt/Waibel cross sections occur. As the mathematical expression for Wentzel's screening parameter in (1.17) is less complicated we use Wentzel's screening for the Mott formula in the following.

Figure 1.5 additionally displays relative errors of higher order transport coefficients whose accuracy is important for the generalized Fokker-Planck asymptotics from Section 1.1.1. Three quantities are compared: $\xi^{\text{el}}_{e,n}$ and $\Sigma^{\text{Mott}}_{e,n}$ from eqs. (1.20) where the latter is calculated by different values for Z and ρ_e. On the one hand, physical values for Z and ρ_e are used and, on the other hand, fitting parameters from Table 1.1.

Transport coefficients of order $n = 0$ to $n = 5$ are presented in Figure 1.5. Except for the total cross section $n = 0$, higher order coefficients are all close to ICRU77 data for a wide energy range. Only at high energies, discrepancies occur. The fitted Wentzel-Mott formula shows relative errors below 5 % for energies between 10^{-2} MeV $\leq E \leq 20$ MeV and $n = 1$. For $n \neq 1$ errors are below 10%. Outside the latter energy range and without the least squares fit larger errors can be observed.

1.2.5 Inelastic Cross Sections

Inelastic cross sections play an important role for modelling energy loss in collisions. Our approximate equations in Section 1.1.1 and Section 1.1.2 include energy loss mechanisms by means of the stopping power. Although angular deflections in inelastic processes are of minor importance we still want to present their contribution here.

Analytic Formulas

The study of inelastic collisions requires knowledge of interactions where ionisation processes occur or the final quantum state of the target is changed. In particular, inner shell ionisations have to be computed accurately. Although quantum theory provides the necessary information, the resulting analytical expressions are either too complicated or require numerical look-up tables. Fortunately, approximative models can be developed to describe realistic inelastic collisions. Due to their complexity, their extensive depiction goes beyond the purpose of these notes. Instead, we refer to [155] and references therein for further studies.

However, we point out one crucial formula which is sometimes solely used in literature to include electron-electron collisions and derive the stopping power thereof. Binary collisions where an incident electron of energy ε' interacts with a free electron at rest is characterized by the following expression [128]:

Møller Angular Differential Cross Section

$$\sigma^e_{\text{Møller}}(\varepsilon,\varepsilon',\mu) = \frac{2\pi r_e^2(\varepsilon+1)^2}{\varepsilon(\varepsilon+2)m_ec^2}\delta_M(\mu,\varepsilon,\varepsilon')\cdot\left[\frac{1}{\varepsilon'^2}+\frac{1}{(\varepsilon-\varepsilon')^2}+\frac{1}{(\varepsilon+1)^2} - \frac{2\varepsilon+1}{(\varepsilon+1)^2\varepsilon'(\varepsilon-\varepsilon')}\right], \tag{1.21a}$$

$$\delta_M(\mu,\varepsilon,\varepsilon') = \delta\left(\mu - \sqrt{\frac{(\varepsilon-\varepsilon')(\varepsilon+2)}{\varepsilon(\varepsilon-\varepsilon'+2)}}\right). \tag{1.21b}$$

Although the Møller formula will not be used to compute inelastic transport coefficients here, it will be important in Section 1.2.7 where the analytical stopping power is derived from the Møller cross section. Instead, we will use an approximation which is often implemented in simulation codes to include angular deflections resulting from inelastic collisions [72]: the factor Z^2 in (1.15) is replaced by $Z(Z+1)$.

ICRU77 Database

Similar to elastic cross sections, we extract inelastic angular differential cross sections from the ICRU77 database. Again, the trapezoidal integration rule yields the n-th inelastic transport coefficient defined by:

$$\xi^{\text{inel}}_{e,n}(E) := 2\pi\,\mathcal{N}_{H_2O}\int_{-1}^{1}(1-\mu)^n\sigma^{\text{inel}}(E,\mu)d\mu \quad \text{with} \quad \mu=\cos(\theta). \tag{1.22}$$

Figure 1.6 reveals that, in contrast to the elastic case, the difference between consecutive higher order transport coefficients is much larger. However, except for the total cross section ($n=0$), their absolute values are much smaller compared to the elastic contribution (Figure 1.7). Especially for large energies and higher orders, the elastic transport coefficient is several magnitudes larger.

1.2.6 Comparison

The accuracy of the Wentzel-Mott formula, corrected by $Z(Z+1)$, is investigated here. We compare the sum of elastic and inelastic ICRU77 transport coefficients to the analytical formula (1.15) where Z^2 is replaced by $Z(Z+1)$:

$$\text{ICRU77:} \quad \xi^{\text{tot}}_{e,n}(\varepsilon) = \mathcal{N}_{H_2O}\left(\xi^{\text{el}}_{e,n}(\varepsilon)+\xi^{\text{inel}}_{e,n}(\varepsilon)\right) \tag{1.23}$$

$$= 2\pi\,\mathcal{N}_{H_2O}\int_{-1}^{1}(1-\mu)^n(\sigma^{\text{el}}(\varepsilon,\mu)+\sigma^{\text{inel}}(\varepsilon,\mu))d\mu \tag{1.24}$$

$$\text{Analytical:} \quad \Sigma^{\text{tot}}_{e,n}(\varepsilon) = 2\pi\frac{\rho_e}{Z}\int_{-1}^{1}(1-\mu)^n\frac{Z(Z+1)}{Z^2}\sigma^e_{\text{Mott}}(\varepsilon,\mu)d\mu. \tag{1.25}$$

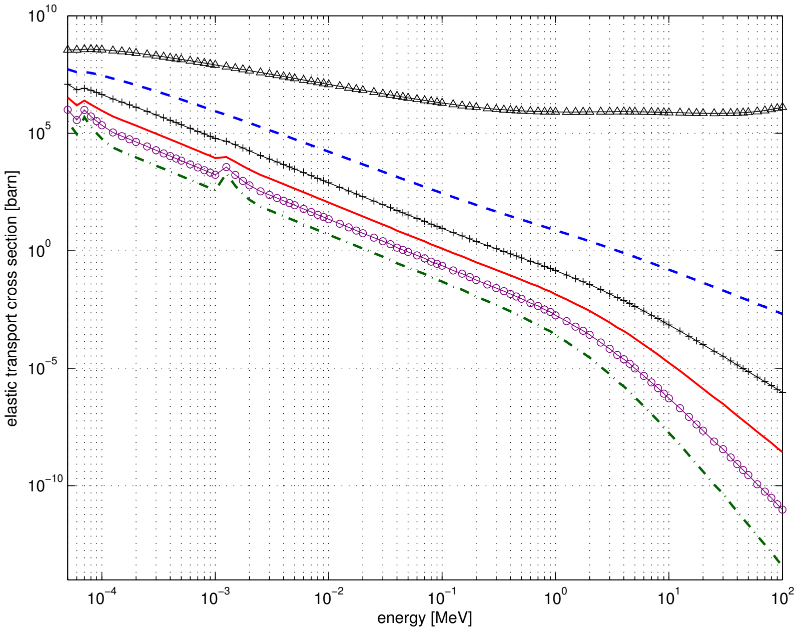


Figure 1.6: Integrated inelastic cross sections $\xi_{e,n}^{\text{inel}}/\mathcal{N}$ in water (extracted from ICRU77): $n = 0$ (black solid triangle line), $n = 1$ (blue dashed line), $n = 2$ (black solid plus line), $n = 3$ (red solid line), $n = 4$ (purple solid circle line), $n = 5$ (dark green dash-dot line).

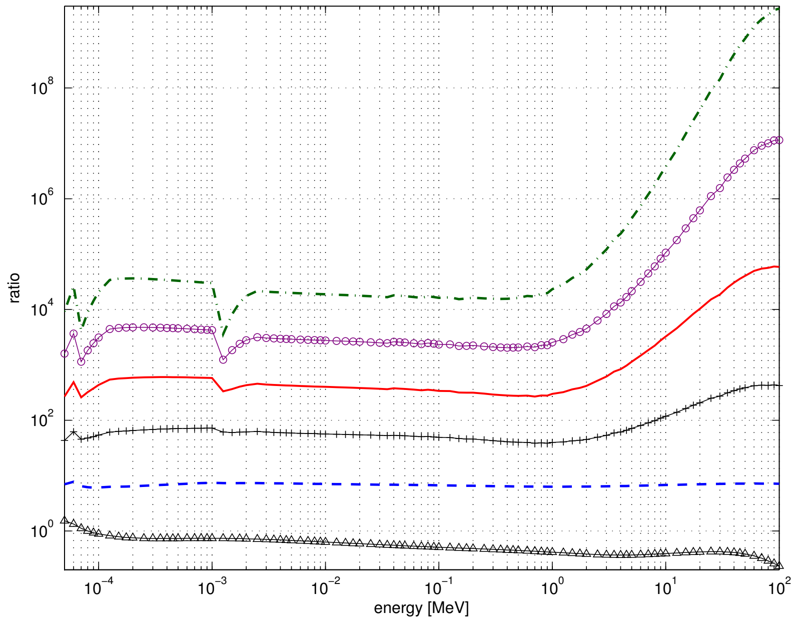


Figure 1.7: Ratio of elastic over inelastic transport coefficient in water (extracted from ICRU77): $n = 0$ (black solid triangle line), $n = 1$ (blue dashed line), $n = 2$ (black solid plus line), $n = 3$ (red solid line), $n = 4$ (purple solid circle line), $n = 5$ (dark green dash-dot line).

Three different values are used for the atomic number Z and electron density ρ_e: physical and fitted values from Table 1.1 and $Z = Z_{\text{eff}}$ combined with $\rho_e = 3.3428847 \cdot 10^{23}$ cm^{-3}. Z_{eff} is referred to as effective atomic number and is often employed in radiation physics.

Relative errors for the functions are plotted in Figure 1.8. Errors for the analytical formulas are throughout very large and not useful for computations. For $n \geq 1$ the total cross section calculated by the Wentzel-Mott (1.25) with $Z = Z_{\text{eff}}$ is always superior to the Wentzel-Mott formula with $Z = 10$. The atomic number $Z = 10$ in the Wentzel-Mott formula only yields better results for the 0th total transport coefficient.

Due to the increase of computational effort, evaluations of tabulated data are not always desirable. Instead, one sometimes prefers to use analytic expressions in codes. Figure 1.7 illustrates that *elastic* transport coefficients in water are largely dominant for $n \geq 1$. For the M1-model, the most accurate analytical formula will therefore be the *elastic* transport coefficient with fitted parameters.

1.2.7 Stopping Power

In their passage through matter, electrons deposit their energy in collision processes. Up to approximately 600 MeV/Z this is, to a big amount, due to interactions with shell electrons which lead to ionisation or excitation of the target [149]. For higher energies a further phenomenon gains in importance: In the electrostatic field of atomic or molecular cores, electrons are accelerated and dispense energy by irradiation of photons. This event is called bremsstrahlung emission.

The average energy loss per unit path length is given by the stopping power. For electrons, one usually distinguishes between the collision and radiative stopping power. The former describes energy loss by interaction of the projectile with shell electrons of the target whereas the latter quantifies the effect of bremsstrahlung processes:

- **collision stopping power:** This is the average energy loss per unit path length caused by ionization or excitation of the target as a consequence of inelastic collisions.
- **radiative stopping power:** This quantity is defined as the average energy loss per unit path length caused by bremsstrahlung quanta emitted in collisions.

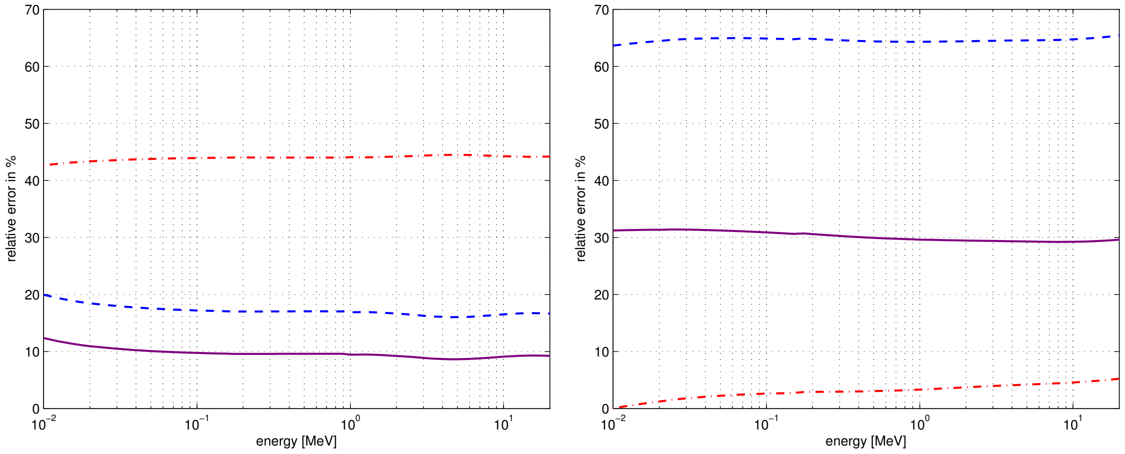
Analytic Formulas

Given an inelastic cross section, we can calculate the stopping power associated with the corresponding interactions by energy-integrating the cross section times energy. As the Møller cross section from eqs. (1.21) describes ionization processes of electrons the collision stopping power is given by:

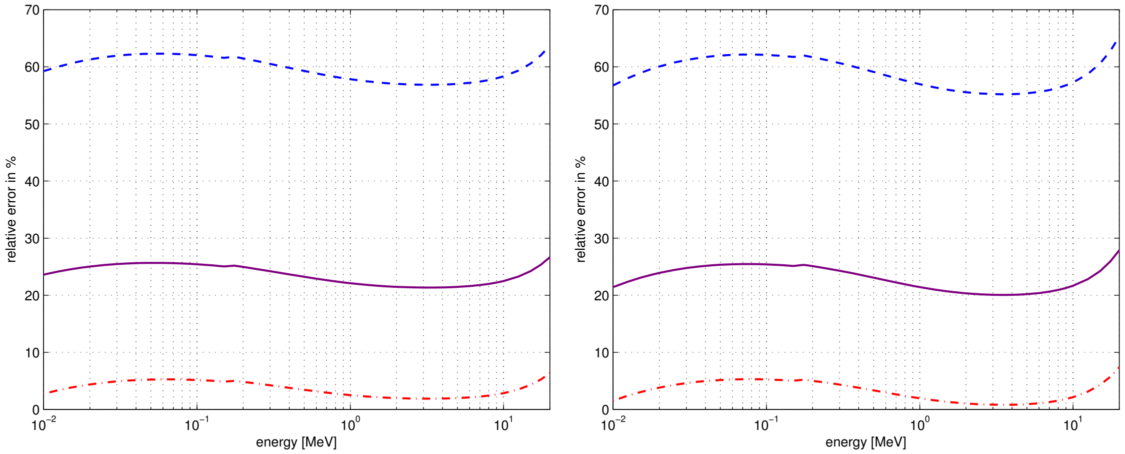
Møller Stopping Power

$$S^e_{\text{Møller}}(\varepsilon) = \rho_e (m_e c^2)^2 \int_{\varepsilon_B}^{(\varepsilon-\varepsilon_B)/2} \varepsilon \sigma^e_{\text{Møller}}(\varepsilon, \varepsilon') d\varepsilon'. \tag{1.26}$$

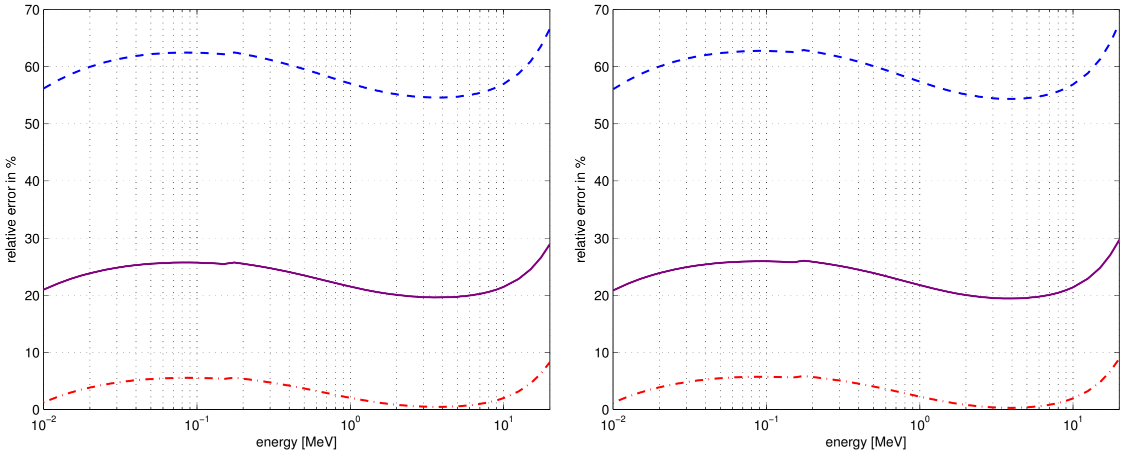
It is important to emphasize that $S^e_{\text{Møller}}$ is in units of MeV/cm. As in many physical publications and databases the stopping power S^e_{ph} is given in units of MeVcm^2/g it



(a) 0th total transport coefficient (b) 1st total transport coefficient



(c) 2nd total transport coefficient (d) 3rd total transport coefficient



(e) 4th total transport coefficient (f) 5th total transport coefficient

Figure 1.8: Relative errors for transport coefficients in water: total Wentzel-Mott (1.25) with $Z = 10$ (blue dashed line), total Wentzel-Mott (1.25) with $Z = Z_{\text{eff}}$ (purple solid line), fitted *elastic* Wentzel-Mott formula with parameters from Table 1.1 (red dash-dot line).

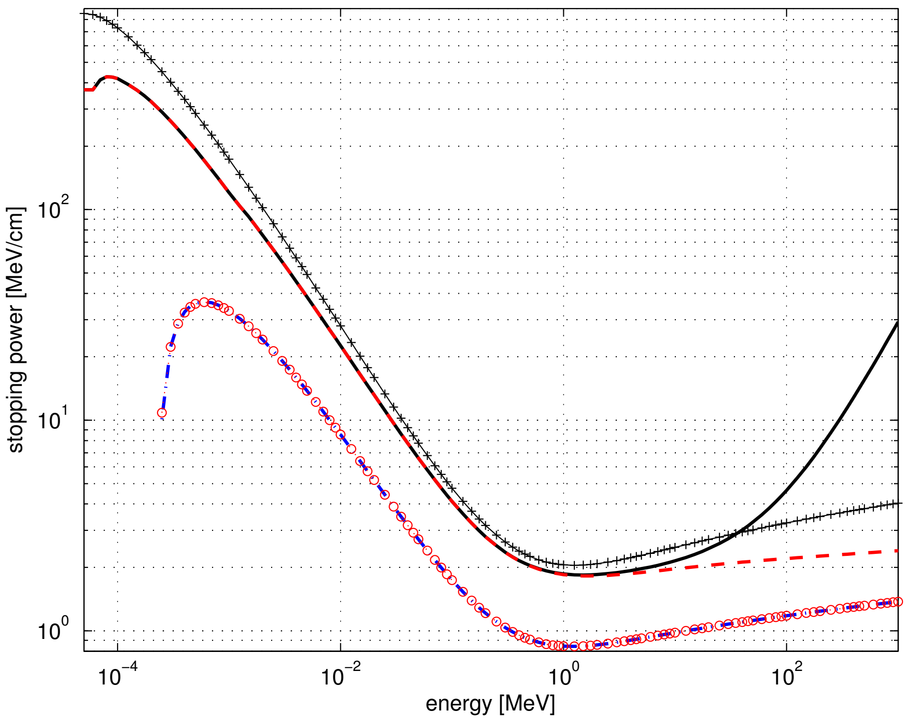


Figure 1.9: Electron stopping power: collision+radiative stopping power (black solid line), collision stopping power (red dashed line), Bethe's formula (black solid plus line), PENELOPE's stopping power from close interactions (red circles).

is necessary to multiply S^e_{ph} by the mass density of the material ρ: $S^e_{\mathrm{Møller}} = \rho S^e_{\mathrm{ph}}$. If the simulation code requires dimensionless scaled energies in units of the electron rest energy $S^e_{\mathrm{Møller}}$ must be additionally divided by $m_e c^2 = 0.511\mathrm{MeV}$.

One method to validate the accuracy of inelastic cross sections is to compare the stopping power calculated by energy-integration to data which is benchmarked against experiments (Figure 1.9). Again, we use tabulated stopping power data provided by the ICRU77 [85] as our benchmark. Figure 1.9 displays the collision (without bremsstrahlung) as well as the total (including bremsstrahlung emission) stopping power. For $E \geq 10^{-4}$ MeV, the total stopping power is monotonically decreasing until it reaches a minimum at ≈ 1.5 MeV. At 6 MeV, bremsstrahlung effects contributes 5% to the total stopping power. The larger the energy the more dominant bremsstrahlung emission becomes and it prevails for $E \geq 100$ MeV.

The basic theory in ICRU77 is derived by Bethe's stopping power equation which is corrected by several effects. For comparison reasons, we include the behavior of Bethe's stopping power in Figure 1.9 which results from the following expression [185]:

$$S_{\mathrm{Bethe}}(r, \varepsilon') = \frac{4\pi r_e^2 \rho_e m_e c^2}{\beta^2} \left[\ln\left(\frac{2 m_e c^2 \beta^2}{I \cdot (1-\beta^2)} \right) - \beta^2 \right], \tag{1.27}$$

where $I = 75$ eV is the mean excitation energy for water.

The difference between the ICRU77-collision stopping power and the aforementioned Møller stopping power is large for the whole energy range. Especially at small energies, the Møller stopping power is at least one magnitude smaller and becomes even negative for $E < 2.5 \cdot 10^{-4}$ MeV.

In the Monte Carlo code PENELOPE [155], inelastic collisions are described by distant and close electron-electron interactions. Distant interactions occur with bound shell electrons. In close collisions, however, target electrons are assumed to be free and at rest which corresponds to the assumptions for the Møller cross section. This is why, no difference can be observed Figure 1.9 between the Møller stopping power and PENELOPE's formula for the stopping power for close interactions:

$$\sigma_{\text{clo}}(\varepsilon) = \frac{2\pi r_e^2 Z \mathcal{N}_{H_2O}\, m_e c^2}{\beta^2} \Big[\ln\left(\frac{\varepsilon}{I}\right) + 1 - \left(1 + \beta^2 + 2\sqrt{1-\beta^2}\right) \ln(2) \\ + \frac{1}{8}\left(1 - \sqrt{1-\beta^2}\right)^2 \Big] \tag{1.28}$$

$$\text{with } I = 75\text{eV}. \tag{1.29}$$

1.2.8 Dose Calculations

Our final goal is to simulate electron transport in matter and compute the deposited dose in tissue as accurate as possible. For this purpose, Monte Carlo system codes are considered one of the most accurate simulation tools which are additionally benchmarked against experiments. Comparisons to Monte Carlo simulations therefore provide information about the accuracy of the mathematical model as well as of the included physical cross sections. In [137], a certain choice of deterministic generalized Fokker-Planck dose computations were performed and compared to stochastic Monte Carlo results. Here, we only highlight one test case of electron propagation in pure water with two different initial energy beams to demonstrate the accuracy of cross sections discussed above.

The model equations from (1.9) are solved numerically for GFP_2 by discretizing the following initial boundary value problem:

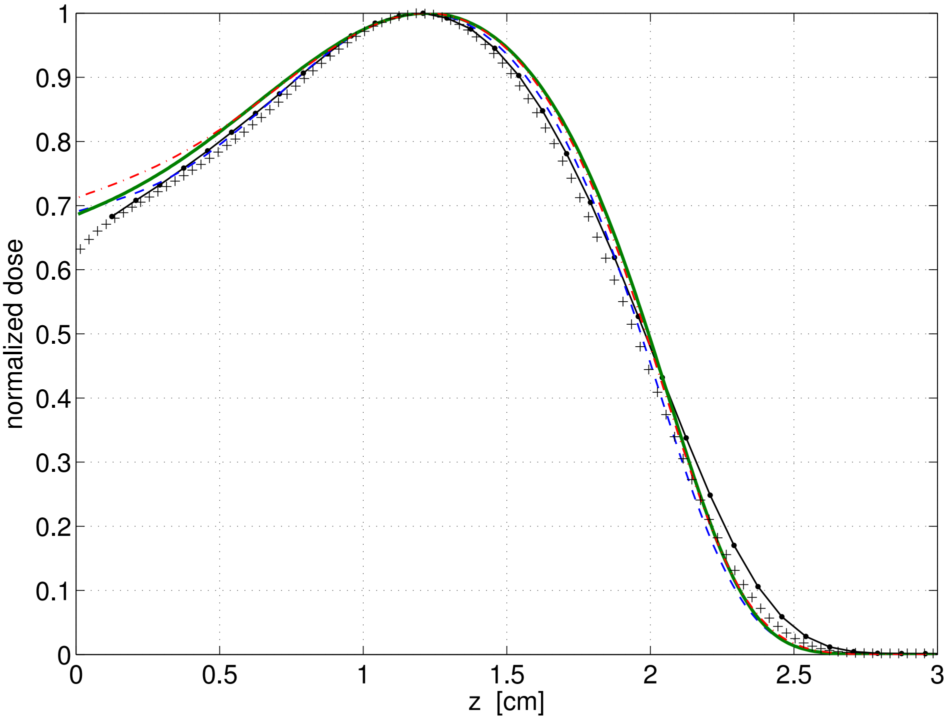
$$\sigma_a \Phi_0(z,E,\mu) + \frac{\partial \Phi_0(z,E,\mu)}{\partial z} \cdot \mu = \alpha L_\mu \Phi_1(z,E,\mu) + \frac{\partial (S(z,E)\Phi_0(z,E,\mu))}{\partial E} \\ (I - \beta L_\mu)\Phi_1(z,\mu,s) = \Phi_0(z,E,\mu) \tag{1.30}$$

$$\begin{array}{lll} \underline{\text{BC}}: & \Phi_0(0,E,\mu) = 10^5 \cdot e^{-200(1-\mu)^2} e^{-50(E_0-E)^2} & 1 \geq \mu > 0, E \in I. \\ & \Phi_0(d,E,\mu) = 0 & -1 \leq \mu < 0, E \in I. \end{array}$$

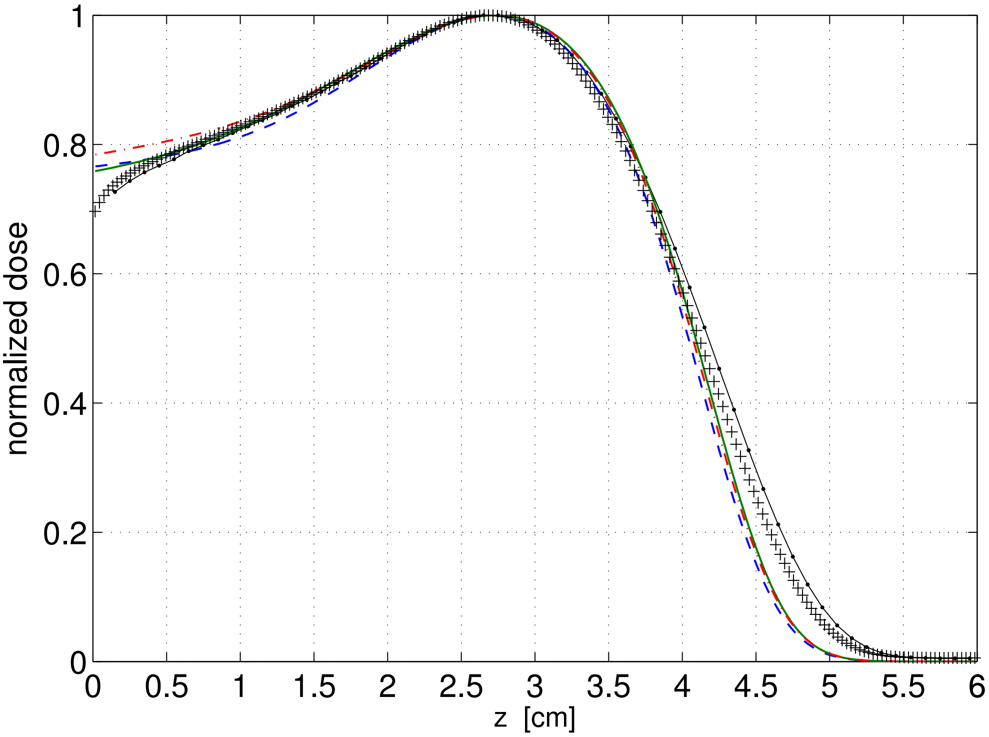
Eq. 1.30 describes the propagation of electrons through matter with a monoenergetic pencil beam of energy E_0 irradiated orthogonally to the boundary surface of the material. This beam is modelled by a product of two narrow Gaussian functions around $\mu = 1$ and $E = E_0$. After computing the solution, one can calculate the absorbed dose by (1.10).

Dose calculations in a semi-infinite water phantom are performed for 5 MeV and 10 MeV beams. As our benchmark, we use solutions of the Monte Carlo code systems GEANT4 (standard physics package) [3,5] and PENELOPE [155]. The following criterion is applied to quantify the accuracy of solutions in a homogeneous geometry [177]: 2%/2mm (pointwise difference within 2% or 2mm horizontal distance-to-agreement).

We implement a semi-discretization to solve (1.30) numerically: First, the angular and spatial variable is discretized with finite differences so that we end up with an



(a) 5 MeV electron beam.



(b) 10 MeV electron beam.

Figure 1.10: Normalized dose in liquid water: FP (blue dashed line), GFP2 (darkgreen solid line), GFP3 (red dash-dot line), GEANT4 (black plus signs), PENELOPE (black solid circle line).

ordinary differential equation in the energy variable E. Second, the ODE-solution is obtained by the embedded 2nd/3rd order Runge-Kutta MATLAB solver `ode23` solving from the initial condition $\Phi_0(z, E_{\max}, \mu) = 0$ backward in energy to $E = 0$. Morel's second-order finite difference discretization [131] is applied for the spherical Laplace-Beltrami operator (32 Gauss-Legendre quadrature points in μ) and the first-order upwind scheme for the spatial variable (350 points in z).

Characteristic electron dose profiles in water first provide a high surface dose, increase to a maximum at a certain depth and drop off with a steep slope afterwards (Figure 1.10). Transport coefficients $\xi_{e,n}^{\text{tot}}$ extracted from ICRU77 are used in our calculations. All approximations are close to each other because transport coefficients for water do not fall off highly enough within our energy interval (Figure 1.3). Solutions for GFP$_4$ and GFP$_5$ are omitted because they overlap with GFP$_3$ in our plots.

All in all, the calculated results agree well with PENELOPE and GEANT4. All dose profiles for a 5 MeV beam satisfy the 2%/2mm criterion. Transport of secondary electrons dominate the built-up region at $z \approx 0$. Hence, discrepancies in the entrance region are mainly formed because our model does not include the simulation of secondary electrons (delta rays). A similar reason also causes differences to Monte Carlo computations at higher penetration depths: This region is referred to as bremsstrahlung tail where photons largely contribute to the deposited dose. As we neglect photon transport our penetration depth is smaller and the fall-off larger. This behavior becomes more significant for 10 MeV because bremsstrahlung effects increasingly gain on importance from $E \approx 6$ MeV (Figure 1.9). In fact, the largest FP and GFP$_2$ distance to PENELOPE and GEANT4 becomes 3mm at $z \approx 5$ cm and hence, they do not meet the criterion for a 10 MeV beam.

1.3 Photon Interactions

Although the interest in heavy (charged) particle beams has drastically increased in recent years, photon beams are still the most widely-spread particle beams for the treatment of cancer nowadays. The most common energy range in radiotherapy is between approximately 1 MeV and 25 MeV. They are called megavoltage X-rays and are most frequently produced by linear particle accelerators.

The upcoming information on the penetration of photons through matter only give a summary of the most important interaction effects. It does not contain extensive descriptions of the occurring mechanisms; nor does it capture all physical effects (detailed information can be found in [26] and [155]). The main goal is rather to provide reliable data which can be incorporated in deterministic electron/photon transport codes. A consistent mathematical model for coupled photon and electron transport for dose calculations in photon radiotherapy is derived in [80]. So far, forthcoming data have not been included in a deterministic code and is an issue for future work. In this way, below cross sections can be viewed as a starting point and are supposed to be refined and improved.

In the energy range of 50 eV – 1 GeV photons dominantly interact with matter by the following processes:

- *Elastic (or Coherent Rayleigh) Scattering:* This process describes the scattering of photons by bound atomic electrons. After the scattering event the incident

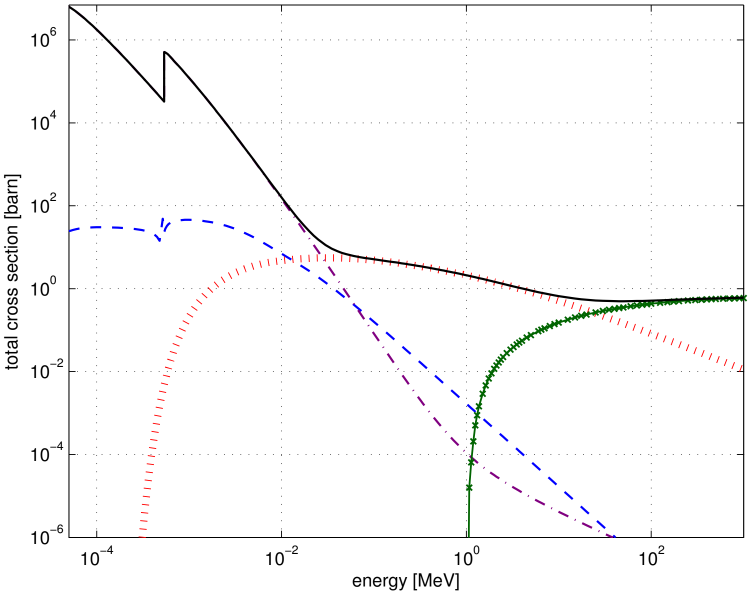


Figure 1.11: Total cross sections for liquid water (extracted from PENELOPE [155]): elastic (blue dashed line), Compton (red dotted line), photoelectric (purple dash-dot line) and pair production (dark green cross line), sum of all contributions (black solid line).

photon leaves the atom in its original state. Consequently, the scattered photon has the same energy as the incident photon.

- *Photoelectric Effect:* An incident photon with kinetic energy E_γ is absorbed by the target atom which is in turn excited to a higher state. It interacts with a bounded shell electron which leaves the atom with the kinetic energy $E_\gamma - E_B$. E_b is the binding or ionisation energy of an individual electron. Then, an outer electron transits to the lower state to fill the formed vacancy in the corresponding shell. This is most likely accompanied by the emission of either a photon or a different outer electron.

- *Compton (or Incoherent) Scattering:* It represents a photon interaction with a bound atomic electron of binding energy E_B. This electron absorbs the incoming photon and re-emits a secondary Compton photon of energy E'_γ. If the incident photon energy E_γ is large enough, the electron is ejected from the atom with energy $E'_e = E_\gamma - E'_\gamma - E_B$ after the Compton interaction.

- *Pair Production:* Electron-positron pairs are produced when photons are absorbed and their energy is transformed to mass. As an electron-positron pair is created out of a photon a minimum photon energy of $2m_ec^2 = 1.02$ MeV is required. Below this threshold no pair production effect can occur. Additionally, for conservation reasons a certain amount of momentum is always transferred to a massive particle (e.g., nucleus or electron) so that this effect is always coupled to matter.

Interaction probabilities for different effects are summarized in Figure 1.11. It shows energy-dependent angle-integrated cross sections for water. The ”total” cross section is hereby referred to as the sum of all scattering events described above. For photon

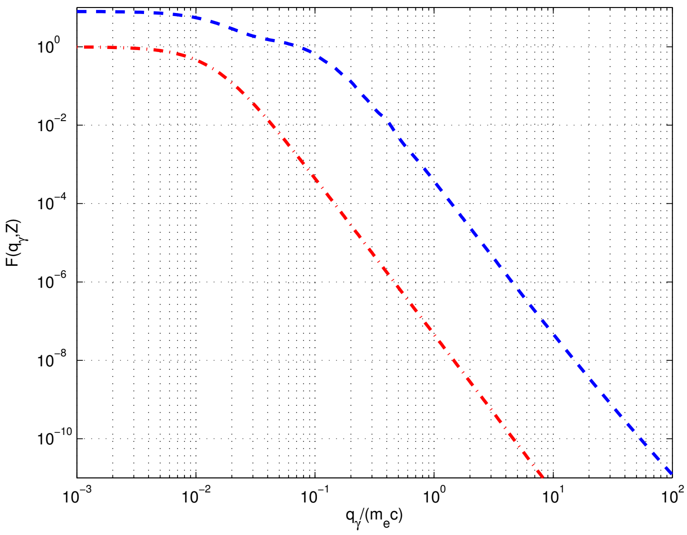


Figure 1.12: Atomic form factor for elastic scattering (extracted from [155]): oxygen O16 (blue dashed line), hydrogen H1 (red dash-dot line).

energies below 0.01 MeV the photoelectric absorption prevails. In the transition regime between approximately 0.01 MeV and 0.1 MeV Compton, elastic and photoelectric effects are non-negligible. Although in this region the probability of photoelectric or Compton effect is always larger than that of elastic scattering, its contribution is nevertheless non-negligible for accurate simulations. In the interval of 0.1 MeV to 10 MeV Compton scattering is dominant. For larger energies, pair production plays an important role and is superior to other effects.

1.3.1 Terminology

The energy of an incoming photon is denoted by E_γ and of an outgoing photon by E'_γ. Energy variables are scaled to electron rest energy $m_e c^2 = 0.51099906$ MeV and denoted by $\varepsilon_\gamma = E_\gamma/(m_e c^2)$ or $\varepsilon'_\gamma = E'_\gamma/(m_e c^2)$. As coupled photon/electron interactions occur, we additionally need energy variables for outgoing electrons. They are denoted by an index e instead of γ, e.g., E'_e or $\varepsilon'_e = E'_e/(m_e c^2)$. The scattering cosine in the laboratory system for the initial photon direction and the direction of the scattered photon is denoted by μ_γ. Similarly, μ_e is the scattering cosine for an incident photon and scattered electron. $c = 2.99792458 \cdot 10^8$ m/s is the speed of light. Remaining variables and constants are defined in the same way as in Section 1.2.

1.3.2 Elastic (Coherent Rayleigh) Scattering

Elastic photon scattering is characterized by small scattering angles. Since no energy is transferred to the target atom the scattering is only essential for the spatial distribution of photon tracks in the medium. The corresponding cross section, depending on energy

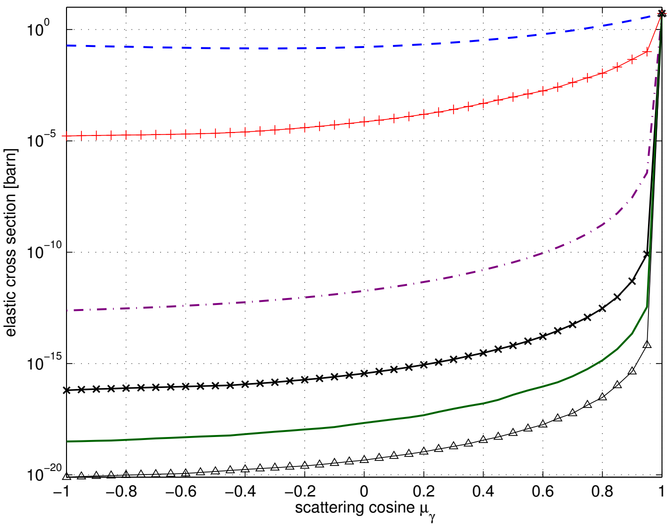


Figure 1.13: Photon elastic angular scattering distribution for water: 10 keV (blue dashed line), 100 keV (red solid plus line), 1 MeV (purple dash dot line), 3 MeV (black solid cross line), 6 MeV (solid dark green line), 10 MeV (black solid triangle line).

and scattering cosine, reads as follows [155]:

$$\sigma_\gamma^{\mathrm{el}}(\varepsilon_\gamma, \mu_\gamma) = r_e^2 \frac{1+\mu_\gamma^2}{2} \cdot [F(q_\gamma(\varepsilon_\gamma, \mu_\gamma), Z)]^2, \tag{1.31a}$$

$$q_\gamma(\varepsilon_\gamma, \mu_\gamma) = \varepsilon_\gamma \, m_e c \sqrt{2(1-\mu_\gamma)}. \tag{1.31b}$$

where $F(q_\gamma(\varepsilon_\gamma, \mu_\gamma), Z)$ is called atomic form factor which is tabulated in [37] and q_γ is the magnitude of the momentum transfer.

It is a monotonically decreasing function for increasing values of q_γ (Figure 1.12). Here, the form factor was extracted from PENELOPE's files `pdgraZZ.p08` [155]. It varies from $F(0, Z) = Z$ to $\lim_{q_\gamma \to \infty} F(q_\gamma, Z) = 0$. Note that the magnitude of the momentum transfer q_γ has a unit of the quantity $m_e c$. In literature and some databases (e.g. ENDF) it is common to use the dimensionless variable $x = 20.6074 \cdot q_\gamma/(m_e c)$ instead of q_γ.

Above formula is an approximation for photons with an energy higher than the ionization energy of the K-shell. The mean ionization energy of electrons in water molecules is $\varepsilon_B = 1.4677 \cdot 10^{-4}$ (in units of the electron rest energy). As the typical energy range for X-ray radiotherapy is from about 1 MeV to 25 MeV this assumption is fulfilled for simulation purposes. However, if lower energies are applied an additional correction is needed. It is known as the anomalous scattering factor which can also be found in [37].

Figure 1.13 displays the angular behavior of photons being elastically scattered at water molecules. We apply the additivity approximation for the calculation, i.e., the corresponding cross section for water $\sigma_{\gamma,\mathrm{H_2O}}^{\mathrm{el}}$ is obtained by a simple linear combination of the cross section for oxygen $\sigma_{\gamma,\mathrm{O16}}^{\mathrm{el}}$ and hydrogen $\sigma_{\gamma,\mathrm{H1}}^{\mathrm{el}}$:

$$\sigma_{\gamma,\mathrm{H_2O}}^{\mathrm{el}}(\varepsilon_\gamma, \mu_\gamma) = \sigma_{\gamma,\mathrm{O16}}^{\mathrm{el}}(\varepsilon_\gamma, \mu_\gamma) + 2\sigma_{\gamma,\mathrm{H1}}^{\mathrm{el}}(\varepsilon_\gamma, \mu_\gamma). \tag{1.32}$$

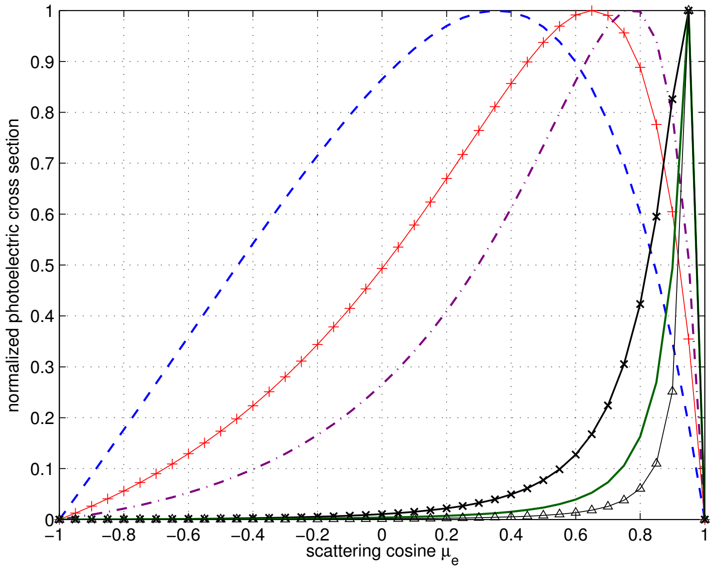


Figure 1.14: Normalized photoelectric cross section for water: 10 keV (blue dashed line), 50 keV (red solid plus line), 100 keV (purple dash dot line), 500 keV (black solid cross line), 1 MeV (solid dark green line), 10 MeV (black solid triangle line).

It is striking that the scattering is strongly in the forward direction for large energies above approximately 10 keV. However, at lower energies there is a non-negligible amount of large angle scattering. Note that at $q_\gamma(\mu_\gamma = 1) = 0$ which implies that the form factor F equals the atomic number Z and hence, (1.31) reduces to one value

$$r_e^2 \cdot 66 \approx 5.24 \text{ barn}$$

for all energies.

1.3.3 Photoelectric Effect

When a photon of energy E_γ is absorbed by an atom this atom is lifted to a higher state. If the photon energy exceeds the corresponding shell ionisation energy the electron is emitted with an energy given by the incident photon energy E_γ minus its binding energy. *Carron* describes an important condition for the photon-electron interaction as follows [26]: "Kinematically, a free electron cannot absorb a photon, but an electron bound in an atom can. The less tightly bound it is, the less likely it is to absorb." The closest shell to the nucleus is called K-shell where electrons are tightly bound. Hence, the probability for photoelectric absorption becomes large when ionisations of K-shell electrons occur. This is exactly what can be observed in Figure 1.11: The sharp increase at $E \approx 5 \cdot 10^{-4}$ is the photoionisation of the K-shell when the photon energy slightly exceeds the corresponding ionisation energy. It is often called the characteristic K-edge.

For increasing energies, the probability for photoelectric absorption drops because binding energies of atomic electrons become very small relatively to the incident photon energy. Consequently, shell electrons act like almost free targets.

The initial direction of photoelectrons can be described by the angular differential cross section derived by Sauter. Although Sauter's formula is only exact for K-shell

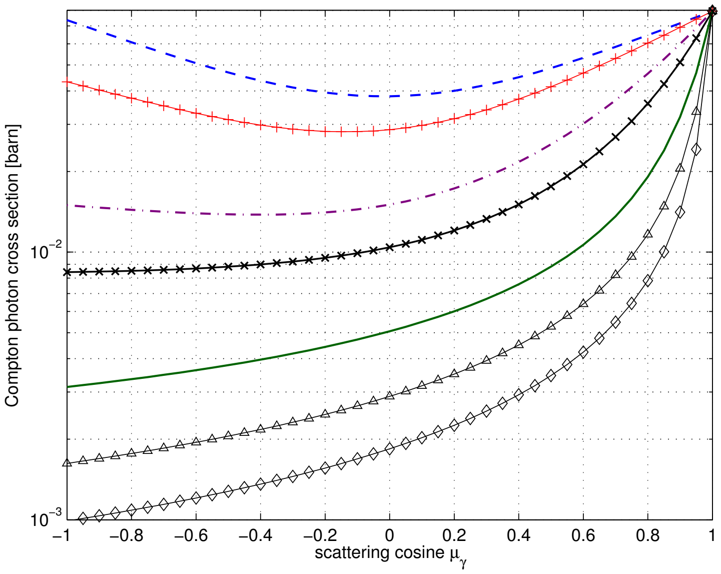


Figure 1.15: Compton photon cross section (Klein-Nishina): 10 keV (blue dashed line), 100 keV (red solid plus line), 500 keV (purple dash dot line), 1 MeV (black solid cross line), 3 MeV (solid dark green line), 6 MeV (black solid triangle line), 10 MeV (black solid diamond line).

ionisation it turns out to be a good approximation for any photoionisation event [155]:

$$\sigma_\gamma^{\text{phel}}(\varepsilon_\gamma, \varepsilon'_e, \mu_e) = \frac{\alpha^4 r_e^2 \beta_e'^3}{\gamma'_e} \left[\frac{Z}{\varepsilon_\gamma}\right]^5 \frac{1-\mu_e^2}{(1-\beta'_e\mu_e)^4} \cdot \left[1 + \frac{1}{2}\gamma'_e(\gamma'_e - 1)(\gamma'_e - 2)(1-\beta'_e\mu_e)\right] \quad (1.33)$$

where $\alpha = 1/137.036$, $\gamma'_e = 1+\varepsilon'_e$ and $\beta_e'^2 = \varepsilon'_e(\varepsilon'_e+2)/(\varepsilon'_e+1)^2$. Moreover, $\varepsilon'_e = \varepsilon_\gamma - \varepsilon_B$ where ε_B is the ionization energy of the K-shell in electron rest energy.

In Figure 1.14 the normalized photoelectric cross section from (1.33) is presented for water. The approximation for water molecules is achieved by a linear combination of oxygen and hydrogen from (1.32). The corresponding values for Z and ε_B are given in Table 1.2. The emission of photoelectrons, which serve as an additional source in the transport equation, is not significantly forward-peaked. At $\mu_e = 1$, it always drops to zero. However, above a few hundreds of keV the angular distribution forms sharp peaks close to $\mu_e = 1$. Consequently, a large amount of photoelectrons are ejected close to the original incident photon direction.

1.3.4 Compton (Incoherent) Scattering

Photon Scattering

In Compton scattering, an incident photon of energy ε_γ is absorbed by an electron, re-emitting a secondary photon of energy ε'_γ in a new direction characterized by the scattering cosine μ_γ. Assuming that the photon interacts with a *free* electron *at rest*

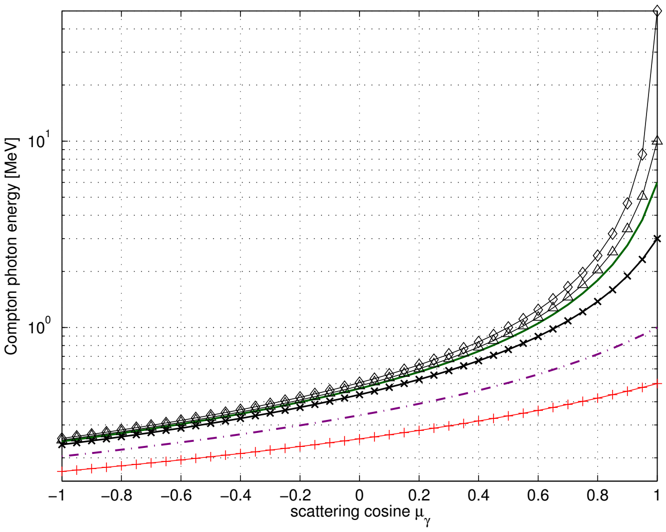


Figure 1.16: Scattered photon energy vs. outgoing scattering cosine μ_γ: 100 keV (blue dashed line), 500 keV (red solid plus line), 1 MeV (purple dash dot line), 3 MeV (black solid cross line), 6 MeV (solid dark green line), 10 MeV (black solid triangle line), 50 MeV (black solid diamond line).

the differential cross section is given by the Klein-Nishina formula [155]:

$$\sigma_\gamma^{\mathrm{KN}}(\varepsilon_\gamma,\mu_\gamma) = \frac{r_e^2}{2(1+\varepsilon_\gamma(1-\mu_\gamma))^2}\left(\frac{\varepsilon'_\gamma}{\varepsilon_\gamma}+\frac{\varepsilon_\gamma}{\varepsilon'_\gamma}+\mu_\gamma^2-1\right) \tag{1.34a}$$

$$= \frac{r_e^2}{2}\frac{1}{(1+\varepsilon_\gamma(1-\mu_\gamma))^2}\left(1+\mu_\gamma^2+\frac{\varepsilon_\gamma^2(1-\mu_\gamma)^2}{1+\varepsilon_\gamma(1-\mu_\gamma)}\right). \tag{1.34b}$$

The behavior of (1.34) is shown in Figure 1.15 for several photon energies. At $\mu_\gamma = 1$ the expression reduces to $\sigma_\gamma^{\mathrm{KN}} = r_e^2 \approx 0.08$ barn which implies that cross sections for all energies hit this value. Although at large incident energies ε_γ photons are preferably emitted in the forward direction there is still a non-negligible amount of large scattering. At $\varepsilon_\gamma = 10$ keV the distribution is almost axially symmetric such that the probability for forward and backward scattering is nearly the same.

The energy of the scattered photon follows from the conservation of energy and momentum:

$$\varepsilon'_\gamma = \frac{\varepsilon_\gamma}{1+\varepsilon_\gamma(1-\mu_\gamma)}.$$

The energy of the scattered photon ε'_γ as a function of its outgoing scattering cosine μ_γ is displayed in Figure 1.16. It illustrates that photons with large energies are scattered in the forward direction.

Integrating (1.34) over the scattering cosine $\mu_\gamma = -1,\ldots,1$ yields the total Klein-Nishina cross section [26]:

$$\Sigma_{\gamma,0}^{\mathrm{KN}}(\varepsilon_\gamma) = 2\pi r_e^2\left[\frac{1+\varepsilon_\gamma}{\varepsilon_\gamma^2}\left(\frac{2(1+\varepsilon_\gamma)}{1+2\varepsilon_\gamma}-\frac{\ln(1+2\varepsilon_\gamma)}{\varepsilon_\gamma}\right)\right.$$
$$\left.+\frac{\ln(1+2\varepsilon_\gamma)}{2\varepsilon_\gamma}-\frac{1+3\varepsilon_\gamma}{(1+2\varepsilon_\gamma)^2}\right] \tag{1.35}$$

Material	f_i	ε_{B_i}	J_i^0 in $\hbar/(m_e c^2)$
hydrogen $Z = 1$	1	$2.6614 \cdot 10^{-5}$	$8.49 \cdot 10^{-1}$
oxygen $Z = 8$	2	$1.0528 \cdot 10^{-3}$	$1.13 \cdot 10^{-1}$
	2	$5.5733 \cdot 10^{-5}$	$5.79 \cdot 10^{-1}$
	2	$2.6653 \cdot 10^{-5}$	$3.50 \cdot 10^{-1}$
	2	$2.6653 \cdot 10^{-5}$	$3.50 \cdot 10^{-1}$

Table 1.2: Quantities needed for the calculation of $S(\varepsilon_\gamma, \mu_\gamma)$ from (1.36).

In reality, atomic electrons are bound and move with a certain momentum. Taking this into consideration (1.34) can be corrected by the incoherent scattering function [155]:

$$S(\varepsilon_\gamma, \mu_\gamma) = \sum_{i=1}^{N} f_i \, \Theta(\varepsilon_\gamma - \varepsilon_{B_i}) \, n_i(p_i), \tag{1.36}$$

where

$$p_i(\varepsilon_\gamma, \mu_\gamma) = m_e c \, \frac{\varepsilon_\gamma(\varepsilon_\gamma - \varepsilon_{B_i})(1-\mu_\gamma) - \varepsilon_{B_i}}{\sqrt{2\varepsilon_\gamma(\varepsilon_\gamma - \varepsilon_{B_i})(1-\mu_\gamma) + \varepsilon_{B_i}^2}}, \tag{1.37}$$

$$n_i(p_i) = \begin{cases} \frac{1}{2} \exp\left[2\, J_i^0\, p_i\, (1 - J_i^0\, p_i)\right] & \text{if} \quad p_i < 0 \\ 1 - \frac{1}{2} \exp\left[-2\, J_i^0\, p_i\, (1 + J_i^0\, p_i)\right] & \text{if} \quad p_i > 0 \end{cases}. \tag{1.38}$$

Above quantities are defined as follows:

$$N \equiv \text{total number of shells,} \tag{1.39a}$$

$$f_i \equiv \text{number of electrons in the i-th shell,} \tag{1.39b}$$

$$\Theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else} \end{cases}, \qquad \text{Heaviside function,} \tag{1.39c}$$

$$\varepsilon_{B_i} \equiv \text{shell ionization energy in electron rest energy,} \tag{1.39d}$$

$$n_i \equiv \text{fraction of electrons in the i-th shell which can be excited in a Compton event,} \tag{1.39e}$$

$$J_i^0 \equiv \text{one-electron profile of i-th shell evaluated at 0.} \tag{1.39f}$$

The Compton profile function J_i^0 evaluated at 0 for all shells can be found in PENELOPE [155] in the file `pdatconf.p06`. For hydrogen and oxygen, all quantities are tabulated Table 1.2. Again, one can use the additivity approximation from (1.32) to compute the cross section for water. Altogether, the final formula for Compton scattering of photons is then given by

$$\sigma_\gamma^{\text{Co}}(\varepsilon_\gamma, \mu_\gamma) = \sigma_\gamma^{\text{KN}}(\varepsilon_\gamma, \mu_\gamma) \cdot S(\varepsilon_\gamma, \mu_\gamma). \tag{1.40}$$

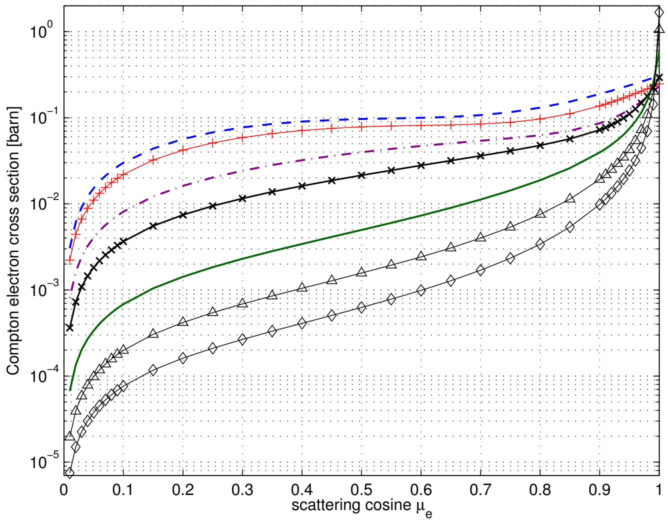


Figure 1.17: Compton electron cross section: 10 keV (blue dashed line), 100 keV (red solid plus line), 500 keV (purple dash dot line), 1 MeV (black solid cross line), 3 MeV (solid dark green line), 6 MeV (black solid triangle line), 10 MeV (black solid diamond line).

Electron Scattering

If the incoming photon has enough energy ε_γ to free the bounded electron, the corresponding atom is ionized. The angular differential cross section for scattering of Compton electrons reads [80]:

$$\sigma_e^{\mathrm{Co}}(\varepsilon_\gamma,\mu_e) = \frac{4r_e^2(1+\varepsilon_\gamma)^2}{\mu_e^3(a(\varepsilon_\gamma,\mu_e)+2\varepsilon_\gamma)^2}\left[1-\frac{2}{a(\varepsilon_\gamma,\mu_e)}+\frac{2}{a^2(\varepsilon_\gamma,\mu_e)} + \frac{2\varepsilon_\gamma^2}{a(\varepsilon_\gamma,\mu_e)(a(\varepsilon_\gamma,\mu_e)+2\varepsilon_\gamma)}\right], \tag{1.41a}$$

$$\text{where}\quad a(\varepsilon_\gamma,\mu_e) = (1+\varepsilon_\gamma)^2\frac{1-\mu_e^2}{\mu_e^2}+1. \tag{1.41b}$$

After collision the emitted electron has the kinetic energy

$$\varepsilon_e' = \frac{2\varepsilon_\gamma^2}{2\varepsilon_\gamma + a(\varepsilon_\gamma,\mu_e)}. \tag{1.42}$$

In contrast to photon emission, Compton electrons move in the forward direction (Figure 1.18). As μ_e approaches zero, both electron energies as well as cross sections drastically decrease to zero (Figure 1.17). The reason is that Compton electrons cannot be scattered backward which can be confirmed by kinematics calculations. Their cross section is strongly forward-peaked for incident photon energies above roughly 1 MeV. Nevertheless, electrons emerging from smaller photon energies are also scattered at large angles.

1.3.5 Comparison

The validation of above formulas is performed in this section. First, we calculate cross sections by numerical integration of approximate analytic expressions. And second,

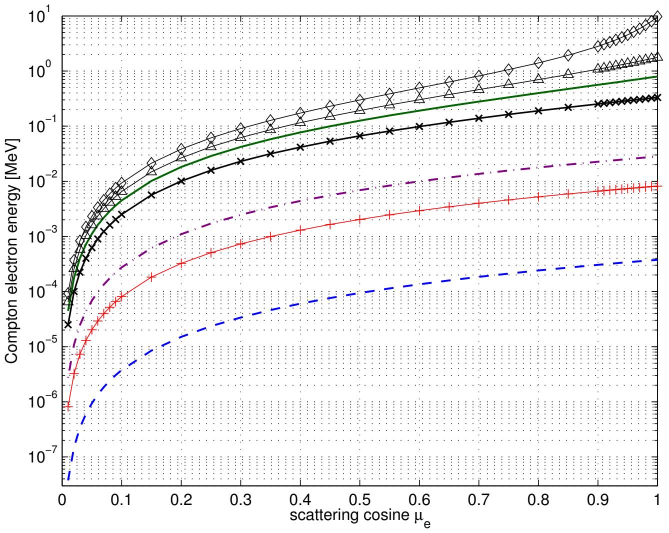


Figure 1.18: Scattered electron energy vs. outgoing scattering cosine μ_e: 10 keV (blue dashed line), 50 keV (red solid plus line), 100 keV (purple dash dot line), 500 keV (black solid cross line), 1 MeV (solid dark green line), 2 MeV (black solid triangle line), 10 MeV (black solid diamond line).

tabulated data from PENELOPE [155] is extracted and serve as a benchmark. It is important to emphasize that only approximate equations are explicitly computed whereas PENELOPE includes more corrections leading to more accurate cross sections.

Elastic (Coherent Rayleigh) Scattering

Elastic scattering is fairly well captured by (1.31) which is approximately valid for photon energies above the K-shell ionization energy. Figure 1.19 illustrates the differences between $\xi^{\mathrm{el}}_{\gamma,0}$ and $\Sigma^{\mathrm{el}}_{\gamma,0}/(2\pi)$ defined by:

$$\text{PENELOPE [155]: } \xi^{\mathrm{el}}_{\gamma,0}(\varepsilon_\gamma) = \mathcal{N}_{H_2O} \cdot \sigma^{\mathrm{el}}_{\mathrm{Pen,tot}}(\varepsilon_\gamma) \tag{1.43a}$$

$$\text{Analytic: } \Sigma^{\mathrm{el}}_{\gamma,0}(\varepsilon_\gamma) = 2\pi\,\rho_e \int_{-1}^{1} \left[\sigma^{\mathrm{el}}_{\gamma,\mathrm{O16}}(\varepsilon_\gamma,\mu_\gamma) + 2\sigma^{\mathrm{el}}_{\gamma,\mathrm{H1}}(\varepsilon_\gamma,\mu_\gamma)\right] d\mu_\gamma. \tag{1.43b}$$

For comparison reasons, the factor of 2π has to be neglected in (1.43b). Although, to the eye, both functions are close to each other for $E \geq 10^{-2}$ MeV, an almost constant large relative error of 60 % occur. Hence, corrections by a least squares fit are needed here. For photon energies near the K-edge and below, the difference to PENELOPE becomes larger because the approximation assumption is not fulfilled in this energy region. More accurate results can be obtained by an additional correction to the form factor (anomalous factor).

Photoelectric Effect

Since the photoelectric effect generates electrons, the corresponding cross section can be included by an additional source term in the electron transport equation. At the same time, photons are removed from the system and can be accounted for in the

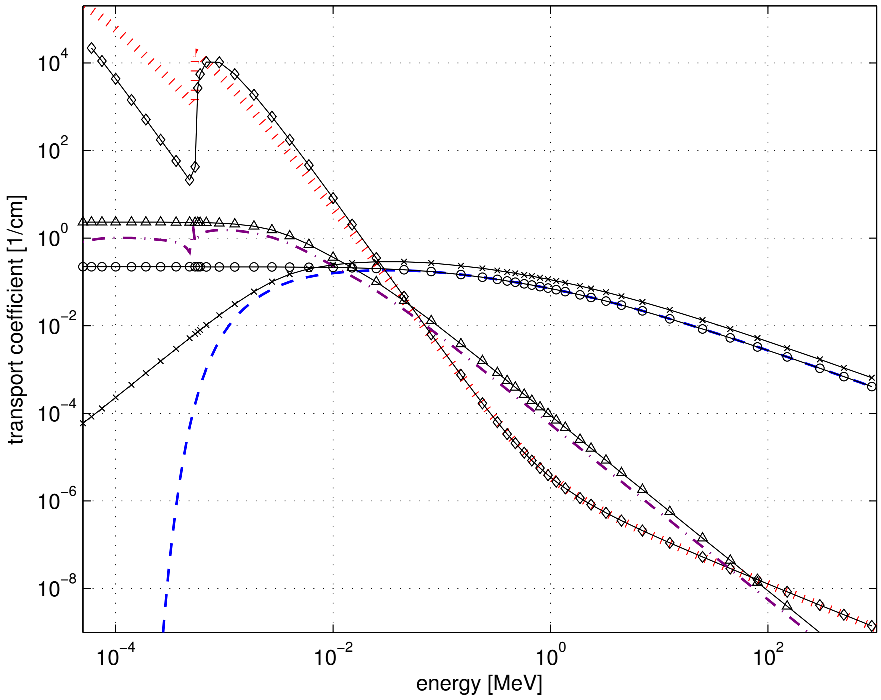


Figure 1.19: Transport coefficient for liquid water: PENELOPE-Compton (blue dashed line), total Klein-Nishina (black solid circle line), corrected Klein-Nishina (black solid cross line) PENELOPE-photoelectric (red dotted line), analytic photoelectric (black solid diamond line), PENELOPE-elastic (purple dash-dot line), analytic elastic (black solid triangle line).

absorption term of the photon transport equation. It is the total absorption cross section approximated by integrating (1.33) with respect to μ. This result is compared to the reference $\xi^{\text{phel}}_{\gamma,0}$:

$$\text{PENELOPE [155]: } \xi^{\text{phel}}_{\gamma,0}(\varepsilon_\gamma) = \mathcal{N}_{H_2O} \cdot \sigma^{\text{phel}}_{\text{Pen,tot}}(\varepsilon_\gamma), \tag{1.44a}$$

$$\text{Analytic: } \Sigma^{\text{phel}}_{\gamma,0}(\varepsilon_\gamma) = 2\pi\, \rho_e \int_{-1}^{1} \left[\sigma^{\text{phel}}_{\gamma,\text{O16}}(\varepsilon_\gamma, \mu_\gamma) + 2\sigma^{\text{phel}}_{\gamma,\text{H1}}(\varepsilon_\gamma, \mu_\gamma)\right] d\mu_\gamma. \tag{1.44b}$$

Figure 1.19 shows $\Sigma^{\text{phel}}_{\gamma,0}/(2\pi)$ and $\xi^{\text{phel}}_{\gamma,0}$. Again, to the eye, both quantities are hardly to distinguish in a large energy range. For $E > 10^{-1}$ MeV relative errors below 5 % are established. At larger energies, errors increase and corrections to the analytic formula are required. Especially for photon energies below the K-absorption edge, Sauter's cross section from (1.33) becomes inaccurate because it is only exact for ionization of a single K-shell electron by high-energy photons.

Compton Scattering

To estimate the accuracy of the Klein-Nishina formula we compare (1.35) to

$$\text{PENELOPE [155]: } \quad \xi^{\text{Co}}_{\gamma,0}(\varepsilon_\gamma) = \mathcal{N}_{H_2O} \cdot \sigma^{\text{Co}}_{\text{Pen,tot}}(\varepsilon_\gamma). \tag{1.45}$$

Eq. 1.35 is in a very good agreement to $\xi^{\text{Co}}_{\gamma,0}$ for energies above 10^{-1} MeV (Figure 1.19) where relative errors of less than 1 % occur. However, smaller photon energies yield poor results. Eq. 1.35 describes free electrons at rest. Realistic Compton cross sections include Doppler broadening which is a consequence of moving atomic electrons. Moreover, binding effects also play an important role. Both phenomena are accounted for by the incoherent scattering function $S(\varepsilon_\gamma, \mu_\gamma)$ described in Section 1.3.4. To demonstrate the improvement, we approximate S by the Waller-Hartree incoherent scattering function S_{WH} which is tabulated in PENELOPE [155] in the file `pdaffZZ.p08`. Although it does not completely capture both corrections Figure 1.19 shows the behavior of $\Sigma^{\mathrm{Co}}_{\gamma,0}/(2\pi)$ which is more accurate at lower energies and is defined by

$$\Sigma^{\mathrm{Co}}_{\gamma,0}(\varepsilon_\gamma) = 2\pi\, \rho_e \int_{-1}^{1} \Big[S^{\mathrm{O16}}_{\mathrm{WH}}(\varepsilon_\gamma, \mu_\gamma) \sigma^{\mathrm{phel}}_{\gamma,\mathrm{O16}}(\varepsilon_\gamma, \mu_\gamma) + 2\, S^{\mathrm{H1}}_{\mathrm{WH}}(\varepsilon_\gamma, \mu_\gamma) \sigma^{\mathrm{phel}}_{\gamma,\mathrm{H1}}(\varepsilon_\gamma, \mu_\gamma) \Big] \, d\mu_\gamma. \tag{1.46}$$

1.4 High Energy Proton Interactions

A monoenergetic proton beam forms a characteristic dose profile which makes proton radiotherapy superior to conventional X-rays: The entrance region is an almost constant (or only slowly rising) dose which, up to a certain point near the end of the range, abruptly increases to a much larger value followed by a steep drop-off up to almost zero. This sharp dose distribution is often referred to as Bragg peak which makes up the most significant advantage of proton beams in radiotherapy. A superposition of multiple beams of different energies provides dose profiles consisting of three segments: a relatively low initial value, an uniformly high dose plateau within the tumor region and a zero dose beyond the target.

This behavior enables to distribute the energy in tissue more precisely. Consequently, the average dose to healthy tissues can be reduced, a smaller number of beams and fewer treatments are necessary. Especially in critical regions, like the brain, spinal cord or eye, protons have already been applied successfully (see [88] and references therein). The Particle Therapy Co-Operative Group [2] regularly publishes statistics and information about recent developments in proton, light ion and heavy charged particle radiotherapy. The total number of patients being treated with protons is steadily increasing. At the end of 2010 more than 70.000 patients were already treated and more and more particle therapy facilities are planed or under construction [2]. Hence, the simulation of accurate dose distributions, which should be fast enough for clinical applications, is a major contribution to the success of cancer treatments.

1.4.1 Terminology

To keep our notation simple we develop equations similar to those for electrons. Hence, all kinetic energies of protons are scaled to proton rest energies of $m_p c^2 = 938,27198$ MeV. The kinetic energy of an incident particle is denoted by τ' whereas the outgoing particle is set to be τ. Again, $\beta = v'/c$ is the scaled velocity of the proton and it holds: $\beta^2 = \tau'(\tau'+2)/(\tau'+1)^2$. Later on we need the molecular density for water ($\mathcal{N}_{H_2O} = 3.3428847 \cdot 10^{22}\mathrm{cm}^{-3}$), its electron density $\rho_e = 3.3428847 \cdot 10^{23}\mathrm{cm}^{-3}$, the density of atomic cores in water $\rho_c = \rho_e/Z = 3.3428847 \cdot 10^{22}\mathrm{cm}^{-3}$ and its atomic number $Z = 10$. The Debye length is defined by $\lambda_D = 2.0439 \cdot 10^{-10}\mathrm{cm}$ and $h = 4.135667516 \cdot 10^{-21}\mathrm{MeV{\cdot}s}$ is Planck's constant. We also use the unit barn: 1 barn $= 10^{-24}$ cm^2.

1.4.2 Elastic Cross Sections

Since the main goal in radiotherapy is to kill tumour cells it is necessary to destroy the atomic structure of the material. In proton-nucleus interactions this is caused by a momentum transfer from the incident proton to the corresponding nucleus. Qualitatively, two different types of collisions can be classified: In a soft collision, the momentum transfer is small and the scattering is forward-peaked, i.e., protons are hardly deflected and travel in a almost straight line. Especially for high energy protons, collisions of this type dominate. On the other hand, in hard collisions a large momentum transfer is observed and an incident proton can be significantly altered in direction.

For the major part, nuclear recoils are induced by elastic scattering processes whose behavior can be, to some extent, determined by the analytic Rutherford cross section formula. Particularly for low energies and small angles the Rutherford cross section becomes more important (see discussion below). It is derived based on the assumption that scattering occurs between point masses. It only takes the interaction between projectile and target in their Coulomb potential into account. The physical correct shape of the particles is neglected. However, high energetic incident ions can overcome the repulsive Coulomb potential (known as Coulomb barrier) and reach the attractive potential of the strong nuclear force. This effect is referred to as nuclear elastic scattering and is not negligible as will be shown later.

Analytic Formulas

One classic way to describe elastic collisions between a proton and a target nucleus of charge Ze are mechanics trajectory calculations. The basic idea is to multiply the Coulomb potential for bare nuclei by an additional screening factor. It is introduced to model the influence of atomic electrons. Using this scattering kernel *Nikjoo et al.*, however, mention unrealistic results in Monte Carlo simulations with high-energy protons [72]. Instead, they propose a modified Mott-scattering formula:

$$\sigma^p_{\mathrm{Mott}}(\tau,\mu) = \left(\frac{Zr_e m_e}{m_p}\right)^2 \left(\frac{1+\tau}{\tau(\tau+2)}\right)^2 \frac{1}{(1-\mu)^2}\left[1-\frac{\tau(\tau+2)}{2(1+\tau)^2}(1-\mu)\right], \qquad (1.47a)$$

where

$$\theta_{\mathrm{cut}} = Z^{1/3}\frac{m_e}{137 m_p}\frac{1}{\sqrt{\tau(\tau+2)}} \quad \text{and} \quad \mu = \cos(\theta). \qquad (1.47b)$$

θ_{cut} is introduced to avoid singularities in integration. Above Mott-formula consists of the relativistic Rutherford cross section multiplied by a correction term which takes spin interactions of projectile and target into account [149]. This term contributes significantly at relativistic energies. However, structure functions (i.e., volume expansion of particles) and screening by atomic electrons are not included.

To calculate the n-th transport coefficient it is necessary to integrate (1.47):

$$\xi^{\mathrm{Mott}}_{p,n}(\tau) = 2\pi\rho_c \int_{-1}^{\mu_{\mathrm{cut}}} \sigma^p_{\mathrm{Mott}}(\tau,\mu)(1-\mu)^n d\mu \quad \text{with} \quad \mu_{\mathrm{cut}} = \cos(\theta_{\mathrm{cut}}). \qquad (1.48)$$

A different possibility is shown in [114] where interactions of high-energy electrons with plasma are modelled. Although only the first transport coefficient is explicitly derived therein, we describe this formula in detail because it can be used for the M1 model and for validation reasons.

The basic idea is to sum up the elastic cross section for electrons being scattered at a bare nucleus and for electrons scattering off the atomic electrons. Electron-ion interactions in plasma are modelled by the relativistic Rutherford cross section. Therefore, this transport coefficient can be adapted to protons in a similar way as described in [72] which leads to the following result:

$$\begin{aligned}\xi_{p,1}^{\text{Li/Petrasso}}(\tau) = 4\pi\rho_c Z \left(\frac{r_e m_e}{m_p}\right)^2 \left(\frac{1+\tau}{\tau(\tau+2)}\right)^2 \Bigg[& Z\ln(\Lambda^{\text{p-I}}) \\ & + \frac{4(\tau+2)^2}{\left(2\sqrt{(\tau+2)/2}\right)^4}\ln(\Lambda^{\text{p-e}})\Bigg],\end{aligned} \tag{1.49a}$$

where

$$\Lambda^{\text{p-I}}(\tau) = \frac{\lambda_D}{b_{\text{min}}^{\text{p-I}}(\tau)}, \qquad b_{\text{min}}^{\text{p-I}}(\tau) = \min\left\{\lambda_{\text{Broglie}}(\tau), b_\perp^{\text{p-I}}(\tau)\right\}, \tag{1.49b}$$

and

$$\Lambda^{\text{p-e}}(\tau) = \frac{\lambda_D}{b_{\text{min}}^{\text{p-e}}(\tau)}, \qquad b_{\text{min}}^{\text{p-e}}(\tau) = \min\left\{\lambda_{\text{Broglie}}(\tau), b_\perp^{\text{p-e}}(\tau)\right\}. \tag{1.49c}$$

Moreover, the de Broglie wave length can be written as

$$\lambda_{\text{Broglie}}(\tau) = \frac{h}{p(\tau)} = \frac{hc}{p(\tau)c} = \frac{h}{m_p c}\frac{1}{\sqrt{\tau(\tau+2)}} \approx 1.3214\cdot 10^{-13}\frac{1}{\sqrt{\tau(\tau+2)}}\ \text{cm}. \tag{1.50}$$

where $p(\tau)$ is the momentum of the proton. The remaining variables are defined by

$$b_\perp^{\text{p-I}}(\tau) = \frac{Z m_e r_e}{m_p}\frac{\tau+1}{\tau(\tau+2)}, \qquad b_\perp^{\text{p-e}}(\tau) = \frac{2 m_e r_e}{m_p}\frac{(\tau+1)}{\tau\, 2^{\sqrt{2(\tau+2)}}}. \tag{1.51}$$

ENDF/B-VII Database

We want to use the ENDF/B-VII library [126] to compute the elastic differential cross sections for collisions of protons with various elements. As we are mainly interested in radiotherapy applications we focus on gaining cross sections for proton-water collisions.

The raw ENDF-files are text-files which require certain programs to gain the stored information. We extract this information by two approaches:

(a) The LISTEF-6.13 program [49] is executed to generate text-files from original ENDF/B-VII data. This program is designed to extract specific information, explicitly chosen by the user, from raw ENDF-files.

(b) Unfortunately, in certain cases LISTEF does not provide sufficient descriptions about the output data. In these cases, cross sections are used from

http://t2.lanl.gov/data/proton7.html.

Quantity	**Description**	**Unit**
$\sigma_{\mathrm{cd}}(\mu, E)$	Rutherford scattering cross section for distinguishable particles (e.g. p-O16 collision)	barn $\cdot$ ster^{-1}
$\sigma_{\mathrm{ci}}(\mu, E)$	Rutherford scattering cross section for identical particles (e.g. p-H1 collision)	barn $\cdot$ ster^{-1}
μ	cosine of the scattering angle {*center-of-mass (CM) system*}	dimensionless
m	incident particle mass	AMU
Z_1, Z_2	charge numbers of incident particle and target	dimensionless
s	spin (identical particles only, s = 0, 1/2, 1, 3/2,...)	dimensionless
A	ratio of target to projectile mass	dimensionless
k	particle wave number	barn$^{-\frac{1}{2}}$
η	Coulomb parameter	dimensionless
E	energy of the incident particle {*laboratory (lab) system*}	eV

Table 1.3: Elastic cross section for protons: Quantities and their units.

> This website provides access to the ENDF/B-VII library of evaluated incident-proton data for different elements. This information consists of raw and interpreted views of the ENDF/B-VII file as well as plots of the cross sections.

After required data from the ENDF library is available all forthcoming transformations and manipulations for the desired quantities are performed by algorithms written in MATLAB.

Since the library does not contain information on cross sections for protons being scattered at water molecules we use the additivity approximation: Differential cross sections for elastic scattering of projectiles by molecules are computed by adding the cross sections of the isolated atoms. Clearly, this approach discards physical phenomena like binding effects or interference between atoms. Although being purely heuristic, this ansatz is often applied in this context. Hence, we approximate the elastic differential cross section for water molecules by a linear combination of one oxygen nucleus and two hydrogen nuclei, i.e.,

$$\sigma^p_{\mathrm{H_2O}}(\tau, \mu) = \sigma^p_{\mathrm{O16}}(\tau, \mu) + 2\sigma^p_{\mathrm{H1}}(\tau, \mu). \tag{1.52}$$

In the ENDF library, elastic cross sections for protons scattered at individual atoms are represented by three components:

- Coulomb scattering (analytic Rutherford formula without electronic screening),
- nuclear scattering and
- the interference between them.

Above quantities are obtained by either theoretical calculations or experimental data for two-body interactions. Depending on the target nucleus, the database provides several formula. To avoid confusion we stick to the notation in the ENDF manual [126] and use the same quantities and units as tabulated in Table 1.3. In the following, the procedure for computing elastic cross sections of oxygen and hydrogen is explicitly described.

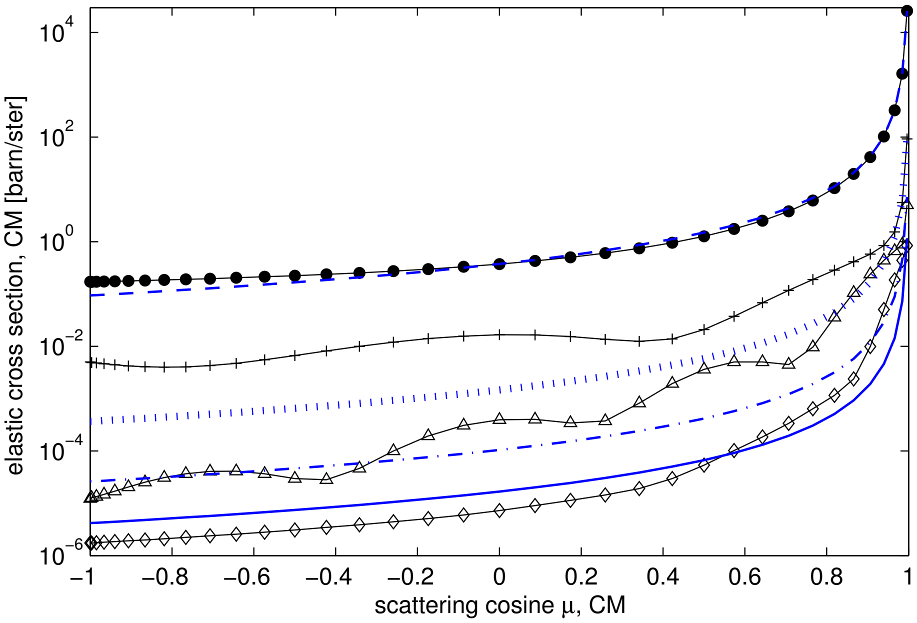


Figure 1.20: Elastic cross sections in p-O16 collisions: Black lines with markers are the corrected cross sections σ_{e} from (1.55); blue lines without markers are the Coulomb contribution from (1.53a). 1 MeV (solid circle and dashed line), 16 MeV (solid plus and dotted line), 60 MeV (solid triangle and dash-dot line), 150 MeV (solid diamond and solid line).

A. O16-Nucleus:

The Coulomb scattering is determined by

$$\sigma_{\mathrm{cd}} = \frac{\eta^2}{k^2(1-\mu^2)} \tag{1.53a}$$

where

$$k \approx \frac{A}{1+A}\sqrt{4.78453 \cdot 10^{-6}\, m\, E\ (\mathrm{AMU\ eV})^{-1}}, \tag{1.53b}$$

$$\eta \approx Z_1 Z_2 \sqrt{2.48058 \cdot 10^{4}\, \frac{m}{E}\ \mathrm{eV\ AMU}^{-1}}. \tag{1.53c}$$

The numerical values in (1.53b) and (1.53c) result from evaluating fundamental physical constants whose representation is much easier in this way.

In an energy range of 1MeV to 150MeV only experimental data is provided for various values of the scattering cosine between −1 and 0.9966. The domain for energy is subdivided into 74 points and the scattering cosine is tabulated for 36 values; both grids are non-equidistant.

In case of oxygen nuclei the representative quantity is given by the ”nuclear plus interference” cross section $\sigma_{\mathrm{ni}}(E)$. According to [126] this quantity is given by

$$\sigma_{\mathrm{ni}}(E) = \int_{\mu_{\min}}^{\mu_{\max}} [\sigma_{\mathrm{e}}(\mu, E) - \sigma_{\mathrm{cd}}(\mu, E)]d\mu \tag{1.54}$$

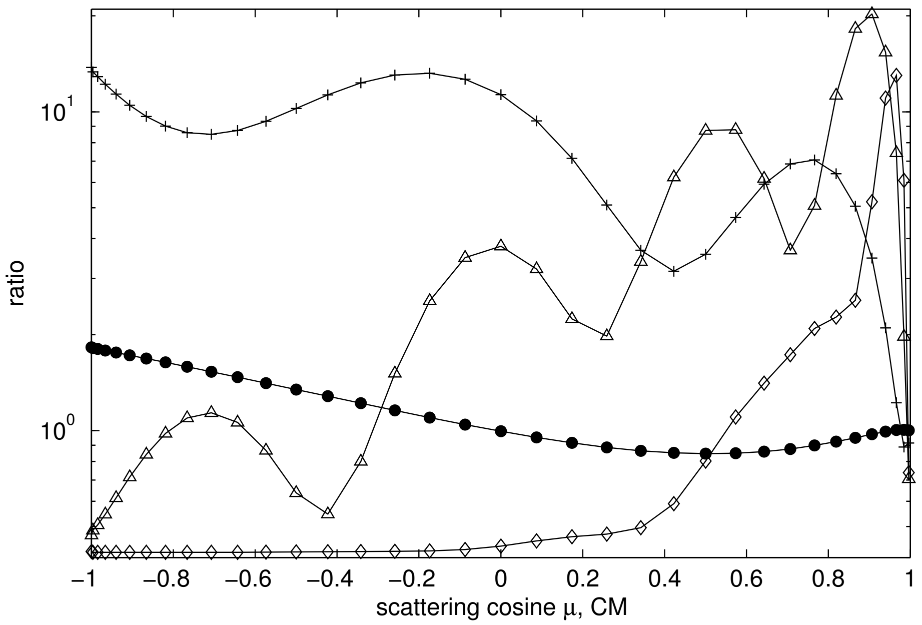


Figure 1.21: Ratios of $\sigma_{\mathrm{ed}}/\sigma_{\mathrm{cd}}$: 1 MeV (solid circle line), 16 MeV (solid plus line), 60 MeV (solid triangle line), 150 MeV (solid diamond line)

which is explicitly tabulated in the CM system as a function of μ and E. Hereby, $\sigma_{\mathrm{ed}}(\mu, E)$ is the desired elastic scattering cross section and $\sigma_{\mathrm{cd}}(\mu, E)$ the Coulomb term from (1.53).

Additionally, the ENDF data base also contains the auxiliary variable

$$p_{ni} = \begin{cases} \frac{\sigma_{\mathrm{ed}}(\mu,E)-\sigma_{cd}(\mu,E)}{\sigma_{\mathrm{ni}}(E)}, & \mu_{\min} \leq \mu \leq \mu_{\max} \\ 0, & \text{otherwise,} \end{cases} \tag{1.55}$$

which can be interpreted as the relative deviation of the elastic cross section σ_{ed} from its Coulomb contribution σ_{cd}. With tabulated values for σ_{ni} above expression can be easily solved for the elastic cross section σ_{ed}.

Remark 1. *Although we use the LISTEF output file to calculate our elastic cross sections it is important to emphasize that some files contain confusing information about the represented data. According to the description therein, some files are supposed to contain the 'Residual Cross Section' with 'Barns/Sr' as its unit. However, it turns out that, in fact, these quantities are the dimensionless p_{ni} from (1.55).*

For various energies, Figure 1.20 shows two different elastic cross sections: one including nuclear terms and one neglecting them. The larger the energy of the incident proton becomes the lower gets the corresponding elastic cross section. It is striking that for $\mu \approx 1$ the values drastically increase, i.e., the scattering is dominantly in the forward-direction. However, scattering also occurs for larger angles. This behavior is not captured well by the Rutherford cross section, i.e., the nuclear contribution must not be neglected. At 16 MeV, e.g., the corrected cross section for $\mu < 0$ is one magnitude larger than the Coulomb term (Figure 1.21). Nevertheless, for low energies and small deflection angles the difference between them decreases. Consequently, differential elastic cross sections which only describe Coulomb scattering are inaccurate.

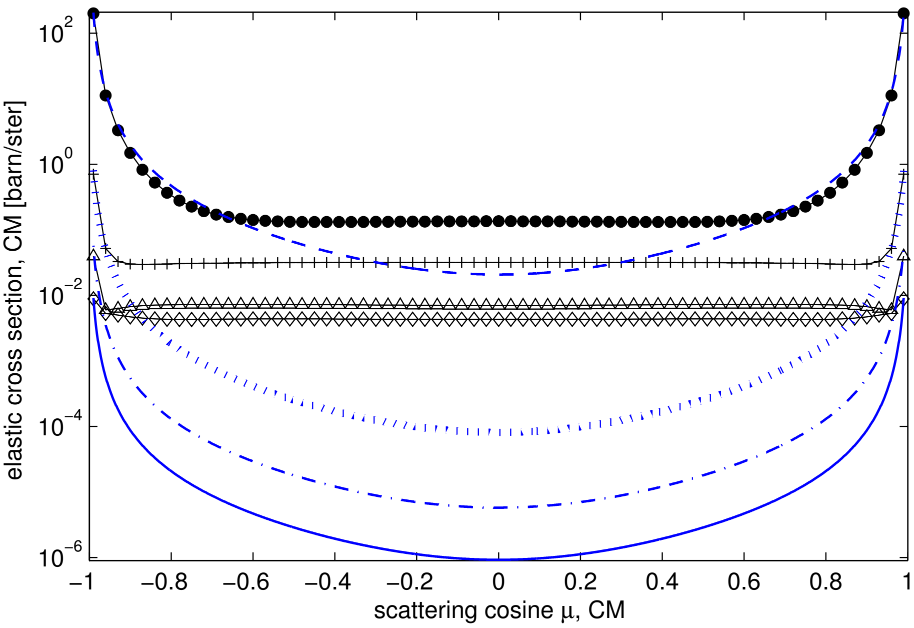


Figure 1.22: Elastic cross sections in p-p collisions: Black lines with markers are the corrected cross sections $\sigma_{\rm ei}$ from (1.57); blue lines without markers are the Coulomb contribution from (1.56). 1 MeV (solid circle and dashed line), 16 MeV (solid plus and dotted line), 60 MeV (solid triangle and dash-dot line), 150 MeV (solid diamond and solid line).

Calculations, where material damages by proton beams are studied and the angular dependence of scattering probabilities are important, should include both the Coulomb and nuclear interactions of protons with the material.

B. H1-Nucleus:

The situation for H1-nuclei is very different from collisions with O16-nuclei. As the hydrogen core only consists of one proton we, in fact, study proton-proton interactions, i.e., elastic collisions of identical particles. In this case, the Coulomb scattering cross section can also be computed analytically and reads:

$$\sigma_{\rm ci}(\mu, E) = \frac{2\eta^2}{k^2(1-\mu^2)}\left[\frac{1+\mu^2}{1-\mu^2} + \frac{(-1)^{2s}}{2s+1}\cos\left(\eta\,\ln\left(\frac{1+\mu}{1-\mu}\right)\right)\right], \quad s = 1/2, \tag{1.56}$$

where the particle wave number k and Coulomb parameter η are defined in eqs. (1.53).

The elastic p-p cross section is derived analytically by an R-matrix analysis [29, 74]. It is therefore possible to write it as an analytic formula. However, the ENDF database uses the expansion in a series of orthogonal Legendre polynomials which is truncated at some point and tabulates the coefficients of the truncated series.

If we denote the Legendre polynomials by $P_l(\mu), l = 0, 1, \ldots$, then the elastic cross section for p-p collisions is calculated by [126]:

$$\sigma_{\rm ei}(\mu, E) = \sigma_{\rm ci}(\mu, E) + \sigma_{\rm nc}(\mu, E) - \sigma_{\rm if}(\mu, E), \quad s = 1/2. \tag{1.57}$$

$\sigma_{\rm ci}$ is hereby the Coulomb term from (1.56), $\sigma_{\rm nc}$ represents the nuclear term and $\sigma_{\rm if}$ models the interaction between the latter and is called interference term.

First, we consider the nuclear contribution which is determined by

$$\sigma_{\mathrm{nc}}(\mu, E) = \sum_{l=0}^{NL} \frac{4l+1}{2} \mathbf{b_l}(E) P_{2l}(\mu). \tag{1.58}$$

NL is a natural number at which the series is truncated and the coefficients $\mathbf{b_l}(E) \in \mathbb{R}$ are real-valued functions. The expression for the interference contribution is the bottleneck of the representation because it contains singularities at $\mu \pm 1$ which show numerical instabilities:

$$\begin{aligned}\sigma_{\mathrm{if}}(\mu, E) = \frac{2\eta}{1-\mu^2} \Re \Bigg\{ \sum_{l=0}^{NL} \frac{2l+1}{2} \mathbf{a_l}(E) P_l(\mu) \Big[(1+\mu) e^{i\eta \ln\left(\frac{1-\mu}{2}\right)} \\ + (-1)^l (1-\mu) e^{i\eta \ln\left(\frac{1+\mu}{2}\right)} \Big] \Bigg\},\end{aligned} \tag{1.59a}$$

or equivalently,

$$\begin{aligned}\sigma_{\mathrm{if}}(\mu, E) = 2\eta \sum_{l=0}^{NL} \frac{2l+1}{2} P_l(\mu) \Bigg\{ \frac{\Re(\mathbf{a_l}(E)) \cos(\phi_1) - \Im(\mathbf{a_l}(E)) \sin(\phi_1)}{1-\mu} \\ + (-1)^l \frac{\Re(\mathbf{a_l}(E)) \cos(\phi_2) - \Im(\mathbf{a_l}(E)) \sin(\phi_2)}{1+\mu} \Bigg\},\end{aligned} \tag{1.59b}$$

where

$$\phi_1 = \eta \ln\left(\frac{1-\mu}{2}\right) \quad \text{and} \quad \phi_2 = \eta \ln\left(\frac{1+\mu}{2}\right). \tag{1.59c}$$

$\mathbf{a_l}(E) \in \mathbb{C}$ is a *complex-valued* function with $\Re(\mathbf{a_l}(E))$ as its real and $\Im(\mathbf{a_l}(E))$ as its imaginary part.

Altogether, to evaluate (1.57) one only needs the coefficients of the series $\Re(\mathbf{a_l}(E))$, $\Im(\mathbf{a_l}(E))$ and $\mathbf{b_l}(E)$. These coefficients are tabulated up to $NL = 6$.

Remark 2. *The interference term in eqs. (1.59) includes fractions where the numerator oscillates (due to sine and cosine functions) and the denominator tends to zero whenever* $\mu \to \pm 1$. *Indeed, this fact causes numerical instabilities and leads to negative cross sections* σ_{ei} *close to* $\mu \approx \pm 1$ *when making use of (1.59a) which is stated in the manual [126]. However, the coefficients in (1.59b) are grouped in such a way that the evaluation near* $\mu \approx \pm 1$ *remains stable.*

Remark 3. *Note that we do not use the output of the LISTEF program here because many coefficients of* σ_{if} *are either missing or not uniquely defined. Instead, the already mentioned website*

http://t2.lanl.gov/data/proton7.html

comes into play. It is well documented and contains the full ENDF data for protons. We use their tables to compute the desired quantities.

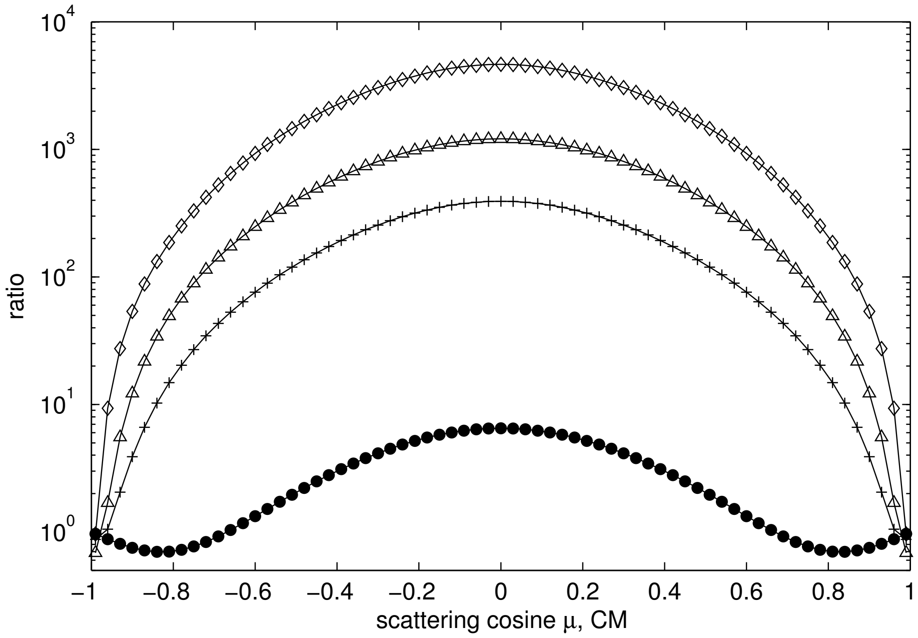


Figure 1.23: Ratios of $\sigma_{\mathrm{ei}}/\sigma_{\mathrm{ci}}$: 1 MeV (solid circle line), 16 MeV (solid plus line), 60 MeV (solid triangle line), 150 MeV (solid diamond line)

p-p elastic cross sections are provided for energies between 10^{-3} MeV and 150 MeV. This energy range is subdivided into 131 points whereas the scattering cosine is kept continuous.

Kinematics calculations show that in the center-of-mass system $\sigma_{ei}(\mu, E) = \sigma_{ei}(-\mu, E)$ which can also be observed in Figure 1.22. Like for $O16$ nuclei, the Rutherford cross section becomes less important for large energies. The nuclear contribution even dominates up to several orders of magnitude for large enough energies and μ away from ± 1 (Figure 1.23).

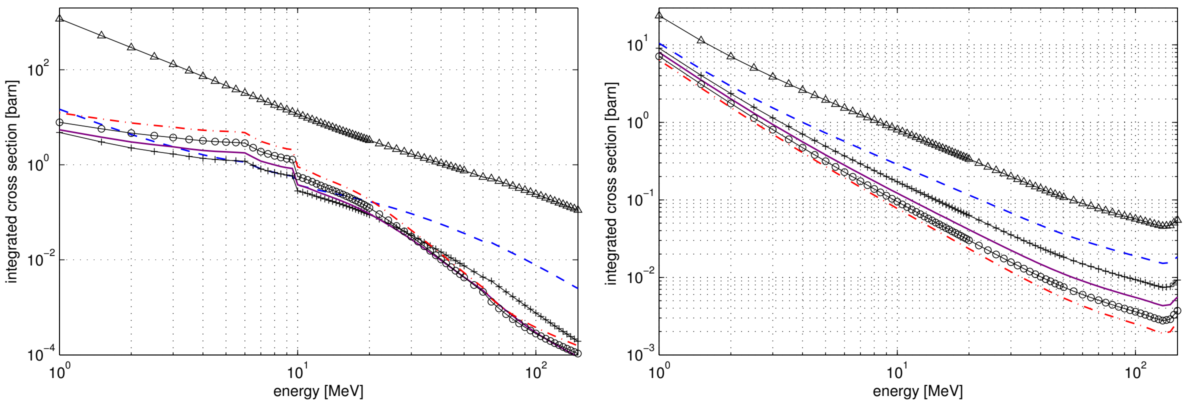
Remark 4. *Comparing p-O16 to p-p collisions, it might appear that there is a large probability for backward-scattering in p-p scattering. However, it should be emphasized that all data are presented in the center-of-mass system so far. Quantities in the Boltzmann equation are given in the laboratory system so that a transformation is still needed. This transformation has a smaller impact on p-O16 cross sections because the difference between the particle masses is large; but in case of p-p scattering this difference is significant.*

Transformation: Center-of-Mass to Laboratory System

We restrict our discussion to elastic scattering where the transformation for distinguishable particles can be computed by [17]:

$$\mu_{\mathrm{lab}} = \frac{\mu_{\mathrm{CM}} + \tau}{(1 + 2\tau\mu_{\mathrm{CM}} + \tau^2)^{1/2}} \quad \tau = \frac{m_P}{m_T}, \tag{1.60a}$$

$$\sigma_{\mathrm{lab}}(\mu_{\mathrm{lab}}, E) = \frac{(1 + 2\tau\mu_{\mathrm{CM}} + \tau^2)^{3/2}}{|1 + \tau\mu_{\mathrm{CM}}|}\, \sigma_{\mathrm{CM}}(\mu_{\mathrm{CM}}, E) \tag{1.60b}$$



(a) O16-nucleus from (1.62b)

(b) H1-nucleus from (1.62a)

Figure 1.24: Integrated cross sections for hydrogen and oxygen from (1.62): $n = 0$ (black solid triangle line), $n = 1$ (blue dashed line), $n = 2$ (black solid plus line), $n = 3$ (purple solid line), $n = 4$ (black solid circle line), $n = 5$ (red dash-dot line).

where μ_{lab} and μ_{CM} are the scattering cosines in the laboratory and CM system. Similarly, σ_{lab} denotes the elastic scattering cross section in the laboratory and σ_{CM} in the CM system. m_P is the projectile mass and m_T the mass of the target. For identical particles $\tau = 1$ and the formula simplifies to

$$\mu_{\text{lab}} = \sqrt{\frac{1+\mu_{\text{CM}}}{2}}, \tag{1.61a}$$

$$\sigma_{\text{lab}}(\mu_{\text{lab}}, E) = 4\sqrt{\frac{1+\mu_{\text{CM}}}{2}}\,\sigma_{\text{CM}}(\mu_{\text{CM}}, E) \tag{1.61b}$$

In particular, (1.61a) implies that scattering angles larger than 90° are not possible in p-p collisions.

For comparison reasons, we compute the following integrated cross sections in the laboratory system for oxygen and hydrogen separately (which up to a certain factor are related to the transport coefficients):

$$\int_{-1}^{1} \sigma^p_{H1}(\tau, \mu)(1-\mu)^n d\mu, \quad n = 0, ..., 5, \tag{1.62a}$$

$$\int_{-1}^{1} \sigma^p_{O16}(\tau, \mu)(1-\mu)^n d\mu, \quad n = 0, ..., 5. \tag{1.62b}$$

Figure 1.24 shows the results for different $n = 0, ..., 5$. For p-p interactions, it is striking that for increasing order n the quantities are strictly decreasing and the difference becomes significantly larger in regions of large energies. In the framework of generalized Fokker-Planck asymptotics (briefly discussed in Section 1.1.1), this behavior implies that, on the one hand, the scattering is mostly in the forward direction but, on the other hand, there is also an important amount of large-angle scattering. In this case, one can expect that high-order generalized Fokker-Planck calculations will give improved results in contrast to the standard low order Fokker-Planck approximation.

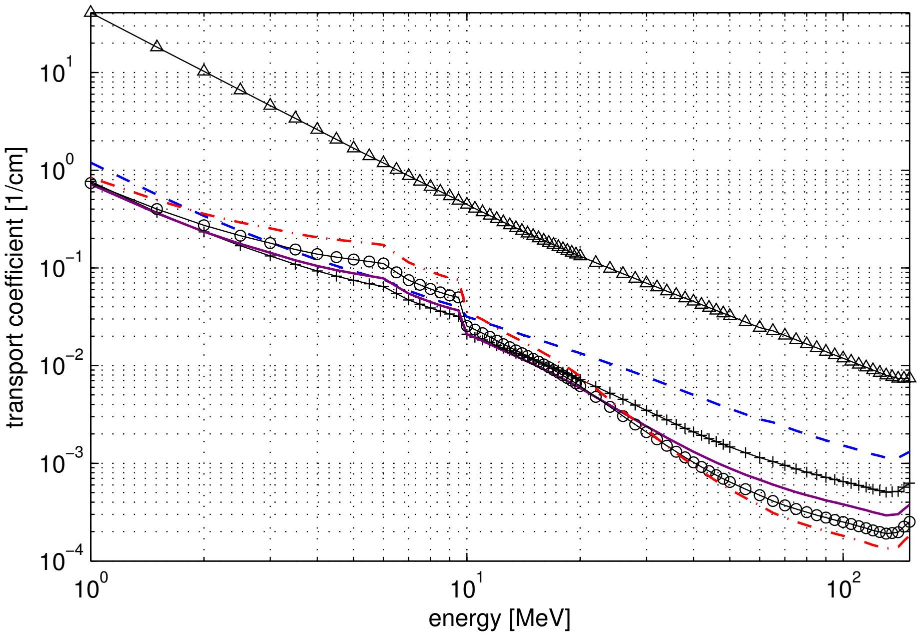


Figure 1.25: Elastic transport coefficients for proton-water scattering collisions: $n = 0$ (black solid triangle line), $n = 1$ (blue dashed line), $n = 2$ (black solid plus line), $n = 3$ (purple solid line), $n = 4$ (black solid circle line), $n = 5$ (red dash-dot line).

In p-O16 collisions we can only observe a large gap of one order of magnitude between the quantities for $n = 0$ and $n = 1$. However, larger orders of $n \geq 2$ are more or less grouped and a difference is only established for large energies.

Comparing H1- to O16-nuclei, the total cross section for $n = 0$, which gives the probability that the incident proton will undergo a scattering event, is much larger for p-O16 events.[1] However, for the first order $n = 1$ the behavior is vice versa: Particularly at high energies, p-H1 events dominate.

After performing the transformation to laboratory system we are now also able to compute the desired quantities from (1.52) which is numerically integrated over μ by the trapezoidal rule to obtain the n-th transport coefficient:

$$\xi_{p,n}^{\text{ENDF}}(\tau) = 2\pi \mathcal{N}_{H_2O} \int_{-1}^{1} \sigma_{H_2O}^{p}(\tau, \mu)(1-\mu)^n d\mu. \tag{1.63}$$

The transport coefficients for water are displayed in Figure 1.25. The 0-th coefficient is dominated by p-O16 collisions to a large extent. For $n \geq 1$, the behavior can be subdivided into two parts: Up to energies of approximately 20 MeV the coefficient is dominated by scatterings at O16-nuclei which become less important for larger energies where p-H1 events prevail.

---

[1] A geometrical interpretation for the total cross section is the area of a plane surface, orthogonal to the direction of the incident beam, which particles have to hit in order to be scattered at all.

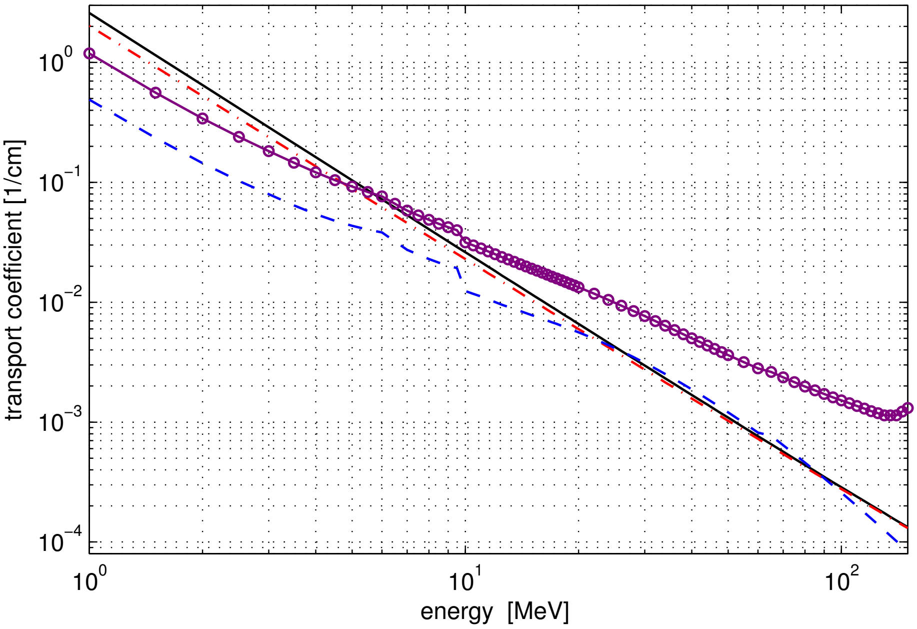


Figure 1.26: Different models for the 1st elastic transport coefficient of protons scattered at water molecules: Li/Petrasso (black solid line), Mott (red dash-dot line), ENDF: oxygen only (blue dashed line), ENDF: linear combination of oxygen and hydrogen (purple solid circle line).

Comparison

We compare above approximations for the first transport coefficient of protons scattered at water molecules. Figure 1.26 illustrates the different approaches. Although small deviations appear for smaller energies, the analytic formulas of Mott (eqs. (1.48)) and Li/Petrasso (eqs. (1.49)) are close to each other. This is due to the fact that the largest contribution in both models is the Coulomb interaction of protons with an atom of atomic number $Z = 10$. As cross sections from the ENDF database additionally include important nuclear interactions they show significant differences. The transport coefficient for oxygen, computed by simply neglecting the H1-term σ^p_{H1} in (1.52), is much closer to the analytic expressions for large energies. On the contrary, taking p-p collisions into account by a linear combination of hydrogen and oxygen cross sections leads to the opposite behavior: Values for growing energies above 10 MeV increasingly deviate from results calculated by the Mott- or Li/Petrasso-formula.

Remark 5. *We want to emphasize that above transport coefficients represent several possibilities to approach the realistic transport coefficients in water. To distinguish which model describes reality best, dose calculations must be performed and compared to Monte Carlo results which are benchmarked against physical experiments.*

1.4.3 Stopping Power

The process of proton stopping when passing through matter has been intensively studied. Although ion-nucleus collisions are dominated by elastic scattering the energy transfer is relatively small. The nucleus is recoiled but, due to the comparatively large mass of the target, the projectile only transfers little energy to the nucleus. However, the largest amount of energy is lost in ion-electron interactions: After colliding with ions, atomic electrons can be excited to higher states or are even ionized which leads to an absorption of a large amount of energy. Therefore, energy loss due to proton-nucleus collisions is divided into two crucial phenomena:

- **electronic stopping**: Interactions with atomic electrons in the electromagnetic field induce energy losses due to Coulomb interactions which result in the ionization and excitation of the atom. This is the dominant contribution for a large energy range.
- **nuclear stopping**: In elastic collisions of projectiles with atomic cores energy is transferred to the recoiling cores.

The stopping power is defined as the average energy loss per unit path length. It strongly depends on the material and energy of the projectile.

Analytic formula

The fundamental equation for the electronic stopping power is Bethe's formula [14] which has been modified and improved step-by-step to approach realistic results gained from experiments:

$$S_{\mathrm{Bethe}}(\tau') = \frac{4\pi r_e^2 \rho_e m_e c^2}{\beta^2} \left[\ln\left(\frac{2 m_e c^2 \beta^2}{I \cdot (1-\beta^2)} \right) - \beta^2 \right], \tag{1.64}$$

where I is the mean excitation energy. This quantity plays a central role in Bethe's stopping power formula and has been determined empirically for a large number of materials. We use tabulated data from [155] for I.

NIST-Database

The National Institute of Standards and Technology provides a program called PSTAR which calculates stopping powers for protons in various materials in the energy range from 10^{-3} MeV to 10^5 MeV. For protons, methods and their underlying theory are taken from [83].

At high energies, values for electronic stopping powers are evaluated using Bethe's stopping-power formula with the following corrections:

- shell correction (contributes mostly below ≈ 1 MeV),
- density-effect correction (important for energies above several hundred MeV) and
- Barkas and Bloch correction (small impact on the stopping power form small energies ≈ 1 MeV).

A detailed description and comparison of the contributions of the single correction terms can be found in [185].

Experimental information and fitting-formulas for the electronic stopping power is used at low energies. For protons, the boundary between the high- and low-energy regions is set to roughly 0.5 MeV.

The nuclear stopping power, however, is obtained by a classical-mechanics orbit calculation [50] where the screened potential is assumed to be the Thomas-Fermi potential.

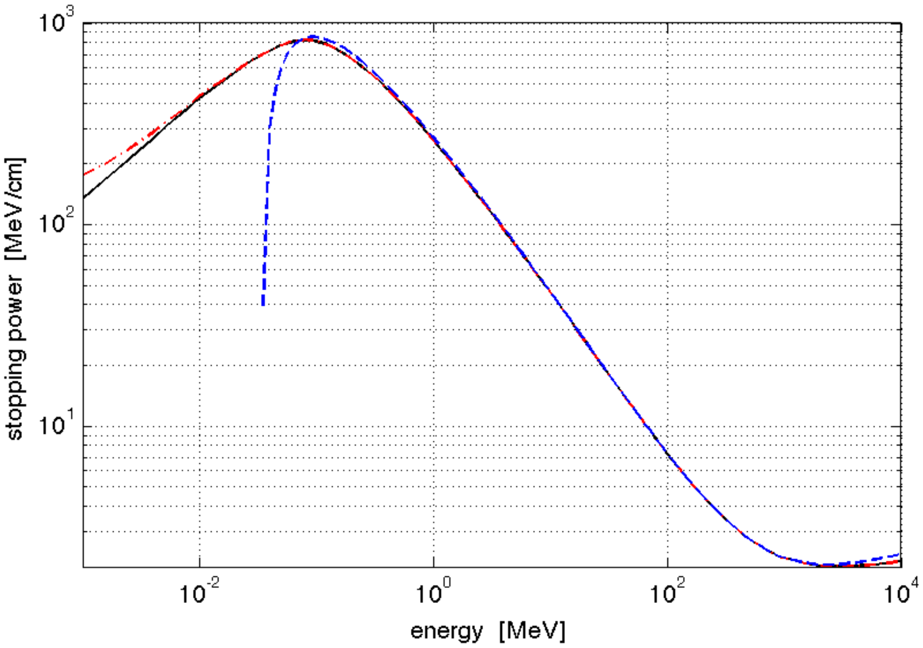


Figure 1.27: Proton stopping power for water: electronic (black solid line), total (red dash-dot line), Bethe-formula (blue dashed line)

Comparison

Figure 1.27 displays the stopping power for water computed with different formulas. For growing energy, the function increases until a maximum at approximately 0.1 MeV is reached. Then the stopping power is monotonically decreasing and a minimum is achieved at $2 \cdot 10^3$ MeV. It is striking that (1.64) is very close to NIST's database down to fairly small energies, of the order of 0.1 MeV. Nevertheless, for smaller energies values according to Bethe's formula drop and become highly inaccurate.

Chapter 2

Time-Dependent Simplified P_N Equations

2.1 Introduction

The mathematical equation describing linear transport problems is the linear Boltzmann equation. Its large dimensionality and the integro-differential structure makes this equation difficult to solve analytically and numerically. This is why a lot of effort has been made to develop approximations. The spherical harmonics (P_N) equations have been a standard approximation already known at the beginning of the 20th century [30, 39, 87]. A big drawback of the P_N equations in 3-D, which has made them unattractive for practical applications, is their complicated form and the large number of equations growing as $(N + 1)^2$. In view of low computational resources in those days, it was inevitable to come up with simpler equations for the solution of realistic problems: Gelbard [63–65] therefore proposed the steady-state *simplified* P_N (SP_N) equations which are simpler to implement and whose number increases in general geometries only linearly as $(N + 1)$ (versus quadratically as $(N + 1)^2$). However, his derivation in 3-D geometries was purely ad-hoc (by taking the 1-D P_N equations and replacing the 1-D spatial operators by its 3-D generalizations, i.e., gradients and divergence operators). Due to the lack of a theoretical foundation, the SP_N equations had not been accepted as an approximation to the transport equation for a long time until first theoretical justifications were presented (asymptotic and variational analysis in [18, 102, 148, 175]). In the framework of a Galerkin finite element method, the well-posedness of the *steady-state* SP_N equations is shown in [184] for $N = 1, 3, 5, 7$ where a proof for existence and uniqueness is provided. A detailed review of the SP_N equations can be found in [120].

Originally intended for applications in nuclear engineering, the SP_N equations are, indeed, implemented and used for neutron transport problems nowadays [12,31,43,71]. A wide range of additional applications mainly developed in the past decade after first theoretical foundations for the SP_N method has been given, e.g., radiative cooling of glass [57, 105], radiative transfer in tissue [93], fluorescence tomography [94], design of combustion chambers for gas turbines [157, 159], crystal growth of semitransparent materials [10] and photon and electron radiotherapy [95].

The majority of previous investigations focused on *steady-state* SP_N equations. One of the ideas for deriving SP_N equations is to explicitely solve for odd moments and introduce them into equations with even moments. This is the reason why the *steady-state* SP_N equations reduce to a hierarchy of diffusion equations. In the *time-dependent* case, this procedure cannot be applied in the same way because of the additional time derivative.

To the author's knowledge, the first formal asymptotic derivation for *time-dependent* SP_N equations is developed in [56] and Finite Element numerical solutions of this system are computed in [58]. An alternative strategy for the derivation of moment methods for the time-dependent radiative transfer equation is the method of optimal prediction [59, 160]. It turns out that this formalism yields existing moment models such as P_N. Additionally, it is shown in [160] that this ansatz can be used to derive variations of the parabolic SP_N equations from [56].

Here, we present a different asymptotic analysis for *time-dependent* SP_N equations and explain how these equations can be derived in an ad-hoc way. Our analysis makes use of a different scaling leading to final equations which are not equivalent to those investigated in [56]. We want to highlight the differences and similarities between the approach therein and the work presented here:

- Guided by the fact that, in *steady-state*, SP_N approximations are diffusion equations, the authors in [56] derive *time-dependent* SP_N equations which are *parabolic* PDEs. We drop this goal in our asymptotic analysis here which, indeed, allows us to derive a system of *hyperbolic* PDEs for the time-dependent SP_N equations.

- The analysis in [56] is performed by a parabolic scaling where the time-derivative is scaled by ε^2. As the final SP_N equations are only accurate for $\varepsilon \approx 0$ this assertion implies that temporal changes of the solution should be small. The forthcoming asymptotic theory does not require a scaling of the time derivative operator. Hence, numerical solutions of problems with large time-derivatives should also be accurate. However, both asymptotic theories assume that space-derivatives (scaled by ε) are small.

- The derivation in [56] unfolds an ambiguity of how to define the ϕ_2 unknown. This ambiguity is only partly captured by an introduced free parameter α in the approximate system. Although leading to more flexibility, different choices of this parameter α can only give results which differ in the magnitude of $\mathcal{O}(\varepsilon^6)$. Moreover, for a well-posed system α is bounded by $0 < \alpha < 0.9$. It turns out that approaching the lower or upper bound, numerical solutions of the regarding system diverge from the true solution and develop spurious shapes.

  Our analysis discussed here is similar to the asymptotic derivation of the steady-state SP_N equations in the sense that auxiliary variables can be defined in a similar way. This is why there is less ambiguity and no free parameter.

- The asymptotics in both papers is performed only up to SP_3. However, as it is additionally presented here how the classic derivation by Gelbard [63–65] can yield exactly the same SP_N equations as asymptotically derived up to $N = 3$, it is straight-forward to obtain higher-order SP_N approximations. In this way, one can avoid high-order asymptotic analysis which becomes even more lengthy and complicated. Of course, this is only possible on condition that high-order asymptotics would give the same results.

As the first asymptotic derivation for the time-dependent SP_N equations has not been developed until 2007 [56], it had still been necessary to perform time-dependent

SP_N simulations before. In practice, to keep the second-order form for the *time-dependent* generalization, some simplifications were performed or time-derivatives were dropped in certain equations. A different possibility, which is already included in some codes nowadays (e.g., PARCS [43]), was to simply add the partial time derivative to each of the steady-state equations in first-order form. It will be shown in Section 2.2 that the SP_3 equations, gained heuristically in this way, are equivalent to those derived in this work. A similar approach can be found in [130] where the additional time derivative in the SP_N equations is first discretized while treating the other variables continuously. The standard procedure to obtain the *steady-state* SP_N approximation is then applied to the generated system.

The chapter is organized as follows: In Section 2.2, a derivation of the time-dependent 3-D SP_3 equations is presented which basically follows the lines of the classic derivation given by Gelbard [63–65]. In particular, this approach is purely ad-hoc. A theoretical foundation is then given in Section 2.3 for the time-dependent method up to order $N = 3$. It is important to stress that our asymptotic analysis in Section 2.3 is only correct for a homogeneous medium. In Section 2.2, heterogeneous media are considered in an ad-hoc derivation. This procedure is also generalized to SP_N equations of arbitrary order and numerical SP_N solutions are compared to diffusion and P_N results in Section 2.4.

2.2 Classic (Ad-Hoc) Derivation

In 1-D slab geometry, the P_N equations have a simple structure and their number of unknowns is only $(N + 1)$. However, extending them to multi-dimensional geometries implies an expansion of the angular flux in spherical harmonics. Many extra degrees of freedom are added and an additional coupling occurs. Consequently, their original simplicity is lost and the number of equations increases quadratically. To keep the pleasing form of the 1-D slab geometry case one can formally replace the terms in the P_N equations in a proper way to get the simplified P_N approximation.

The time-dependent, monoenergetic, isotropically scattering linear Boltzmann equation reads as follows:

$$\begin{aligned}\frac{1}{v}\frac{\partial\Psi}{\partial t}(\underline{x},\underline{\Omega},t) + \underline{\Omega}\cdot\underline{\nabla}\Psi(\underline{x},\underline{\Omega},t) + \Sigma_t(\underline{x},t)\Psi(\underline{x},\underline{\Omega},t)&\\ = \frac{\Sigma_s(\underline{x},t)}{4\pi}\int_{S^2}\Psi(\underline{x},\underline{\Omega}',t)d\Omega' + \frac{1}{4\pi}Q(\underline{x},t)& \qquad (2.1)\end{aligned}$$

$$\begin{aligned}\Sigma_s(\underline{x},t) &\equiv \text{scattering cross section},\\ \Sigma_a(\underline{x},t) &\equiv \text{absorption cross section},\\ \Sigma_t(\underline{x},t) &= \Sigma_s(\underline{x},t) + \Sigma_a(\underline{x},t) \equiv \text{total cross section},\\ Q(\underline{x},t) &\equiv \text{internal source}.\end{aligned}$$

The angular flux $\Psi(\underline{x},\underline{\Omega},t)$ describes the particle density at position $\underline{x}\in\mathbb{R}^3$ and time t traveling in direction $\underline{\Omega}\in S^2$ at velocity v. Penetrating the background medium, particles interact with the material which is specified by scattering and absorption probabilities $\Sigma_s(\underline{x},t)$ and $\Sigma_a(\underline{x},t)$. However, they are assumed not to interact with themselves. The first term of (2.1) is the temporal rate of change in Ψ, the second is the leakage or drift term and the third quantifies the loss of particles due to out-scattering and absorption by the medium. The right-hand side of (2.1) characterizes the gain in particles: isotropic in-scattering processes are modeled by $\Sigma_s(\underline{x}, t)$ times the integral of Ψ over the unit sphere (all possible outgoing directions). Additional particles can also be inserted into the system by an internal isotropic source $Q(\underline{x}, t)$.

To derive the SP_N approximation to above Boltzmann equation, (2.1) is first restricted to 1-D planar geometry:

$$\begin{aligned}\frac{1}{v}\frac{\partial \Psi}{\partial t}(x,\mu,\varphi,t) + \mu\frac{\partial \Psi}{\partial x}(x,\mu,\varphi,t) + \Sigma_t(x,t)\Psi(x,\mu,\varphi,t) \\ = \frac{\Sigma_s(x,t)}{4\pi}\int_0^{2\pi}\int_{-1}^{1}\Psi(x,\mu',\varphi',t)d\mu' d\varphi' + \frac{1}{4\pi}Q(x,t)\end{aligned} \tag{2.2}$$

Operating on (2.2) by $\int_0^{2\pi}\cdot\, d\varphi$ and defining

$$\psi(x,\mu,t) := \int_0^{2\pi}\Psi(x,\mu,\varphi,t)d\varphi,$$

we obtain the 1-D azimuthally-symmetric transport equation in slab geometry:

$$\begin{aligned}\frac{1}{v}\frac{\partial \psi}{\partial t}(x,\mu,t) + \mu\frac{\partial \psi}{\partial x}(x,\mu,t) + \Sigma_t(x,t)\psi(x,\mu,t) \\ = \frac{\Sigma_s(x,t)}{2}\int_{-1}^{1}\psi(x,\mu',t)d\mu' + \frac{1}{2}Q(x,t)\end{aligned} \tag{2.3}$$

Introducing the orthogonal Legendre polynomials $P_n(\mu), n \geq 0$, which satisfy

$$\mu P_n(\mu) = \frac{n+1}{2n+1}P_{n+1}(\mu) + \frac{n}{2n+1}P_{n-1}(\mu) \tag{2.4a}$$

$$\int_{-1}^{1}P_n(\mu)P_m(\mu)d\mu = \frac{2}{2n+1}\delta_{n,m}, \tag{2.4b}$$

and defining the Legendre moments of $\psi(x,\mu,t)$:

$$\phi_n(x,t) = \int_{-1}^{1}P_n(\mu)\psi(x,\mu,t)d\mu, \quad n \geq 0, \tag{2.5}$$

we can write (2.3) as

$$\frac{1}{v}\frac{\partial \psi}{\partial t}(x,\mu,t) + \mu\frac{\partial \psi}{\partial x}(x,\mu,t) + \Sigma_t(x,t)\psi(x,\mu,t) = \frac{\Sigma_s(x,t)}{2}\phi_0(x,t) + \frac{Q(x,t)}{2}. \tag{2.6}$$

Multiplying (2.6) by $P_n(\mu)$ and using (2.4), we get

$$\begin{aligned}\frac{1}{v}\frac{\partial}{\partial t}P_n(\mu)\psi(x,\mu,t) + \frac{\partial}{\partial x}\left(\frac{n+1}{2n+1}\psi(x,\mu,t) + \frac{n}{2n+1}P_{n-1}(\mu)\psi(x,\mu,t)\right) \\ + \Sigma_t(x,t)P_n(\mu)\psi(x,\mu,t) = P_n(\mu)\left(\frac{\Sigma_s(x,t)}{2}\phi_0(x,t) + \frac{Q(x,t)}{2}\right).\end{aligned} \tag{2.7}$$

Integrating this equation over $-1 \leq \mu \leq 1$ and using (2.4) and (2.5), we get

$$\begin{aligned} \frac{1}{v}\frac{\partial \phi_n}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(\frac{n+1}{2n+1}\phi_{n+1}(x,t) + \frac{n}{2n+1}\phi_{n-1}(x,t)\right) + \Sigma_t(x,t)\phi_n(x,t) \\ = \delta_{n,0}\left(\Sigma_s(x,t)\phi_0(x,t) + Q(x,t)\right) \end{aligned} \tag{2.8}$$

This result holds, and is exact, for all integers $n \geq 0$. Unfortunately, it never yields a closed system of equations; there is always one more unknown function than there are equations. For example, the first four equations (corresponding to $n = 0, 1, 2, 3$) are:

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(x,t) + \frac{\partial \phi_1}{\partial x}(x,t) + \Sigma_a(x,t)\phi_0(x,t) = Q(x,t), \tag{2.9a}$$

$$\frac{1}{v}\frac{\partial \phi_1}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(\frac{2}{3}\phi_2(x,t) + \frac{1}{3}\phi_0(x,t)\right) + \Sigma_t(x,t)\phi_1(x,t) = 0, \tag{2.9b}$$

$$\frac{1}{v}\frac{\partial \phi_2}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(\frac{3}{5}\phi_3(x,t) + \frac{2}{5}\phi_1(x,t)\right) + \Sigma_t(x,t)\phi_2(x,t) = 0, \tag{2.9c}$$

$$\frac{1}{v}\frac{\partial \phi_3}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(\frac{4}{7}\phi_4(x,t) + \frac{3}{7}\phi_2(x,t)\right) + \Sigma_t(x,t)\phi_3(x,t) = 0. \tag{2.9d}$$

These are four equations containing the five unknown functions $\phi_n(x,t), 0 \leq n \leq 4$.

The standard P_N approximation is simply to set the highest Legendre moment of $\psi(x,\mu,t)$ equal to zero. In the case of (2.9), one sets:

$$\phi_4(x,t) = 0. \tag{2.10}$$

From now on, to keep the discussion simple, we will work with the specific system of (2.9) and (2.10). However, everything which is done in the following can also be generalized to more (or less) than four angular moments of (2.3).

After invoking the closure of (2.10), eqs. (2.9) become the following classic planar geometry time-dependent P_3 equations:

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(x,t) + \frac{\partial \phi_1}{\partial x}(x,t) + \Sigma_a(x,t)\phi_0(x,t) = Q(x,t), \tag{2.11a}$$

$$\frac{1}{v}\frac{\partial \phi_1}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(\frac{2}{3}\phi_2(x,t) + \frac{1}{3}\phi_0(x,t)\right) + \Sigma_t(x,t)\phi_1(x,t) = 0, \tag{2.11b}$$

$$\frac{1}{v}\frac{\partial \phi_2}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(\frac{3}{5}\phi_3(x,t) + \frac{2}{5}\phi_1(x,t)\right) + \Sigma_t(x,t)\phi_2(x,t) = 0, \tag{2.11c}$$

$$\frac{1}{v}\frac{\partial \phi_3}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(\frac{3}{7}\phi_2(x,t)\right) + \Sigma_t(x,t)\phi_3(x,t) = 0. \tag{2.11d}$$

Solving for the odd moments of eqs. (2.11) we divide (2.11b) and (2.11d) by $\Sigma_t(x)$ and get

$$\mathcal{T}\phi_1(x,t) + \frac{1}{3}\mathcal{X}(2\phi_2(x,t) + \phi_0(x,t)) + \phi_1(x,t) = 0, \tag{2.12a}$$

$$\mathcal{T}\phi_3(x,t) + \frac{1}{7}\mathcal{X}(3\phi_2(x,t)) + \phi_3(x,t) = 0, \tag{2.12b}$$

where

$$\mathcal{T} := \frac{1}{v\Sigma_t(x,t)}\frac{\partial}{\partial t}, \tag{2.13a}$$

$$\mathcal{X} := \frac{1}{\Sigma_t(x,t)}\frac{\partial}{\partial x} \tag{2.13b}$$

are two dimensionless operators. Hence, we obtain:

$$\phi_1(x,t) = -\frac{1}{3}(I+\mathcal{T})^{-1}\mathcal{X}(2\phi_2(x,t)+\phi_0(x,t)), \tag{2.14a}$$

$$\phi_3(x,t) = -\frac{1}{7}(I+\mathcal{T})^{-1}\mathcal{X}(3\phi_2(x,t)). \tag{2.14b}$$

Introducing these expressions into the first and third of eqs. (2.11), we get

$$\frac{1}{v}\frac{\partial\phi_0}{\partial x}(x,t) - \frac{1}{3}\frac{\partial}{\partial x}(I+\mathcal{T})^{-1}\mathcal{X}(2\phi_2(x,t)+\phi_0(x,t)) + \Sigma_a(x,t)\phi_0(x,t) = Q(x,t) \tag{2.15}$$

and

$$\frac{1}{v}\frac{\partial\phi_2}{\partial x}(x,t) - \frac{\partial}{\partial x}\left[\frac{9}{35}(I+\mathcal{T})^{-1}\mathcal{X}\phi_2(x,t) + \frac{2}{15}(I+\mathcal{T})^{-1}\mathcal{X}(2\phi_2(x,t)+\phi_0(x,t))\right] \\ + \Sigma_t(x,t)\phi_2(x,t) = 0. \tag{2.16}$$

Equivalently,

$$\frac{1}{v}\frac{\partial\phi_0}{\partial t}(x,t) - \frac{1}{3}\left[\frac{\partial}{\partial x}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t(x,t)}\frac{\partial}{\partial x}\right]\Big(\phi_0(x,t)+2\phi_2(x,t)\Big) \\ + \Sigma_a(x,t)\phi_0(x,t) = Q(x,t), \tag{2.17a}$$

$$\frac{1}{v}\frac{\partial\phi_2}{\partial t}(x,t) - \frac{1}{3}\left[\frac{\partial}{\partial x}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t(x,t)}\frac{\partial}{\partial x}\right]\left(\frac{2}{5}\phi_0(x,t)+\frac{11}{7}\phi_2(x,t)\right) \\ + \Sigma_t(x,t)\phi_2(x,t) = 0. \tag{2.17b}$$

Remark 6. *Eqs. (2.17) are the time-dependent planar geometry P_3 equations, written in second-order form. These equations contain the operator $(I+\mathcal{T})^{-1}$ which depends parametrically on x but acts only on t.*
Assuming the most general possible situation, in which $\Sigma_t = \Sigma_t(x,t)$ is a function of both x and t we can solve the first-order ODE

$$\left(1+\frac{1}{v\Sigma_t(x,t)}\frac{\partial}{\partial t}\right)g(x,t) = f(x,t),$$

for

$$g(t) = \left(1+\frac{1}{v\Sigma_t(x,t)}\frac{\partial}{\partial t}\right)^{-1} f(x,t) \tag{2.18}$$

$$= \left[\left(1+\frac{1}{v\Sigma_t(x,t)}\frac{\partial}{\partial t}\right)^{-1} f(x,t_0)\right] e^{-\int_{t_0}^{t} v\Sigma_t(x,t')dt'} \\ + \int_{t_0}^{t} v\Sigma_t(x,t')f(x,t')e^{-\int_{t'}^{t} v\Sigma_t(x,t'')dt''}dt'. \tag{2.19}$$

Thus, $(I+\mathcal{T})^{-1}$ is a relatively simple operator and later we will show that it can be made to disappear from the final result.

To derive the SP$_3$ equations, we formally replace the 1-D operator

$$\frac{\partial}{\partial x}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\frac{\partial}{\partial x} \tag{2.20}$$

by the 3-D operator:

$$\begin{aligned}\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\underline{\nabla} = \frac{\partial}{\partial x}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t(\underline{x},t)}\frac{\partial}{\partial x} + \frac{\partial}{\partial y}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t(\underline{x},t)}\frac{\partial}{\partial y}\\ + \frac{\partial}{\partial z}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t(\underline{x},t)}\frac{\partial}{\partial z}.\end{aligned} \tag{2.21}$$

This step is purely ad-hoc; yet it is exactly what is (or rather, was) done in the original (1960) derivation of the steady-state SP_N equations [63–65]. The time-dependent planar geometry P$_3$ equations then become the time-dependent 3-D SP$_3$ equations:

$$\begin{aligned}&\frac{1}{v}\frac{\partial\phi_0}{\partial t}(\underline{x},t) + \Sigma_a(\underline{x},t)\phi_0(\underline{x},t)\\ &\qquad = \underline{\nabla}\cdot(I+\mathcal{T})^{-1}\frac{1}{3\Sigma_t(\underline{x},t)}\underline{\nabla}\Big(\phi_0(\underline{x},t)+2\phi_2(\underline{x},t)\Big) + Q(\underline{x},t),\end{aligned} \tag{2.22a}$$

$$\begin{aligned}&\frac{1}{v}\frac{\partial\phi_2}{\partial t}(\underline{x},t) + \Sigma_t(\underline{x},t)\phi_2(\underline{x},t)\\ &\qquad = \underline{\nabla}\cdot(I+\mathcal{T})^{-1}\frac{1}{3\Sigma_t(\underline{x},t)}\underline{\nabla}\Big(\frac{2}{5}\phi_0(\underline{x},t)+\frac{11}{7}\phi_2(\underline{x},t)\Big).\end{aligned} \tag{2.22b}$$

These equations directly reduce to the standard steady-state SP$_3$ equations when

$$\mathcal{T} = \frac{1}{v\Sigma_t(\underline{x},t)}\frac{\partial}{\partial t} = 0.$$

To summarize, eqs. (2.22) are obtained by the following procedure:

(1) Restrict (2.1) to 1-D planar geometry [(2.2)].

(2) Take the first four Legendre moments of the 1-D transport equation [eqs. (2.9)].

(3) Set $\phi_4(x,t)=0$ [(2.10)] to obtain the 1-D P$_3$ eqs. (2.11).

(4) Eliminate the odd moments $\phi_1(x,t)$ and $\phi_3(x,t)$ to obtain two second-order equations for the even moments $\phi_0(x,t)$ and $\phi_2(x,t)$ [eqs. (2.17)].

(5) Extend the (spatially) 1-D operator

$$\frac{\partial}{\partial x}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t(x,t)}\frac{\partial}{\partial x}$$

to the 3-D operator [eqs. (2.21)]

$$\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t(\underline{x},t)}\underline{\nabla}.$$

Remark 7. *In this procedure, steps 3 and 5 are ad-hoc. Moreover, the restriction from 3-D to 1-D in step 1 and the subsequent generalization from 1-D to 3-D in step 5 both seem artificial. Nonetheless, the non-rigorous derivation of eqs. (2.22) given above follows step-by-step the original derivation of the steady-state SP_N equations by Gelbard [63–65].*

Next, we rewrite eqs. (2.22) to a system of hyperbolic PDEs by defining

$$\underline{J}_0(\underline{x},t) := -(I+\mathcal{T})^{-1}\frac{1}{3\Sigma_t(\underline{x},t)}\underline{\nabla}\Big(\phi_0(\underline{x},t)+2\phi_2(\underline{x},t)\Big), \tag{2.23a}$$

$$\underline{J}_2(\underline{x},t) := -(I+\mathcal{T})^{-1}\frac{1}{3\Sigma_t(\underline{x},t)}\underline{\nabla}\left(\frac{2}{5}\phi_0(\underline{x},t)+\frac{11}{7}\phi_2(\underline{x},t)\right). \tag{2.23b}$$

Equivalently,

$$(I+\mathcal{T})\underline{J}_0 = -\frac{1}{3\Sigma_t(\underline{x},t)}\underline{\nabla}\Big(\phi_0(\underline{x},t)+2\phi_2(\underline{x},t)\Big), \tag{2.24a}$$

$$(I+\mathcal{T})\underline{J}_2 = -\frac{1}{3\Sigma_t(\underline{x},t)}\underline{\nabla}\left(\frac{2}{5}\phi_0(\underline{x},t)+\frac{11}{7}\phi_2(\underline{x},t)\right), \tag{2.24b}$$

or

$$\left(\Sigma_t(\underline{x},t)+\frac{1}{v}\frac{\partial}{\partial t}\right)\underline{J}_0 = -\frac{1}{3}\underline{\nabla}\Big(\phi_0(\underline{x},t)+2\phi_2(\underline{x},t)\Big), \tag{2.25a}$$

$$\left(\Sigma_t(\underline{x},t)+\frac{1}{v}\frac{\partial}{\partial t}\right)\underline{J}_2 = -\frac{1}{3}\underline{\nabla}\left(\frac{2}{5}\phi_0(\underline{x},t)+\frac{11}{7}\phi_2(\underline{x},t)\right). \tag{2.25b}$$

Then, eqs. (2.22) can be written as:

$$\frac{1}{v}\frac{\partial\phi_0}{\partial t}(\underline{x},t)+\underline{\nabla}\cdot\underline{J}_0(\underline{x},t)+\Sigma_a(\underline{x},t)\phi_0(\underline{x},t) = Q(\underline{x},t), \tag{2.26a}$$

$$\frac{1}{v}\frac{\partial\phi_2}{\partial t}(\underline{x},t)+\underline{\nabla}\cdot\underline{J}_2(\underline{x},t)+\Sigma_t(\underline{x},t)\phi_2(\underline{x},t) = 0, \tag{2.26b}$$

and

$$\frac{1}{v}\frac{\partial\underline{J}_0}{\partial t}(\underline{x},t)+\frac{1}{3}\underline{\nabla}\Big(\phi_0(\underline{x},t)+2\phi_2(\underline{x},t)\Big)+\Sigma_t(\underline{x},t)\underline{J}_0(\underline{x},t) = 0, \tag{2.26c}$$

$$\frac{1}{v}\frac{\partial\underline{J}_2}{\partial t}(\underline{x},t)+\frac{1}{3}\underline{\nabla}\left(\frac{2}{5}\phi_0(\underline{x},t)+\frac{11}{7}\phi_2(\underline{x},t)\right)+\Sigma_t(\underline{x},t)\underline{J}_2(\underline{x},t) = 0. \tag{2.26d}$$

This is a coupled system of first-order PDEs, and in this form, the equations will be much easier to discretize, spatially and temporally. This will be discussed again later. We note that in eqs. (2.26) all quantities (cross sections, source term and fluxes) can be functions of $\underline{x}$ and t.

Remark 8. *It is straight forward to see that eqs. (2.26) can formally be derived directly from eqs. (2.11). If we expand the functions of x and t in eqs. (2.11) into functions of*

$\underline{x} = (x, y, z)$ *and* t *by*

$$\begin{aligned} \phi_0(x,t) &\quad\rightarrow\quad \phi_0(\underline{x},t) \quad \textit{scalar}, &(2.27\text{a})\\ \phi_1(x,t) &\quad\rightarrow\quad \underline{J}_0(\underline{x},t) \quad \textit{vector}, &(2.27\text{b})\\ \phi_2(x,t) &\quad\rightarrow\quad \phi_2(\underline{x},t) \quad \textit{scalar}, &(2.27\text{c})\\ \frac{2}{5}\phi_1(x,t) + \frac{3}{5}\phi_3(x,t) &\quad\rightarrow\quad \underline{J}_2(\underline{x},t) \quad \textit{vector}. &(2.27\text{d}) \end{aligned}$$

Additionally, if we replace

$$\begin{aligned} \frac{\partial}{\partial x} \text{ in (2.11a) and (2.11c)} &\quad\rightarrow\quad \underline{\nabla}\,\cdot\,, &(2.28\text{a})\\ \frac{\partial}{\partial x} \text{ in (2.11b) and (2.11d)} &\quad\rightarrow\quad \underline{\nabla}\,, &(2.28\text{b}) \end{aligned}$$

then eqs. (2.11) directly become eqs. (2.26).

It is interesting to note that the steady-state SP_N equations are problematic in systems containing void regions (in which $\Sigma_t(\underline{x},t) = 0$), but eqs. (2.26) do not seem to have an issue with voids.

Remark 9. *Adding the partial time derivative to each of the steady-state SP_3 equations in first-order form formally implies [120]:*

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot\underline{\phi}_1(\underline{x},t) + \Sigma_a(\underline{x},t)\phi_0(\underline{x},t) = Q(\underline{x},t), \tag{2.29a}$$

$$\frac{1}{v}\frac{\partial \phi_2}{\partial t}(\underline{x},t) + \frac{1}{5}\underline{\nabla}\cdot\Big(2\underline{\phi}_1(\underline{x},t) + 3\underline{\phi}_3(\underline{x},t)\Big) + \Sigma_t(\underline{x},t)\phi_2(\underline{x},t) = 0, \tag{2.29b}$$

and

$$\frac{1}{v}\frac{\partial \underline{\phi}_1}{\partial t}(\underline{x},t) + \frac{1}{3}\underline{\nabla}(\phi_0(\underline{x},t) + 2\phi_2(\underline{x},t)) + \Sigma_t(\underline{x},t)\underline{\phi}_0(\underline{x},t) = 0, \tag{2.29c}$$

$$\frac{1}{v}\frac{\partial \underline{\phi}_3}{\partial t}(\underline{x},t) + \underline{\nabla}\left(\frac{3}{7}\phi_2(\underline{x},t)\right) + \Sigma_t(\underline{x},t)\underline{\phi}_3(\underline{x},t) = 0, \tag{2.29d}$$

where the three-dimensional vectors $\underline{\phi}_1(\underline{x},t)$ and $\underline{\phi}_3(\underline{x},t)$ are obtained by formally replacing the scalars $\phi_1(x,t)$ and $\phi_3(x,t)$ in eqs. (2.11) by vectors.

Although eqs. (2.29) differ from eqs. (2.26) it is, however, possible to transform them into each other by a similarity transformation acting on one of the PDE systems. Suppose that we rewrite eqs. (2.26) to

$$\begin{aligned} \frac{1}{v}\frac{\partial \vec{u}}{\partial t}(\underline{x},t) + \frac{\partial}{\partial x}M_x\vec{u}(\underline{x},t) + \frac{\partial}{\partial y}M_y\vec{u}(\underline{x},t) + \frac{\partial}{\partial z}M_z\vec{u}(\underline{x},t) + \Sigma_t(\underline{x},t)\vec{u}(\underline{x},t) \qquad (2.30)\\ = \Sigma_s(\underline{x},t)\phi_0(\underline{x},t)\begin{bmatrix}1\\ \vdots\\ 0\end{bmatrix} + Q(\underline{x},t)\begin{bmatrix}1\\ \vdots\\ 0\end{bmatrix} \end{aligned}$$

where $M_x, M_y, M_z \in \mathbb{R}^{8\times 8}$ *and* $\vec{u} = [\phi_0, \phi_2, \underline{J}_0, \underline{J}_2]^T$.

Similarly, eqs. (2.29) can be written as

$$\frac{1}{v}\frac{\partial \hat{\vec{u}}}{\partial t}(\underline{x},t) + \frac{\partial}{\partial x}\hat{M}_x\hat{\vec{u}}(\underline{x},t) + \frac{\partial}{\partial y}\hat{M}_y\hat{\vec{u}}(\underline{x},t) + \frac{\partial}{\partial z}\hat{M}_z\hat{\vec{u}}(\underline{x},t) + \Sigma_t(\underline{x},t)\hat{\vec{u}}(\underline{x},t) \tag{2.31}$$
$$= \Sigma_s(\underline{x},t)\phi_0(\underline{x},t)\begin{bmatrix}1\\ \vdots\\ 0\end{bmatrix} + Q(\underline{x},t)\begin{bmatrix}1\\ \vdots\\ 0\end{bmatrix}$$

where $\hat{M}_x, \hat{M}_y, \hat{M}_z \in \mathbb{R}^{8\times 8}$ *and* $\hat{\vec{u}} = [\phi_0, \phi_2, \underline{\phi}_1, \underline{\phi}_3]^T$.

Introducing the transformation matrix

$$T = \begin{bmatrix} & & & 0 & \dots & 0 \\ & \mathbb{I}_5 & & \vdots & & \vdots \\ & & & 0 & \dots & 0 \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & \frac{2}{5}\mathbb{I}_3 & \frac{3}{5}\mathbb{I}_3 & \\ 0 & 0 & 0 & & & \end{bmatrix}, \tag{2.32}$$

where $\mathbb{I}_n$ *denotes an* $n \times n$ *identity matrix, implies that* $\vec{u} = T\,\hat{\vec{u}}$. *If this is plugged into eqs. (2.30) and additionally, eqs. (2.30) is multiplied by* T^{-1} *from the left, we immediately end up with eqs. (2.31) where*

$$\hat{M}_x = T^{-1}M_xT, \quad \hat{M}_y = T^{-1}M_yT, \quad \hat{M}_z = T^{-1}M_zT.$$

Consequently, eqs. (2.26) and eqs. (2.29) are equivalent and their solutions are either identical or can easily be transformed.

Above remark confirms that numerical codes which already used time-dependent SP_3 equations, gained by simply adding the time-derivative, are equivalent to those developed above. Nevertheless, they are all derived in a purely heuristic way and we now want to provide a theoretical foundation for these equations which are already being solved numerically.

2.3 Formal Asymptotic Derivation

To keep our discussion simple, we restrict the asymptotic analysis to the monoenergetic, 3-D isotropically scattering particle transport in a *homogeneous* medium. However, a similar analysis might also be performed for anisotropic scattering which is already presented in [99] for the steady-state equations. As in Section 2.2, we start our analysis with the linear Boltzmann equation:

$$\frac{1}{v}\frac{\partial \Psi}{\partial t}(\underline{x},\underline{\Omega},t) + \underline{\Omega}\cdot\underline{\nabla}\Psi(\underline{x},\underline{\Omega},t) + \Sigma_t\Psi(\underline{x},\underline{\Omega},t) = \frac{1}{4\pi}(\Sigma_s\phi_0(\underline{x},t) + Q(\underline{x},t)), \tag{2.33a}$$

or, dividing by Σ_t

$$\mathcal{T}\Psi(\underline{x},\underline{\Omega},t)+\underline{\Omega}\cdot\underline{\mathcal{X}}\Psi(\underline{x},\underline{\Omega},t)+\Psi(\underline{x},\underline{\Omega},t)=\frac{1}{4\pi}\left(c\phi_0(\underline{x},t)+\frac{Q(\underline{x},t)}{\Sigma_t}\right), \quad (2.33b)$$

where

$$\phi_0(\underline{x},t)=\int_{S^2}\Psi(\underline{x},\underline{\Omega},t)d\Omega\ \equiv\ \text{scalar flux}, \quad (2.34a)$$

$$\mathcal{T}=\frac{1}{v\Sigma_t}\frac{\partial}{\partial t}, \quad (2.34b)$$

$$\underline{\mathcal{X}}=\frac{1}{\Sigma_t}\underline{\nabla}, \quad (2.34c)$$

$$c=\frac{\Sigma_s}{\Sigma_t}\ \equiv\ \text{scattering ratio}. \quad (2.34d)$$

If we define the angular projection operator

$$(\mathcal{P}\psi)(\underline{x},t):=\frac{1}{4\pi}\int_{S^2}\Psi(\underline{x},\underline{\Omega},t)d\Omega, \quad (2.35a)$$

then

$$(\mathcal{P}\psi)(\underline{x},t)=\frac{1}{4\pi}\phi_0(\underline{x},t)\quad\text{and}\quad(\mathcal{P}\underline{\Omega}\psi)(\underline{x},t)=\frac{1}{4\pi}\underline{J}(\underline{x},t), \quad (2.35b)$$

where

$$\underline{J}(\underline{x},t)=\int_{S^2}\underline{\Omega}\Psi(\underline{x},\underline{\Omega},t)d\Omega\ \equiv\ \text{current}. \quad (2.35c)$$

Operating on (2.33b) by $4\pi\mathcal{P}$, we obtain

$$\mathcal{T}\phi_0(\underline{x},t)+\underline{\mathcal{X}}\cdot\underline{J}(\underline{x},t)+\phi_0(\underline{x},t)=\left(c\phi_0(\underline{x},t)+\frac{Q(\underline{x},t)}{\Sigma_t}\right). \quad (2.36)$$

We also operate on (2.33b) by $(I-\mathcal{P})$ to get the additional equation

$$(I-\mathcal{P})\mathcal{T}\Psi(\underline{x},\underline{\Omega},t)+(I-\mathcal{P})\underline{\Omega}\cdot\underline{\mathcal{X}}\Psi(\underline{x},\underline{\Omega},t)+(I-\mathcal{P})\Psi(\underline{x},\underline{\Omega},t)=0, \quad (2.37)$$

or

$$\begin{aligned}(I+\mathcal{T})\Psi(\underline{x},\underline{\Omega},t)+(I-\mathcal{P})\underline{\Omega}\cdot\underline{\mathcal{X}}\Psi(\underline{x},\underline{\Omega},t)&=\mathcal{P}(I+\mathcal{T})\Psi(\underline{x},\underline{\Omega},t) && (2.38a)\\ &=(I+\mathcal{T})\mathcal{P}\Psi(\underline{x},\underline{\Omega},t) && (2.38b)\\ &=\frac{1}{4\pi}(I+\mathcal{T})\phi_0(\underline{x},t). && (2.38c)\end{aligned}$$

This implies

$$\Psi(\underline{x},\underline{\Omega},t)+\left[(I+\mathcal{T})^{-1}(I-\mathcal{P})\underline{\Omega}\cdot\underline{\mathcal{X}}\right]\Psi(\underline{x},\underline{\Omega},t)=\frac{1}{4\pi}\phi_0(\underline{x},t). \quad (2.39)$$

If we define the operator $\mathcal{M}$ by

$$\mathcal{M} := (I + \mathcal{T})^{-1}(I - \mathcal{P}), \tag{2.40}$$

then (2.39) can be written as

$$\Psi(\underline{x}, \underline{\Omega}, t) + \mathcal{M}\,\underline{\Omega} \cdot \underline{\mathcal{X}}\Psi(\underline{x}, \underline{\Omega}, t) = \frac{1}{4\pi}\phi_0(\underline{x}, t), \tag{2.41}$$

or

$$\Psi(\underline{x}, \underline{\Omega}, t) = \frac{1}{4\pi}(I + \mathcal{M}\,\underline{\Omega} \cdot \underline{\mathcal{X}})^{-1}\phi_0(\underline{x}, t). \tag{2.42}$$

Introducing (2.42) into the definition of the current in (2.35c):

$$\begin{aligned} \underline{J}(\underline{x}, t) &= \int_{S^2} \underline{\Omega}\Psi(\underline{x}, \underline{\Omega}, t)d\Omega && (2.43)\\ &= \frac{1}{4\pi}\int_{S^2} \underline{\Omega}(I + \mathcal{M}\,\underline{\Omega} \cdot \underline{\mathcal{X}})^{-1}\phi_0(\underline{x}, t)d\Omega, && (2.44) \end{aligned}$$

and then introducing this result into (2.36), we end up with:

$$\begin{aligned} \mathcal{T}\phi_0(\underline{x}, t) + \frac{1}{4\pi}\int_{S^2} \underline{\mathcal{X}} \cdot \underline{\Omega}(I + \mathcal{M}\,\underline{\Omega} \cdot \underline{\mathcal{X}})^{-1}\phi_0(\underline{x}, t)d\Omega \qquad\qquad && (2.45)\\ + \phi_0(\underline{x}, t) = \left(c\phi_0(\underline{x}, t) + \frac{Q(\underline{x}, t)}{\Sigma_t}\right), \end{aligned}$$

or multiplying by Σ_t,

$$\frac{1}{v}\frac{\partial\phi_0}{\partial t}(\underline{x}, t) + \frac{1}{4\pi}\int_{S^2} \underline{\Omega} \cdot \underline{\nabla}(I + \mathcal{M}\,\underline{\Omega} \cdot \underline{\mathcal{X}})^{-1}\phi_0(\underline{x}, t)d\Omega + \Sigma_a\phi_0(\underline{x}, t) = Q(\underline{x}, t). \tag{2.46}$$

This equation for $\phi_0(\underline{x}, t)$ is formally exact but the integral term is a complicated operator which is not expressed in a useful form. We introduce a small, positive, dimensionless parameter ε into (2.46) such that the operator $\frac{1}{\Sigma_t}\underline{\nabla}$ becomes small, i.e.,

$$\underline{\mathcal{X}} = \varepsilon\frac{1}{\Sigma_t}\underline{\nabla}. \tag{2.47}$$

Remark 10. *It should be emphasized that the only assumption for the following asymptotic analysis is that* $\underline{\mathcal{X}} = \mathcal{O}(\varepsilon)$*. Neither the time derivative is supposed to be small nor source terms are scaled. This is purely formal and is chosen to keep the framework as general as possible. To draw a line from the scaling considered in this section to scalings from previous asymptotic* SP_N *derivations in literature, we list some of them which are acceptable for our asymptotics:*

- ***conventional scaling*** *[73, 98, 101, 102, 148]: Here, the system is assumed to be scattering-dominated with its collision rate being much larger than its absorption rate. In this case,* ε *is the ratio of the mean free path (which corresponds to* Σ_t^{-1}*) over a typical length scale for the solution. This scaling is the standard scaling which has been used to gain the diffusion or steady-state* SP_N *equations by performing an asymptotic analysis for* $\varepsilon \approx 0$.

- ***generalized conventional scaling*** *[99]: Larsen introduces an alternate scaling which is physically consistent with the conventional scaling and additionally, includes free parameters. Depending on the choice of these parameters either the standard or modified diffusion and SP_N equations are obtained. From the theoretical point of view, the latter equations are proven to increase the accuracy for deep penetration problems.*

- ***parabolic scaling*** *[56]: An asymptotic approach for the time-dependent SP_3 equations is presented here where the time derivative, source term and absorption cross section are scaled by ε^2 and the space derivative by ε. This scaling can also be achieved by plugging*

$$
\begin{aligned}
v &= \frac{\tilde{v}}{\varepsilon} && \text{(2.48a)}\\
\Sigma_t &= \frac{\sigma_t}{\varepsilon}, && \text{(2.48b)}\\
\Sigma_a &= \varepsilon \sigma_a, && \text{(2.48c)}\\
Q(\underline{x}, t) &= \varepsilon q(\underline{x}, t), && \text{(2.48d)}
\end{aligned}
$$

where $\tilde{v}, \sigma_t, \sigma_a, q$ are of $\mathcal{O}(1)$, into (2.1) and dividing by Σ_t. In addition to the physical assertions from the conventional scaling, it also requires that particles travel at high velocities. Combined with a high collision rate, low absorption rate and small source terms, the scaling as a whole implies a slowly varying solution in space and an even smaller variation in time.

Note that all scalings above implicate additional powers of ε in front of the time derivative which introduces only minor changes in the following asymptotic analysis.

Next, we will asymptotically expand the operator

$$
\mathcal{L} = \frac{1}{4\pi} \int_{S^2} \underline{\Omega} \cdot \underline{\nabla} (I + \mathcal{M}\, \underline{\Omega} \cdot \underline{\mathcal{X}})^{-1} d\Omega. \tag{2.49}
$$

Due to the assertion in (2.47) about the dimensionless spatial gradient $\underline{\mathcal{X}} = \varepsilon \frac{1}{\Sigma_t} \underline{\nabla}$ or any scaling which yields $\underline{\mathcal{X}} = \mathcal{O}(\varepsilon)$, we can expand the operator $\mathcal{L}$ in (2.49) in a Neumann series.

Thus, for ε sufficiently small,

$$
\mathcal{L} = \sum_{n=0}^{\infty} (-1)^n \mathcal{L}_n, \tag{2.50}
$$

where

$$
\mathcal{L}_n = \left[\frac{1}{4\pi} \int_{S^2} \underline{\Omega} \cdot \underline{\nabla} (\mathcal{M}\, \underline{\Omega} \cdot \underline{\mathcal{X}})^n \right] = \mathcal{O}(\varepsilon^n). \tag{2.51}
$$

To achieve an $\mathcal{O}(\varepsilon^7)$ approximation we need the first seven $\mathcal{L}_n$:

$$\mathcal{L}_0 = 0, \tag{2.52a}$$

$$\mathcal{L}_1 = \frac{1}{3}\underline{\nabla} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}} \tag{2.52b}$$

$$\mathcal{L}_2 = 0, \tag{2.52c}$$

$$\mathcal{L}_3 = \frac{4}{45}\left[\underline{\nabla} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right](I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right] \tag{2.52d}$$

$$\mathcal{L}_4 = 0, \tag{2.52e}$$

$$\begin{aligned}\mathcal{L}_5 = \frac{44}{945}\left[\underline{\nabla} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right](I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right] \\ \cdot (I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right]\end{aligned} \tag{2.52f}$$

$$\mathcal{L}_6 = 0, \tag{2.52g}$$

which are calculated in detail in the Appendix.

Remark 11. *It is important to emphasize one crucial assumption made in the derivation of (2.52d) and (2.52f): Both are only exact for either homogeneous media or a system in which Σ_t depends only on one spatial variable. However, the rest of eqs. (2.52) are also exact for heterogeneous media.*

2.3.1 SP_1 Equations

Ignoring terms of $\mathcal{O}(\varepsilon^3)$ in (2.50) we obtain from eqs. (2.52)

$$\mathcal{L} = -\frac{1}{3}\underline{\nabla} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}} + \mathcal{O}(\varepsilon^3).$$

Introducing this approximation for $\mathcal{L}$ in (2.49) and (2.46) we get

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(\underline{x},t) - \frac{1}{3}\underline{\nabla} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}}\phi_0(\underline{x},t) + \Sigma_a\phi_0(\underline{x},t) + \mathcal{O}(\varepsilon^3) = Q(\underline{x},t). \tag{2.53}$$

If we additionally define

$$\underline{J}_0(\underline{x},t) := -(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\phi_0(\underline{x},t)$$

and drop the error term, above equations simplify to the SP_1 equations:

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(\underline{x},t) + \underline{\nabla} \cdot \underline{J}_0(\underline{x},t) + \Sigma_a(\underline{x},t)\phi_0(\underline{x},t) = Q(\underline{x},t), \tag{2.54a}$$

$$\frac{1}{v}\frac{\partial \underline{J}_0}{\partial t}(\underline{x},t) + \frac{1}{3}\underline{\nabla}\phi_0(\underline{x},t) + \Sigma_t(\underline{x},t)\underline{J}_0 = 0. \tag{2.54b}$$

Remark 12. *Whereas the steady-state SP_1 approximation is the standard diffusion equation which requires only one scalar-valued function ϕ_0, eqs. (B.28) are a system of two equations with the scalar variable ϕ_0 and vector $\underline{J}_0$. In contrast to the result in [56], this time-dependent SP_1 approximation is not a parabolic equation, as one might expect by simply adding the time-derivative to the steady-state equation.*

2.3.2 SP_2 Equations

More accurate solutions can be gained systematically by simply taking higher order terms in (2.50) into account. An asymptotically higher order approximation of $\mathcal{O}(\varepsilon^5)$ is given by:

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(\underline{x},t) - \frac{\Sigma_t}{3}L\left[I + \frac{4}{15}(I+\mathcal{T})^{-1}L\right]\phi_0(\underline{x},t) + \Sigma_a\phi_0(\underline{x},t) + \mathcal{O}(\varepsilon^5) = Q(\underline{x},t) \quad (2.55)$$

where the operator L is defined as

$$L := \left[\underline{\mathcal{X}} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right] = \mathcal{O}(\varepsilon^2). \quad (2.56)$$

We approximate the operator in square brackets in (2.55) by

$$\left[I + \frac{4}{15}(I+\mathcal{T})^{-1}L\right] = \left[I - \frac{4}{15}(I+\mathcal{T})^{-1}L\right]^{-1} + \mathcal{O}(\varepsilon^4), \quad (2.57)$$

and set

$$2\phi_2(\underline{x},t) + \phi_0(\underline{x},t) := \left[I - \frac{4}{15}(I+\mathcal{T})^{-1}L\right]^{-1}\phi_0(\underline{x},t), \quad (2.58)$$

which can be rewritten to

$$\frac{4}{15}(I+\mathcal{T})^{-1}L(2\phi_2(\underline{x},t) + \phi_0(\underline{x},t)) = 2\phi_2(\underline{x},t). \quad (2.59)$$

Hence, by applying the operator $(I+\mathcal{T})$ on the left, it follows

$$\frac{1}{v}\frac{\partial}{\partial t}\phi_2(\underline{x},t) + \Sigma_t\phi_2(\underline{x},t) = \frac{2}{15}\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}(\phi_0(\underline{x},t) + 2\phi_2(\underline{x},t)). \quad (2.60)$$

Combining (2.55) and (2.60) as well as discarding the error term, gives the system

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot\underline{J}_0(\underline{x},t) + \Sigma_a(\underline{x},t)\phi_0(\underline{x},t) = Q(\underline{x},t), \quad (2.61a)$$

$$\frac{1}{v}\frac{\partial \phi_2}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot\underline{J}_2(\underline{x},t) + \Sigma_t(\underline{x},t)\phi_2(\underline{x},t) = 0, \quad (2.61b)$$

$$\frac{1}{v}\frac{\partial \underline{J}_0}{\partial t}(\underline{x},t) + \frac{1}{3}\underline{\nabla}\Big(\phi_0(\underline{x},t) + 2\phi_2(\underline{x},t)\Big) + \Sigma_t(\underline{x},t)\underline{J}_0(\underline{x},t) = 0, \quad (2.61c)$$

$$\frac{1}{v}\frac{\partial \underline{J}_2}{\partial t}(\underline{x},t) + \frac{2}{15}\underline{\nabla}\Big(\phi_0(\underline{x},t) + 2\phi_2(\underline{x},t)\Big) + \Sigma_t(\underline{x},t)\underline{J}_2(\underline{x},t) = 0, \quad (2.61d)$$

2.3.3 SP_3 Equations

Having collected all operators, we approximate $\mathcal{L}$ by truncating the series in (2.50) at $n = 6$ and introducing eqs. (2.52) into (2.50):

$$\begin{aligned}
\mathcal{L} =& -\frac{1}{3}\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}} && (2.62a)\\
&-\frac{4}{45}\left[\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right](I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right] && (2.62b)\\
&-\frac{44}{945}\left[\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right]\left((I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right]\right)^2+\mathcal{O}(\varepsilon^7) && (2.62c)\\
=& -\frac{1}{3}\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\left\{I+\frac{4}{15}(I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right]\right. && (2.62d)\\
&\left.+\frac{44}{315}\left((I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right]\right)^2\right\}+\mathcal{O}(\varepsilon^7).
\end{aligned}$$

We rewrite (2.62d) to

$$\begin{aligned}
\mathcal{L} =& -\frac{1}{3}\Sigma_t L\left\{I+\frac{4}{15}(I+\mathcal{T})^{-1}L+\frac{44}{315}\left((I+\mathcal{T})^{-1}L\right)^2\right\}+\mathcal{O}(\varepsilon^7) && (2.63a)\\
=& -\frac{1}{3}\Sigma_t L\left\{I+\left[I+\frac{11}{21}(I+\mathcal{T})^{-1}L\right]\frac{4}{15}(I+\mathcal{T})^{-1}L\right\}+\mathcal{O}(\varepsilon^7) && (2.63b)
\end{aligned}$$

Approximating the term in square brackets of (2.63b) like in (2.57), we conclude

$$\mathcal{L} = -\frac{1}{3}\Sigma_t L\left\{I+\left[I-\frac{11}{21}(I+\mathcal{T})^{-1}L\right]^{-1}\frac{4}{15}(I+\mathcal{T})^{-1}L\right\}+\mathcal{O}(\varepsilon^7) \tag{2.64}$$

Using this approximation for $\mathcal{L}$ in (2.49) and (2.46), we get:

$$\begin{aligned}
&\frac{1}{v}\frac{\partial\phi_0}{\partial t}(\underline{x},t)+\Sigma_a\phi_0(\underline{x},t) && (2.65)\\
&= Q(\underline{x},t)+\frac{\Sigma_t}{3}L\left\{\phi_0(\underline{x},t)+\left[I-\frac{11}{21}(I+\mathcal{T})^{-1}L\right]^{-1}\frac{4}{15}(I+\mathcal{T})^{-1}L\phi_0(\underline{x},t)\right\}+\mathcal{O}(\varepsilon^7).
\end{aligned}$$

If we define

$$2\phi_2(\underline{x},t) := \left[I-\frac{11}{21}(I+\mathcal{T})^{-1}L\right]^{-1}\frac{4}{15}(I+\mathcal{T})^{-1}L\phi_0(\underline{x},t), \tag{2.66}$$

then ϕ_2 satisfies

$$\left[I-\frac{11}{21}(I+\mathcal{T})^{-1}L\right]\phi_2(\underline{x},t) = \frac{2}{15}(I+\mathcal{T})^{-1}L\phi_0(\underline{x},t). \tag{2.67}$$

Operating with $(I+\mathcal{T})$ on the last equation yields

$$\left[I+\mathcal{T}-\frac{11}{21}L\right]\phi_2(\underline{x},t) = \frac{2}{15}L\phi_0(\underline{x},t), \tag{2.68}$$

or

$$
\begin{aligned}
(I+\mathcal{T})\,\phi_2(\underline{x},t) &= L\left(\frac{2}{15}\phi_0(\underline{x},t)+\frac{11}{21}\phi_2(\underline{x},t)\right) &(2.69)\\
&= \frac{1}{3}L\left(\frac{2}{5}\phi_0(\underline{x},t)+\frac{11}{7}\phi_2(\underline{x},t)\right). &(2.70)
\end{aligned}
$$

Equivalently,

$$\left(\frac{1}{v}\frac{\partial}{\partial t}+\Sigma_t\right)\phi_2(\underline{x},t)=\frac{1}{3}\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\underline{\nabla}\left(\frac{2}{5}\phi_0(\underline{x},t)+\frac{11}{7}\phi_2(\underline{x},t)\right). \quad (2.71)$$

Altogether, we get

$$
\begin{aligned}
&\frac{1}{v}\frac{\partial\phi_0}{\partial t}(\underline{x},t)+\Sigma_a\phi_0(\underline{x},t) &(2.72a)\\
&\qquad=\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\frac{1}{3\Sigma_t}\underline{\nabla}\Big(\phi_0(\underline{x},t)+2\phi_2(\underline{x},t)\Big)+Q(\underline{x},t),\\
&\frac{1}{v}\frac{\partial\phi_2}{\partial t}(\underline{x},t)+\Sigma_t\phi_2(\underline{x},t)=\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\frac{1}{3\Sigma_t}\underline{\nabla}\left(\frac{2}{5}\phi_0(\underline{x},t)+\frac{11}{7}\phi_2(\underline{x},t)\right). &(2.72b)
\end{aligned}
$$

In a homogeneous medium, eqs. (2.72) are identical to eqs. (2.22). This essentially proves the desired result and one can continue rewriting these equations to eqs. (2.26) in the same way as it is done in Section 2.2.

2.4 Numerical Results in 2D

We perform 2-D simulations for diffusion, the P_N, and the SP_N equations. The computations for the latter two approximations are done with a version of the code `StaRMAP` by Seibold and Frank [161]. The name `StaRMAP` stands for "`Sta`ggered grid `R`adiation `M`oment method `A`symptotic `P`reserving", which describes the key methodology and properties of the approach. More specifically, it is a second order accurate finite difference method for linear hyperbolic balance laws of the form

$$\partial_t\vec{u}+M_x\cdot\partial_x\vec{u}+M_y\cdot\partial_y\vec{u}+C\cdot\vec{u}=\vec{q}\,, \quad (2.73)$$

where the matrices M_x, M_y, and C possess specific patterns of their nonzero entries, as described below. The numerical method is implemented in a concise MATLAB code that the authors plan to make publicly available upon submission of the corresponding paper [161]. Let the components of the solution vector $\vec{u}$ be indexed by $\{1,2,\dots,S\}$. The requirement on the nonzero entry patterns of M_x, M_y, and C is that the components of $\vec{u}$ can be distributed into four disjoint sets, according to $\{1,2,\dots,S\}=I_{00}\,\dot\cup\, I_{10}\,\dot\cup\, I_{01}\,\dot\cup\, I_{11}$, such that the following properties hold:

$$
\begin{aligned}
(M_x)_{i,j}&=0\ \forall\,(i,j)\notin((I_{00}\times I_{10})\cup(I_{10}\times I_{00})\cup(I_{01}\times I_{11})\cup(I_{11}\times I_{01}))\,,\\
(M_y)_{i,j}&=0\ \forall\,(i,j)\notin((I_{00}\times I_{01})\cup(I_{01}\times I_{00})\cup(I_{10}\times I_{11})\cup(I_{11}\times I_{10}))\,, \quad (2.74)\\
C_{i,j}&=0\ \forall\,(i,j)\notin((I_{00}\times I_{00})\cup(I_{10}\times I_{10})\cup(I_{01}\times I_{01})\cup(I_{11}\times I_{11}))\,.
\end{aligned}
$$

With this distribution of the indices of the solution components, we consider the following four fully staggered sub-grids

$$G_{00} = \{(i\Delta x, j\Delta y) \,|\, i,j \in \mathbb{Z}\} \;, \quad G_{10} = \{((i+\tfrac{1}{2})\Delta x, j\Delta y) \,|\, i,j \in \mathbb{Z}\} \;,$$
$$G_{01} = \{(i\Delta x, (j+\tfrac{1}{2})\Delta y) \,|\, i,j \in \mathbb{Z}\} \;, \quad G_{11} = \{((i+\tfrac{1}{2})\Delta x, (j+\tfrac{1}{2})\Delta y) \,|\, i,j \in \mathbb{Z}\} \;,$$

and assign the components with indices in $I_{k\ell}$ to the corresponding sub-grid $G_{k\ell}$, where $k,\ell \in \{0,1\}$. On these fully staggered grids, any spatial derivative is approximated by a simple central difference stencil:

$$\partial_x w(i\Delta x, j\Delta y) \approx \tfrac{1}{\Delta x}\left(w((i+\tfrac{1}{2})\Delta x, j\Delta y) - w((i-\tfrac{1}{2})\Delta x, j\Delta y)\right) \;\forall\, i,j \in \tfrac{1}{2}\mathbb{Z} \;,$$
$$\partial_y w(i\Delta x, j\Delta y) \approx \tfrac{1}{\Delta y}\left(w(i\Delta x, (j+\tfrac{1}{2})\Delta y) - w(i\Delta x, (j-\tfrac{1}{2})\Delta y)\right) \;\forall\, i,j \in \tfrac{1}{2}\mathbb{Z} \;.$$

Hence, x-derivatives of components on $G_{k\ell}$ live on $G_{1-k,\ell}$, and y-derivatives of components on $G_{k\ell}$ live on $G_{k,1-\ell}$, where $k,\ell \in \{0,1\}$. The nonzero entry patterns (2.74) guarantee that the distribution of the indices of $\vec{u}$ into the sets I_{00}, I_{10}, I_{01}, and I_{11} is identical to the corresponding distribution of the indices of $M_x \cdot \partial_x \vec{u} + M_y \cdot \partial_y \vec{u} + C \cdot \vec{u}$. Both the classic P_N equations as well as the here derived SP_N equations, possess precisely the nonzero entry patterns (2.74), that admit a solution on fully staggered grids.

The time-derivative in (2.73) is resolved by bootstrapping. Having a time step Δt, we associate the components that live on $G_{00} \cup G_{11}$ with the times $T_0 = \{n\Delta t \,|\, n \in \mathbb{Z}\}$, and the components that live on $G_{10} \cup G_{01}$ with the times $T_1 = \{(n+\frac{1}{2})\Delta t \,|\, n \in \mathbb{Z}\}$. A full time step consists of two sub-steps: first, update information on the grid $G_{10} \cup G_{01}$ from time $(n-\frac{1}{2})\Delta t$ to $(n+\frac{1}{2})\Delta t$, where information on $G_{00} \cup G_{11}$ at the mid-time $n\Delta t$ is used; second, update information on the grid $G_{00} \cup G_{11}$ from time $n\Delta t$ to $(n+1)\Delta t$, where information on $G_{10} \cup G_{01}$ at the mid-time $(n+\frac{1}{2})\Delta t$ is used.

Specifically, the sub-step update rule is implemented as follows (here for $G_{00} \cup G_{11}$; the other sub-step works analogously). The terms $\vec{r} = \vec{q} - M_x \cdot \partial_x \vec{u} - M_y \cdot \partial_y \vec{u}$ that come from $G_{10} \cup G_{01}$ at the mid-step time are considered constant over the sub-step. Thus, equation (2.73) becomes the ODE

$$\partial_t \vec{u} + C \cdot \vec{u} = \vec{r} \tag{2.75}$$

with $\vec{r} = \text{const}$. In the special case that C is a diagonal matrix, $C = \operatorname{diag}(c_1, \dots, c_S)$, we can solve (2.75) from $n\Delta t$ to $(n+1)\Delta t$ explicitly:

$$u_k(x,(n+1)\Delta t) = \exp(-c_k(x)\Delta t)u_k(x, n\Delta t) - \tfrac{1}{c_k}(1 - \exp(-c_k(x)\Delta t))r_k \;. \tag{2.76}$$

Note that, if C is time-dependent, we evaluate it at the mid-step time, in which case (2.76) becomes a second-order accurate approximation to the true solution as demonstrated in Appendix B.3.

As P_N as well as SP_N solutions are spatially discontinuous at material interfaces for even N [175], we only present numerical calculations for *odd* N in the following. In some applications the diffusion equation is used to calculate approximations to the Boltzmann equation. Concerning computational effort, this approach is relatively cheap. Unfortunately, its accuracy is not satisfactory. For the sake of comparison, we solve the following time-dependent diffusion equation

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(x,y,t) = \underline{\nabla} \cdot [D(x,y,t)\underline{\nabla}\phi_0(x,y,t)] - \Sigma_a(x,y,t)\phi_0(x,y,t) + Q(x,y,t), \tag{2.77}$$

$$D(x,y,t) = \frac{1}{3\Sigma_t(x,y,t)}, \tag{2.78}$$

which can be interpreted as a low-order approximation to the Boltzmann equation (2.1) [56]. We apply the second order Finite Difference Crank-Nicolson scheme to discretize above equation.

2.4.1 Equivalence of SP_N and P_N equations

The detailed analysis of (simplified) spherical harmonics approximation brought along certain conditions under which the steady-state P_N and SP_N equations are equivalent. In [120] *McClarren* describes some of them and demonstrates this equivalence on a square of dimension $L = 5$ in a homogeneous medium with isotropic material coefficients and sources. We slightly modify the inhomogeneous source from [120] and change it to the following *time-dependent* sinusoidal term

$$Q(x,y,t) = \left(2 + \sin(4\pi t)e^{-t/3}\right) \cdot \begin{cases} 1, & (x,y) \in [1.75, 2.25] \times [1.75, 2.25], \\ 1, & (x,y) \in [2.75, 3.25] \times [1.5, 2.5], \\ 1, & (x,y) \in [1.75, 2.25] \times [2.75, 3.25], \\ 1, & (x,y) \in [3.5, 4.25] \times [3.5, 3.75], \\ 0, & \text{otherwise}. \end{cases} \tag{2.79}$$

All criteria for the equivalence were developed for the *steady-state* equations and there is no obvious reason why they should also be true for the time-dependent case. We implement this test case with $\Sigma_t = 1$ and $\Sigma_a = 0.9$ and investigate the temporal as well as steady-state behavior of various approximations. Periodic boundary conditions are enforced for all methods.

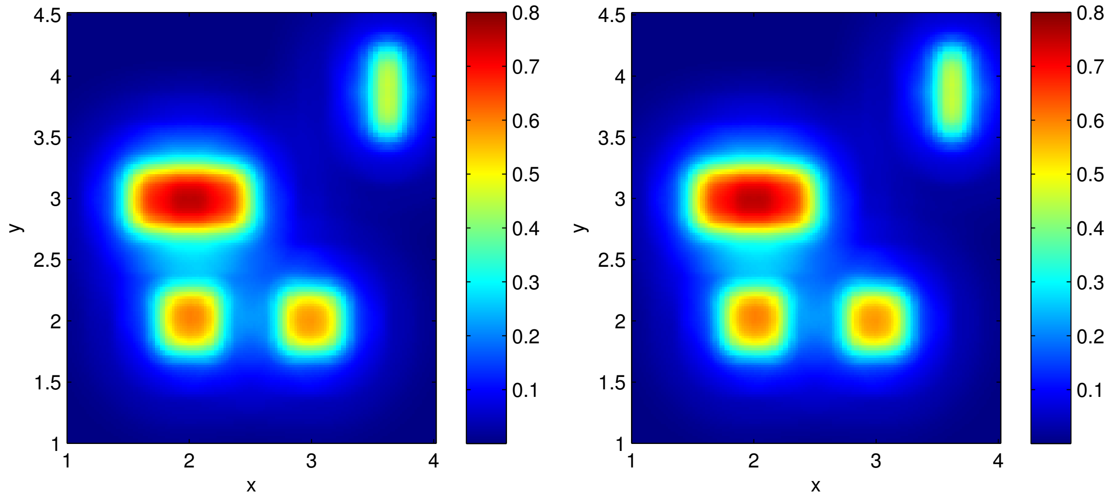
This problem has an absorption coefficient which is nine times larger than the scattering coefficient. Due to the inhomogeneous source, the solution has additionally large spatial gradients. Consequently, an agreement between P_N and SP_N solutions cannot be justified by any of the asymptotic scalings mentioned in Section 2.3.

Figure 2.1 displays the scalar flux ϕ_0 at $t = 1$ from several methods. Note that ϕ_0 is not in steady-state at this time. As the condition of weak spatial derivatives, which underlie the diffusion approximation, is not met in this situation the diffusion solution is inaccurate. SP_9 and P_9 solutions appear to be identical which is also confirmed in Figure 2.2. This figure shows the scalar flux along $x = 2$ for (simplified) spherical harmonics of order $N = 1, 3, 5, 9$. To the eye, there is again no difference between SP_N and P_N. Moreover, we observe a significant improvement from the SP_1 to SP_3 approximation. Although SP_5 still shows significant deviations, an order of $N = 9$ is sufficient for the simplified P_N approximation to be very close to the high-order P_{39} solution which is considered as reference here.

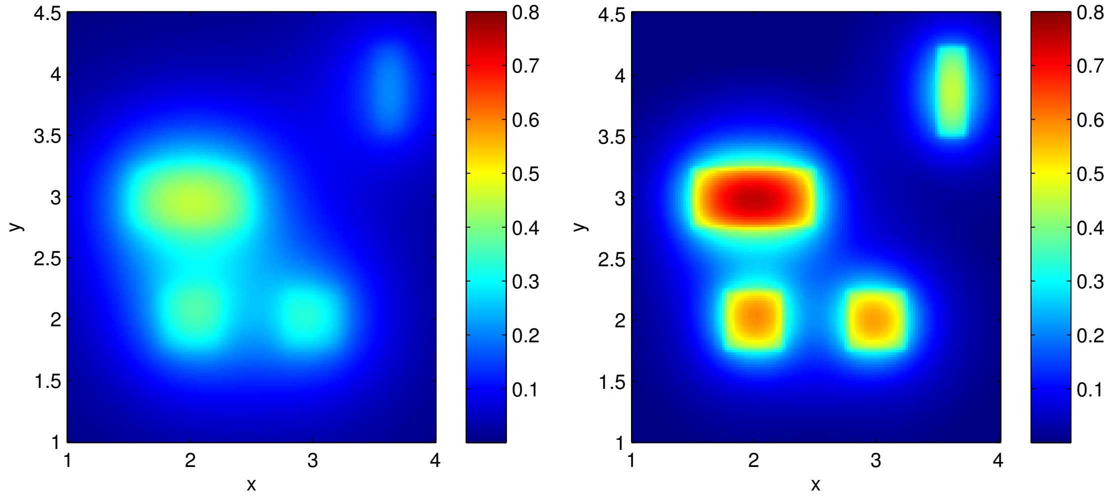
It turns out that all SP_N and P_N solutions are indeed equivalent. To verify this behavior more precisely we calculated the difference in the L^∞-norm

$$\max_{i,j} |\phi_0^{SP_N}(x_i, y_j) - \phi_0^{P_N}(x_i, y_j)|$$

at several times for $N = 1, 3, 5, 7, 11$ and observed that all differences throughout go down to machine precision. Our numerical experiments indicate that the analysis, performed for the steady-state equations, to prove the equivalence of P_N and SP_N equations in a general, homogeneous medium with isotropic cross sections and sources



(a) SP_9 (b) P_9



(c) Diffusion (d) P_{39}

Figure 2.1: Approximations to ϕ_0 at $t_{\text{final}} = 1$, 250 × 250 discretization points.

might be extended to the time-dependent equations including time-dependent sources. This is an issue of future work which goes beyond the purpose of this work.

Note that in this case of a 2-D problem the P_N equations have $\frac{(N+1)(N+2)}{2}$ unknowns whereas our SP_N method consists only of $3(\lfloor \frac{N}{2} \rfloor + 1)$ unknowns. Nevertheless, it is remarkable that P_N and SP_N solutions agree even for the time-dependent case. Hence, there exist situations where high-order SP_N approximations yield very accurate solutions (here for $N = 9$) and the diffusion approximation gives unsatisfactory results.

2.4.2 A Moving Rod

The prediction of the behavior of nuclear reactors is essential for the design and safe operation of nuclear power plants. Apart from many other nuclear and nonnuclear interactions, one challenge is to develop efficient and accurate techniques for the de-

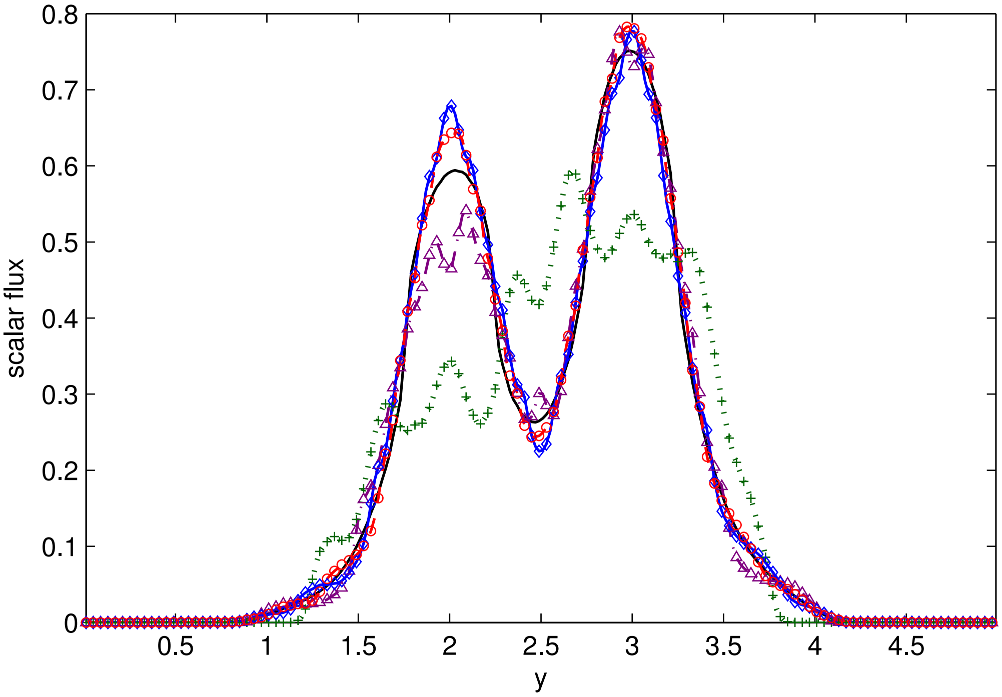


Figure 2.2: ϕ_0 at $t_{\text{final}} = 1$ along $x = 2$, 250 × 250 discretization points: P_1 (green dotted line), SP_1 (green crosses), P_3 (purple dash-dot line), SP_3 (purple triangles), P_5 (blue solid line), SP_5 (blue diamonds), P_9 (red dashed line), SP_9 (red circles), P_{39} (black solid line)

scription of neutron distributions.

We consider the linear transport Boltzmann equation (2.1) with slight modifications as before. Neglecting energy dependence and the coupling to precursors, approximative models to the following equation on $[-\frac{L}{2}, \frac{L}{2}] \times [-\frac{L}{2}, \frac{L}{2}]$ are solved:

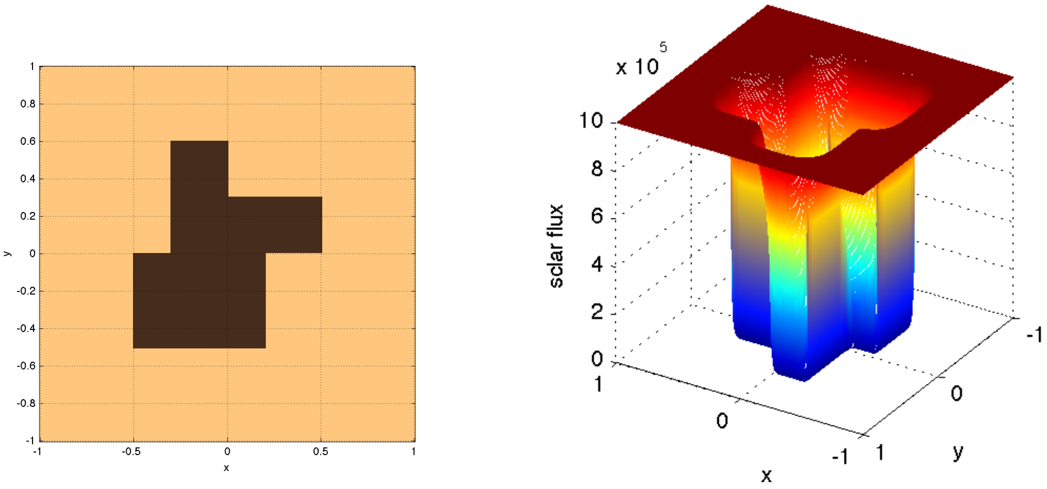
$$\begin{aligned}\frac{1}{v}\frac{\partial \Psi}{\partial t}(x,y,\underline{\Omega},t) + \underline{\Omega}\cdot\underline{\nabla}\Psi(x,y,\underline{\Omega},t) + [\Sigma_s + \Sigma_f + \Sigma_\gamma(x,y,t)]\Psi(x,y,\underline{\Omega},t)&\\ = \frac{\Sigma_s + \nu\Sigma_f}{4\pi}\int_{S^2}\Psi(x,y,\underline{\Omega}',t)d\Omega'.&\qquad(2.80)\end{aligned}$$

In addition to the scattering process (taken into account by Σ_s), above equation includes two relatively frequent interactions: Σ_f is the fission cross section describing the probability that a neutron will initiate a fission event when it collides with a nucleus; $\Sigma_\gamma(x,y,t)$ is the capture cross section characterizing a capture event in which the nucleus gains a neutron. Consequently, the absorption cross section is the sum of both:

$$\Sigma_a(x,y,t) = \Sigma_f + \Sigma_\gamma(x,y,t).$$

In a fission event, the target nucleus splits into two daughter nuclei and ν is usually the mean number of fission neutrons that are released. However, to keep the test case simpler, we use $0 \le \nu \le 1$ as a free parameter and set the capture cross section to

$$\Sigma_\gamma(x,y,t) = (\nu - 1)\Sigma_f + \begin{cases} s(t), & (x,y) \in \Omega_R, \\ 0, & \text{else}, \end{cases} \qquad (2.81)$$



(a) Rod Geometry

(b) P_{39} solution at $t = 0.2$: $s_{\max} = 100, \Delta T = 0.1$

Figure 2.3

where the domain Ω_R is defined as

$$\begin{aligned} \Omega_R = \{(x,y) \in \mathbb{R}^2 : -0.5 \le x \le 0.2,\ -0.5 \le y \le 0, \qquad (2.82)\\ -0.3 \le x \le 0,\ 0 < y \le 0.6, \qquad (2.83)\\ 0 < x \le 0.5,\ 0 < y < 0.3\}. \qquad (2.84) \end{aligned}$$

This definition of Σ_γ models an asymmetric rod with a cross section shown in Figure 2.3a which is moved into or out of the moderator material in a way specified by the function $s(t)$. We choose a sequence of three processes:

- pushing the rod into the material in the time of ΔT,
- keeping it in the moderator for the time of $T - 2\Delta T$ and
- pulling the rod out in the same time of ΔT.

and, hence, define

$$s(t) = \begin{cases} \frac{s_{\max}}{\Delta T}\, t, & 0 < t \le \Delta T, \\ s_{\max}, & \Delta T < t < T - \Delta T, \\ \frac{s_{\max}}{\Delta T}\,(T - t), & T - \Delta T \le t \le T. \end{cases} \qquad (2.85)$$

For large $s_{\max}/\Delta T$ the rod is moved very quickly and large time derivatives are generated. Due to the finite geometry of the rod, large gradients in space additionally occur. On the contrary, for small $s_{\max}/\Delta T$ the problem becomes gentle with weaker space- and time-derivatives.

The main goal of this problem is to analyze the behavior of diffusion, SP_N and P_N solutions in 2D for large time and space gradients in a semi-realistic setting where the absorption cross section depends on all three variables x, y and t.

The initial condition for the scalar flux is set to $\phi_0(x, y, 0) = 10^6$ and for all other variables to zero. Periodic boundary conditions are enforced and we use the following parameters:

$$v = 1, \qquad L = 2, \qquad T = 0.6,$$
$$\nu = 0.9, \qquad \Sigma_f = 2, \qquad \Sigma_s = 1.$$

Figures 2.4-2.5 display a cut of the scalar flux distribution along $y = 0$ for 251×251 discretization points. The P_{39} solution is considered as our reference which can be seen in Figure 2.3b. We distinguish between the two aforementioned cases: When the control rod is moved into the moderator, neutrons are absorbed and hence, their number diminishes. If this is done at a small speed (compared to the velocity of the particles) neutrons will quickly flood the region close to the absorber. Therefore, the scalar flux (which is a quantity for the particle number) is smooth but still forms steep slopes in our setting (Figure 2.4a). Although the solution drastically decreases both SP_3 and P_3 approximations are close to the reference P_{39} result. Moreover, P_3 is so close to SP_3 that they are hardly to distinguish. On the contrary, the diffusion solution is inaccurate in regions where large spatial gradients are formed. For increasing time, after the rod has been pulled out of the system, the neutrons spread in the whole domain and their distribution flattens more and more until, in steady-state, ϕ_0 levels off to a constant value which is smaller than at the beginning of the process. Figure 2.4b shows the distribution at an advanced time where P_3 and SP_3 are still accurate and the diffusion approximation is far off the reference solution.

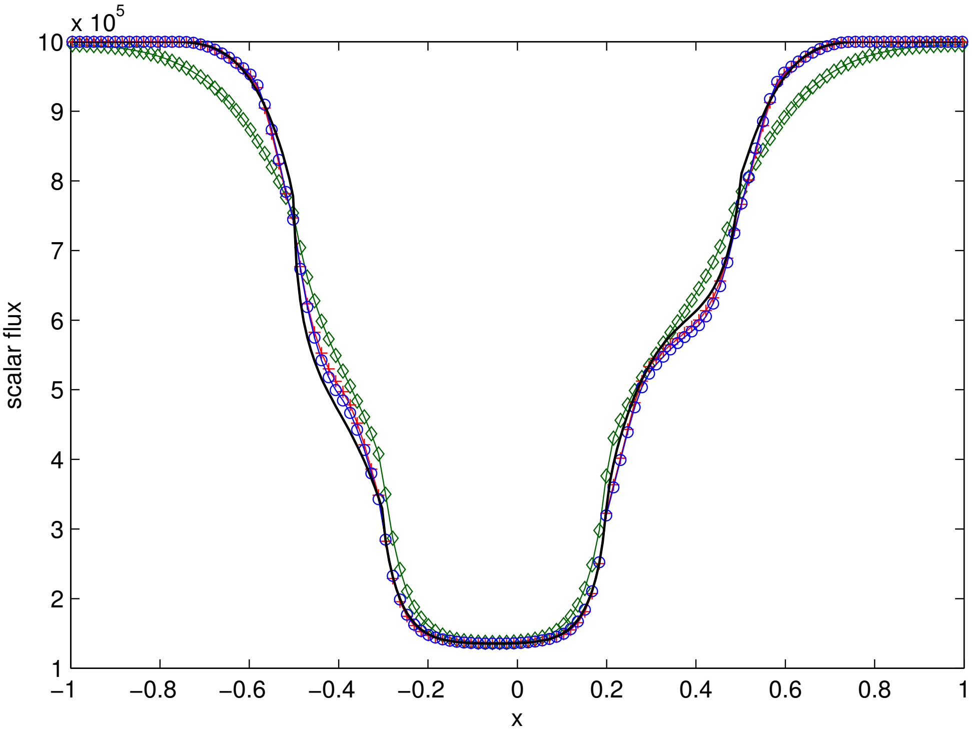
Pushing the rod almost instantly, results in an even more drastic decrease of the scalar flux which becomes almost a step function in some regions (Figure 2.5a). Here, the situation changes: Differences between P_3 and SP_3 become obvious although both solutions still capture larger slopes and more details of the P_{39} result than diffusion. Again, as time goes by the problem becomes more gentle, P_3 and SP_3 approach the reference whereas diffusion is still inaccurate(Figure 2.5b).

Our observations described above coincide with the analysis in Section 2.3. The asymptotics implies that the SP_3 approximation is a higher-order correction to diffusion in cases where steep spatial slopes occur. As long as these slopes are not extremely large, SP_3 and P_3 approximations yield similar results. However, when a certain threshold is reached the problem significantly gets out of the asymptotic limit. As a consequence, differences between SP_3 and P_3 become larger and SP_3 might lose accuracy.

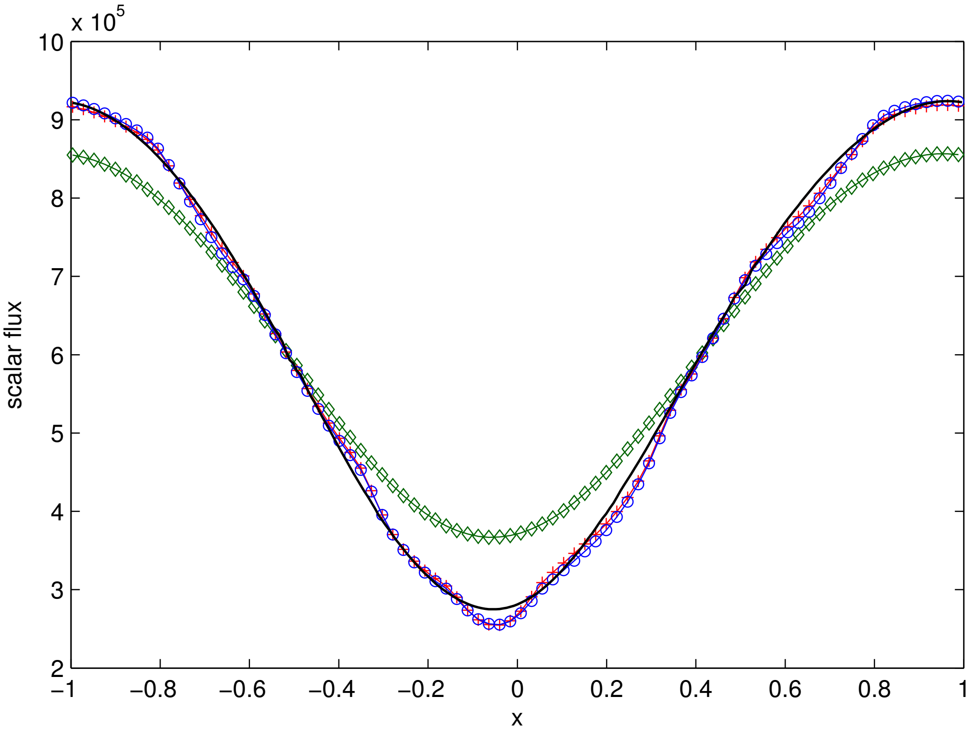
2.4.3 Checkerboard

We consider the checkerboard problem from [20]: It consists of a square $[0, 7] \times [0, 7]$ where the majority of the region is purely scattering. In the middle of the lattice system, there is an isotropic source $Q = 1$ continuously generating particles. Additionally, there are eleven small spots of either

- purely absorbing squares where $\Sigma_a = 10 = \Sigma_t$ or
- highly scattering squares where $\Sigma_a = 1$ and $\Sigma_s = 10$.

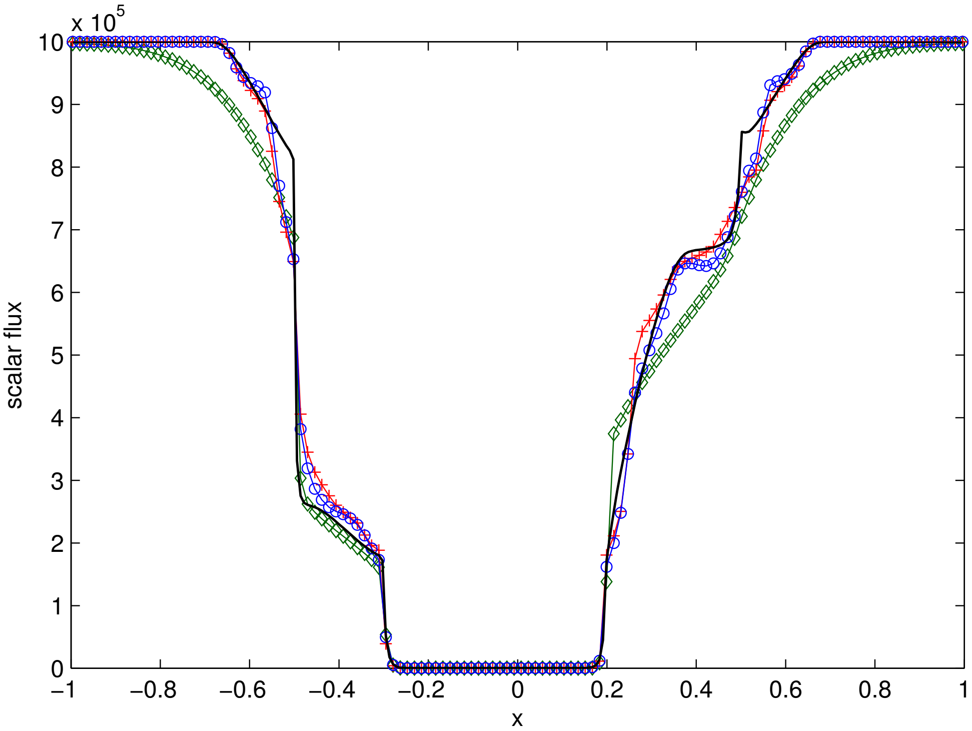


(a) $t = 0.3$

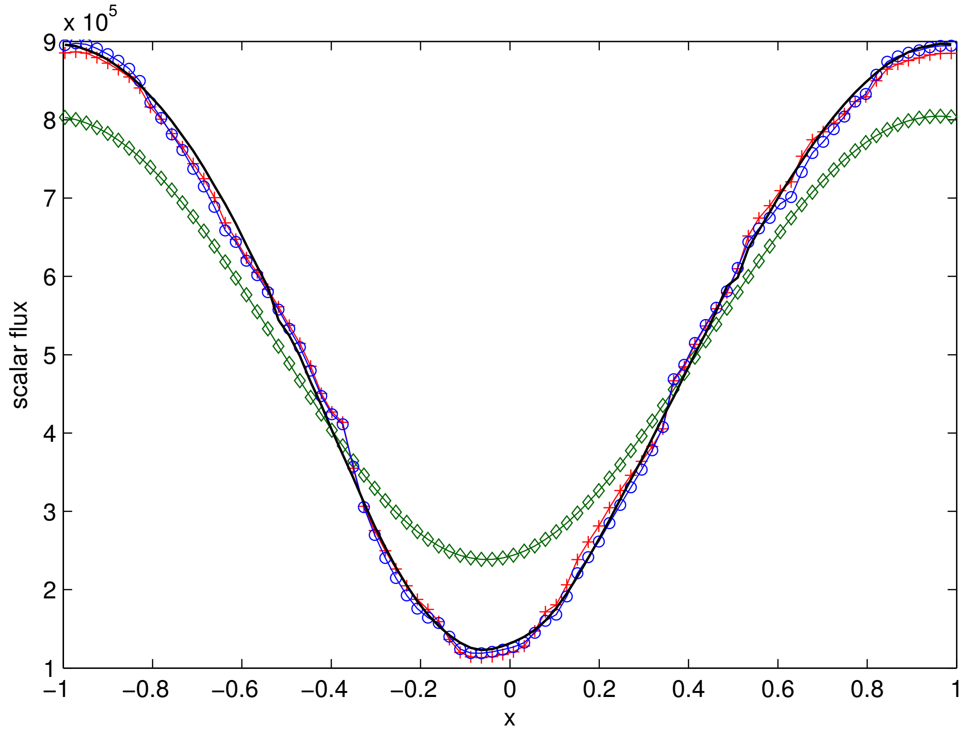


(b) $t = 1$

Figure 2.4: ϕ_0 along $y = 0$, $s_{\max} = 10, \Delta T = 0.2$: Diffusion (green diamond line), SP_3 (red cross line), P_3 (blue circle line), P_{39} (black solid line).

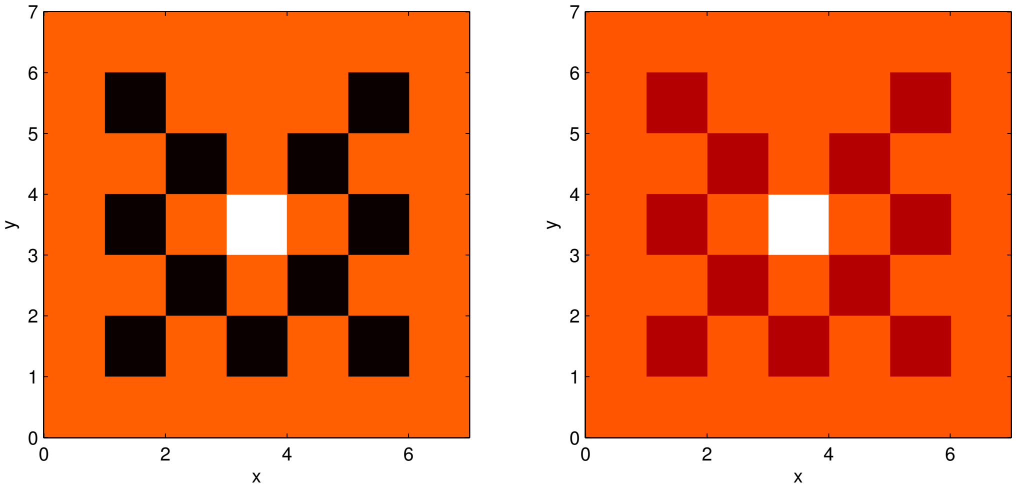


(a) $t = 0.2$



(b) $t = 1$

Figure 2.5: ϕ_0 along $y = 0$, $s_{\max} = 100, \Delta T = 0.1$: Diffusion (green diamond line), SP_3 (red cross line), P_3 (blue circle line), P_{39} (black solid line).



(a) Purely absorbing spots: purely scattering $\Sigma_s = 1 = \Sigma_t$, purely absorbing $\Sigma_a = 10$

(b) Highly scattering spots: purely scattering $\Sigma_s = 2 = \Sigma_t$, highly scattering $\Sigma_s = 10, \Sigma_a = 1$

Figure 2.6: Checkerboard test problem. Material coefficients: isotropic source (white), purely scattering (orange and white), highly scattering (red), purely absorbing (black).

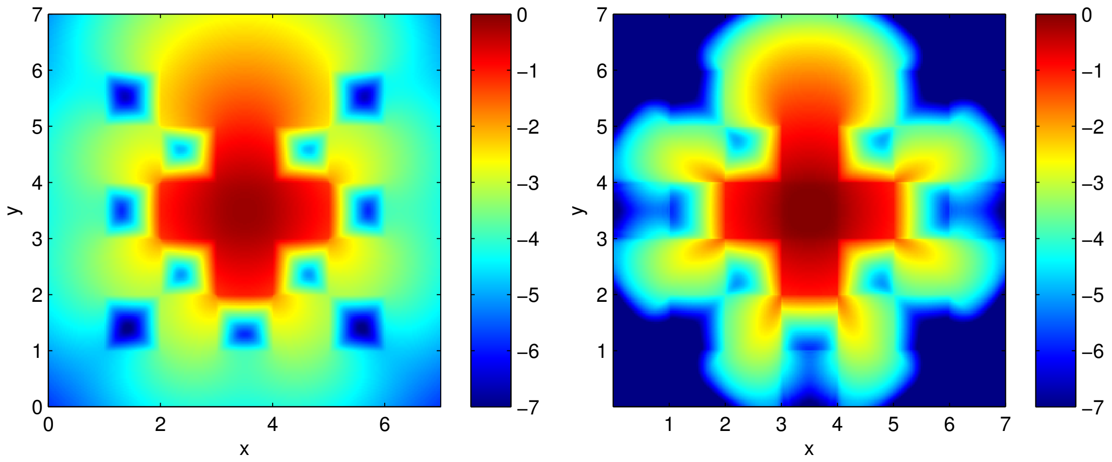
Figure 2.6 illustrates the problem settings more precisely. Vacuum boundary conditions are enforced and all initial quantities at $t = 0$ are zero. We compare the scalar flux using different methods (including high-order P_N and SP_N).

Classic Problem: Purely Absorbing Spots

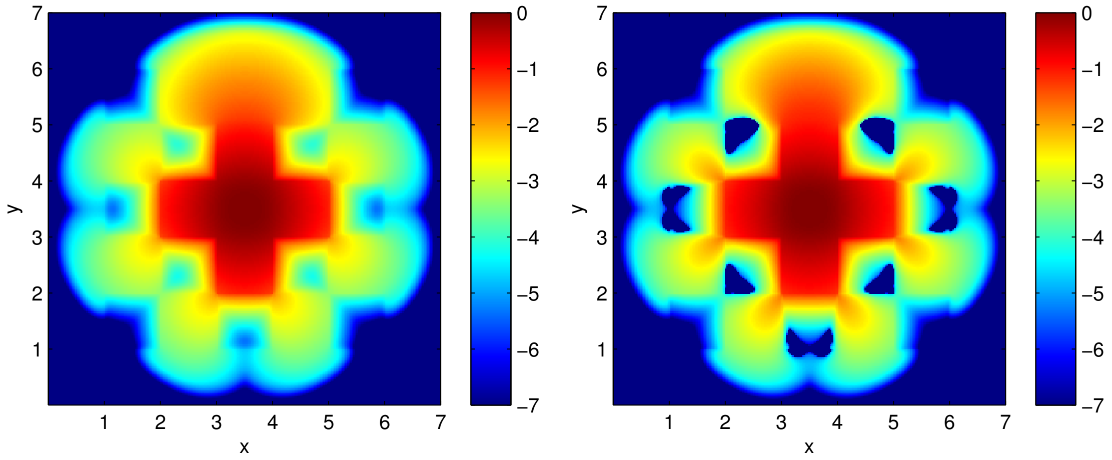
Here, the test case is chosen to be identical to the problem in [20] which we therefore refer to as classic problem. The diffusion solution in Figure 2.7 gives a poor result. Particles are transported from the central source at a much higher speed to the boundaries of the domain. The solution is much smoother and large slopes are washed out.

Although the SP_3 calculation shows large improvements upon the diffusion result, it is still too diffusive compared to the reference P_{39} (Figure 2.7). In some purely absorbing regions away from the center, the P_3 computation lacks particles because the approximation does not allow particle waves to travel at a high speed. Hence, depending on the desired purpose, P_3 is not evidently superior to SP_3 in this case. Nevertheless, the particle beams between the corners of the absorbing regions are well resolved in all P_N solutions.

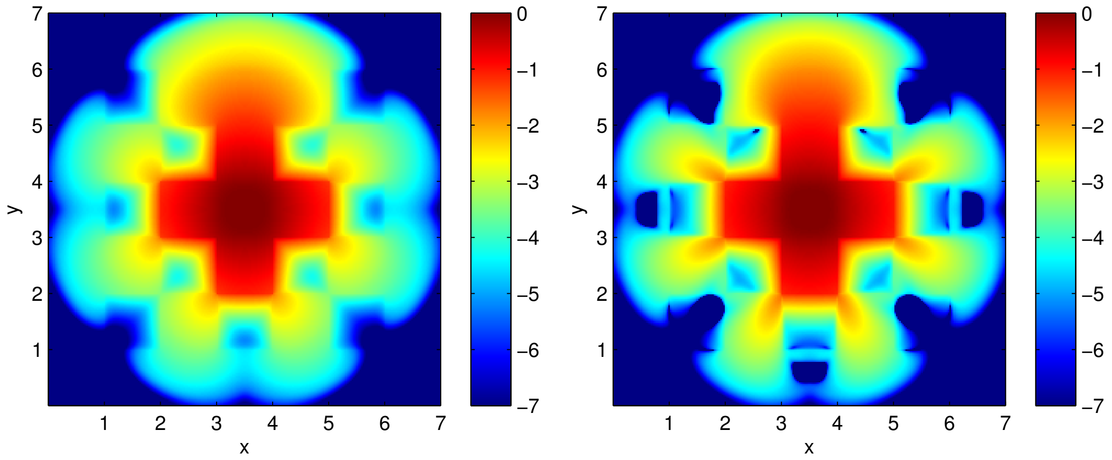
For increasing N, the spherical harmonics solutions show better improvements than SP_N. Although SP_5 calculation yields visible changes it still shows obvious differences from the reference. It is important to emphasize that SP_N solutions of this problem are not supposed to be highly accurate for several reasons: The underlying cross-sections are discontinuous and lead to a transport solution with steep slopes varying over seven orders of magnitude. Hence, our scaling parameter ε from Section 2.3 cannot be small. Moreover, purely absorbing regions have a scattering ratio of $c = \frac{\Sigma_s}{\Sigma_t} = 0$ which is significantly outside any asymptotic limit discussed in Section 2.3. And last, the total cross-section Σ_t depends on both spatial variables x and y which violates an assumption in our asymptotic derivation.



(a) Diffusion (b) P_{39}



(c) SP_3 (d) P_3



(e) SP_5 (f) P_5

Figure 2.7: Purely absorbing spots: Scalar flux ϕ_0 at $t = 3.2$ for from several approximations with 250×250 points. The values are plotted in $\log_{10}(\phi_0)$ and limited to seven orders of magnitude.

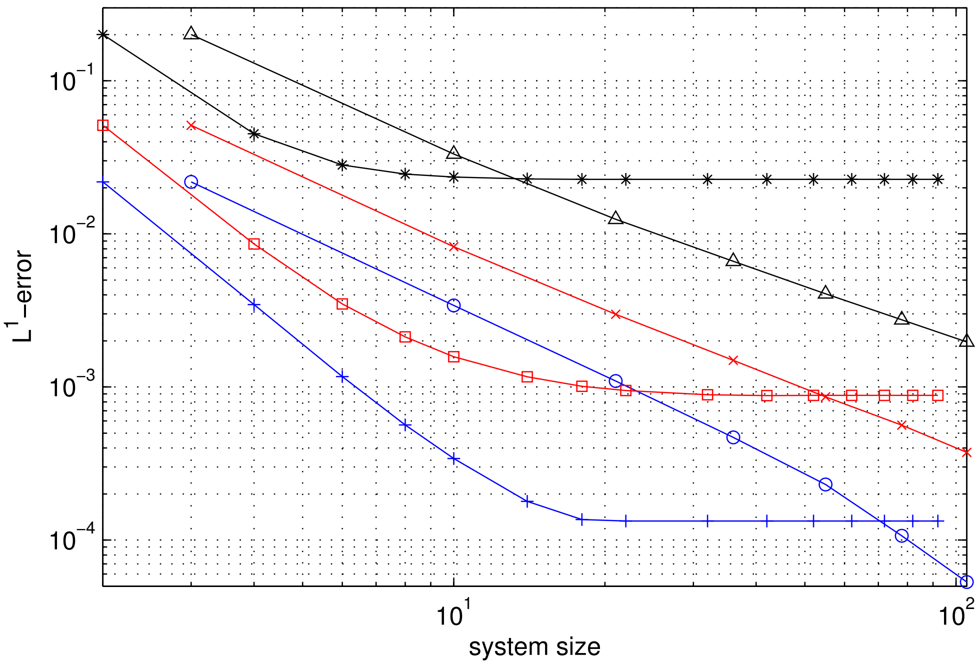


Figure 2.8: L^1-error between SP_N and P_{39} or P_N and P_{39} for 1000×1000 spatial discretization points: $\varepsilon = 0.5 : P_N$ (black solid triangle line), $\varepsilon = 0.5 : SP_N$ (black solid asterisk line), $\varepsilon = 0.1 : P_N$ (red solid cross line), $\varepsilon = 0.1 : SP_N$ (red solid square line), $\varepsilon = 0.04 : P_N$ (blue solid circle line), $\varepsilon = 0.04$ (blue solid plus line).

However, realistic applications often include this kind of geometries where cross-sections are (highly) varying and entail large spatial gradients in solutions. If SP_N approximations are computed for these problems it will be of big interest to know how much we gain for increasing order N. In other words, for growing N, how much does the error between SP_N and transport solution decrease? An answer to this question is necessary to determine the benefit of a high-order SP_N calculation in comparison to the additional computational effort.

To demonstrate the behavior of this error in a numerical way, SP_N approximations are calculated for odd N, from $N = 1$ to $N = 91$. In order to make the numerical discretization error comparabely small we choose a very fine discretization grid of 1000×1000 points. Additionally, a scaling parameter ε is introduced into all equations to achieve the scaling from (2.47). Approximative solutions are then computed for the scaled transport equation (2.36) where the velocity v is set to one. Again, P_{39} is assumed to be the reference solution.

Figure 2.8 shows the L^1-error between SP_N and P_{39} or P_N and P_{39} for different values of $\varepsilon = 0.04, 0.1, 0.5$. The error is plotted against the system size (i.e., the number of equations to be solved) which is $N + 1$ for the SP_N and $(N + 1)(N + 2)/2$ for the P_N approximation. This choice is made to give a reasonable comparison based on computational effort between P_N and simplified P_N calculations.

Indeed, both the P_N- and SP_N-error become smaller for decreasing values of ε. Moreover, for a fixed ε and increasing order N, the error of the P_N and SP_N method decreases monotonically. The P_N-error is even strictly decreasing to zero. In contrast, due to the discontinuous property of the cross sections or a scattering ratio which is outside the asymptotic limit, the SP_N-error approaches a constant non-zero value and saturates at system sizes of roughly $10 - 20$.

For a fixed system size (i.e. the same number of equations to be solved), Figure 2.8 shows SP_N-errors which are smaller than the corresponding P_N-errors. Consequently, to achieve a desired L^1-error above a minimum threshold the SP_N approximation is computationally less costly than P_N calculations. However, this is only valid for small system sizes and errors which are above the saturation limit of the SP_N method.

The asymptotic analysis in Section 2.3 is strictly valid for homogeneous media. Nevertheless, the SP_N error function for a small $\varepsilon = 0.04$ in Figure 2.8 demonstrates that there is a significant error decrease for high order SP_N solutions. Hence, this behavior shows numerically that SP_N computations of large orders $N \geq 5$ can also be used in settings with discontinuous cross sections to achieve a specified error.

Highly Scattering Spots

As above example does not only include discontinuous cross sections but also highly absorbing regions outside the asymptotic limit it can be considered as a torture test case for the SP_N method. Keeping the same heterogeneous lattice, we now enforce scattering ratios inside the asymptotic limit (Figure 2.6b).

Figure 2.9 displays the results at $t = 3.2$: In the middle of the domain with purely scattering regions all solutions are in a good agreement. However, going off the center, diffusion again shows significant differences to the transport solution. SP_N and P_N calculations are throughout close to each other, although the discrepancies become slightly larger at material interfaces. In comparison to the SP_3 solution, SP_5 transports particles to the boundary at a higher velocity which agrees well with the reference solution.

2.4.4 Line Source

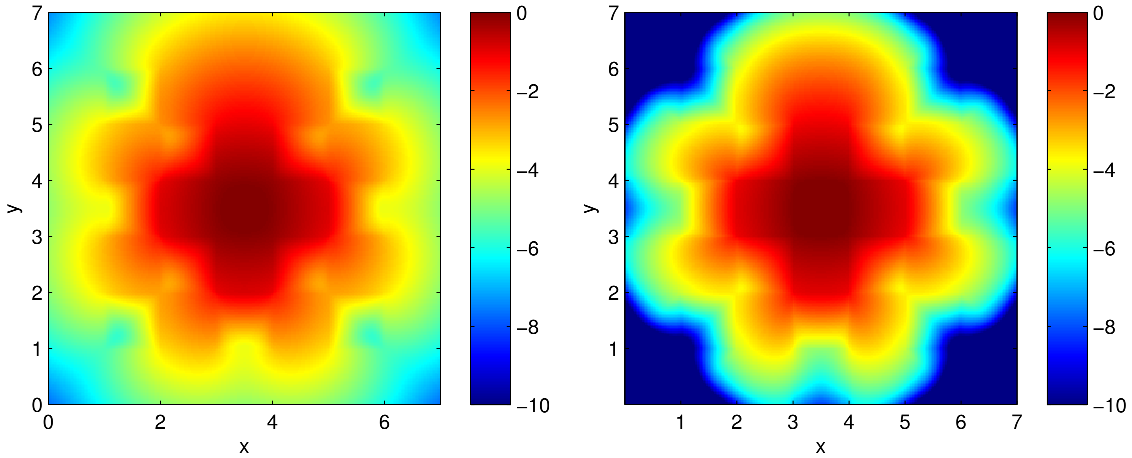
The last example deals with a modification of the standard linesource problem already discussed in several publications [20,75]. The standard version simulates particles being emitted by a pulsed line. This line source is represented by an initial condition which is a Gaussian in the middle of the domain $[0,1] \times [0,1]$:

$$\phi_0(x,y,0) = 10\, e^{-\left(\frac{x-0.5}{0.04}\right)^2 - \left(\frac{x-0.5}{0.04}\right)^2}. \tag{2.86}$$

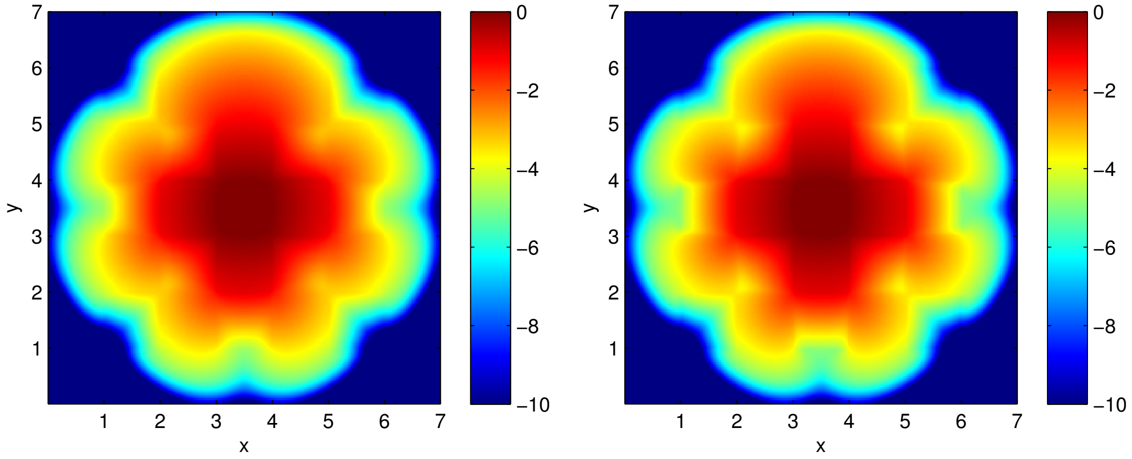
This example is considered a torture test case for the P_N equations because it yields negative particle densities. To overcome this unphysical behavior modifications to the standard P_N closure were formulated which give nonnegative results [75]. Usually, particles travel in either vacuum or a weakly scattering homogeneous medium. However, this implies that P_N and SP_N equations are equivalent as already noticed in Section 2.4.1 above. Indeed, a similar numerical investigation confirms this phenomenon.

To distinguish between both methods we add two orthogonal, purely scattering stripes of width 0.1 to the medium, i.e.,

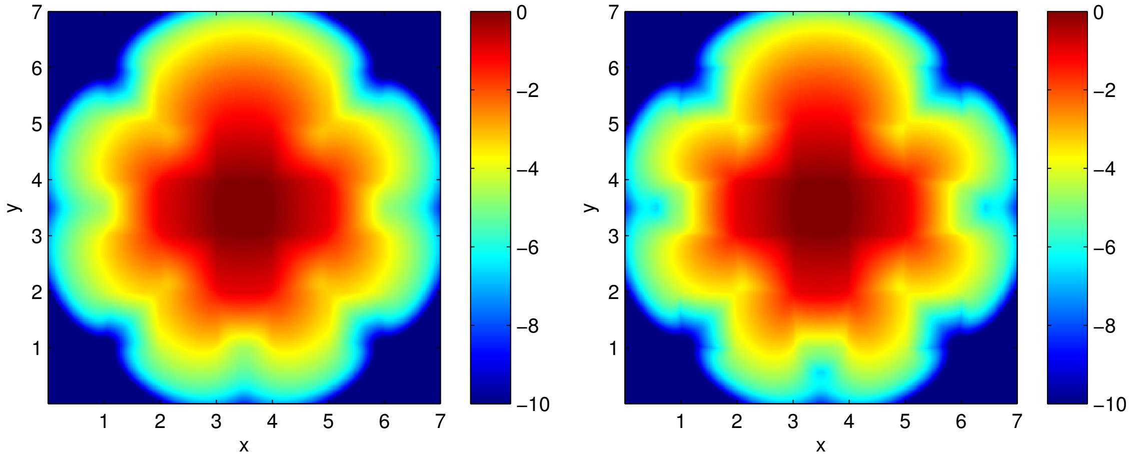
$$\Sigma_s(x,y) = \begin{cases} 10, & (x,y) \in [0.45, 0.55] \times [0,1], \\ 10 & (x,y) \in [0,1] \times [0.45, 0.55], \\ 0, & \text{else}. \end{cases} \tag{2.87}$$



(a) Diffusion (b) P_{39}

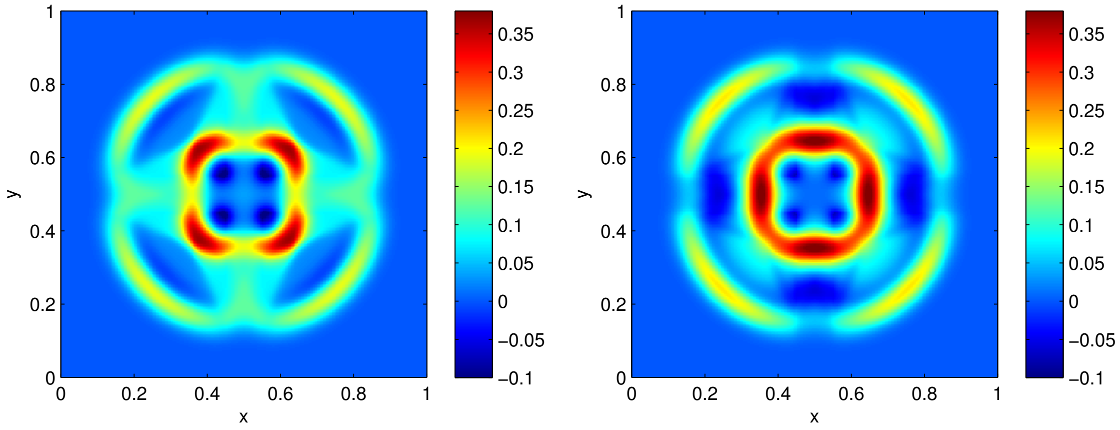


(c) SP_3 (d) P_3

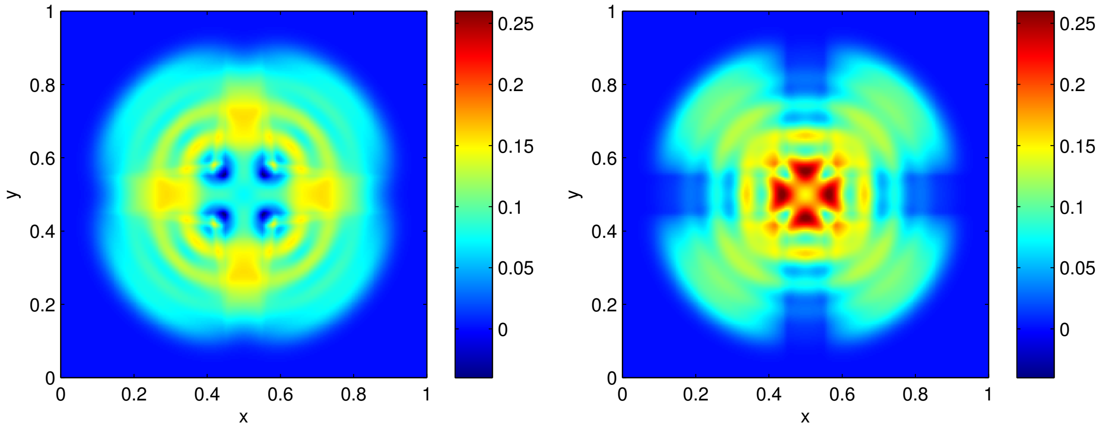


(e) SP_5 (f) P_5

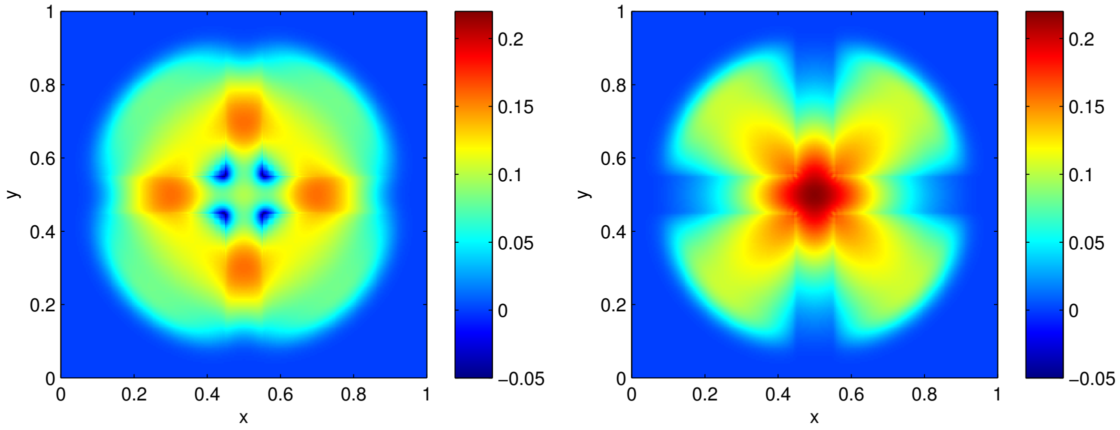
Figure 2.9: Highly scattering spots: Scalar flux ϕ_0 at $t = 3.2$ from several approximations with 250×250 points. The values are plotted in $\log_{10}(\phi_0)$ and limited to ten orders of magnitude.



(a) SP_3 (b) P_3



(c) SP_{11} (d) P_{11}: all values are nonnegative



(e) SP_{59} (f) P_{59}: all values are nonnegative

Figure 2.10: Scalar flux ϕ_0 at $t = 0.4$ with 250×250 points.

The absorption cross-section is set to zero on the whole domain and Neumann boundary conditions are enforced.

Several profiles of the particle distributions are illustrated in Figure 2.10. P_3 shows negative particle densities and artificial wave fronts whose number increases for growing approximation order N. In contrast to the classic linesource problem, there are enough particle collisions that the nonnegativity property is preserved by P_N approximations starting from $N = 11$. Here, we numerically observe that an order of $N = 59$ is needed until the P_N method converges. P_{59} forms four significant rays in corners of highly scattering and vacuum regimes. However, the converged SP_{59} solution cannot resolve these rays and does not converge to same P_{59} result. Like all simplified P_N computations shown above, it still develops negative particle concentrations and gives an inaccurate solution. Similar to P_N calculations, SP_3 and SP_{11} display the characteristic wave fronts.

This example demonstrates that the assumption of a *homogeneous* medium which is made in the asymptotic derivation in Section 2.3 must not be neglected. Indeed, it is not just a purely theoretical issue but can also show practical consequences leading to a degradation of SP_N solutions.

Chapter 3

A Realizability-Preserving DG Method: The M_1 Model

3.1 Introduction

In this chapter, we construct and implement a discontinuous Galerkin method for the following system of partial differential equations:

$$\begin{aligned} \partial_t E + \partial_x F &= -c\sigma_{\mathrm{a}}(x)(E - aT^4)\,, \\ \partial_t F + c^2\partial_x(\chi(E,F)E) &= -c\sigma_{\mathrm{t}}(x)F\,, \end{aligned} \tag{3.1}$$

where the Eddington factor χ is given by

$$\chi(E,F) = \frac{1}{3}\left(5 - 2\sqrt{4 - 3\left(\frac{F}{cE}\right)^2}\right)\,. \tag{3.2}$$

This system, usually referred to as the (grey) M_1 model of radiative transfer, approximates the evolution of the radiation energy $E = E(x,t)$ and the radiation energy flux $F = F(x,t)$ of photons passing through a material medium with slab geometry. Here, the coefficients $\sigma_{\mathrm{a}}(x)$ and $\sigma_{\mathrm{t}}(x)$ are material constants and $T = T(x,t)$ is the temperature of the material. The evolution of T is determined by an equation for the material energy $e(T)$:

$$\partial_t e(T) = c\sigma_{\mathrm{a}}(x)\left(E - aT^4\right)\,. \tag{3.3}$$

The M_1 model is the first member in a hierarchy of models which rely on the physical principle of maximum entropy as a tool for deriving angular moment closures of radiative transfer. Entropy-based models have been studied extensively in the areas of extended thermodynamics [44,132], gas dynamics [70,78,90,91,109,112,158], semiconductors [6–9,76,89,92,110,151], quantum fluids [41,45], radiation transport [21,22,27,28,47,48,54,77,79,127,129,169,179], and phonon transport in solids [45]. Since we are interested in photons, the Eddington factor in (3.2) is derived using the Bose-Einstein entropy. Other entropies (Maxwell-Boltzmann, Fermi-Dirac, etc...) lead to different forms of χ.

In the context of radiative transfer, the M_1 model dates back to [127], where it was first derived using Maxwell-Boltzmann statistics. For problems with Bose-Einstein statistics, theoretical properties such as hyperbolicity and entropy dissipation were first reported in [47] for general entropy-based, or M_N, models. Since then, computational studies have focused primarily on properties of the M_1 model and its extensions, including multigroup equations [176] and partial moment models [48, 54]. In related work, one may find simulations of M_1 models based on other statistics, including Maxwell-Boltzmann and Fermi-Dirac [19, 21, 22]. This attachment to M_1 is due to the fact that the higher order members of the M_N hierarchy require the repeated solution of expensive numerical optimization problems. However, simulations of the M_2 model [129, 179] (the next member in the hierarchy) have been performed for Bose-Einstein statistics and for M_N up to order $N = 15$ for special benchmark problems using Maxwell-Boltzmann statistics [77].

One of the fundamental questions associated with any moment model is the issue of *realizability* (which we define precisely in Section 3.2). In the context of the M_1 model, we say that E and F are realizable if and only if they are the first two moments of an underlying angular distribution. This requirement on E and F is mathematically equivalent to the following condition

$$0 \leq |F| \leq cE, \tag{3.4}$$

which is understood pointwise in x and t. This leads to the following conjecture:

Conjecture 1. *Let E and F solve* (3.1) *with boundary conditions at x_L and x_R and initial conditions at time t_0 that satisfy* (4.1). *Then E and F satisfy* (4.1) *for all $x \in [x_L, x_R]$ and all $t \geq t_0$.*

If this conjecture holds, we say that the set of realizable moments is invariant under the dynamics of (3.1). In multi-dimensional settings, it is known that analytical solutions to the P_1 model (for which $\chi = 1/3$) and for spherical harmonic closures in general can yield values of E that are negative [20, 121, 123, 124, 135, 140]. Moreover, in steady-state cases, the P_1 model predicts that $F = c(3\sigma_{\mathrm{t}})^{-1}\partial_x E$, which near discontinuities is unbounded and thus inconsistent with (4.1). However, spherical harmonic closures are derived assuming an ansatz for the angular distribution which is not necessarily positive. We expect the conjecture to hold because the M_1 model—and M_N closures in general—are derived assuming an ansatz for the angular distribution that is positive. Unfortunately, we know of no rigorous proof of this conjecture, only of partial results found in [36].

In this chapter, we address the problem of realizability from a numerical point of view via the design and implementation of a Runge-Kutta discontinuous Galerkin (RKDG) method, which combines discontinuous, finite-element spatial reconstructions with explicit Runge-Kutta methods for time integration. The spatial reconstructions are typically high-order polynomials defined on local elements. As neighboring reconstructions may not agree at cell edges, numerically stable flux functions must be defined. Like many numerical methods for hyperbolic PDEs, DG methods may produce spurious, possibly unstable oscillations around discontinuities. In such cases limiters are required to alter the local polynomial reconstruction.

Even if the conjecture above holds, numerical simulations may not preserve such a property. To address the issue in the context of a DG method, we extend a recent limiting technique which has been used to preserve maximum principles for scalar conservation laws and, more generally, to preserve invariant convex sets for systems of hyperbolic balance laws [182, 183]. The basic idea of this technique is to write the local polynomial reconstruction as an average plus a perturbation and to dampen the coefficients of the perturbation to ensure, under a more restrictive CFL condition, that the local averages at the next time step satisfy the required convexity condition. We call such a scheme realizability-preserving. For finite volume methods, such schemes have been constructed for the M_1 model in [13, 23, 24]

The chapter is laid out as follows. In Section 3.2, we review the derivation of the M_1 model and some important properties. In Section 3.3, we present the standard RKDG formulation and in Section 3.4, we discuss the new limiting procedure. In Section 3.5, we present numerical results.

3.2 The M_1 Model for Radiative Transfer

In this section, we summarize the derivation of the M_1 model for the radiative transfer equation, restricting ourselves to the case of slab geometries. The reader should note, however, that many of the results below are generally applicable to general three-dimensional geometries and moment models of arbitrarily high order.

3.2.1 The Radiative Transfer Equation

We consider a collection of photons which move at the speed of light c through a static material medium with slab geometry. In engineering and physics applications, the fundamental quantity of interest is the radiation intensity $\psi = \psi(x, \mu, \nu, t)$ which depends on the spatial coordinate $x \in (x_{\mathrm{L}}, x_{\mathrm{R}}) \subset \mathbb{R}$ along the direction perpendicular to the slab, on the cosine $\mu \in [-1, 1]$ of the angle between the x axis and the photon direction of flight, on the photon frequency $\nu \in (0, \infty)$, and on time $t \in (0, \infty)$. If f is the kinetic density of photons—that is, the number density with respect to the Lebesgue measure $dx d\mu d\nu$—then $\psi = h\nu c f$, where h is Planck's constant.

The material medium is characterized by a temperature $T = T(x, t)$, an equation of state for the energy $e = e(T)$, and by scattering, absorption, and total cross-sections: $\sigma_{\mathrm{s}}(x)$, $\sigma_{\mathrm{a}}(x)$, and $\sigma_{\mathrm{t}}(x) = \sigma_{\mathrm{a}}(x) + \sigma_{\mathrm{s}}(x)$ that depend on x directly and also indirectly through the material temperature.

The radiative transfer equation, which approximates the evolution of ψ, is given by

$$\frac{1}{c}\partial_t \psi + \mu \partial_x \psi + \sigma_{\mathrm{t}} \psi = \frac{1}{2}\left(\sigma_{\mathrm{s}} \phi + \sigma_{\mathrm{a}} B(T)\right) \tag{3.5}$$

where ϕ is the angular integral of ψ:

$$\phi := \int_{-1}^{1} \psi d\mu \tag{3.6}$$

and the Planckian[1]

$$B(T) := \frac{2h\nu^3}{c^2} \frac{2\pi}{\exp\left(\frac{h\nu}{kT}\right) - 1} \tag{3.7}$$

---

[1] The reader should note that the factor of 2π is not included in the standard definition of B. It arises from integration over the azimuthal angle of the sphere in $\mathbb{R}^3$ which occurs when reducing from general to slab geometry.

models blackbody radiation from the material. The constant k is Boltzmann's constant. The term $\sigma_{\mathrm{t}}\psi$ on the left-hand side of (3.5) models the loss of photon energy at a particular frequency and angle due to out-scattering and absorption by the material. The right-hand side of (3.5) models the gain in photon energy due to in-scattering from other angles and re-emission by the material.

The evolution of the material energy is determined by a balance of emitted and absorbed photons:

$$\partial_t e(T) = \sigma_{\mathrm{a}}\left(\langle\psi\rangle - acT^4\right) , \tag{3.8}$$

where angle brackets are used as a shorthand notation for integration over angle and frequency:

$$\langle\,\cdot\,\rangle \equiv \int_0^\infty \int_{-1}^1 (\,\cdot\,)\, d\mu d\nu \,, \tag{3.9}$$

and the T^4 term in (3.8) comes from the Stefan-Boltzmann Law:

$$\langle B(T)\rangle = acT^4 \,. \tag{3.10}$$

The constant $a = \frac{8\pi^5 k^4}{15h^3c^3}$ is the *radiation constant*. We assume that the material energy is independent of the material density so that

$$\partial_t e(T) = C_v \partial_t T \,, \tag{3.11}$$

where the specific heat at constant volume $C_v := \partial e/\partial T$ is constant. Though the material equation (3.8) plays an important role, our emphasis here will be on simulating the transport equation (3.5).

3.2.2 The M_1 Closure

The M_1 model is usually expressed in terms of the energy density E and flux density F, given by

$$E := \frac{1}{c}\langle\psi\rangle \quad \text{and} \quad F := \langle\mu\psi\rangle. \tag{3.12}$$

The exact equations for these moments are

$$\partial_t E + \partial_x F = -c\sigma_{\mathrm{a}}(x)(E - aT^4) \tag{3.13}$$

$$\partial_t F + c^2\partial_x\langle\mu^2\psi\rangle = -c\sigma_{\mathrm{t}}(x)F \tag{3.14}$$

Thus, aside from the energy equation, one needs an approximation of the second moment $\langle\mu^2\psi\rangle$ as a function of E and F in order to close the model. Entropy-based closures generate such an approximation by replacing ψ in (3.14) with a generalized Bose-Einstein (BE) distribution:

$$\mathcal{B}(\alpha,\beta) = \frac{2h\nu^3}{c^2}\frac{2\pi}{\exp\left[-\frac{h\nu}{kT}(\alpha+\beta\mu)\right] - 1} \,, \tag{3.15}$$

where α and β are related to E and F via the moment conditions

$$E = \frac{1}{c}\langle\mathcal{B}(\alpha,\beta)\rangle \,, \quad \text{and} \quad F = \langle\mu\mathcal{B}(\alpha,\beta)\rangle \,. \tag{3.16}$$

After some calculation, the generalized BE distribution gives

$$\langle\mu^2\psi\rangle \simeq \langle\mu^2\mathcal{B}(\alpha,\beta)\rangle = \chi(E,F)E \,, \tag{3.17}$$

where the Eddington factor χ is given in (3.2).

3.2.3 Realizability and Properties of the M_1 Model

The well-posedness of the M_1 closure relies entirely on the assumption that for each (x, t), there exist coefficients $\alpha(x, t)$ and $\beta(x, t)$ such that (3.16) holds. Furthermore, as discussed in the introduction, such coefficients will exist only if $E(x, t)$ and $F(x, t)$ satisfy the conditions in (3.4). This brings us back to the notion of realizability, which we define more precisely now.[2]

Definition 1. *A vector* $[\Psi_0, \Psi_1, \dots, \Psi_N]^T \in \mathbb{R}^N$ *is called realizable with respect to* $[1, \mu, \dots, \mu^N]$ *if there exists a non-negative measure on* $d\mu d\nu$ *with density* $\Psi(\mu, \nu)$ *such that* $\Psi_k = \langle \mu^k \Psi \rangle$ *for* $k = 1, \dots, N$. *The set* $\mathcal{R}_N$ *of all such vectors is called the realizable set.*

We also collect some properties that are necessary for the numerical analysis in Section 3.4. The first of these is equivalent to (3.4).

Lemma 1. *The vector* $[\Psi_0, \Psi_1]^T \in \mathbb{R}^2$ *is realizable if and only if*

$$|\Psi_1| \leq \Psi_0. \tag{3.18}$$

We call (3.18) *the realizability condition(s) (for* $\mathcal{R}_2$*).*

Proof. Let Ψ_0 and Ψ_1 be the moments of a non-negative measure Ψ. Because $|\mu| \leq 1$,

$$0 \leq |\Psi_1| = |\langle \mu \Psi \rangle| \leq |\langle \Psi \rangle| = \Psi_0. \tag{3.19}$$

Conversely, let $[\Psi_0, \Psi_1]$ satisfy the realizability conditions. We only need to construct a measure Ψ which generates Ψ_0 and Ψ_1. If $\Psi_0 = 0$, then let $\Psi = 0$. Otherwise, let

$$\Psi(\mu, \nu) = \eta(\mu)\gamma(\nu) \tag{3.20}$$

where γ is any probability distribution on $(0, \infty)$ and η is a weighted delta function:

$$\eta = \Psi_0 \delta \left(\mu - \frac{\Psi_1}{\Psi_0} \right) \tag{3.21}$$

A short calculation shows that this measure has moments Ψ_0 and Ψ_1. □

An immediate consequence of the realizability conditions for the set $\mathcal{R}_2$ is the following:

Lemma 2. *The set* $\mathcal{R}_2$ *is a closed, convex cone.*

Lemma 3. *If* $[\Psi_0, \Psi_1, \Psi_2]^T \in \mathcal{R}_3$, *then*

$$\Psi_2 \leq \Psi_0 \tag{3.22}$$

and

$$|\Psi_1 \pm \Psi_2| \leq \Psi_0 \pm \Psi_1. \tag{3.23}$$

---

[2] The careful reader will note that we define realizability independently of the speed of light c. However, in order to be consistent with the physics literature, we include a factor of c in the definition of E. This factor of c must be carefully carried through all subsequent calculations .

Proof. Let ψ be the density of the measure that generates $[\Psi_0, \Psi_1, \Psi_2]^T$. Since $|\mu|^2 \le 1$, the bound in (3.22) is immediate. To show the bound in (3.23), one may observe that

$$|\Psi_1 \pm \Psi_2|^2 = \left|\langle \mu\psi \pm \mu^2\psi\rangle\right|^2 = |\langle \mu(1 \pm \mu)\psi\rangle|^2 \tag{3.24}$$

and that $1 \pm \mu \ge 0$ for all $\mu \in [-1, 1]$. Hence, by the Cauchy-Schwarz inequality,

$$\begin{aligned} |\Psi_1 \pm \Psi_2|^2 &\le \left\langle \mu^2(1 \pm \mu)\psi\right\rangle \langle (1 \pm \mu)\psi\rangle \\ &\le \langle (1 \pm \mu)\psi\rangle \langle (1 \pm \mu)\psi\rangle = \langle (1 \pm \mu)\psi\rangle^2 = |\Psi_0 \pm \Psi_1|^2. \end{aligned} \tag{3.25}$$

The assertion follows by taking the square root on both sides of (3.25). □

Having defined realizability, we can state the following [47, 109].

Theorem 1. *For each* $[\Psi_0, \Psi_1] \in \operatorname{int}\mathcal{R}_2$ *(in the interior of* $\mathcal{R}_2$*), there exists a unique vector* $[\Lambda_0, \Lambda_1]^T$ *such that*

$$\begin{pmatrix} \Psi_0 \\ \Psi_1 \end{pmatrix} = \left\langle \begin{pmatrix} 1 \\ \mu \end{pmatrix} \mathcal{B}(\Lambda_0, \Lambda_1) \right\rangle . \tag{3.26}$$

Hence, when $[\Psi_0, \Psi_1]$ *is restricted to* $\operatorname{int}\mathcal{R}_2$*, the* M_1 *model is strictly hyperbolic and, furthermore, when expressed in terms of the variables* α *and* β*, the left hand side of* (3.1) *takes the symmetric form:*

$$H(\alpha, \beta)\, \partial_t \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + cJ(\alpha, \beta)\, \partial_x \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{3.27}$$

where

$$H(\alpha, \beta) := \left\langle \begin{pmatrix} 1 & \mu \\ \mu & \mu^2 \end{pmatrix} \mathcal{B}(\alpha, \beta)[\mathcal{B}(\alpha, \beta) + 1] \right\rangle \tag{3.28}$$

is a positive definite, symmetric matrix and

$$J(\alpha, \beta) := \left\langle \begin{pmatrix} \mu & \mu^2 \\ \mu^2 & \mu^3 \end{pmatrix} \mathcal{B}(\alpha, \beta)[\mathcal{B}(\alpha, \beta) + 1] \right\rangle \tag{3.29}$$

is symmetric.

Remark 13. *It is straight forward to show that*

$$\operatorname{int}\mathcal{R}_2 = \left\{[\Psi_0, \Psi_1]^T \in \mathbb{R}^2 \colon |\Psi_1| < \Psi_0\right\} \tag{3.30}$$

can be generated by moments of non-negative L^1 *functions. However, boundary elements of* $\mathcal{R}_2$ *can be generated only by the zero function or a delta function at* ± 1*. The Bose-Einstein ansatz cannot generate these moments with any finite values* α *and* β*.*

A consequence of this theorem is the following

Lemma 4. *The characteristic velocities for the* M_1 *model in modulus are bounded by* c*.*

Proof. Due to (3.27), any characteristic speed λ of the M_1 model satisfies $J\mathbf{v} = (\lambda/c)H\mathbf{v}$ for some eigenvector $\mathbf{v} = [v_0, v_1]^T \in \mathbb{R}^2$. Let $p(\mu) := (v_0 + v_1\mu)^2$. Then it follows from the definitions of H and J in (3.28) and (3.29),

$$\frac{|\lambda|}{c} = \frac{|\mathbf{v}^T J\mathbf{v}|}{\mathbf{v}^T H\mathbf{v}} = \frac{|\langle \mu p(\mu)\mathcal{B}(\alpha, \beta)[\mathcal{B}(\alpha, \beta) + 1]\rangle|}{\langle p(\mu)\mathcal{B}(\alpha, \beta)[\mathcal{B}(\alpha, \beta) + 1]\rangle} \le 1 . \tag{3.31}$$

□

3.3 DG Formulation

In this section, we describe the Runge-Kutta, discontinuous Galerkin (RKDG) method as applied to (3.1) – (3.3). The RKDG method is a method of lines: the DG discretization is only applied to spatial variables while time discretization is achieved by explicit Runge-Kutta time integrators. Our presentation follows closely the standard formulation, which can be found, for example, in [33, 34].

The one-dimensional M_1 model can be written as a system

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathcal{S}(\mathbf{u}, T), \quad (x,t) \in (x_L, x_R) \times (0, t_{\text{final}}), \tag{3.32}$$

where

$$\mathbf{u} = \begin{bmatrix} cE \\ F \end{bmatrix}, \quad \mathbf{f}(\mathbf{u}) = \begin{bmatrix} cF \\ c^2\chi(E,F)E \end{bmatrix}, \quad \mathcal{S}(\mathbf{u}, T) = \begin{bmatrix} c^2\sigma_{\text{a}}(x)(aT^4 - E) \\ -c\sigma_{\text{t}}(x)F \end{bmatrix}. \tag{3.33}$$

Additionally, we have the material energy equation from (4.6) which, by using the relation in (3.11), can be reformulated as

$$\partial_t T = \frac{c\sigma_{\text{a}}(x)}{C_v}(E - aT^4). \tag{3.34}$$

Initial and boundary conditions are imposed on $T(x,t)$ and $\mathbf{u}(x,t)$:

$$T(x,0) = T_0(x), \quad \mathbf{u}(x,0) = \mathbf{u}_0(x), \quad x \in (x_L, x_R), \tag{3.35}$$

$$T(x_L,t) = T_L(t), \quad \mathbf{u}(x_L,t) = \mathbf{u}_L(t), \quad t > 0, \tag{3.36}$$

$$T(x_R,t) = T_R(t), \quad \mathbf{u}(x_R,t) = \mathbf{u}_R(t). \tag{3.37}$$

Note that boundary conditions for the hyperbolic system in (3.32) should be carefully chosen. To guarantee well-posedness, it is necessary to distinguish between in- and outgoing characteristics at each boundary. A detailed description of how initial and boundary values should be implemented numerically can be found in [34].

3.3.1 Spatial Discretization

We divide the computational domain $[x_L, x_R]$ into J cells:

$$x_L = x_{1/2} < x_{3/2} < \ldots < x_{J+1/2} = x_R,$$

where x_j is the center of each cell $I_j = (x_{j-1/2}, x_{j+1/2})$. Let $h_j := x_{j+1/2} - x_{j-1/2}$ be the length of the interval I_j and $h := \max_j h_j$. Denote the finite-dimensional approximation space by

$$V_h^k = \{v \in L^1(x_L, x_R) : v_{|I_j} \in \mathcal{P}^k(I_j),\, j = 1, \ldots, J\},$$

where $\mathcal{P}^k(I_j)$ is the space of polynomials of degree at most k on the interval I_j.

The semidiscrete DG scheme is derived from a weak formulation of (3.32) and (3.34). Multiplying these equations by an arbitrary smooth test function φ and integrating over each cell I_j gives

$$\begin{aligned} &\int_{I_j} \varphi(x)\partial_t \mathbf{u}(x,t)dx - \int_{I_j} \mathbf{f}(\mathbf{u}(x,t))\partial_x\varphi(x)dx \\ &\quad + \mathbf{f}(\mathbf{u}(x_{j+1/2},t))\varphi(x_{j+1/2}) - \mathbf{f}(\mathbf{u}(x_{j-1/2},t))\varphi(x_{j-1/2}) = \int_{I_j} \mathcal{S}(\mathbf{u}(x,t), T(x,t))\varphi(x)dx, \end{aligned} \tag{3.38}$$

$$\int_{I_j} \varphi(x)\partial_t T(x,t)dx = \int_{I_j} \varphi(x)\frac{c\sigma_\mathrm{a}(x)}{C_v}(E(x,t) - aT^4(x,t))dx. \tag{3.39}$$

To define the DG finite element method, the exact solutions $\mathbf{u}(.,t)$ and $T(.,t)$ are replaced by approximations $\mathbf{u}_h(.,t) \in V_h^k \times V_h^k$ and $T_h(.,t) \in V_h^k$. The resulting set of equations is then required to hold for all test functions $\varphi_h \in V_h^k$.[3] The task is then to find $T_h(.,t)$ and $\mathbf{u}_h(.,t)$ such that for all test functions $\varphi_h \in V_h^k$, the following holds:

$$\begin{aligned} \int_{I_j} \varphi_h(x)\partial_t \mathbf{u}_h(x,t)dx - \int_{I_j} \mathbf{f}(\mathbf{u}_h(x,t))\partial_x\varphi_h(x)dx + \left[\!\!\left[\hat{\mathbf{f}}\varphi_h(x)\right]\!\!\right]_j \\ = \int_{I_j} \mathcal{S}(\mathbf{u}_h(x,t), T_h(x,t))\varphi_h(x)dx, \end{aligned} \tag{3.40}$$

$$\int_{I_j} \varphi_h(x)\partial_t T_h(x,t)dx = \int_{I_j} \varphi_h(x)\frac{c\sigma_\mathrm{a}(x)}{C_v}(E_h(x,t) - aT_h^4(x,t))dx, \tag{3.41}$$

where

$$\left[\!\!\left[\hat{\mathbf{f}}\varphi_h(x)\right]\!\!\right]_j = \hat{\mathbf{f}}_{j+1/2}\varphi_h(x_{j+1/2}^-) - \hat{\mathbf{f}}_{j-1/2}\varphi_h(x_{j-1/2}^+) \tag{3.42}$$

and

$$\varphi_h(x_{j+1/2}^-) = \lim_{\substack{\varepsilon\to 0\\ \varepsilon>0}} \varphi_h(x_{j+1/2} - \varepsilon), \quad \varphi_h(x_{j-1/2}^+) = \lim_{\substack{\varepsilon\to 0\\ \varepsilon>0}} \varphi_h(x_{j-1/2} + \varepsilon) \tag{3.43}$$

are the right and left limits of φ_h at the cell interfaces $x_{j\pm1/2}$.

Since $\mathbf{u}_h(.,t)$ is a piecewise polynomial, $\mathbf{f}(\mathbf{u}_h(x_{j+1/2},t))$ is not strictly defined. Thus the nonlinear flux function f is replaced by a numerical flux $\hat{\mathbf{f}}$ which depends on the pointwise limits of $\mathbf{u}_h$ on either side of the edge at $x_{j+1/2}$:

$$\hat{\mathbf{f}}_{j-1/2} = \hat{\mathbf{f}}(\mathbf{u}_h(x_{j-1/2}^-,t), \mathbf{u}_h(x_{j-1/2}^+,t)) \quad \text{and} \quad \hat{\mathbf{f}}_{j+1/2} = \hat{\mathbf{f}}(\mathbf{u}_h(x_{j+1/2}^-,t), \mathbf{u}_h(x_{j+1/2}^+,t)).$$

The definition of the semi-discrete scheme is completed by the choice of a numerical flux $\hat{\mathbf{f}}$. In order to maintain desirable properties like stability and convergence to the entropy solution for conservation laws, one typically chooses numerical fluxes associated with monotone schemes satisfying certain properties [142]. Here, we follow common convention and use the Lax-Friedrichs flux:

$$\hat{\mathbf{f}}_{j\pm1/2} = \frac{1}{2}[\mathbf{f}(\mathbf{u}_{j\pm1/2}^-) + \mathbf{f}(\mathbf{u}_{j\pm1/2}^+) - \lambda(\mathbf{u}_{j\pm1/2}^+ - \mathbf{u}_{j\pm1/2}^-)]\,, \tag{3.44}$$

where λ is the largest magnitude of any eigenvalue of the flux Jacobian. For the M_1 system, we invoke Lemma 4 and set $\lambda = c$.

The DG solution $\mathbf{u}_h$ is expanded in terms of local basis functions $\{\phi_l^j\}_{l=0}^k$ for $\mathcal{P}^k(I_j)$ in each cell I_j:

$$\mathbf{u}_h^j(x,t) = \sum_{l=0}^{k} \mathbf{u}_l^j(t)\phi_l^j(x), \quad \text{for } x \in I_j. \tag{3.45}$$

---

[3] For all equations we use the same test functions $\varphi_h(x)$ in our weak formulation which is not always necessary. One could also multiply each equation in (3.32) and (3.34) by a different test function belonging to V_h^k.

Substituting this expansion into (3.40) yields a system of $3J(k+1)$ equations for the unknowns $\mathbf{u}_l^j(t)$. For all $j = 1, \dots, J$ and $m = 0, \dots, k$,

$$\begin{aligned}\sum_{l=0}^{k} \partial_t \mathbf{u}_l^j(t) \int_{I_j} \phi_m^j(x)\phi_l^j(x)dx - \int_{I_j} \mathbf{f}(\mathbf{u}_h^j(x,t))\, \partial_x \phi_m^j(x)dx \\ + \left[\!\left[\hat{\mathbf{f}}\phi_m^j(x) \right]\!\right]_j = \int_{I_j} \mathcal{S}(\mathbf{u}_h^j(x,t), T_h^j(x,t))\phi_m^j(x)dx.\end{aligned} \tag{3.46}$$

$$\sum_{l=0}^{k} \partial_t T_l^j(t) \int_{I_j} \phi_m^j(x)\phi_l^j(x)dx = \int_{I_j} \phi_m^j(x) \frac{c\sigma_{\mathrm{a}}(x)}{C_v}(E_h^j(x,t) - a{T_h^j}^4(x,t))dx. \tag{3.47}$$

The standard choice of basis for $\mathcal{P}(I_j)$ is generated by Legendre polynomials P_l, defined on the reference cell $[-1,1]$:

$$\phi_l^j(x) = P_l\left(\frac{2(x-x_j)}{h_j}\right). \tag{3.48}$$

Due to the L_2-orthogonality,

$$\int_{-1}^{1} P_l(y)P_m(y)dy = \frac{2}{2m+1}\delta_{l,m} \tag{3.49}$$

the mass matrices in (3.46) and (3.47) become diagonal. With $\xi_j(y) := x_j + yh_j/2$, this gives a formulation defined on the reference cell for all $m = 0, \dots, k$,:

$$\begin{aligned}\frac{h_j}{2m+1}\partial_t \mathbf{u}_m^j(t) - \int_{-1}^{1} \mathbf{f}(\mathbf{u}_h^j(\xi_j(y),t))\partial_y P_m(y)dy \\ + \hat{\mathbf{f}}_{j+1/2} - (-1)^m \hat{\mathbf{f}}_{j-1/2} = \frac{h_j}{2}\int_{-1}^{1} \mathcal{S}(\mathbf{u}_h^j(\xi_j(y),t), T_h^j(\xi_j(y),t))P_m(y)dy,\end{aligned} \tag{3.50}$$

$$\begin{aligned}\frac{h_j}{2m+1}\partial_t T_m^j(t) = \frac{ch_j}{2C_v}\sum_{l=0}^{k} E_l^j(t)\left(\int_{-1}^{1} P_m(y)P_l(y)\sigma_{\mathrm{a}}(\xi_j(y))\right)dy \\ - \frac{ach_j}{2C_v}\int_{-1}^{1} P_m(y){T_h^j}^4(\xi_j(y),t)\sigma_{\mathrm{a}}(y)dy\,,\end{aligned} \tag{3.51}$$

where the remaining integrals are calculated by a quadrature rule. For polynomials of degree at most k, the order of accuracy will be maintained if quadrature formulas are implemented which are exact for polynomials of degree $2k+1$ [32].

Equations (3.50) and (3.51) form a system of ODEs for the coefficients $\mathbf{u}_m^j(t)$ and $T_m^j(t)$. For all $j = 1, \dots, J$ and $m = 0, \dots, k$, we write this system in the abstract form:

$$\partial_t \mathbf{u}_m^j(t) = \mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_0^{j-1}, \dots, \mathbf{u}_k^{j-1}, \mathbf{u}_0^j, \dots, \mathbf{u}_k^j, \mathbf{u}_0^{j+1}, \dots, \mathbf{u}_k^{j+1}), \tag{3.52}$$

$$\partial_t T_m^j(t) = \mathcal{L}_{T,m}^j(E_0^j, \dots, E_k^j, T_0^j, \dots, T_k^j). \tag{3.53}$$

Here, $\mathcal{L}_{\mathbf{u},m}^j$ and $\mathcal{L}_{T,m}^j$ are the respective right-hand sides of the ODEs. Numerical integration of the ODE system is discussed in Section 3.3.3.

3.3.2 Slope Limiter

Various types of limiters have been developed to improve the stability of DG methods and to reduce nonphysical oscillations in the solution. We refer to [150] for a summary of popular alternatives. Most slope limiters are constructed by first marking so-called "troubled cells" which need to be limited and then replacing the original reconstruction in those cells by a new polynomial with modified coefficients. In order for the overall method to remain conservative, the limiting procedure must maintain the original cell averages.

In our code, we implement the moment limiter of *Burbeau et al.* [25], which is a modification of the original limiter of *Biswas et al.* [15]. The main advantage of the approach in [15] is that the limiting does not include solution-dependent parameters. This is not the case for the often used TVB limiter of *Cockburn et al.* [34].

The limiting procedure is based on the use of an indicator which compares coefficients $\mathbf{u}_l^j$ with weighted differences of neighboring coefficients with index $l-1$ via the standard minmod function:

$$\operatorname{minmod}(a_1, a_2, \dots, a_n) := \begin{cases} s \cdot \min_{1 \le i \le n} |a_i|, & \text{if } \operatorname{sign}(a_1) = \dots = \operatorname{sign}(a_n) = s \\ 0, & \text{otherwise.} \end{cases} \tag{3.54}$$

For a given j and l, the indicator takes the form

$$\mathbf{u}_l^{j,\min} := \frac{1}{2l-1} \operatorname{minmod}\left\{(2l-1)\mathbf{u}_l^j, \mathbf{u}_{l-1}^{j+1} - \mathbf{u}_{l-1}^j, \mathbf{u}_{l-1}^j - \mathbf{u}_{l-1}^{j-1}\right\} . \tag{3.55}$$

A cell is declared as troubled if the minmod function in (3.55) for the *highest-order* coefficient returns: $\mathbf{u}_k^{j,\min} \neq \mathbf{u}_k^j$. The procedure then continues in recursive fashion. First, the highest-order coefficient $\mathbf{u}_k^j$ is limited. Then, quoting [15], *"the limiter is applied to successively lower-order coefficients when the next higher coefficient on the interval has been changed by the limiting. Once the lower-order coefficients are limited, the higher-order coefficients are re-limited using the updated low-order coefficients."* To clarify this procedure we illustrate the limiting in a flow chart in Figure 3.1 for the special case of $k = 3$.

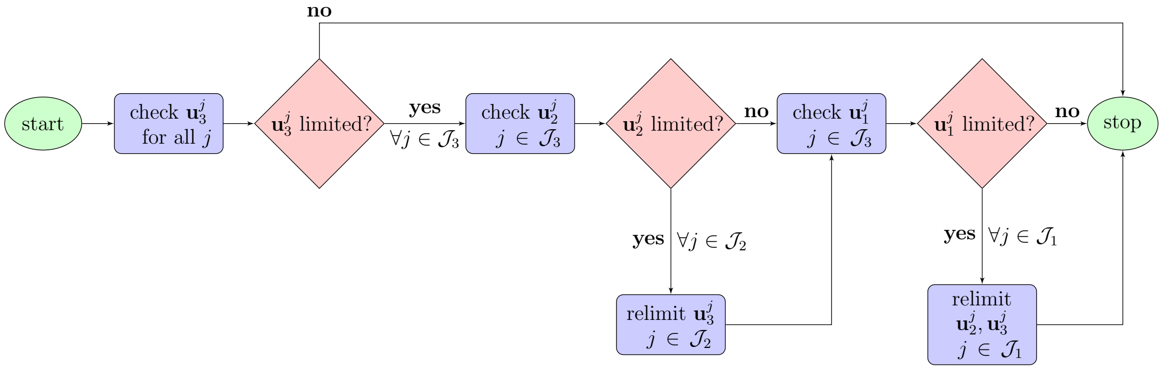

no
start
check $\mathbf{u}_3^j$ for all j
$\mathbf{u}_3^j$ limited?
yes $\forall j \in \mathcal{J}_3$
check $\mathbf{u}_2^j$ $j \in \mathcal{J}_3$
$\mathbf{u}_2^j$ limited?
no
check $\mathbf{u}_1^j$ $j \in \mathcal{J}_3$
$\mathbf{u}_1^j$ limited?
no
stop
yes $\forall j \in \mathcal{J}_2$
relimit $\mathbf{u}_3^j$ $j \in \mathcal{J}_2$
yes $\forall j \in \mathcal{J}_1$
relimit $\mathbf{u}_2^j, \mathbf{u}_3^j$ $j \in \mathcal{J}_1$

Figure 3.1: Limiting procedure for polynomials of degree $k = 3$. Here $\mathcal{J}_l$ denotes the set of all cells j which require limiting for the coefficient $\mathbf{u}_l^j$.

The application of the limiter to the coefficient $\mathbf{u}_l^j$ for a given j and l is as follows: If the minmod function returns the value $\mathbf{u}_l^j$, then the coefficient does not change.

Otherwise, it is redefined as

$$\mathbf{u}_l^j = \text{maxmod}(\mathbf{u}_l^{j,\min}, \mathbf{u}_l^{j,\max}), \tag{3.56}$$

where

$$\begin{aligned}
\mathbf{u}_l^{j,\max} &= \frac{1}{2l-1}\text{minmod}\left\{(2l-1)\mathbf{u}_l^j\ ,\ \hat{\mathbf{u}}_{l,-}^{j+1} - \mathbf{u}_{l-1}^j\ ,\ \mathbf{u}_{l-1}^j - \hat{\mathbf{u}}_{l,+}^{j-1}\right\}, && (3.57)\\
\hat{\mathbf{u}}_{l,-}^{j+1} &= \mathbf{u}_{l-1}^{j+1} - (2l-1)\mathbf{u}_l^{j+1}, && (3.58)\\
\hat{\mathbf{u}}_{l,+}^{j-1} &= \mathbf{u}_{l-1}^{j-1} + (2l-1)\mathbf{u}_l^{j-1}, && (3.59)
\end{aligned}$$

and

$$\text{maxmod}(a_1, a_2, \ldots, a_n) = \begin{cases} s \cdot \max_{1\le i\le n} |a_i|, & \text{if } \text{sign}(a_1) = \cdots = \text{sign}(a_n) = s \\ 0, & \text{otherwise.} \end{cases}$$

Whereas the minmod function tends to flatten smooth extrema, the maxmod function is introduced to relax the limiting and allow larger slopes.

Usually limiters are developed for scalar equations. For systems, the limiting is most robust when applied to each component of the local characteristic variables. Neglecting this transformation, one may observe non-physical oscillations around an otherwise monotonic solution [34]. However, the transformation to characteristic variables is expensive and not always needed. Thus, unless such a transformation is needed, we apply the limiter to the components of $\mathbf{u}$. A detailed description of the limiting procedure when applied to the characteristic variables is given in the appendix.

3.3.3 Time Discretization: Explicit SSP Runge-Kutta Schemes

The purpose of high-order, strong stability property (SSP) Runge-Kutta time integration methods is to achieve high-order accuracy in time while preserving desirable properties of the forward Euler method. We only use explicit schemes, which compute values of the unknowns at several intermediate stages. Each stage is a convex combination of forward Euler operators which usually lead to modified CFL restrictions.

Let $\{t^n\}_{n=0}^N$ be an equidistant partition of $[0, t_{\text{final}}]$ and set $\Delta t := t_{\text{final}}/N$. Let Λ denote the application of a generic slope limiter. The algorithm for the optimal third-order SSP Runge-Kutta (SSPRK(3,3)) method [66] reads as follows: :

- For all $j \in \{1, \ldots, J\}$ and $m \in \{0, \ldots, k\}$, set

$$\mathbf{u}_m^{j,0} = \Lambda\{\pi_{V_h^m}(\mathbf{u}_0)\}\ . \tag{3.60}$$

- For all $n \in \{0, \ldots, N-1\}$, $j \in \{1, \ldots, J\}$, and $m \in \{0, \ldots, k\}$,

  (1) Compute the intermediate stages

$$\begin{aligned}
\mathbf{u}_m^{j,(1)} &= \Lambda\left\{\mathbf{u}_m^{j,n} + \Delta t \mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_h^{j,(0)})\right\}\\
\mathbf{u}_m^{j,(2)} &= \Lambda\left\{\frac{3}{4}\mathbf{u}_m^{j,n} + \frac{1}{4}\mathbf{u}_m^{j,(1)} + \frac{1}{4}\Delta t \mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_h^{j,(1)})\right\}\\
\mathbf{u}_m^{j,(3)} &= \Lambda\left\{\frac{1}{3}\mathbf{u}_m^{j,n} + \frac{2}{3}\mathbf{u}_m^{j,(2)} + \frac{2}{3}\Delta t \mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_h^{j,(2)})\right\}
\end{aligned} \tag{3.61}$$

(2) Set $\mathbf{u}_m^{j,n+1} = \mathbf{u}_m^{j,(3)}$.

In the initial step, $\pi_{V_h^m}(\mathbf{u}_0)$ is the projection of the initial condition $\mathbf{u}_0$ on the finite dimensional space V_h. Note that Λ is applied at every Runge-Kutta stage.

For the sake of completeness we also state the optimal fourth order scheme SSPRK(5,4):

$$\begin{aligned}
\mathbf{u}_m^{j,(1)} &= \Lambda\{\mathbf{u}_h^{(n)} + 0.391752226571890\,\Delta t\,\mathcal{L}^j_{\mathbf{u},m}(\mathbf{u}_m^{j,(n)})\} \\
\mathbf{u}_m^{j,(2)} &= \Lambda\{0.444370493651235\,\mathbf{u}_m^{j,(n)} + 0.555629506348765\,\mathbf{u}_m^{j,(1)} \\
&\quad + 0.368410593050371\,\Delta t\,\mathcal{L}^j_{\mathbf{u},m}(\mathbf{u}_m^{j,(1)})\} \\
\mathbf{u}_m^{j,(3)} &= \Lambda\{0.620101851488403\,\mathbf{u}_m^{j,(n)} + 0.379898148511597\,\mathbf{u}_m^{j,(2)} \\
&\quad + 0.251891774271694\,\Delta t\,\mathcal{L}^j_{\mathbf{u},m}(\mathbf{u}_m^{j,(2)})\} \\
\mathbf{u}_m^{j,(4)} &= \Lambda\{0.178079954393132\,\mathbf{u}_m^{j,(n)} + 0.821920045606868\,\mathbf{u}_m^{j,(3)} \\
&\quad + 0.544974750228521\,\Delta t\,\mathcal{L}^j_{\mathbf{u},m}(\mathbf{u}_m^{j,(3)})\} \\
\mathbf{u}_h^{(n+1)} &= \Lambda\{0.517231671970585\,\mathbf{u}_m^{j,(2)} + 0.096059710526147\,\mathbf{u}_m^{j,(3)} \\
&\quad + 0.386708617503269\,\mathbf{u}_m^{j,(4)} + 0.063692468666290\,\Delta t\,\mathcal{L}^j_{\mathbf{u},m}(\mathbf{u}_m^{j,(3)}) \\
&\quad + 0.226007483236906\,\Delta t\,\mathcal{L}^j_{\mathbf{u},m}(\mathbf{u}_m^{j,(4)})\}.
\end{aligned} \tag{3.62}$$

Note that SSPRK(3,3) permits a timestep of the same size as forward Euler, while the SSPRK(5,4) method is less restrictive, allowing for a time step that is 1.508 times larger the forward Euler scheme.

3.4 The Realizability-Preserving Limiter

To achieve physically meaningful results, it is not only important to reduce oscillations near discontinuities, but also to enforce crucial properties like positivity of mass or energy. In the case of the M_1 model, the realizability condition (3.18) reflects the fact that energy density E is positive and that photons cannot move faster than the speed of light. From a mathematical point of view, it also provides a sufficient condition for hyperbolicity of the system (Theorem 1). Motivated by approaches in [182, 183], we develop an additional limiter which enforces the realizability condition by modifying higher-order coefficients of the polynomial reconstructions in each cell. Our aim is to modify the reconstruction of $\mathbf{u}$ so that

(a) the limiter does not destroy the accuracy for smooth solutions;

(b) one forward Euler step keeps the cell averages in the realizability region.

An essential ingredient in both [183] and [182] is the choice of the Gauss-Lobatto quadrature set

$$\{x_{j-1/2} = \hat{x}_j^1, \hat{x}_j^2, \dots, \hat{x}_j^{M-1}, \hat{x}_j^M = x_{j+1/2}\} \subset I_j, \tag{3.63}$$

where, for a spatial reconstruction of order k, M is the largest integer such that $2M - 3 \geq 2k + 1$. This condition on M ensures accuracy of the scheme [32]. The weaker condition $2M - 3 \geq k$ ensures that the quadrature integrates elements of the approximation space V_h^k exactly. In general, the limiter is defined in order to ensure that the values of the DG reconstruction at these points lies inside the convex set of interest, which in the case of the M_1 model, is the realizable set $\mathcal{R}_2$. However, we will enforce the convexity condition indirectly by requiring the positivity of the intermediate quantities[4]

$$Q := \frac{cE + F}{2} \quad \text{and} \quad R := \frac{cE - F}{2}\,. \tag{3.64}$$

The inverse transformation that maps $[Q, R]^T \mapsto [cE, F]^T$ is given by

$$E = \frac{Q + R}{c} \quad \text{and} \quad F = Q - R\,. \tag{3.65}$$

Lemma 5. *Let* $\varepsilon \geq 0$ *be given. The moment* $\mathbf{u} = [cE, F]^T$ *is an element of* $\mathcal{R}_2^\varepsilon := \mathcal{R}_2 + [\varepsilon, 0]^T$ *if and only if* $Q \geq \varepsilon/2$ *and* $R \geq \varepsilon/2$.

Proof. The proof is a simple application of (3.64) and (3.65). □

We now proceed to define the limiter. Let $\mathbf{u}_h^{j,n} = [cE_h^{j,n}, F_h^{j,n}]^T$ and $T_h^{j,n}$ be the approximation of $\mathbf{u}$ and T in cell I_j at time t^n, and let $\hat{\mathbf{u}}_h^{j,n}$ and $\hat{T}_h^{j,n}$ denote the modifications of $\mathbf{u}_h^{j,n}$ and $T_h^{j,n}$ that are generated by the limiting. We assume that the cell average of $\mathbf{u}_h^{j,n}$, which we denote by $\overline{\mathbf{u}}_h^{j,n}$, is realizable, i.e. $\overline{\mathbf{u}}_h^{j,n} \in \mathcal{R}_2$. We also assume that the cell average of $T_h^{j,n}$, which we denote by $\overline{T}_h^{j,n}$, is positive. Let $Q_h^{j,n}(x)$ and $R_h^{j,n}(x)$ be the approximations of Q and R, respectively, and define the limited variables by

$$\hat{Q}_h^{j,n}(x) = \theta_Q^{j,n} Q_h^{j,n}(x) + (1 - \theta_Q^{j,n})\overline{Q}_h^{j,n}\,, \tag{3.66}$$

$$\hat{R}_h^{j,n}(x) = \theta_R^{j,n} R_h^{j,n}(x) + (1 - \theta_R^{j,n})\overline{R}_h^{j,n}\,, \tag{3.67}$$

$$\hat{T}_h^{j,n}(x) = \theta_T^{j,n} T_h^{j,n}(x) + (1 - \theta_T^{j,n})\overline{T}_h^{j,n}\,, \tag{3.68}$$

where

$$\theta_Q^{j,n} := \min\left\{ \frac{\overline{Q}_h^{j,n} - \varepsilon/2}{\overline{Q}_h^{j,n} - Q_{\min}^{j,n}}, 1 \right\}, \quad Q_{\min}^{j,n} := \min_{\ell=1,\dots,M} Q_h^{j,n}(\hat{x}_j^\ell)\,, \tag{3.69a}$$

$$\theta_R^{j,n} := \min\left\{ \frac{\overline{R}_h^{j,n} - \varepsilon/2}{\overline{R}_h^{j,n} - R_{\min}^{j,n}}, 1 \right\}, \quad R_{\min}^{j,n} := \min_{\ell=1,\dots,M} R_h^{j,n}(\hat{x}_j^\ell)\,, \tag{3.69b}$$

$$\theta_T^{j,n} := \min\left\{ \frac{\overline{T}_h^{j,n}}{\overline{T}_h^{j,n} - T_{\min}^{j,n}}, 1 \right\}, \quad T_{\min}^{j,n} := \min_{\ell=1,\dots,M} T_h^{j,n}(\hat{x}_j^\ell)\,, \tag{3.69c}$$

and the parameter $\varepsilon > 0$ is chosen to maintain numerical stability with finite precision arithmetic. The value of ε should be small relative to the magnitude of the variables in a given problem. The components of $\hat{\mathbf{u}}_h^{j,n}$ are then defined using (3.65).

Item (a) above has essentially been proven in [183].

Theorem 2 ([183])**.** *For smooth solutions of the* M_1 *model, the* k^{th}*-order discontinuous Galerkin Runge-Kutta method with the above limiter is* k^{th}*-order accurate.*

---

[4] The meaning of all subsequent subscripts, superscripts and adornments of Q and R will be inherited from analogous definitions for E and F.

Proof. The triangle inequality gives

$$||\mathbf{u}^{j,n} - \hat{\mathbf{u}}_h^{j,n}|| \leq ||\mathbf{u}_h^{j,n} - \hat{\mathbf{u}}_h^{j,n}|| + ||\mathbf{u}^{j,n} - \mathbf{u}_h^{j,n}|| \,, \tag{3.70}$$

where the second term on the right is $O(h^{k+1})$ by the accuracy assumption of the theorem. Thus to verify the result, one simply has to show that for some constant $C > 0$,

$$||\mathbf{u}_h^{j,n} - \hat{\mathbf{u}}_h^{j,n}|| \leq Ch^{k+1} \,. \tag{3.71}$$

Such an inequality is shown in [183] and applies directly here since the transformation from $\mathbf{u}$ to $[Q, R]^T$ is linear and bounded. The error for the temperature equation follows a similar argument. □

We now address item (b). The modification in (3.66) does not change the cell averages of Q_h or R_h (or the cell averages of $\mathbf{u}$). However, it does ensure that their pointwise values at the quadrature points are bounded below by $\varepsilon/2$.

Lemma 6. *If* $\overline{Q}_h^{j,n} \geq 0$ *(respectively:* $\overline{R}_h^{j,n} \geq 0$, $\overline{T}_h^{j,n} \geq 0$*), then* $Q_h^{j,n}(\hat{x}_j^\ell) \geq \varepsilon/2$ *(respectively:* $R_h^{j,n}(\hat{x}_j^\ell) \geq \varepsilon/2$, $T_h^{j,n}(\hat{x}_j^\ell) \geq 0$*) for* $\ell = 1, \ldots, M$.

Proof. We show the results for Q only. The proofs for R and T follows the same argument. From (3.66) and the definition of $Q_{\min}^{j,n}$ in (3.69a), it follows that

$$\hat{Q}_h^{j,n}(\hat{x}_j^\ell) \geq \theta_Q^{j,n} Q_{\min}^{j,n} + (1 - \theta_Q^{j,n})\overline{Q}_h^{j,n} \,, \quad \ell = 1, \ldots, M \,. \tag{3.72}$$

By definition, $\theta_Q^{j,n} = 1$ only if $Q_{\min}^{j,n} \geq \varepsilon/2$. Hence, if $\theta_Q^{j,n} = 1$, the statement of the lemma follows immediately from (3.72). If $\theta_Q^{j,n} < 1$, then we can simply plug in the definition of $\theta_Q^{j,n}$ to find that

$$\theta_Q^{j,n} Q_{\min}^{j,n} + (1 - \theta_Q^{j,n})\overline{Q}_h^{j,n} = \frac{\varepsilon}{2} \,, \quad \ell = 1, \ldots, M, \tag{3.73}$$

and again the result is immediate. □

To prove the main theorem of this chapter, we introduce the following notation

$$\mathbf{u}_\ell^{j,n} := \mathbf{u}_h^{j,n}(\hat{x}_j^\ell) \,, \quad T_\ell^{j,n} := T_h^{j,n}(\hat{x}_j^\ell) \,, \quad \sigma_{\mathrm{t},\ell} := \sigma_{\mathrm{t}}(\hat{x}_j^\ell) \,. \tag{3.74}$$

Theorem 3. *Assume that* $2M - 3 \geq k$ *and for each* $\ell = 1, \ldots, M$,

$$\mathbf{u}_\ell^{j,n} \in \mathcal{R}_2 \quad \text{and} \quad T_\ell^{j,n} \geq 0. \tag{3.75}$$

Assume further that Δt *satisfies the following conditions:*

(A1) $\Delta t < \min_{\ell=1,\ldots,M} \left\{ \frac{1}{c\sigma_{\mathrm{t},\ell}} \right\}$,

(A2) $\Delta t < \min_{\ell=1,\ldots,M} \left\{ \frac{w_\ell \Delta x}{c(1 + w_\ell \sigma_{\mathrm{t},\ell} \Delta x)} \right\}$,

(A3) $\Delta t \leq \min_{\ell=1,\ldots,M} \left\{ \frac{C_v}{\sigma_{\mathrm{a},\ell} ac (T_\ell^{j,n})^3} \right\}$.

Then for the forward Euler time step,

$$\overline{\mathbf{u}}_h^{j,n+1} \in \mathcal{R}_2 \quad \textit{and} \quad \overline{T}_h^{j,n+1} \geq 0. \tag{3.76}$$

Proof. We prove that the algorithm preserves the positivity of the cell averages for Q and R and then invoke Lemma (5) with $\varepsilon = 0$. For convenience, we assume a normalized quadrature set: $\sum w_\ell = 1$. Then the algorithm for one Euler step applied to the cell averages $\overline{E}_h^{j,n}$, $\overline{F}_h^{j,n}$, and $\overline{T}_h^{j,n}$ is:

$$\overline{E}_h^{j,n+1} = \overline{E}_h^{j,n} - \frac{\Delta t}{\Delta x}(\hat{f}^{E,n}_{j+1/2} - \hat{f}^{E,n}_{j-1/2}) - \Delta t c \sum_{\ell=1}^{M} w_\ell \left[\sigma_{\mathrm{a},\ell} E_\ell^{j,n} - \sigma_{\mathrm{a},\ell} a (T_\ell^{j,n})^4\right], \tag{3.77a}$$

$$\overline{F}_h^{j,n+1} = \overline{F}_h^{j,n} - \frac{\Delta t}{\Delta x}(\hat{f}^{F,n}_{j+1/2} - \hat{f}^{F,n}_{j-1/2}) - \Delta t c \sum_{\ell=1}^{M} w_\ell \sigma_{\mathrm{t},\ell} F_\ell^{j,n}, \tag{3.77b}$$

$$C_v \overline{T}_h^{j,n+1} = C_v \overline{T}_h^{j,n} + \Delta t c \sum_{\ell=1}^{M} w_\ell \left[\sigma_{\mathrm{a},\ell} E_\ell^{j,n} - \sigma_{\mathrm{a},\ell} a (T_\ell^{j,n})^4\right], \tag{3.77c}$$

where the corresponding numerical fluxes at time t^n are the same as in (4.81):

$$\hat{f}^{E,n}_{j\pm1/2} = \frac{1}{2}\left[F^{+,n}_{j\pm1/2} + F^{-,n}_{j\pm1/2} - c\,(E^{+,n}_{j\pm1/2} - E^{-,n}_{j\pm1/2})\right], \tag{3.78}$$

$$\hat{f}^{F,n}_{j\pm1/2} = \frac{1}{2}\left[c^2 P^{+,n}_{j\pm1/2} + c^2 P^{-,n}_{j\pm1/2} - c\,(F^{+,n}_{j\pm1/2} - F^{-,n}_{j\pm1/2})\right] \tag{3.79}$$

and

$$P^{\pm,n}_{j\pm1/2} = \chi\left(E^{\pm,n}_{j\pm1/2}, F^{\pm,n}_{j\pm1/2}\right) E^{\pm,n}_{j\pm1/2}\,. \tag{3.80}$$

For a fixed interval I_j, we additionally denote for all $\ell = 1, \ldots, M-1$:

$$\hat{f}^{E,n}_{j,\ell+1/2} = \hat{f}^{E,n}(\mathbf{u}_h^j(\hat{x}_j^\ell), \mathbf{u}_h^j(\hat{x}_j^{\ell+1})) \tag{3.81}$$

and for the fluxes at the cell boundaries ($\ell = 0, M$):

$$\hat{f}^{E,n}_{j,1/2} = \hat{f}^{E,n}(\mathbf{u}_h^{j-1}(\hat{x}_{j-1}^M), \mathbf{u}_h^j(\hat{x}_j^1)) \quad \text{and} \quad \hat{f}^{E,n}_{j,M+1/2} = \hat{f}^{E,n}(\mathbf{u}_h^j(\hat{x}_j^M), \mathbf{u}_h^{j+1}(\hat{x}_{j+1}^1)),$$

with similar definitions for $\hat{f}^{F,n}_{j,\ell+1/2}$ and $\hat{f}^{F,n}_{j,\ell-1/2}$. Note that

$$\hat{f}^{E,n}_{j+1/2} - \hat{f}^{E,n}_{j-1/2} = \hat{f}^{E,n}_{j,M+1/2} - \hat{f}^{E,n}_{j,1/2} = \sum_{\ell=1}^{M}\left[\hat{f}^{E,n}_{j,\ell+1/2} - \hat{f}^{E,n}_{j,\ell-1/2}\right]. \tag{3.82}$$

As $E_h^{j,n}$, $F_h^{j,n}$, and $T_h^{j,n}$ are polynomial expansions, the cell averages in (3.77) can be written as exact quadrature formulas. This gives for E

$$\begin{aligned}\overline{E}_h^{j,n+1} &= \sum_{\ell=1}^{M} w_\ell E_\ell^{j,n} - \sum_{\ell=1}^{M}(1 - c\Delta t\sigma_{\mathrm{t},\ell}) w_\ell \left[\frac{\Delta t}{(1 - c\Delta t\sigma_{\mathrm{t},\ell}) w_\ell \Delta x}(\hat{f}^{E,n}_{j,\ell+1/2} - \hat{f}^{E,n}_{j,\ell-1/2})\right] \\ &\quad + \Delta t c \sum_{\ell=1}^{M} w_\ell \left(-\sigma_{\mathrm{a},\ell} E_\ell^{j,n} + \sigma_{\mathrm{a},\ell} a\,(T_\ell^{j,n})^4\right),\end{aligned} \tag{3.83}$$

Making use of $\sigma_{\mathrm{t},\ell} = \sigma_{\mathrm{s},\ell} + \sigma_{\mathrm{a},\ell}$ we get

$$\begin{aligned}\overline{E}_h^{j,n+1} &= \sum_{\ell=1}^{M} w_\ell (1 - c\Delta t \sigma_{\mathrm{t},\ell}) \left[E_\ell^{j,n} - \frac{\Delta t}{(1 - c\Delta t \sigma_{\mathrm{t},\ell}) w_\ell \Delta x} (\hat{f}^{E,n}_{j,\ell+1/2} - \hat{f}^{E,n}_{j,\ell-1/2}) \right] \\ &\quad + \Delta t c \sum_{\ell=1}^{M} w_\ell \left(\sigma_{\mathrm{s},\ell} E_\ell^{j,n} + \sigma_{\mathrm{a},\ell} a \, (T_\ell^{j,n})^4 \right). \end{aligned} \tag{3.84}$$

Similarly,

$$\overline{F}_h^{j,n+1} = \sum_{\ell=1}^{M} w_\ell (1 - c\Delta t \sigma_{\mathrm{t},\ell}) \left[F_\ell^{j,n} - \frac{\Delta t}{(1 - c\Delta t \sigma_{\mathrm{t},\ell}) w_\ell \Delta x} (\hat{f}^{F,n}_{j,\ell+1/2} - \hat{f}^{F,n}_{j,\ell-1/2}) \right] \tag{3.85}$$

$$C_v \overline{T}_h^{j,n+1} = \sum_{\ell=1}^{M} w_\ell \left(C_v - \Delta t \sigma_{\mathrm{a},\ell} a c (T_\ell^{j,n})^3 \right) T_\ell^{j,n} + \Delta t c \sum_{\ell=1}^{M} w_\ell \sigma_{\mathrm{a},\ell} E_\ell^{j,n}. \tag{3.86}$$

In terms of the variables Q_h and R_h,

$$\begin{aligned}\overline{Q}_h^{j,n+1} &= \sum_{\ell=1}^{M} w_\ell (1 - c\Delta t \sigma_{\mathrm{t},\ell}) \left[Q_{h,\ell}^{j,n} - \frac{\Delta t}{(1 - c\Delta t \sigma_{\mathrm{t},\ell}) w_\ell \Delta x} (\hat{f}^{Q,n}_{j,\ell+1/2} - \hat{f}^{Q,n}_{j,\ell-1/2}) \right] \\ &\quad + \sum_{\ell=1}^{M} \frac{w_\ell c^2 \Delta t}{2} (\sigma_{\mathrm{s},\ell} E_\ell^{j,n} + \sigma_{\mathrm{a},\ell} a (T_\ell^{j,n})^4) \end{aligned} \tag{3.87}$$

$$\begin{aligned}\overline{R}_h^{j,n+1} &= \sum_{\ell=1}^{M} w_\ell (1 - c\Delta t \sigma_{\mathrm{t},\ell}) \left[R_{h,\ell}^{j,n} - \frac{\Delta t}{(1 - c\Delta t \sigma_{\mathrm{t},\ell}) w_\ell \Delta x} (\hat{f}^{R,n}_{j,\ell+1/2} - \hat{f}^{R,n}_{j,\ell-1/2}) \right] \\ &\quad + \sum_{\ell=1}^{M} \frac{w_\ell c^2 \Delta t}{2} (\sigma_{\mathrm{s},\ell} E_\ell^{j,n} + \sigma_{\mathrm{a},\ell} a (T_\ell^{j,n})^4) \end{aligned} \tag{3.88}$$

where

$$\hat{f}^{Q,n}_{j,\ell\pm1/2} = \frac{c \hat{f}^{E,n}_{j,\ell\pm1/2} + \hat{f}^{F,n}_{j,\ell\pm1/2}}{2} \quad \text{and} \quad \hat{f}^{R,n}_{j,\ell\pm1/2} = \frac{c \hat{f}^{E,n}_{j,\ell\pm1/2} - \hat{f}^{F,n}_{j,\ell\pm1/2}}{2}. \tag{3.89}$$

Plugging (3.81) into (3.89) gives

$$\begin{aligned}\hat{f}^{Q,n}_{j,\ell+1/2} - \hat{f}^{Q,n}_{j,\ell-1/2} &= \frac{c}{2} \left(\hat{f}^{E,n}_{j,\ell+1/2} - \hat{f}^{E,n}_{j,\ell-1/2} \right) + \frac{1}{2} \left(\hat{f}^{F,n}_{j,\ell+1/2} - \hat{f}^{F,n}_{j,\ell-1/2} \right) \\ &= \frac{c}{4} \left[F_{\ell+1}^{j,n} - F_{\ell-1}^{j,n} - c \left(E_{\ell+1}^{j,n} - 2E_\ell^{j,n} + E_{\ell-1}^{j,n} \right) \right] \\ &\quad + \frac{1}{4} \left[c^2 \left(P_{\ell+1}^{j,n} - P_{\ell-1}^{j,n} \right) - c \left(F_{\ell+1}^{j,n} - 2F_\ell^{j,n} + F_{\ell-1}^{j,n} \right) \right]. \end{aligned} \tag{3.90}$$

and

$$\begin{aligned}\hat{f}^{R,n}_{j,\ell+1/2} - \hat{f}^{R,n}_{j,\ell-1/2} &= \frac{c}{2} \left(\hat{f}^{E,n}_{j,\ell+1/2} - \hat{f}^{E,n}_{j,\ell-1/2} \right) + \frac{1}{2} \left(\hat{f}^{F,n}_{j,\ell+1/2} - \hat{f}^{F,n}_{j,\ell-1/2} \right) \\ &= \frac{c}{4} \left[F_{\ell+1}^{j,n} - F_{\ell-1}^{j,n} - c \left(E_{\ell+1}^{j,n} - 2E_\ell^{j,n} + E_{\ell-1}^{j,n} \right) \right] \\ &\quad + \frac{1}{4} \left[c^2 \left(P_{\ell+1}^{j,n} - P_{\ell-1}^{j,n} \right) - c \left(F_{\ell+1}^{j,n} - 2F_\ell^{j,n} + F_{\ell-1}^{j,n} \right) \right]. \end{aligned} \tag{3.91}$$

Substituting (3.90) and (3.91) into (3.87) and (3.88), respectively, yields

$$\overline{Q}_h^{j,n+1} = \sum_{\ell=1}^{M} w_\ell(1-c\Delta t\sigma_{\mathrm{t},\ell})C_{j,\ell}^{Q,n} + \sum_{\ell=1}^{M} \frac{w_\ell c^2\Delta t}{2}\left(\sigma_{\mathrm{s},\ell}E_\ell^{j,n} + \sigma_{\mathrm{a},\ell}a(T_\ell^{j,n})^4\right) \tag{3.92}$$

$$\overline{R}_h^{j,n+1} = \sum_{\ell=1}^{M} w_\ell(1-c\Delta t\sigma_{\mathrm{t},\ell})C_{j,\ell}^{R,n} + \sum_{\ell=1}^{M} \frac{w_\ell c^2\Delta t}{2}\left(\sigma_{\mathrm{s},\ell}E_\ell^{j,n} + \sigma_{\mathrm{a},\ell}a(T_\ell^{j,n})^4\right) \tag{3.93}$$

where

$$C_{j,\ell}^{Q,n} := (1-c\delta_\ell)Q_{h,\ell}^{j,n} + \frac{c^2\delta_\ell}{4}\left[E_{\ell+1}^{j,n} - P_{\ell+1}^{j,n}\right] + \frac{c\delta_\ell}{4}\left[cE_{\ell-1}^{j,n} + 2F_{\ell-1}^{j,n} + cP_{\ell-1}^{j,n}\right] , \tag{3.94}$$

$$C_{j,\ell}^{R,n} := (1-c\delta_\ell)R_{h,\ell}^{j,n} + \frac{c^2\delta_\ell}{4}\left[E_{\ell-1}^{j,n} - P_{\ell-1}^{j,n}\right] + \frac{c\delta_\ell}{4}\left[cE_{\ell+1}^{j,n} - 2F_{\ell+1}^{j,n} + cP_{\ell+1}^{j,n}\right] . \tag{3.95}$$

and

$$\delta_\ell := \frac{\Delta t}{(1-c\Delta t\sigma_{\mathrm{t},\ell})w_\ell\Delta x} > 0\,. \tag{3.96}$$

The positivity of δ_ℓ follows from assumption *(A1)*.

It is clear from (3.92) and (3.93) that positivity of $C_{j,\ell}^{Q,n}$ and $C_{j,\ell}^{R,n}$ for $\ell = 1,\dots,M$ implies that of $\overline{Q}_h^{j,n+1}$ and $\overline{R}_h^{j,n+1}$. The first terms in (3.94) and (3.95) are positive by the assumption *(A2)*. The second terms are positive by (3.22) in Lemma 3. The third terms are positive by (3.23) in Lemma 3. We conclude that $\overline{Q}_h^{j,n+1}$ and $\overline{R}_h^{j,n+1}$ are positive and hence $\overline{\mathbf{u}}_h^{j,n+1} \in \mathcal{R}_2$. Finally, by assumption *(A3)*, we have that $C_v - \Delta t\sigma_{\mathrm{a},\ell}ac(T_\ell^{j,n})^3 \geq 0$, so that $\overline{T}_h^{j,n+1}$ is also positive. This completes the proof. □

Remark 14. *The limiting of the material temperature T is not a critical part of the proof. Indeed, it is needed only to ensure that the cross-sections are well-defined: depending on the formula for σ_{a} and σ_{s}, a negative temperature may result in a negative cross-section. However, the cross-sections considered in our numerical experiments are smooth and bounded well away from zero. As a practical matter, the limiting of T is not necessary in these cases.*

Corollary 1. *The Runge-Kutta discontinuous Galerkin scheme which combines the space discretization in* (4.84a) *and* (3.51)*, the limiting in* (3.69)*, and a strong-stability-preserving Runge Kutta time integrator perserves the realizability of the moments and the positivity of the material temperature in the sense of cell averages. In particular, if the time step conditions (A1)-(A3) in the statement of Theorem 3 hold and if*

$$\overline{\mathbf{u}}_h^{j,n} \in \mathcal{R}_2 \quad \textit{and} \quad \overline{T}_h^{j,n} \geq 0\,, \tag{3.97}$$

then

$$\overline{\mathbf{u}}_h^{j,n+1} \in \mathcal{R}_2 \quad \textit{and} \quad \overline{T}_h^{j,n+1} \geq 0\,. \tag{3.98}$$

Proof. Application of the limiters in (3.69) ensures that the conditions of Theorem 3 holds at each stage in the SSP-RK scheme. Each successive stage is an application of the forward Euler operator to the current stage with an appropriately modified time step. Thus the conclusions of Theorem 3 apply at the next stage, including the final stage, which gives $\mathbf{u}_h^{j,n+1}$ and $T_h^{j,n+1}$. □

In order to obtain both slope limited and realizable states we first apply the slope limiter for stability (from Section 3.3.2) and then apply the realizability-preserving limiter. Numerical experiments indicate that our realizability limiter does not produce larger slopes than those allowed by the stability slope limiter, but we have no proof of this fact. However, we can at least show that the realizability limiter does not increase the L^p norm of the limited quantities and is therefore stable:

Proposition 1. *For any p, $1 \leq p \leq \infty$, the mappings from $Q_h^{j,n} \mapsto \hat{Q}_h^{j,n}$, $R_h^{j,n} \mapsto \hat{R}_h^{j,n}$, and $T_h^{j,n} \mapsto \hat{T}_h^{j,n}$ are stable in the L^p norm, i.e.,*

$$||\hat{Q}_h^{j,n}||_{L^p} \leq ||Q_h^{j,n}||_{L^p}\,, \quad ||\hat{R}_h^{j,n}||_{L^p} \leq ||R_h^{j,n}||_{L^p} \quad \textit{and} \quad ||\hat{T}_h^{j,n}||_{L^p} \leq ||T_h^{j,n}||_{L^p}\,. \tag{3.99}$$

Proof. We give the proof for Q; the proofs for R and T use the same argument. Applying the triangle inequality to the definition of $\hat{Q}_h^{j,n}$ in (3.66) gives

$$||\hat{Q}_h^{j,n}||_{L^p} \leq \theta_Q^{j,n}||Q_h^{j,n}(x)||_{L^p} + (1-\theta_Q^{j,n})||\overline{Q}_h^{j,n}||_{L^p} \tag{3.100}$$

Meanwhile, Jensen's inequality gives

$$||\overline{Q}_h^{j,n}||_{L^p} \leq ||Q_h^{j,n}||_{L^p}. \tag{3.101}$$

Combining (3.100) and (3.101) gives the desired result. □

3.5 Numerical Experiments

In this section several numerical results are presented to confirm the analytically derived predictions and explore the behavior of the developed realizability limiter in challenging test cases. For the sake of simplicity, we use equidistant space discretizations of meshsize h. Moreover, the M-point Legendre Gauss-Lobatto quadrature on $[-1,1]$ is chosen for numerical integration. It is exact for the integral of polynomials of degree up to $2M-3$, i.e., M is the smallest integer such that $2M-3 \geq 2k+1$ [32]. Our algorithms are implemented in MATLAB.

Whenever a DG space discretization of polynomials up to degree $k=2$ is applied, the SSPRK(3,3) method from (4.88) is used. For $k=3$ we extend the order of the time-marching scheme and use the fourth order SSPRK(5,4) method from (4.89).

Our examples in Sections 3.5.1–3.5.5 are designed to solve the M_1 system (3.32) for $a=0$. Until we couple to the temperature equation in Section 3.5.6, we additionally set $c=1$ and the numerical stability parameter needed for the realizability limiter to

$$\varepsilon = \min_{j=1,\dots,N}\{10^{-12}, \overline{Q}_h^{j,n}, \overline{R}_h^{j,n}\}.$$

The restrictions on the time step from theorem 3 in section 3.4 are realized in our algorithm in the following way: We denote

$$w_{\max} = \max_{\ell=1,\dots,M} w_\ell = \begin{cases} \frac{4}{3}, & \text{SSPRK(3,3)} \\ \frac{32}{45}, & \text{SSPRK(5,4)} \end{cases}, \quad w_{\min} = \min_{\ell=1,\dots,M} w_\ell = \begin{cases} \frac{1}{3}, & \text{SSPRK(3,3)} \\ \frac{1}{10}, & \text{SSPRK(5,4)} \end{cases}$$

and set the time step to

$$\Delta t < \min\{c_1, c_2, c_3\}\,, \tag{3.102}$$

where

$$c_1 = \frac{1}{c\sigma_{\mathrm{t,max}}}\,, \quad c_2 = \frac{\Delta x w_{\min}}{c(1+w_{\max}\sigma_{\mathrm{t,max}}\Delta x)}\,, \quad \text{and} \quad c_3 = \frac{C_v}{ac\tau_{\max}}\,. \tag{3.103}$$

The values $\sigma_{\mathrm{t,max}}$ and $\tau_{\max}$ are the maximum values of $\sigma_{\mathrm{t},\ell}$, and $\sigma_{\mathrm{a},\ell}{T_\ell^{j,n}}^3$, respectively.

To diminish computational costs, the transformation to characteristic variables in the slope limiting procedure will only be performed if explicitly stated. Otherwise it is neglected.

When our realizability limiter is turned off, it often occurs that we leave the realizability region. However, one still needs to evaluate the Eddington factor in (3.2) which might become imaginary. To push the states into the realizability domain, we replace χ by a cut-off value $\tilde{\chi}(E,F)$:

$$\tilde{\chi}(E,F) = \begin{cases} \chi(E,F), & \text{if } E \text{ and } F \text{ are realizable} \\ \chi_{\max}, & \text{if } E \text{ or } F \text{ are not realizable,} \end{cases} \tag{3.104}$$

The value of $\chi_{\max}$ is set to either one (the maximum physical value of χ) or 5/3 (the maximum mathematical value of χ). If not explicitly mentioned we set $\chi_{\max} = 1$.

3.5.1 Accuracy Tests

We begin with accuracy tests of the method for orders $k = 1, 2, 3$. The convergence analysis in [32] shows that, under certain conditions, the described RKDG schemes can reach convergence orders of $k + 1$ when polynomials of degree at most k are used to approximate the solution. We neglect the material energy equation (3.34) and solve the M_1 system (3.32) for $c = 1$, $\sigma_a = 0$ and $\sigma_t = 5$, and $T = 0$. Smooth initial conditions:

$$\mathbf{u}_0(x) = [2(\sin(2\pi x) + 2), \sin(2\pi x) + 2]^T, \quad \text{on } x \in I = (-1/2, 1/2),$$

as well as periodic boundary conditions are enforced. As reference solution we always use an unlimited solution calculated with $J = 1000$ discretization cells and polynomials of degree $k = 3$. To study the effects and interaction between slope and realizability limiters, we compare the convergence rates for different combinations of limiters being turned on or off and applied with or without transformation to characteristic variables.

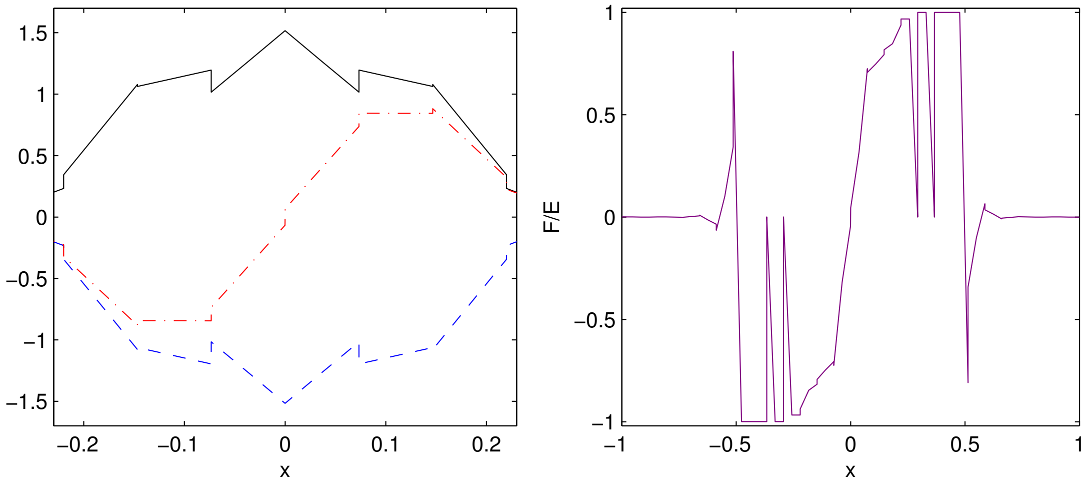
Tables 3.1 and 3.2 contain numerical order of accuracies calculated in the L^1- and L^∞-norm at $t_{\mathrm{final}} = 0.1$. The unlimited case, ignoring any limiting procedure, shows the expected accuracy order for both norms. Switching on only the slope limiter in the state variables, the L^1-error also achieves the analytically predicted order (except for negligible deviations at $h = 1/20$). Limiting the slopes in the characteristic system keeps the L^1-order and does not generate significant changes. However, we observe a loss of convergence order in the L^∞-norm as soon as the slope limiter is turned on which coincides with results in [25]. Note that the initial condition $\mathbf{u}_0(x)$ is chosen to keep the moment system far away from the realizability boundary. Hence, the realizability limiter is not supposed to turn on at all. Indeed, we observe that in all cases, the realizability limiter does not affect the convergence rate in the L^1- or L^∞-norm.

		$\mathcal{P}^1$	$\mathcal{P}^2$	$\mathcal{P}^3$
	h	order	order	order
unlimited	1/10	–	–	–
	1/20	2.05	3.15	4.00
	1/40	2.02	2.96	4.07
	1/80	2.01	2.93	4.01
	1/160	2.01	2.96	4.12
	1/320	2.00	2.98	4.00
realiz. limited	1/10	–	–	–
	1/20	2.05	3.15	4.00
	1/40	2.02	2.96	4.07
	1/80	2.01	2.93	4.01
	1/160	2.01	2.96	4.12
	1/320	2.00	2.98	4.00
slope limited	1/10	–	–	–
	1/20	2.70	2.70	3.88
	1/40	2.11	3.23	4.24
	1/80	1.94	3.20	4.08
	1/160	2.20	3.17	4.22
	1/320	2.13	3.24	4.06
slope + realiz. limited	1/10	–	–	–
	1/20	2.70	2.70	3.88
	1/40	2.11	3.23	4.24
	1/80	1.94	3.20	4.08
	1/160	2.20	3.17	4.22
	1/320	2.13	3.24	4.06
charact. slope limited	1/10	–	–	–
	1/20	2.39	2.75	4.19
	1/40	1.97	2.94	4.04
	1/80	2.24	3.01	4.10
	1/160	2.17	3.17	4.12
	1/320	2.17	3.37	4.01
charact. slope + realiz. limited	1/10	–	–	–
	1/20	2.39	2.75	4.19
	1/40	1.97	2.94	4.04
	1/80	2.24	3.01	4.10
	1/160	2.17	3.17	4.12
	1/320	2.17	3.37	4.01

Table 3.1: L^1-Order of Accuracy for E.

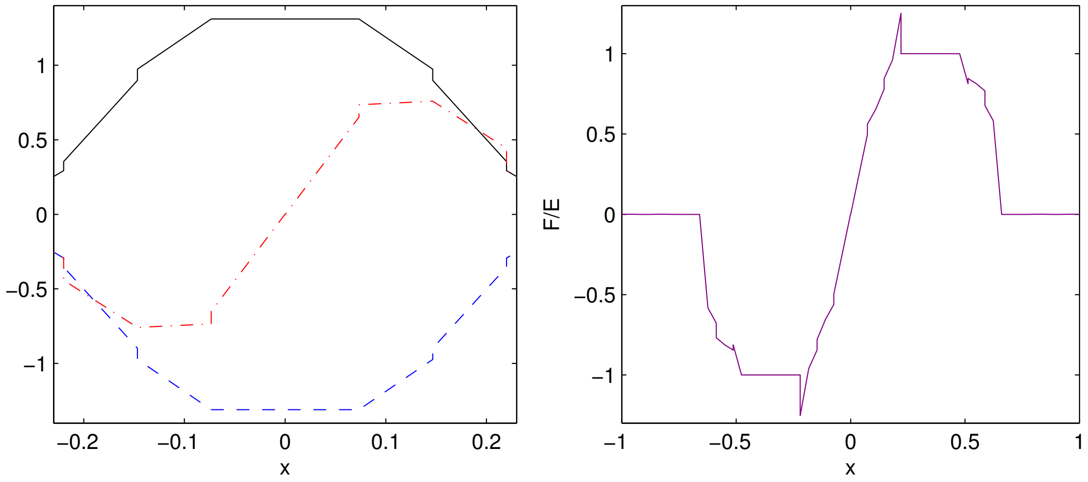
		$\mathcal{P}^1$	$\mathcal{P}^2$	$\mathcal{P}^3$
	h	order	order	order
unlimited	1/10	–	–	–
	1/20	2.00	3.15	4.17
	1/40	2.00	2.98	4.07
	1/80	2.00	2.94	3.98
	1/160	2.00	2.96	3.97
	1/320	2.00	2.98	4.00
realiz. limited	1/10	–	–	–
	1/20	2.00	3.15	4.17
	1/40	2.00	2.98	4.07
	1/80	2.00	2.94	3.98
	1/160	2.00	2.96	3.97
	1/320	2.00	2.98	4.00
slope limited	1/10	–	–	–
	1/20	2.53	1.64	2.95
	1/40	1.73	2.88	4.61
	1/80	1.75	2.95	3.37
	1/160	1.94	3.18	4.25
	1/320	2.01	3.01	3.75
slope + realiz. limited	1/10	–	–	–
	1/20	2.53	1.64	2.95
	1/40	1.73	2.88	4.61
	1/80	1.75	2.95	3.37
	1/160	1.94	3.18	4.25
	1/320	2.01	3.01	3.75
charact. slope limited	1/10	–	–	–
	1/20	2.16	1.59	4.10
	1/40	1.36	2.86	3.26
	1/80	1.88	1.95	3.70
	1/160	1.50	2.92	3.40
	1/320	1.65	2.69	3.27
charact. slope + realiz. limited	1/10	–	–	–
	1/20	2.16	1.59	4.10
	1/40	1.36	2.86	3.26
	1/80	1.88	1.95	3.70
	1/160	1.50	2.92	3.40
	1/320	1.65	2.69	3.27

Table 3.2: L^∞-Order of Accuracy for E.



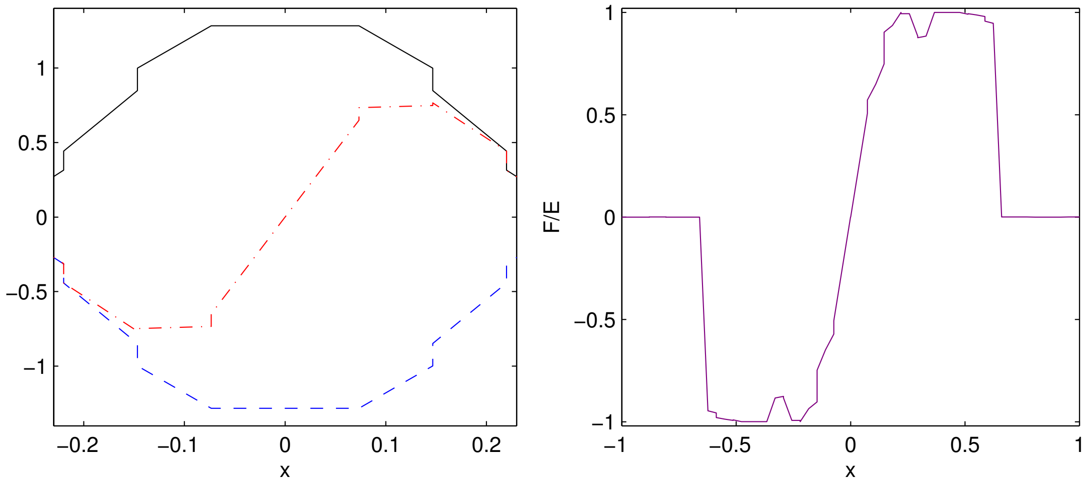
(a) Clipping: Realiz. limited E, -E, F

(b) Realiz. limited F/E



(c) Clipping: Slope limited E, -E, F

(d) Slope limited F/E



(e) Clipping: Slope + realiz. limited E, -E, F

(f) Slope + realiz. limited F/E

Figure 3.2: Under-resolved plane source solution at $t_{\text{final}} = 0.3$, polynomial degree $k = 1$, $J = 30$: E (black solid line), $-E$ (blue dashed line), F (red dash-dot line), F/E (purple solid line).

3.5.2 Under-resolved Plane Source

When particles are emitted from an initial strong and locally restricted source into an infinite medium, discontinuities are formed at the boundary of the source. Extremely sharp sources are often used to analyze the behavior of approximation models or numerical methods near severe discontinuities. Such problems serve as a tough test of whether the DG method can preserve realizability. Very steep slopes arise where the DG slope limiter changes the solution to avoid oscillations. Without the new realizability limiter it is therefore likely that the scheme violates the realizability conditions in these regions.

We solve (3.32) with $c = 1$ and $T = 0$ in a purely absorbing medium with $\sigma_a = 0.5, \sigma_s = 0$. We represent the delta function in the initial condition:

$$\mathbf{u}_0(x) = \begin{cases} [\frac{1}{2h}, 0]^T , & x \in [-h, h] , \\ [2\varepsilon, 0]^T & \text{otherwise} . \end{cases}$$

Despite the infinite medium, boundary conditions are needed for the numerical scheme. We enforce periodic boundary conditions and compute the solution until $t_{\text{final}} = 0.3$ on $[-2, 2]$ which is large enough to make sure that no particles from the boundaries incorrectly affect the solution in the domain.

Figure 3.2 shows the comparison between different computations at $t = 0.3$, with polynomial degree $k = 1$ and $J = 30$ cells. In each cell j the polynomial reconstruction is evaluated at both end points $x^{\mp}_{j\pm1/2}$ as well as the mid point x_j. We plot both E and its negative counterpart $-E$ to indicate the boundaries of the realizability region. We also plot the ratio F/E which, according to (4.1), should lie in the interval $[-1, 1]$.

Running the calculations without the realizability limiter (even with limited slopes) immediately produces results away from the realizability domain. Precisely, in *all* time steps it is necessary to evaluate the Eddington factor as mentioned in (3.104), which we do with $\chi_{\max} = 1$ Although slope limited, results in Figures 3.2(c)-(d) still demonstrate solutions violating (4.1). When the realizability limiter is turned on, no violations occur.

3.5.3 Riemann Problem

A Riemann problem is designed to additionally test our method and compare it to an analytical solution worked out in [36]. Approximative solutions for (3.32) are calculated where $c = 1$, $a = 0$ and $\sigma_s = 0 = \sigma_a$. The set-up for initial and boundary conditions is determined by

$$\mathbf{u}_0(x) = \begin{cases} [1, 0.9999]^T, & x \in (-0.05, 0], \\ [0.5, 0]^T, & x \in (0, 0.1], \end{cases}$$
$$\mathbf{u}(0, t) = [1, 0.9999]^T, \quad \mathbf{u}(1, t) = [0.5, 0]^T, \quad t > 0.$$

Figure 3.3 illustrates realizability and slope + realizability limited results for piecewise affine linear reconstructions on each cell. Although the purely realizability limited result remains within the realizability region, spurious oscillations can be observed. In fact, the solution with both realizability and slope limiters turned on is very close to the reference and captures both shocks pretty well. One can also see that no wriggles

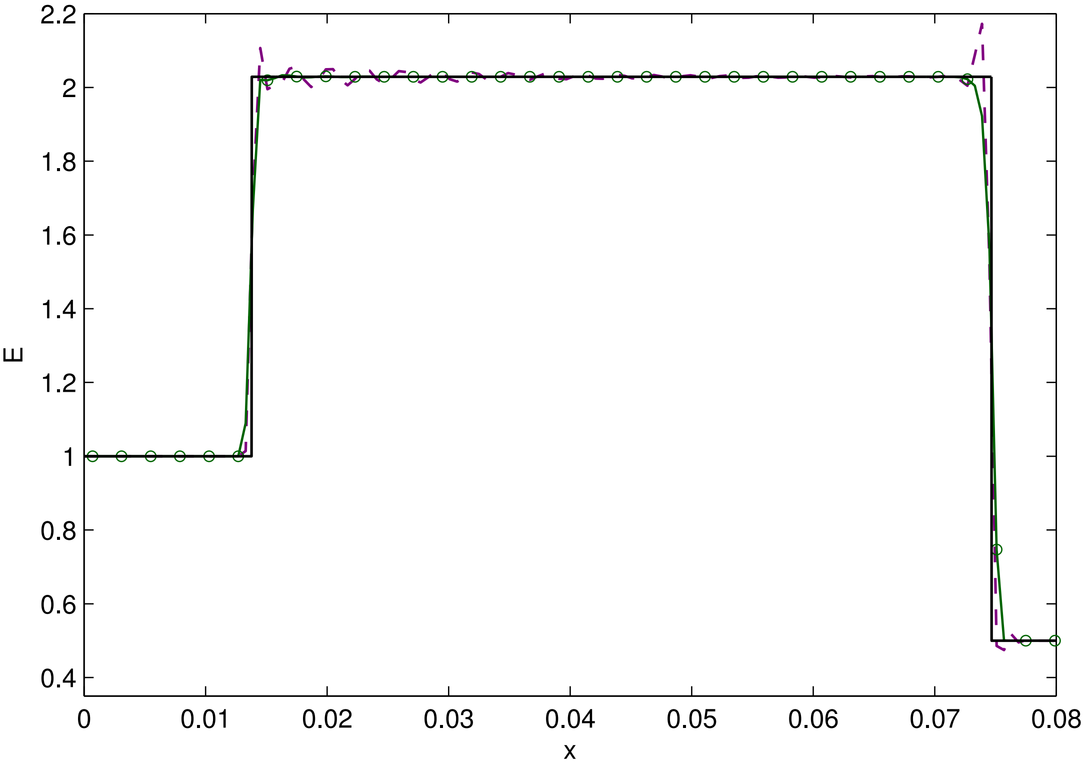


Figure 3.3: Riemann problem at $t_{\text{final}} = 0.1$, $J = 250$, $k = 1$: realiz. limited (purple dashed line), slope + realiz. limited (green circle line), exact (black solid line)

appear because we limit the slope in the characteristic variables which coincides with observations in [34].

The completely unlimited solution is shown in Figure 3.4: It oscillates at the right discontinuity, stops at $x \approx 0$ and is not moving to the right. The reason might be that we manipulate the transport term in the PDE by introducing the cutoff in (3.104) with $\chi_{\max} = 1$. A similar behavior is developed by the merely slope limited result which additionally forms wriggles where it should be a straight line. This is due to the fact that the slope limiter is applied to state variables. As its solution is not realizable the transformation matrix becomes either complex or undefined and hence, the transformation to characteristic variables is not possible.

However, extending the admissibility of E and F to the domain of χ where it remains real, i.e., $F/(cE) \leq 2/\sqrt{3}$, saves at least the slope limited result: By choosing the cutoff in (3.104) as $\chi_{\max} = 5/3$ and performing the limiting in characteristic variables, the solution is transported further to the right and gets close to the analytic solution. In fact, we observed that limiting in characteristic variables was the key to such a good slope limited solution. In this case, evaluating χ at $F/(cE) \leq 2/\sqrt{3}$ improves the behavior of the solution in the sense that both eigenvalues of the system remains real and distinct. That makes the transformation to characteristic variables possible. Although in this special case this cutoff for χ is favorable, there is no guarantee that it works in general because the hyperbolicity of the system cannot be ensured and the M_1 model is ill-posed.

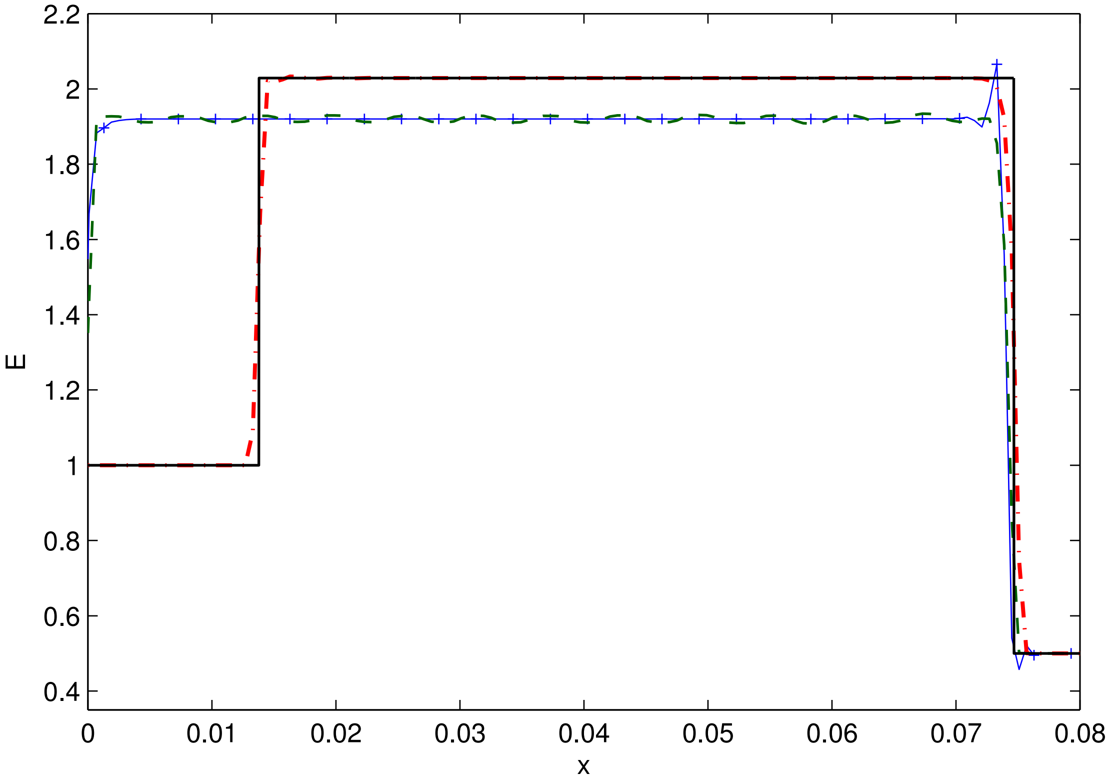


Figure 3.4: Riemann problem at $t_{\text{final}} = 0.1$, $J = 250$, $k = 1$: unlimited (blue plus line), slope limited with $\chi_{\max} = 1$ (green dashed line), slope limited with $\chi_{\max} = 5/3$ (red dash-dot line), exact (black solid line)

3.5.4 Heated Wall

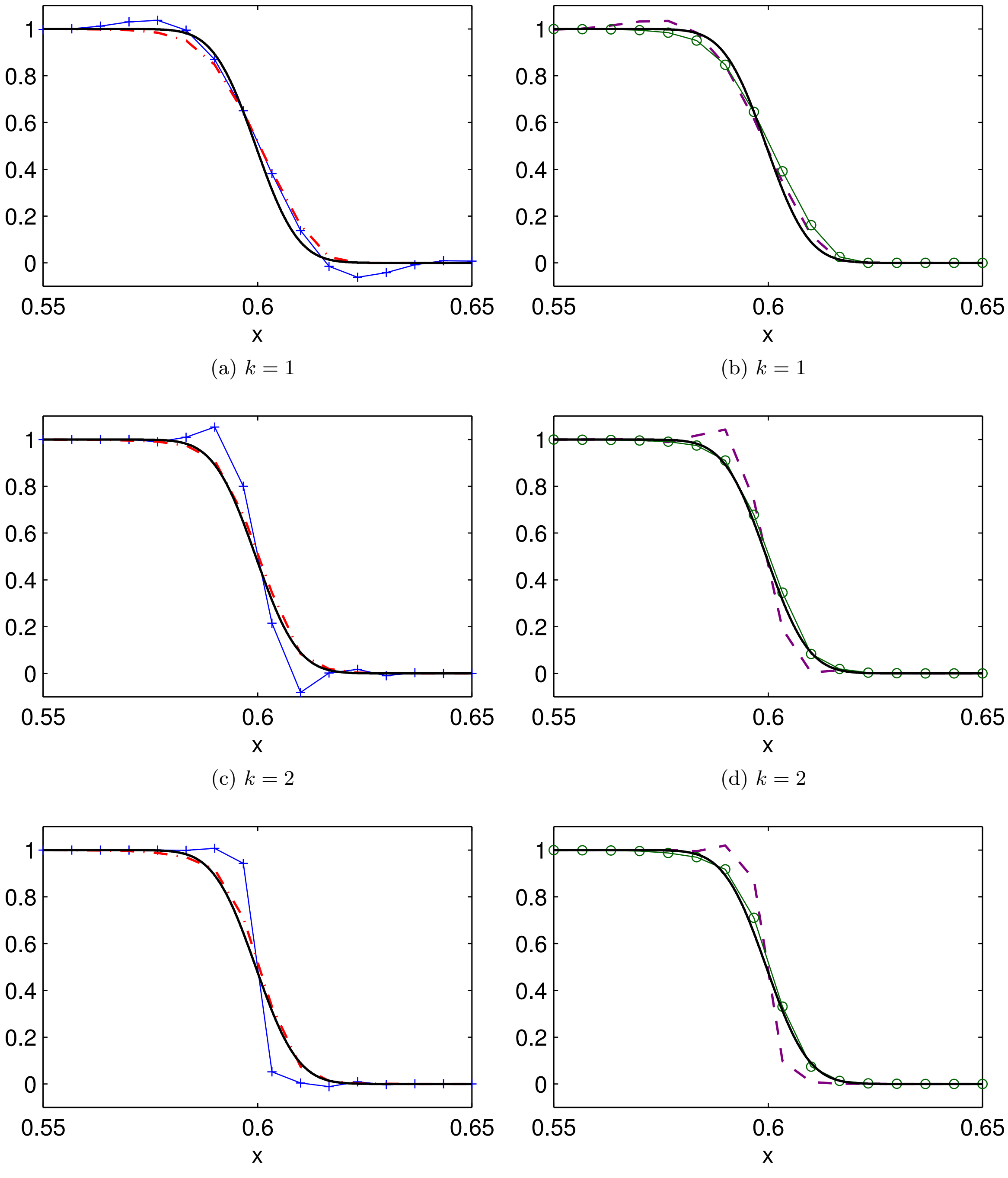
In the next test problem, photons are emitted at the left boundary in a right-ingoing beam and propagate through vacuum with $\sigma_a = 0 = \sigma_s$. At the right boundary no particles are entering the domain. Taking into account the stability parameter ε, we implement the following boundary conditions:

$$\begin{aligned} \mathbf{u}_0(x) &= [2\varepsilon, 0]^T, \quad x \in (0,1), \\ \mathbf{u}(0,t) &= [1, 1-2\varepsilon]^T, \quad \mathbf{u}(1,t) = [2\varepsilon, 0]^T, \quad t > 0. \end{aligned}$$

Again, the corresponding equation is (3.32) where $c = 1$. As a comparison, a reference solution for the transport equation is calculated by a discrete ordinates method with 10^5 discretization points in space and 256 in the angular variable μ. Values on the left boundary for each ordinate μ are taken from an extremely sharp Gaussian distribution, in order to mimic the forward-peaked incoming beam of photons.

Figures 3.5(a)-(c) show the scalar flux E for different reconstructions $k = 1, 2, 3$ evaluated at x_j in each cell. In vacuum, all emitted particles simply propagate at finite speed $c = 1$ without any interaction which leads to the characteristic step function form of the scalar flux. The slopes at $x \approx 0.6$ steepen with increasing k and, except for the unlimited results, $k = 2$ and $k = 3$ capture the shock very well.

When neglecting the slope limiter, both solutions with and without the realizability limiter overshoot the transport solution and become bigger than 1. Whereas the



(a) $k = 1$

(b) $k = 1$

(c) $k = 2$

(d) $k = 2$

(e) $k = 3$

(f) $k = 3$

Figure 3.5: Particles propagating through vacuum – zoomed plots of E at $t_{\text{final}} = 0.6$, $J = 150$: unlimited (blue plus line), slope limited (red dash-dot line), realiz. limited (purple dashed line), slope + realiz. limited (green circle line), transport solution (black solid line)

completely unlimited function allows the energy density E to become negative and to highly oscillate for $k = 3$, the realizability limited solution stays realizable. Although forming small jags near $x \approx 0.6$, solutions with merely the realizability limiter always show steeper slopes. It should be mentioned here that one can eliminate the large oscillations caused by the unlimited computations for $k = 3$ by increasing the number of quadrature points by one. This might be due to the fact that our cutoff in (3.104) diminishes the regularity of the flux function $f(\mathbf{u}(x, t))$ so that the assumption in [32] is not fulfilled.

For all k, the slope limited and slope + realizability limited solutions are indistinguishable to the eye; both closely approximate the transport solution. However, it should be emphasized that all simulations without realizability limiter do violate the realizability condition (4.1) throughout the whole computation.

3.5.5 Two Opposing Beams

A major drawback of the minimum-entropy closure is the inability to simulate particles moving in opposing directions in a physically realistic manner [77]. At steady-state, the M_1 solution, (Figure 3.6) generates an unphysical shock in the profile of E. Except for the unlimited solution, which again produces oscillations, the remaining results are very close to each other.
The precise parameters in our implementation are:

$$\begin{aligned} \sigma_a &= 4, \quad \sigma_s = 0, \\ \mathbf{u}_0(x) &= [2\varepsilon, 0]^T, \quad x \in (-0.5, 0, 5), \\ \mathbf{u}(0,t) &= [1, 0.9999]^T, \quad \mathbf{u}(1,t) = [1, -0.9999]^T, \quad t > 0. \end{aligned}$$

For comparison, a transport solution is calculated by the discrete ordinates method with 256 discretization points in angle and 1000 points in space.

3.5.6 Coupling to the Material Energy Equation

Our next example focuses on problems coupling to the temperature equation. We consider first a contrived example and then a modified version of the so-called Marshak wave problem (first studies in [119, 146]). Originally derived to produce a semi-analytical solution, the Marshak wave is based on the assumption that the material and radiation fields are in equilibrium. Here, solutions are computed allowing a non-equilibrium between the fields. The radiative transfer is modelled as a grey process in (3.32) and is coupled to the material energy balance equation in (3.34). We solve (3.32) and (3.34) with the physical constants

$$\begin{aligned} c &= 3 \cdot 10^{10}\,\mathrm{cm/s} && \text{speed of light,} \\ a &= 1.372 \cdot 10^{14}\,\mathrm{erg/(cm^3 keV^4)} && \text{radiation constant,} \\ C_v &= 3 \cdot 10^{15}\,\mathrm{erg/(cm^3 keV)} && \text{heat capacity.} \end{aligned}$$

Due to a different scale, the numerical stability parameter for the realizability limiter is set to

$$\varepsilon = \min_{j=1,\dots,N} \{10^{-12} ca, \overline{Q}_h^{j,n}, \overline{R}_h^{j,n}\}.$$

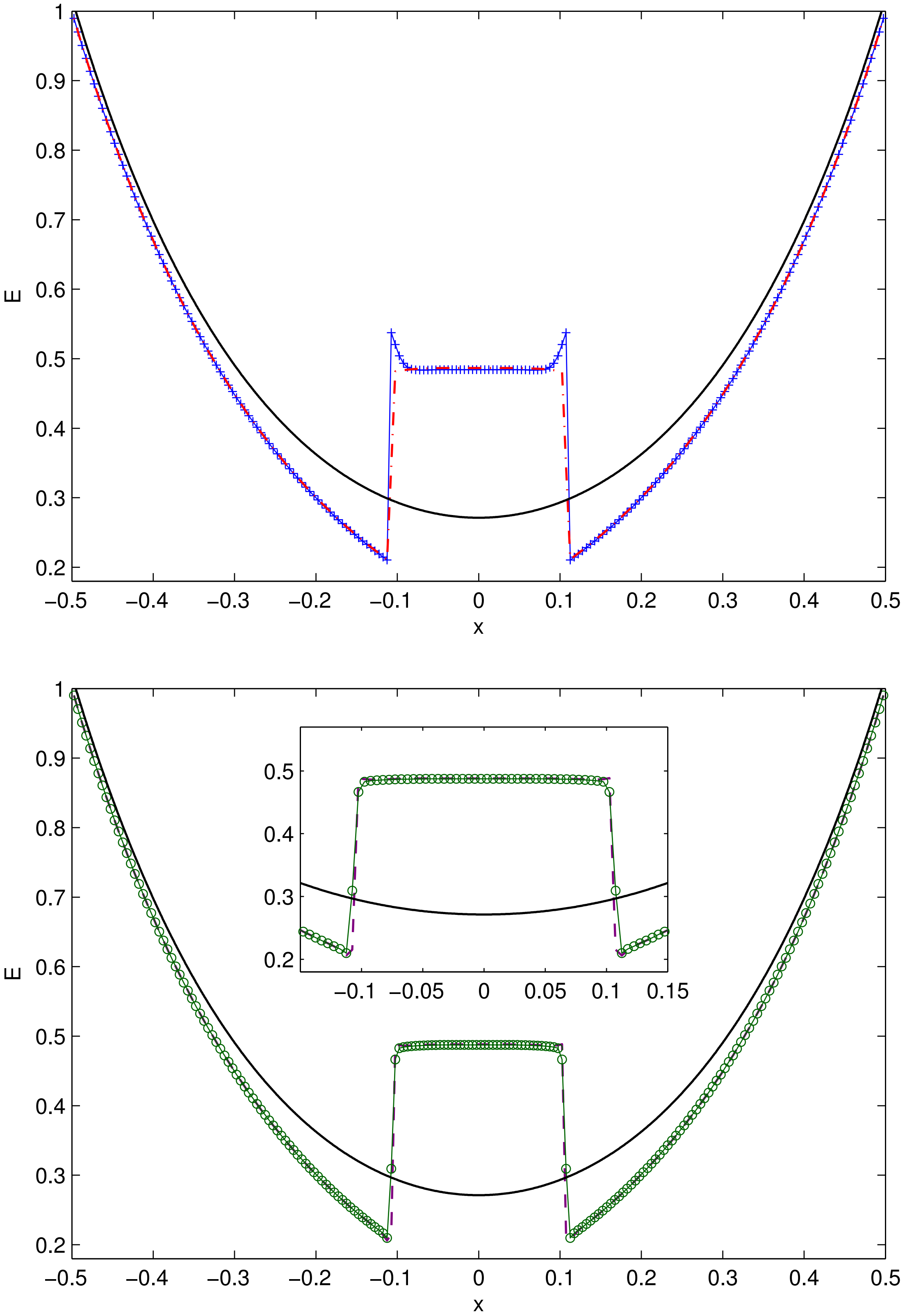


Figure 3.6: Two beams: $t_{\text{final}} = 3$, $J = 200$, $k = 2$: unlimited (blue plus line), slope limited (red dash-dot line), realiz. limited (purple dashed line), slope + realiz. limited (green circle line), transport solution (black solid line)

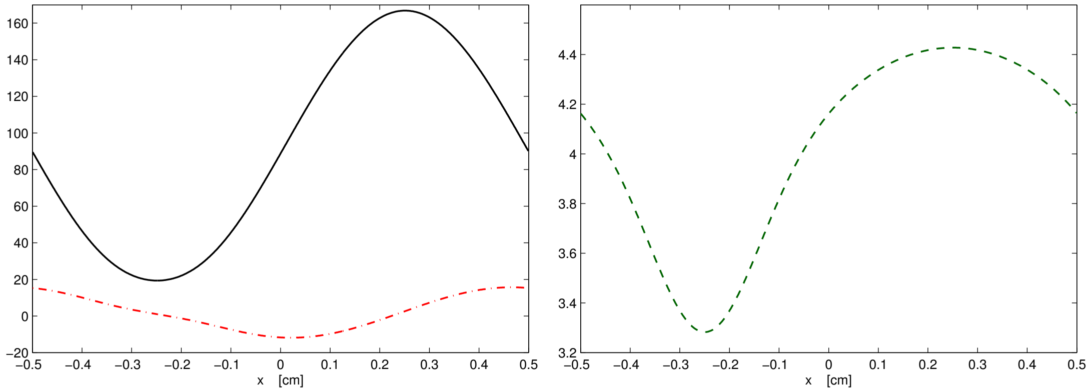


Figure 3.7: Accuracy test including energy equation: $t_{\text{final}} = 0.1\text{ns}$, $k = 3$, $J = 800$, slope+realiz. limited: $a^{-1}E$ (black solid line), $(ac)^{-1}F$ (red dash-dot line), T (green dashed line)

Accuracy Tests

In Section 3.5.1 expected convergence rates up to fourth order could be reached for the M_1 system (3.1). It is now important to make sure that this is also true when coupling to the energy equation (3.3), which introduces a nonlinear source term T^4.

To set up an accuracy test, we choose a spatial domain $[-1/2, 1/2]$ and impose periodic boundary conditions. Smooth initial conditions are chosen as follows:

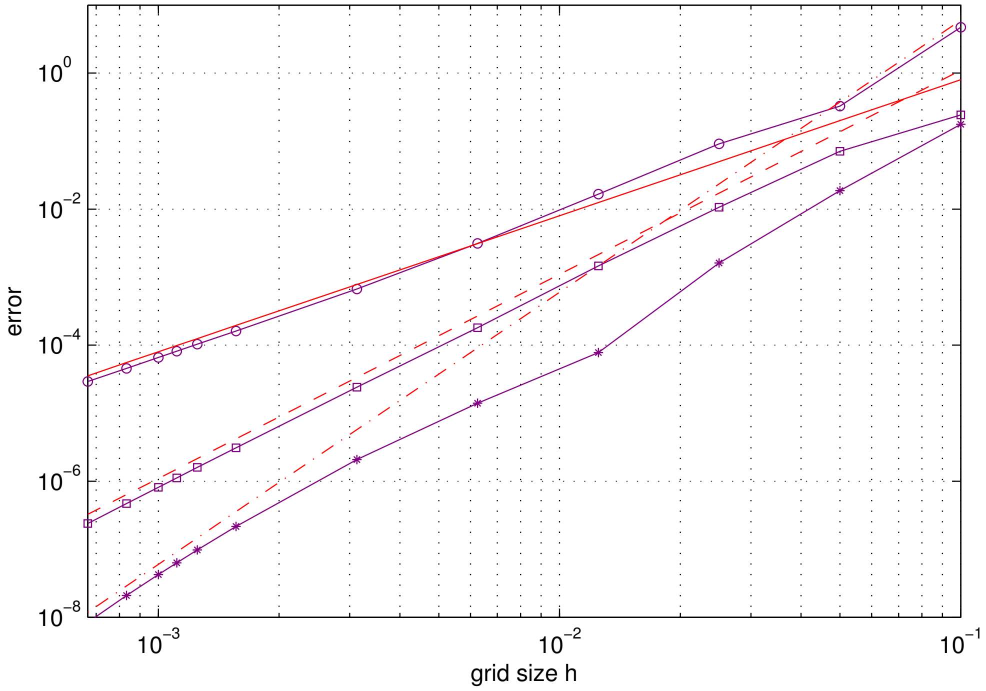
$$\begin{aligned} T_0(x) &= 4(\sin(2\pi x) + 2), && x \in (-1/2, 1/2), \\ \mathbf{u}_0(x) &= [2a(\sin(2\pi x) + 2),\ ac(\sin(2\pi x) + 2)]^T, && x \in (-1/2, 1/2). \end{aligned}$$

Particles travel through a homogeneous and absorbing medium where cross sections are set to $\sigma_{\text{s}} = 0$ and $\sigma_{\text{a}} = 1$. Solutions for $k = 3$ are displayed in Figure 3.7 for different quantities. They are all smooth and realizable, with no steep slopes or discontinuities. Thus, none of the limiters should modify the solution.

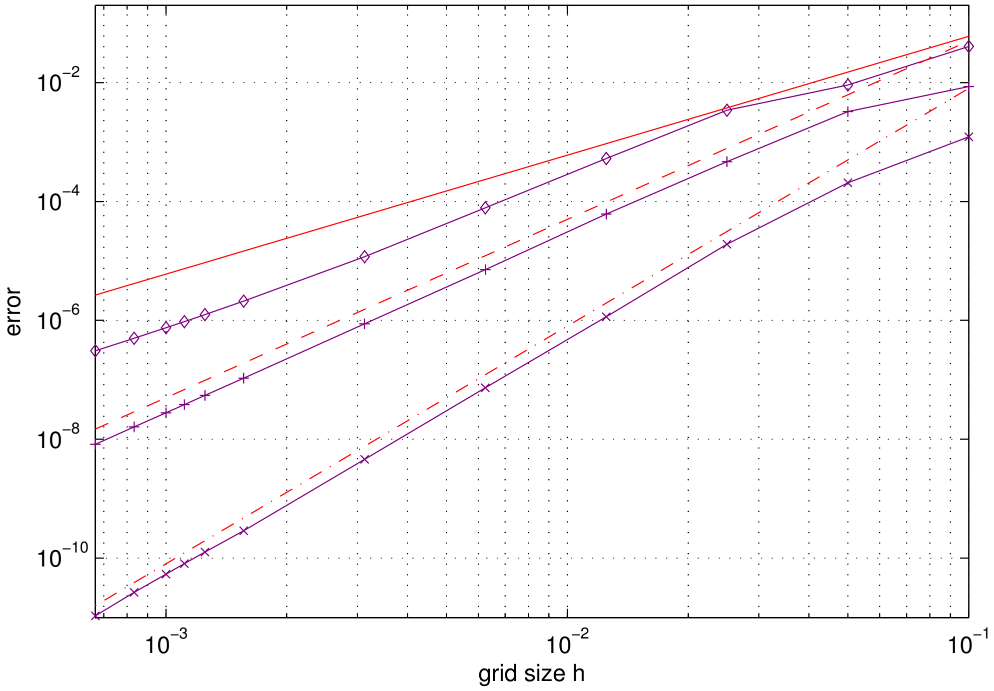
Convergence rates for E and T are shown in Figure 3.8 for different polynomial degrees k, where all solutions are both slope and realizability limited. The reference solution is computed with $J = 3000$ intervals and, as usual, polynomial reconstructions of order $k = 3$. Although both limiters are turned on, our DG algorithm reaches the theoretically expected convergence order for the appropriate polynomial degree. A detailed list of the convergence orders for E and T in the L^∞-norm are displayed in Tables 3.3 and 3.4. Even in this norm, all expected convergence rates are reached which again confirms a correct behavior of the applied limiters as well as of our DG code.

Thin Marshak Wave

In our last test case, a setting similar to the thin Marshak wave in [122] is considered. We use cm^{-1} as the unit for cross sections and keV for T. Our semi-infinite medium



(a) radiation energy E: $k = 1$ (circle line), $k = 2$ (square line), $k = 3$ (asterisk line).



(b) temperature T: $k = 1$ (diamond line), $k = 2$ (plus line), $k = 3$ (cross line).

Figure 3.8: L^1-error convergence at $t_{\text{final}} = 0.1\text{ns}$ for slope + realiz. limited solutions: slope two (red solid line), slope three (red dashed line), slope four (red dash-dot line).

		$\mathcal{P}^1$	$\mathcal{P}^2$	$\mathcal{P}^3$
	h	order	order	order
unlimited	1/20	3.42	2.13	3.13
	1/40	1.87	2.40	3.77
	1/80	2.17	3.04	3.06
	1/160	2.40	2.98	2.30
	1/320	2.26	2.95	2.52
	1/640	2.09	2.95	3.10
	1/800	2.02	2.99	3.44
	1/900	2.01	2.99	3.59
	1/1000	2.00	3.04	3.67
	1/1200	2.00	3.00	3.78
	1/1500	1.99	3.01	3.94
realiz. limited	1/20	3.67	1.79	3.49
	1/40	1.62	2.70	3.59
	1/80	2.17	2.90	3.06
	1/160	2.40	3.05	2.29
	1/320	2.26	2.93	2.58
	1/640	2.09	2.97	3.14
	1/800	2.02	2.97	3.47
	1/900	2.01	3.02	3.64
	1/1000	2.00	3.02	3.72
	1/1200	2.00	3.01	3.80
	1/1500	1.99	3.02	4.01
slope limited	1/20	4.28	1.77	1.32
	1/40	1.65	2.71	3.80
	1/80	2.37	2.90	6.02
	1/160	2.32	3.05	2.29
	1/320	2.21	2.93	2.58
	1/640	2.00	2.97	3.14
	1/800	1.91	2.97	3.47
	1/900	2.49	3.02	3.64
	1/1000	1.27	3.02	3.71
	1/1200	1.94	3.01	3.80
	1/1500	1.87	3.02	4.01
slope + realiz. limited	1/20	4.28	1.77	1.32
	1/40	1.65	2.71	3.80
	1/80	2.37	2.90	6.02
	1/160	2.32	3.05	2.29
	1/320	2.21	2.93	2.58
	1/640	2.00	2.97	3.14
	1/800	1.91	2.97	3.47
	1/900	2.49	3.02	3.64
	1/1000	1.27	3.02	3.71
	1/1200	1.94	3.01	3.80
	1/1500	1.87	3.02	4.01

Table 3.3: L^∞-convergence order for E.

		$\mathcal{P}^1$	$\mathcal{P}^2$	$\mathcal{P}^3$
	h	order	order	order
unlimited	1/20	3.35	1.27	1.81
	1/40	1.15	2.27	3.25
	1/80	2.26	3.05	3.84
	1/160	2.93	3.03	4.19
	1/320	2.25	3.05	4.71
	1/640	2.00	3.03	3.49
	1/800	2.00	3.02	3.61
	1/900	2.00	3.01	3.74
	1/1000	2.00	3.06	3.81
	1/1200	2.00	3.01	3.93
	1/1500	2.00	3.01	4.13
realiz. limited	1/20	3.49	1.02	2.08
	1/40	1.01	2.58	3.16
	1/80	2.26	2.91	3.96
	1/160	2.93	3.11	4.22
	1/320	2.25	3.03	4.67
	1/640	2.00	3.04	3.47
	1/800	2.00	3.00	3.63
	1/900	2.00	3.03	3.78
	1/1000	2.00	3.03	3.86
	1/1200	2.00	3.01	3.96
	1/1500	2.00	3.01	4.22
slope limited	1/20	3.52	1.02	2.05
	1/40	1.01	2.58	3.23
	1/80	2.57	2.91	3.95
	1/160	2.87	3.11	4.22
	1/320	2.00	3.03	4.67
	1/640	2.00	3.04	3.47
	1/800	2.00	3.00	3.63
	1/900	2.00	3.03	3.78
	1/1000	2.00	3.03	3.86
	1/1200	2.01	3.01	3.96
	1/1500	2.00	3.01	4.22
slope + realiz. limited	1/20	3.52	1.02	2.05
	1/40	1.01	2.58	3.23
	1/80	2.57	2.91	3.95
	1/160	2.87	3.11	4.22
	1/320	2.00	3.03	4.67
	1/640	2.00	3.04	3.47
	1/800	2.00	3.00	3.63
	1/900	2.00	3.03	3.78
	1/1000	2.00	3.03	3.86
	1/1200	2.01	3.01	3.96
	1/1500	2.00	3.01	4.22

Table 3.4: L^∞-convergence order for T.

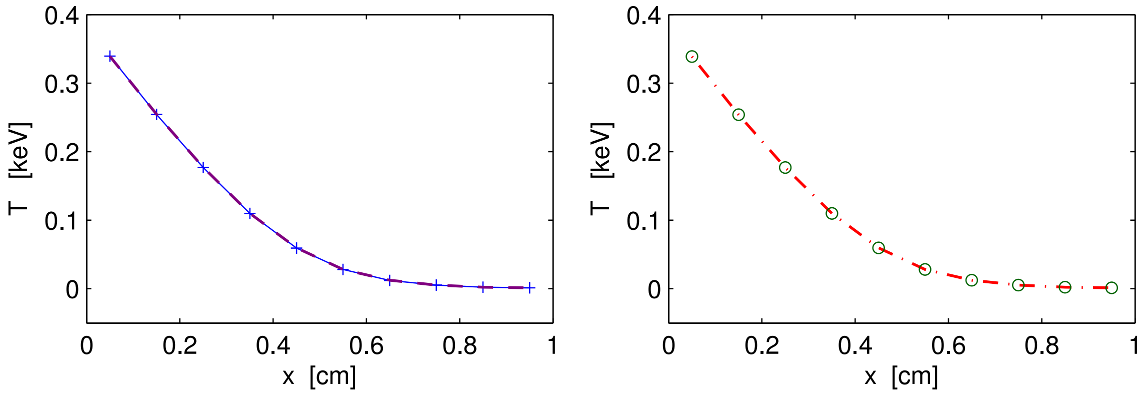


Figure 3.9: Thin Marshak wave: $t_{\text{final}} = 0.1\text{ns}$, $k = 1$, $J = 800$: unlimited (blue cross-line), slope limited (red dash-dot line), realiz. limited (purple dashed line), slope + realiz. limited (green circle-line)

is assumed to be purely absorbing, with absorption cross-section[5]

$$\sigma_{\text{a}}(T) = \frac{1}{(T+0.5)^3}\frac{\text{keV}^3}{\text{cm}}, \qquad \sigma_s = 0.$$

Ingoing radiation is prescribed on the left boundary whereas there is no radiation incoming on the right boundary,

$$\begin{aligned} &T(0,t) = 1, \quad T(1,t) = 0, \quad t > 0, \\ &\mathbf{u}(0,t) = [T(0,t)^4 a, 0.8 \cdot T(0,t)^4 ac]^T, \quad \mathbf{u}(1,t) = [2\varepsilon, 0]^T, \quad t > 0. \end{aligned}$$

These boundary conditions are chosen in a way that the eigenvalues of the Jacobian of the system in (3.32) evaluated at the boundary points are positive. Hence, well-posed boundary conditions are ensured according to [34].

The radiation will therefore propagate through the medium from left to the right starting at an initially small radiation

$$\begin{aligned} &T_0(x) = 5 \cdot 10^{-4}, \quad x \in (0,1), \\ &\mathbf{u}_0(x) = [T_0(x)^4 a, 0]^T, \quad x \in (0,1). \end{aligned}$$

Figure 3.9 displays the material temperature T with various kind of limiters turned on. There is a large decay of radiation near the left boundary at the beginning. As time increases, more and more photons enter the material and form a wavefront which builds up and moves to the right until an equilibrium is reached. To the eye, all solutions are close to each other although both results without the realizability limiter establish unrealizable solutions so that the cutoff in (3.104) with $\chi_{\max} = 1$ must be applied.

Numerical convergence is analyzed and the result for $k = 1$ is shown in Figure 3.10. The reference solution is slope + realizability limited and computed with $J = 3000$ intervals and polynomial degree up to $k = 3$. We observe only second order convergence

---

[5] In [122], $\sigma_{\text{a}} = T^{-3}$. We have modified this formula slightly to avoid the stiffness that arises in the source terms when T is very small. We do not promote the use of a fully explicit scheme in such cases. At a minimum, the source terms should be treated implicitly. This approach will be the subject in future work.

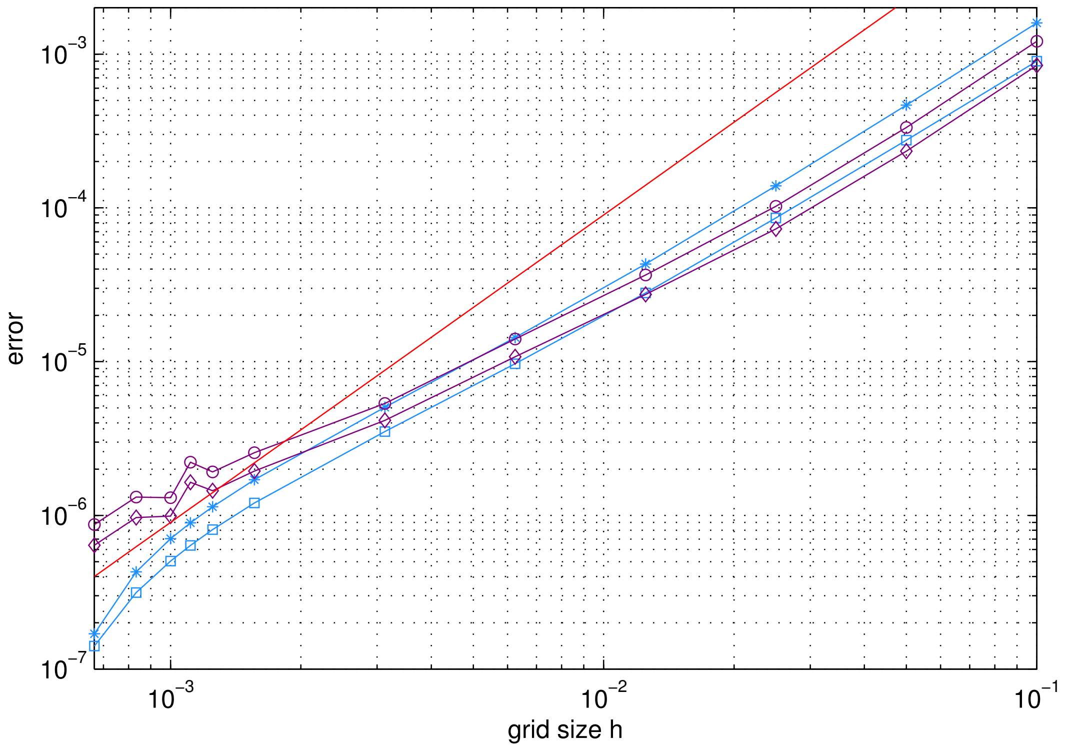


Figure 3.10: L^1-error convergence: $t_{\text{final}} = 0.1$ ns, $k = 1$: realiz. limited $a^{-1}E$ (purple line), slope + realiz. limited $a^{-1}E$ (blue asterisk line), realiz. limited T (purple diamond line), slope + realiz. limited T (blue square line), slope two (red solid line)

for all polynomial degrees $k = 1, 2, 3$. (The larger slopes of the blue lines at the two smallest grid sizes in Figure 3.10 are only artifacts of the reference solution which, in fact, needs a higher number of discretization points.) The second order convergence suggests that this problem has a non-smooth solution which, for consistency reasons, does not allow the DG method to achieve higher order convergence.

Chapter 4

Perturbed, Entropy-Based Closure

4.1 Introduction

In this chapter, we derive a new hierarchy of kinetic moment models in the context of frequency integrated (grey) photon transport. These new models are perturbations of well known entropy-based models; we therefore refer to them as *perturbed entropy-based* or *PEB models*. We present numerical simulations for the simplest member of the hierarchy, the *perturbed* M_1 or PM_1 *model*, in one spatial dimension. In this setting, the PM_1 model approximates the evolution of the photon radiation energy E and radiation flux F through a material medium with slab geometry. The photons interact with the material through scattering and emission/absorption processes.

Entropy-based (EB) models have been studied extensively in areas such as extended thermodynamics [44, 132], gas dynamics [70, 78, 90, 91, 109, 112, 158], semiconductors [6–9, 76, 89, 92, 110, 151], quantum fluids [41, 45], radiation transport [21, 22, 27, 28, 47, 48, 54, 77, 79, 127, 129, 169, 179], and phonon transport in solids [45]. In the context of radiative transfer, entropy models are commonly referred to as M_N *models*, where N is order of the expansion. The M_1 model dates back to [127], where it was first derived using Maxwell-Boltzmann statistics. For problems with Bose-Einstein statistics, formal theoretical properties such as hyperbolicity and entropy dissipation were first reported in [47] for arbitrary N. However, computational studies have focused primarily on properties of the M_1 model and its extensions, including multigroup equations [176] and partial moment models [48, 54]. In related work, one may find simulations of M_1 models based on other statistics, including Maxwell-Boltzmann [19, 21, 22] and Fermi-Dirac [16, 168]. This attachment to M_1 is due to the fact that the higher order members of the M_N hierarchy require the repeated solution of expensive numerical optimization problems. However, simulations of the M_2 model [129, 179] (the next member in the hierarchy) have been performed for Bose-Einstein statistics and for M_N up to order $N = 15$ for special benchmark problems using Maxwell-Boltzmann statistics [4, 77].

There are several reasons to consider perturbative modifications to EB models. First, the defining optimization problem must be solved at each point in a space-time mesh. Thus, it is economical to improve the model with perturbative corrections rather than increasing the number of moments, which, in turn, enlarges the complexity of the optimization. For frequency integrated photon transport, the minimization problem has an analytical solution. In this case, the argument for remaining in the M_1 framework, rather than increasing N, is especially compelling. A second reason is that perturbations add (among other things) diffusive terms to the EB model. It is hoped that these terms will smooth out non-physical shocks which are known to exist in EB models. The shocks are an artifact of the modeling procedure which results from approximating linear transport in phase space by a nonlinear hyperbolic balance law. A third reason is that the specification of boundary conditions for moment equations, which are consistent with the underlying kinetic boundary conditions, is an open problem. For linear moment equations, recent efforts [111] have shown the potential for well-posed boundary conditions for models with perturbative corrections.

One of the fundamental questions associated with any moment model is the issue of *realizability*. In the context of the M_1 and PM_1 models, we say that E and F are realizable if and only if they are the first two moments of an underlying kinetic distribution. This requirement on E and F is mathematically equivalent to the condition

$$|F| \leq cE \tag{4.1}$$

which must be satisfied point-wise in space and time. Here, c is the speed of light. We expect the solutions of the M_1 model to satisfy (4.1) because it (like all EB models) is derived assuming an ansatz for the kinetic distribution which is positive. However, the underlying ansatz for the PM_1 model is a perturbation of the EB ansatz and not necessarily positive. Therefore, a modification of the PEB ansatz is needed which controls the contribution of the perturbative term.

Even for the M_1 model, the realizability condition (4.1) can be destroyed by a numerical method unless special care is taken to enforce it. In this chapter, we build on previous work on the M_1 model from Chapter 3, using a Runge-Kutta discontinuous Galerkin (RKDG) method that is equipped with a special slope limiter [182, 183] in the spatial variable. For implementation of the PM_1 model, this special limiter must be applied in combination with the perturbation limiter in the underlying ansatz. The RKDG method [11] is a natural discretization here because we deal with a hyperbolic system of equations that is augmented by a diffusive term.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce the radiative transfer equation and moment model framework. In Section 4.3, we derive perturbed entropy-based closures and give explicit expressions for the perturbed M_1 model. In Section 4.4, we give details of the discontinuous Galerkin method used for simulation. In Section 4.5, we present numerical results.

4.2 Radiative Transfer and Moment Equations

We consider a collection of photons which move at the speed of light c through a static material medium. In engineering and physics applications, the fundamental quantity of interest is the radiation intensity $\psi = \psi(x, \Omega, \nu, t)$ which is a function of position $x \in K \subset \mathbb{R}^3$, direction $\Omega \in \mathbb{S}^2$, frequency $\nu \in (0, \infty)$, and time $t \in (0, \infty)$. Roughly speaking, ψ is the flux of energy through a surface. If f is the kinetic density of photons—that is, the number density with respect to the Lebesgue measure $dx d\Omega d\nu$—then $\psi = h\nu c f$, where h is Planck's constant.

The material is characterized by a temperature $T = T(x)$, an equation of state for the energy $e = e(T)$, and by scattering, absorption, and total cross-sections: Σ_s, Σ_a, and $\Sigma_t = \Sigma_a + \Sigma_s$ that depend on x directly and also indirectly through the material temperature.

4.2.1 The Radiative Transfer Equation

The radiative transfer equation, which approximates the evolution of ψ, is given by

$$\frac{1}{c}\partial_t\psi + \Omega\cdot\nabla_x\psi = \mathcal{C}(\psi;T)\,. \tag{4.2}$$

The collision operator $\mathcal{C}$ models interactions of photons with the medium. For our purposes, we assume $\mathcal{C}$ has the form

$$\mathcal{C}(\psi;T) := -\Sigma_t\psi + \frac{1}{4\pi}\left(\Sigma_s\phi + \Sigma_a B(T) + S\right)\,, \tag{4.3}$$

where ϕ is the angular integral of ψ:

$$\phi := \int_{4\pi}\psi d\Omega. \tag{4.4}$$

S is an external source, and the Planckian

$$B(T) := \frac{2h\nu^3}{c^2}\frac{1}{\exp\left(\frac{h\nu}{kT}\right)-1} \tag{4.5}$$

models blackbody radiation from the material. The constant k is Boltzmann's constant. The first term in $\mathcal{C}$ accounts for the loss of photons at a particular frequency and angle due to out-scattering and absorption by the material. The second term gives the gain of photons due to in-scattering from other angles and re-emission by the material.

The evolution of the material temperature is determined by a balance of emitting and absorbed photons:

$$\partial_t e(T) = \Sigma_a\left(\langle\psi\rangle - acT^4\right)\,, \tag{4.6}$$

where angle brackets are used as a shorthand notation for integration over angle and frequency:

$$\langle\,\cdot\,\rangle \equiv \int_0^\infty\int_{\mathbb{S}^2}(\,\cdot\,)\,d\Omega d\nu\,, \tag{4.7}$$

and the T^4 term in the first equation comes from the Stefan-Boltzmann Law:

$$\int_0^\infty B(T)d\nu = acT^4\,. \tag{4.8}$$

The constant a is the *radiation constant.* Though the material equation (4.6) plays an important role, we will focus here on simulating the transport equation (4.2).

4.2.2 Moment Equations

The large phase space on which (4.2) is defined makes direct numerical simulation prohibitively expensive. Thus, approximate models are needed to reduce the size of the system. A common and well-known approach is the method of moments, for which full resolution of the angular and/or frequency dependency of ψ is replaced by a finite number of weighted averages.

Derivation of any moment system begins with the choice of a vector-valued function $\mathbf{m}:\mathbb{S}^2\to\mathbb{R}^n$, $\Omega\mapsto[\mathbf{m}_0(\Omega),...,\mathbf{m}_{n-1}(\Omega)]^T$, whose n components are linearly independent functions of Ω. Evolution equations for the moments $\mathbf{u}(x,t) := \langle\mathbf{m}\psi(x,\Omega,t)\rangle$ are found by multiplying the transport equation by $\mathbf{m}$ and integrating over all angles to give

$$\frac{1}{c}\partial_t\mathbf{u} + \nabla_x \cdot \langle \Omega\mathbf{m}\psi\rangle = \langle \mathbf{m}\mathcal{C}(\psi;T)\rangle \,. \tag{4.9}$$

The system (4.9) is not closed. A recipe, or *closure*, must be prescribed to express unknown quantities in terms of the given moments. Often this is done via an approximation for ψ in (4.9) that depends on $\mathbf{u}$,

$$\psi(x,\Omega,t) \simeq \mathcal{E}(\mathbf{u}(x,t))\,, \tag{4.10}$$

and satisfies the consistency relation

$$\langle \mathbf{m}\mathcal{E}(\mathbf{u})\rangle = \mathbf{u}\,. \tag{4.11}$$

The resulting moment system is

$$\frac{1}{c}\partial_t\mathbf{u} + \nabla_x \cdot \langle \Omega\mathbf{m}\mathcal{E}(\mathbf{u})\rangle = \langle \mathbf{m}\mathcal{C}(\mathcal{E}(\mathbf{u});T)\rangle \,. \tag{4.12}$$

In general, a closure is required to evaluate both the flux terms and the collision terms. However, the collision operator given in (4.3) requires no closure. Indeed, it is straightforward to show that $\langle \mathbf{m}\mathcal{C}(\mathcal{E}(\mathbf{u});T)\rangle = \langle \mathbf{m}\mathcal{C}(\psi;T)\rangle$ for any reconstruction that satisfies the consistency relation. Thus, we will be focused on closure of the flux term. As one might expect, the behavior of a moment system—and in particular its ability to capture fundamental features of the kinetic description—depends heavily on the form of the reconstruction.

4.3 Entropy-Based and Perturbed Entropy-Based Closures

In this section, we briefly review the theory of entropy-based closures for radiative transfer [21,22,47,48,54,77,127,129,179] and introduce our new perturbative model.

4.3.1 Entropy-Based Closures

A general strategy for prescribing a closure is to use the solution of a constrained optimization problem

$$\min_{g\in\mathrm{Dom}(\mathcal{H})} \quad \mathcal{H}(g) \tag{4.13}$$

$$\text{s.t.} \quad \langle \mathbf{m}g\rangle = \langle \mathbf{m}\psi\rangle \tag{4.14}$$

where $\mathcal{H}(g) := \langle \eta(g)\rangle$ and $\eta : \mathbb{R} \to \mathbb{R}$ is a strictly convex function that is related to the entropy of the system. For photons, the physically relevant entropy comes from Bose-Einstein statistics and is given by [141,152]

$$\eta(g) = \frac{2k\nu^2}{c^3}\left[n_g \log(n_g) - (n_g+1)\log(n_g+1)\right]\,, \tag{4.15}$$

where the n_g is the occupation number associated with g:

$$n_g := \frac{c^2}{2h\nu^3}g\,. \tag{4.16}$$

The solution of (4.13) is expressed in terms of the Legendre dual η_*. Let

$$\mathcal{B}(\boldsymbol{\alpha}) := \eta_*' \left(\frac{h\nu c}{k} \boldsymbol{\alpha}^T \mathbf{m} \right) = \frac{2h\nu^3}{c^2} \frac{1}{\exp\left(-\frac{h\nu c}{k}\boldsymbol{\alpha}^T\mathbf{m}\right) - 1}, \tag{4.17}$$

then we have the following:

Theorem 4. *The solution of* (4.13) *is given by* $\mathcal{E}(\mathbf{u}) = \mathcal{B}(\hat{\boldsymbol{\alpha}})$, *where* $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\mathbf{u})$ *solves the dual problem*

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left\{ \left\langle \eta_* \left(\frac{h\nu c}{k} \boldsymbol{\alpha}^T \mathbf{m} \right) \right\rangle - \boldsymbol{\alpha}^T \mathbf{u} \right\}. \tag{4.18}$$

It is also the Legendre dual variable of $\mathbf{u}$ *with respect to the strictly convex entropy* $h(\mathbf{u}) := \mathcal{H}(\mathcal{B}(\hat{\boldsymbol{\alpha}}(\mathbf{u})))$, *i.e.,*

$$\hat{\boldsymbol{\alpha}}(\mathbf{u}) = \left[\frac{\partial h}{\partial \mathbf{u}}(\mathbf{u}) \right]^T. \tag{4.19}$$

The moment system derived by substituting (4.17) *into* (4.12) *is hyperbolic and symmetric when expressed in the* $\hat{\boldsymbol{\alpha}}$ *variables and its solution formally dissipates* h. *Moreover,* $\mathcal{E}$ *is an inherently positive quantity.*

Proof. The form of the minimizer in (4.17) can be derived formally using standard Lagrange multiplier techniques. However, a rigorous proof requires more technical arguments, which can be found, for example in [91] for the Maxwell-Boltzmann entropy and applied directly to the current setting. Once the existence of a minimizer is found, the other properties can be verified, as is done in [47, 109]. □

4.3.2 Perturbed Entropy-Based (PEB) Closures

Perturbations to standard P_N closures[1] have been derived for $N = 3$ in [136] for general N in [156] (see also [172] and [76]). The idea behind the derivation in [156] is to write $\psi = \psi_{\text{pn}} + \tilde{\psi}$, where ψ_{pn} is the standard P_N expansion. The perturbation $\tilde{\psi}$ satisfies its own kinetic equation, which can be then used to approximate $\tilde{\psi}$ in terms of ψ_{pn}. The resulting "D_N" models gain a diffusive term in the equations for the highest order moments. Such an approach need not be restricted to the P_N equations. Indeed, following this exact strategy, we define

(1) The moment map $\mathcal{M} : g \mapsto \mathbf{u} := \langle \mathbf{m} g \rangle$;

(2) The expansion map $\mathcal{E} : \mathbf{u} \mapsto \eta_*'(\frac{h\nu c}{k}\hat{\boldsymbol{\alpha}}(\mathbf{u})^T\mathbf{m})$;

(3) The reconstruction $\mathcal{R} = \mathcal{E} \circ \mathcal{M}$;

(4) The kinetic perturbation $\tilde{\psi} = \psi - \mathcal{R}(\psi)$.

The kinetic equation for $\tilde{\psi}$ is

$$\partial_t\tilde{\psi} = \partial_t\psi - \partial_t\mathcal{R}(\psi) = \partial_t\psi - \partial_t\mathcal{E}(\mathbf{u}) = \partial_t\psi - \mathcal{E}'(\mathbf{u})\partial_t\mathbf{u} \tag{4.20}$$

---

[1] These closures are based on a spherical harmonic expansion in angle and can be formulated as an entropy-based closure with an L^2 cost functional [79, 145].

where

$$\mathcal{E}'(\mathbf{u}) = \mathcal{B}'(\hat{\boldsymbol{\alpha}})\frac{\partial\hat{\boldsymbol{\alpha}}}{\partial\mathbf{u}} = \mathbf{m}^T\mathcal{W}(\mathbf{u})\left\langle\mathbf{m}\mathbf{m}^T\mathcal{W}(\mathbf{u})\right\rangle^{-1} . \tag{4.21}$$

and

$$\mathcal{W}(\mathbf{u}) := \eta''_*\left(\frac{h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m}\right) = \frac{2h^2\nu^4}{kc}\frac{\exp(-\frac{h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m})}{\left[\exp(-\frac{h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m})-1\right]^2} > 0. \tag{4.22}$$

We have used the relation

$$\mathcal{I} = \langle\mathbf{m}\mathcal{E}'(\mathbf{u})\rangle = \langle\mathbf{m}\mathcal{B}'(\hat{\boldsymbol{\alpha}})\rangle\frac{\partial\hat{\boldsymbol{\alpha}}}{\partial\mathbf{u}} = \frac{h\nu c}{k}\left\langle\mathbf{m}\mathbf{m}^T\mathcal{W}(\mathbf{u})\right\rangle\frac{\partial\hat{\boldsymbol{\alpha}}}{\partial\mathbf{u}} \tag{4.23}$$

to compute the matrix $\frac{\partial\hat{\boldsymbol{\alpha}}}{\partial\mathbf{u}}$ in (4.21). By operating with $\tilde{\mathcal{P}}_{\mathbf{u}} := \mathcal{I} - \mathcal{P}_{\mathbf{u}}$ on (4.2) , where $\mathcal{P}_{\mathbf{u}} := \mathcal{E}'(\mathbf{u})\mathcal{M}$, we can write (4.20) as

$$\frac{1}{c}\partial_t\tilde{\psi} + \tilde{\mathcal{P}}_{\mathbf{u}}(\Omega\cdot\nabla_x\psi) = \tilde{\mathcal{P}}_{\mathbf{u}}\mathcal{C}(\psi;T). \tag{4.24}$$

It should be noted, for future use, that the projection $\mathcal{Q}_{\mathbf{u}}$, given by

$$\mathcal{Q}_{\mathbf{u}}g := \frac{1}{\mathcal{W}(\mathbf{u})}\mathcal{P}_{\mathbf{u}}(\mathcal{W}(\mathbf{u})g) , \tag{4.25}$$

is self-adjoint in L^2 with respect to the positive weight $\mathcal{W}(\mathbf{u})$.

Equation (4.24) for the perturbation is exact. To derive a closure, we neglect the time derivative and perturbative component of the flux to arrive at the following approximate balance equation

$$\tilde{\mathcal{P}}_{\mathbf{u}}(\Omega\cdot\nabla_x\mathcal{E}(\mathbf{u})) \simeq \tilde{\mathcal{P}}_{\mathbf{u}}\mathcal{C}(\psi;T), \tag{4.26}$$

where

$$\tilde{\mathcal{P}}_{\mathbf{u}}\mathcal{C}(\psi;T) = -\Sigma_t\left[\tilde{\mathcal{P}}_{\mathbf{u}}\mathcal{E}(\mathbf{u}) + \tilde{\psi}\right] + \frac{1}{4\pi}\left[\Sigma_s\tilde{\mathcal{P}}_{\mathbf{u}}\phi + \Sigma_a\tilde{\mathcal{P}}_{\mathbf{u}}B(T) + \tilde{\mathcal{P}}_{\mathbf{u}}S\right] . \tag{4.27}$$

In Lemma 7 below, we show that,

$$\tilde{\mathcal{P}}_{\mathbf{u}}\mathcal{E}(\mathbf{u}) = 0. \tag{4.28}$$

Therefore,

$$\tilde{\mathcal{P}}_{\mathbf{u}}\mathcal{C}(\psi;T) = -\Sigma_t\tilde{\psi} + \frac{1}{4\pi}\left[\Sigma_s\tilde{\mathcal{P}}_{\mathbf{u}}\phi + \Sigma_a\tilde{\mathcal{P}}_{\mathbf{u}}B(T) + \tilde{\mathcal{P}}_{\mathbf{u}}S\right] , \tag{4.29}$$

and we can solve (4.26) for $\tilde{\psi}$ in terms of a convective component $\tilde{\psi}^{\mathrm{c}}$ and a diffusive component $\tilde{\psi}^{\mathrm{d}}$:

$$\tilde{\psi} \simeq \frac{1}{4\pi}\left[r_{\mathrm{s}}\tilde{\mathcal{P}}_{\mathbf{u}}\phi + r_{\mathrm{a}}\tilde{\mathcal{P}}_{\mathbf{u}}B(T) + \frac{1}{\Sigma_t}\tilde{\mathcal{P}}_{\mathbf{u}}S\right] - \frac{1}{\Sigma_t}\tilde{\mathcal{P}}_{\mathbf{u}}(\Omega\cdot\nabla_x\mathcal{E}(\mathbf{u})) =: \tilde{\psi}^{\mathrm{c}} + \tilde{\psi}^{\mathrm{d}}, \tag{4.30}$$

where r_{s} and r_{a} are the scattering and absorption ratios, respectively:

$$r_{\mathrm{s}} = \frac{\Sigma_s}{\Sigma_t} \quad \text{and} \quad r_{\mathrm{a}} = \frac{\Sigma_a}{\Sigma_t} . \tag{4.31}$$

Inserting (4.30) back into the flux term of the moment equation (4.12) gives

$$\langle \Omega \mathbf{m} \psi \rangle \simeq \langle \Omega \mathbf{m} \mathcal{E}(\mathbf{u}) \rangle + \langle \Omega \mathbf{m} \tilde{\psi}^{\mathrm{c}} \rangle + \langle \Omega \mathbf{m} \tilde{\psi}^{\mathrm{d}} \rangle =: \mathbf{f}^{\mathcal{E}} + \mathbf{f}^{\mathrm{C}} + \mathbf{f}^{\mathrm{D}}. \tag{4.32}$$

At this point, it is not clear whether the PEB system dissipates an entropy or if the convective flux is always hyperbolic. However, the perturbed M_1 model is hyperbolic (see Proposition 3 in the next section). In addition, the diffusive flux does, in general, satisfy a local dissipation law.

Proposition 2. *The diffusion term* $\mathbf{f}^{\mathrm{D}}$ *dissipates the entropy* $h(\mathbf{u}) := \mathcal{H}(\mathcal{E}(\mathbf{u}))$ *locally in space.*

Proof. A dissipation law for h is found by multiplying the closed moment system (4.12) by $\hat{\boldsymbol{\alpha}}^T \equiv \frac{\partial h}{\partial \mathbf{u}}$. Multiplying $\nabla_x \cdot \mathbf{f}^{\mathrm{D}}$ on the right by $\hat{\boldsymbol{\alpha}}^T$ gives

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}^T \left(\nabla_x \cdot \mathbf{f}^{\mathrm{D}}\right) &= -\hat{\boldsymbol{\alpha}}^T \left[\nabla_x \cdot \left\langle \Omega \mathbf{m} \Sigma_t^{-1} \tilde{\mathcal{P}}_{\mathbf{u}} \left(\Omega \cdot \nabla_x \mathcal{E}(\mathbf{u})\right)\right\rangle\right] \\
&= -\nabla_x \cdot \left\langle \Omega (\hat{\boldsymbol{\alpha}}^T \mathbf{m}) \Sigma_t^{-1} \tilde{\mathcal{P}}_{\mathbf{u}} \left(\Omega \cdot \nabla_x \mathcal{E}(\mathbf{u})\right)\right\rangle \\
&\qquad + \left(\nabla_x \hat{\boldsymbol{\alpha}}^T\right) \cdot \left\langle \Omega \mathbf{m} \Sigma_t^{-1} \tilde{\mathcal{P}}_{\mathbf{u}} \left(\Omega \cdot \nabla_x \mathcal{E}(\mathbf{u})\right)\right\rangle,
\end{aligned}$$

where ∇_x acts on the components of Ω and the Lagrange multiplier $\hat{\boldsymbol{\alpha}}^T$ on $\mathbf{m}$. We only need to work with the term that is not in divergence form. We use the fact that $\mathcal{B}(\hat{\boldsymbol{\alpha}}) = (h\nu c/k)\mathbf{m}^T \mathcal{W}$ and that $\tilde{\mathcal{Q}}_{\mathbf{u}} := \mathrm{Id} - \mathcal{Q}_{\mathbf{u}}$ to compute

$$\begin{aligned}
\left(\nabla_x \hat{\boldsymbol{\alpha}}^T\right) \cdot \left\langle \Omega \mathbf{m} \Sigma_t^{-1} \tilde{\mathcal{P}}_{\mathbf{u}} \left(\Omega \cdot \nabla_x \mathcal{E}(\mathbf{u})\right)\right\rangle &= \left(\nabla_x \hat{\boldsymbol{\alpha}}^T\right) \cdot \left\langle \Omega \mathbf{m} \Sigma_t^{-1} \mathcal{W} \tilde{\mathcal{Q}}_{\mathbf{u}} \left(\frac{\Omega \cdot \nabla_x \mathcal{E}(\mathbf{u})}{\mathcal{W}}\right)\right\rangle \\
&= \frac{h\nu c}{k} \left\langle \Omega \cdot \nabla_x (\hat{\boldsymbol{\alpha}}^T \mathbf{m}) \Sigma_t^{-1} \mathcal{W} \tilde{\mathcal{Q}}_{\mathbf{u}} \left(\Omega \cdot \nabla_x (\hat{\boldsymbol{\alpha}}^T \mathbf{m})\right)\right\rangle \\
&= \frac{h\nu c}{k} \left\langle \Sigma_t^{-1} \mathcal{W} \left[\tilde{\mathcal{Q}}_{\mathbf{u}} (\nabla_x \cdot (\Omega \hat{\boldsymbol{\alpha}}^T \mathbf{m})\right]^2 \right\rangle \geq 0.
\end{aligned}$$

□

Lemma 7. *For the grey equations and Bose Einstein entropy, the operator* $\mathcal{P}_{\mathbf{u}} = \mathcal{E}'(\mathbf{u})\mathcal{M}$ *acts on the quantity* $\mathcal{E}(\mathbf{u})$ *as the identity operator:*

$$\mathcal{P}_{\mathbf{u}} \mathcal{E}(\mathbf{u}) = \mathcal{E}(\mathbf{u}).$$

Proof. We first calculate two frequency integrals. Let $\kappa := \frac{hc}{k} \hat{\boldsymbol{\alpha}}^T \mathbf{m}$ and $\theta := -\kappa\nu$. Then

$$\begin{aligned}
\int_0^\infty \eta_*'\left(\frac{-h\nu c}{k} \hat{\boldsymbol{\alpha}}^T \mathbf{m}\right) d\nu &= \int_0^\infty \frac{2h\nu^3}{c^2} \frac{1}{\exp(-\kappa\nu) - 1} d\nu \\
&= -\frac{2h}{c^2 \kappa^4} \int_0^\infty \frac{\theta^3}{\exp(\theta) - 1} d\theta = -\frac{2\pi^4 h}{15 c^2 \kappa^4}
\end{aligned} \tag{4.33}$$

and

$$\begin{aligned}
\int_0^\infty \eta_*''\left(\frac{-h\nu c}{k} \hat{\boldsymbol{\alpha}}^T \mathbf{m}\right) d\nu &= \int_0^\infty \frac{2h^2\nu^4}{kc} \frac{\exp(-\kappa\nu)}{[\exp(-\kappa\nu) - 1]^2} d\nu \\
&= \frac{2h^2}{kc\kappa^5} \int_0^\infty \frac{\theta^4 \exp(\theta)}{[\exp(\theta) - 1]^2} d\theta = \frac{8\pi^4 h^2}{15 kc\kappa^5}.
\end{aligned} \tag{4.34}$$

With these two integrals, it is easy to show that $-\hat{\boldsymbol{\alpha}}/4$ is the unique solution to the linear system

$$\left\langle \mathbf{m}\mathbf{m}^T \eta_*''\left(\frac{-h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m}\right)\right\rangle \boldsymbol{\beta} = \left\langle \mathbf{m}\eta_*'\left(\frac{-h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m}\right)\right\rangle, \tag{4.35}$$

so that

$$\left\langle \mathbf{m}\mathbf{m}^T \eta_*''\left(\frac{-h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m}\right)\right\rangle^{-1} \left\langle \mathbf{m}\eta_*'\left(\frac{-h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m}\right)\right\rangle = -\frac{\hat{\boldsymbol{\alpha}}}{4}. \tag{4.36}$$

Using the definition of $\mathcal{P}_{\mathbf{u}}$,

$$\mathcal{P}_{\mathbf{u}}\mathcal{E}(\mathbf{u}) = \frac{\hat{\boldsymbol{\alpha}}^T\mathbf{m}}{4}\eta_*''\left(\frac{-h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m}\right), \tag{4.37}$$

and again the integrals above:

$$\begin{aligned}\int_0^\infty \mathcal{P}_{\mathbf{u}}\mathcal{E}(\mathbf{u})\,d\nu &= \frac{-\hat{\boldsymbol{\alpha}}^T\mathbf{m}}{4}\frac{8\pi^4h^2}{15kc\kappa^5} = \frac{-k\kappa}{4hc}\frac{8\pi^4h^2}{15kc\kappa^5}\\ &= -\frac{2\pi^4h}{15c^2\kappa^4} = \int_0^\infty \eta_*'\left(\frac{-h\nu c}{k}\hat{\boldsymbol{\alpha}}^T\mathbf{m}\right)\,d\nu = \int_0^\infty \mathcal{E}(\mathbf{u})\,d\nu \end{aligned} \tag{4.38}$$

□

4.3.3 The Perturbed M_1 (PM_1) model

The perturbed M_1 model is based on the moments

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_0 \\ \mathbf{u}_1 \end{pmatrix} = \begin{pmatrix} cE \\ F \end{pmatrix} := \begin{pmatrix} \langle\psi\rangle \\ \langle\Omega\psi\rangle \end{pmatrix}. \tag{4.39}$$

where E is the photon energy density and F is the energy flux density. It approximates the evolution of E and F by the following system:

$$\partial_t E + \nabla_x \cdot F = -\Sigma_a(cE - acT^4) + \frac{1}{c}S, \tag{4.40a}$$

$$\partial_t F + c^2\nabla_x \cdot \Pi(E,F) = -c\Sigma_t F\,, \tag{4.40b}$$

where the closure for the pressure term is

$$\Pi(E,F) := \frac{1}{c}\langle(\Omega\vee\Omega)(\mathcal{E}(\mathbf{u}) + \tilde{\psi}^{\mathrm{c}} + \tilde{\psi}^{\mathrm{d}})\rangle =: \Pi^{\mathrm{M}_1}(E,F) + \Pi^{\mathrm{C}}(E,F) + \Pi^{\mathrm{D}}(E,F)\,. \tag{4.41}$$

Here $\Pi^{\mathrm{M}_1}(\mathbf{u})$ is the term that comes from the entropy ansatz (the *entropy-based term*). The term $\Pi^{\mathrm{C}}(\mathbf{u})$ is the *convective correction* and $\Pi^{\mathrm{D}}(\mathbf{u})$ is the *diffusive correction.* These corrections can be expressed in terms of Π^{M_1} and

$$Q^{\mathrm{M}_1} := \langle\Omega^{\vee 3}\mathcal{E}(\mathbf{u})\rangle \tag{4.42}$$

which, in turn, can be expressed in terms of the unit vector $\mathbf{n} := F/|F|$ and the scalars

$$\chi_k = \frac{\langle(\Omega\cdot\mathbf{n})^k\mathcal{E}\rangle}{cE}\,. \tag{4.43}$$

Lemma 8. *The correction terms* Π^{C} *and* Π^{D} *are given by*

$$\Pi^{\mathrm{D}} = \frac{1}{c\Sigma_t}\left[-\nabla_x \cdot Q^{\mathrm{M}_1} + \frac{\partial \Pi^{\mathrm{M}_1}}{\partial E}(\nabla_x \cdot F) + c^2 \frac{\partial \Pi^{\mathrm{M}_1}}{\partial F}(\nabla_x \cdot \Pi^{\mathrm{M}_1})\right], \tag{4.44a}$$

$$\Pi^{\mathrm{C}} = \eta \cdot \left(r_{\mathrm{s}} E + r_{\mathrm{a}} a T^4 + \frac{S}{c\Sigma_t}\right), \quad \textit{where} \quad \eta = \left(\frac{1}{3}\mathrm{Id} - \frac{\partial \Pi^{\mathrm{M}_1}}{\partial E}\right). \tag{4.44b}$$

Proof. The proof is a straight-forward calculation. It turns out to be more efficient to calculate Π^{D} and Π^{C} without directly using (4.21). Instead for any function g, we compute

$$\langle (\Omega \vee \Omega)\tilde{\mathcal{P}}_{\mathbf{u}} g \rangle = \langle (\Omega \vee \Omega) g \rangle - \sum_{k=0}^{1} \frac{\partial \langle (\Omega \vee \Omega)\mathcal{E} \rangle}{\partial \mathbf{u}_k} \langle \mathbf{m}_k g \rangle = \langle (\Omega \vee \Omega) g \rangle - \sum_{k=0}^{1} \frac{\partial \Pi^{\mathrm{M}_1}}{\partial \mathbf{u}_k} \langle \mathbf{m}_k g \rangle.$$

Using above equation, we find for the diffusive correction,

$$\begin{aligned} \Pi^{\mathrm{D}} &= \frac{1}{c\Sigma_t}\left\langle (\Omega \vee \Omega)\tilde{\psi}^{\mathrm{d}} \right\rangle \\ &= -\frac{1}{c\Sigma_t}\left\langle (\Omega \vee \Omega)\tilde{\mathcal{P}}_{\mathbf{u}}(\Omega \cdot \nabla_x \mathcal{E}(\mathbf{u})) \right\rangle \\ &= -\frac{1}{c\Sigma_t}\nabla_x \cdot \left\langle (\Omega^{\vee 3}\mathcal{E}(\mathbf{u}) \right\rangle + \frac{1}{c\Sigma_t}\sum_{k=0}^{1} \frac{\partial \Pi^{\mathrm{M}_1}}{\partial \mathbf{u}_k} \nabla_x \cdot \langle \mathbf{m}_k \Omega \mathcal{E} \rangle \\ &= -\frac{1}{c\Sigma_t}\nabla_x \cdot Q^{\mathrm{M}_1} + \frac{1}{c\Sigma_t}\frac{\partial \Pi^{\mathrm{M}_1}}{\partial E}(\nabla_x \cdot F) + \frac{1}{\Sigma_t}\frac{\partial \Pi^{\mathrm{M}_1}}{\partial F}(\nabla_x \cdot \Pi^{\mathrm{M}_1}). \end{aligned} \tag{4.45}$$

For the convection correction, we use the fact that ϕ, B and S are independent of Ω. This implies that $\langle \mathbf{m}_1 \phi \rangle \equiv \langle \Omega \phi \rangle = 0$ and similarly for B and S. We also use the Stefan-Boltzmann-Law (4.8) and the identity $\mathbf{u}_0 = cE = \int_0^\infty \phi \, d\nu$. This gives

$$\begin{aligned} \Pi^{\mathrm{C}} &= \frac{1}{c}\left\langle (\Omega \vee \Omega)\tilde{\psi}^{\mathrm{c}} \right\rangle \\ &= \frac{r_{\mathrm{s}}}{4\pi c}\left\langle (\Omega \vee \Omega)\tilde{\mathcal{P}}_{\mathbf{u}}\phi \right\rangle + \frac{r_{\mathrm{a}}}{4\pi c}\left\langle (\Omega \vee \Omega)\tilde{\mathcal{P}}_{\mathbf{u}}B \right\rangle + \frac{1}{4\pi c\Sigma_t}\left\langle (\Omega \vee \Omega)\tilde{\mathcal{P}}_{\mathbf{u}}S \right\rangle \\ &= \frac{r_{\mathrm{s}}}{4\pi c}\left(\langle (\Omega \vee \Omega)\phi \rangle - \frac{\partial \Pi^{\mathrm{M}_1}}{\partial E}\langle \phi \rangle\right) + \frac{r_{\mathrm{a}}}{4\pi c}\left(\langle (\Omega \vee \Omega)B \rangle - \frac{\partial \Pi^{\mathrm{M}_1}}{\partial E}\langle B \rangle\right) \\ &\quad + \frac{1}{4\pi c\Sigma_t}\left(\langle (\Omega \vee \Omega)S \rangle - \frac{\partial \Pi^{\mathrm{M}_1}}{\partial E}\langle S \rangle\right) \\ &= \left(\frac{1}{3}\mathrm{Id} - \frac{\partial \Pi^{\mathrm{M}_1}}{\partial E}\right)\left(r_{\mathrm{s}} E + r_{\mathrm{a}} a T^4 + \frac{S}{c\Sigma_t}\right). \end{aligned} \tag{4.46}$$

□

Remark 15. *The formula for the convective correction is independent of the particular form of* $\mathcal{E}(\mathbf{u})$. *In particular for the* P_1 *model, the pressure term is* $\Pi^{\mathrm{P}_1} = \frac{1}{3}E$ *so that* $\frac{\partial \Pi^{\mathrm{P}_1}}{\partial E} = \frac{1}{3}\mathrm{Id}$ *and* $\Pi^{\mathrm{C}} = 0$. *This is consistent with the fact that the "*D_N*" models in [156] contain only diffusive corrections.*

Lemma 9. *The entropy-based terms* Π^{M_1} *and* Q^{M_1} *are given by*

$$\Pi^{\mathrm{M}_1} = \frac{E}{2}[(1-\chi_2)\mathrm{Id} + (3\chi_2 - 1)(\mathbf{n}\vee\mathbf{n})]\,, \tag{4.47a}$$

$$Q^{\mathrm{M}_1} = \frac{3cE}{2}[(\chi_1 - \chi_3)(\mathrm{Id}\vee\mathbf{n}) + (5\chi_3 - 3\chi_1)\mathbf{n}^{\vee 3}]\,, \tag{4.47b}$$

where the scalars χ_1*,* χ_2*, and* χ_3 *are defined in* (4.43)*.*

Proof. Let $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ be any orthogonal basis for $\mathbb{R}^3$. Then

$$\Omega = \sum_{i=1}^{3}\Omega_i\mathbf{e}_i\,, \quad \Omega_i := (\Omega\cdot\mathbf{e}_i)\,, \quad \sum_{i=1}^{3}\Omega_i^2 = 1, \tag{4.48}$$

and

$$\left\langle(\Omega^{\vee k})\mathcal{E}(\mathbf{u})\right\rangle = \left\langle\left(\sum_{i=1}^{3}\Omega_i e_i\right)^{\vee k}\mathcal{E}(\mathbf{u})\right\rangle. \tag{4.49}$$

Now set $\mathbf{e}_3 = \mathbf{n} = F/|F|$ and note that, according to Lemma 10 below, $\mathcal{E}(\mathbf{u})$ depends on Ω only through Ω_3. Thus, only the terms with even powers of Ω_1 and Ω_2 will survive. For $k = 2$, this means

$$c\Pi^{\mathrm{M}_1} = \left\langle\Omega_1^2\mathcal{E}(\mathbf{u})\right\rangle\mathbf{e}_1\vee\mathbf{e}_1 + \left\langle\Omega_2^2\mathcal{E}(\mathbf{u})\right\rangle\mathbf{e}_2\vee\mathbf{e}_2 + \left\langle\Omega_3^2\mathcal{E}(\mathbf{u})\right\rangle\mathbf{n}\vee\mathbf{n}, \tag{4.50}$$

and for $k = 3$,

$$Q^{\mathrm{M}_1} = 3\left\langle\Omega_1^2\Omega_3\mathcal{E}(\mathbf{u})\right\rangle\mathbf{e}_1\vee\mathbf{e}_1\vee\mathbf{n} + 3\left\langle\Omega_2^2\Omega_3\mathcal{E}(\mathbf{u})\right\rangle\mathbf{e}_2\vee\mathbf{e}_2\vee\mathbf{n} + \left\langle\Omega_3^3\mathcal{E}(\mathbf{u})\right\rangle\mathbf{n}^{\vee 3}. \tag{4.51}$$

The goal then is to write these formulas in terms of Ω_3 only. Let us focus first on Π^{M_1}. Because $\mathcal{E}(\mathbf{u})$ depends only on Ω_3, symmetry arguments can be used to conclude that first two terms in (4.50) are the same. Combined with the far right relation (4.48), this gives

$$\begin{aligned}
c\Pi^{\mathrm{M}_1} &= \left\langle\Omega_1^2\mathcal{E}(\mathbf{u})\right\rangle(\mathbf{e}_1\vee\mathbf{e}_1 + \mathbf{e}_2\vee\mathbf{e}_2) + \left\langle(\Omega_3^2\mathcal{E}(\mathbf{u})\right\rangle\mathbf{n}\vee\mathbf{n}\\
&= \left\langle\Omega_1^2\mathcal{E}(\mathbf{u})\right\rangle(\mathbf{e}_1\vee\mathbf{e}_1 + \mathbf{e}_2\vee\mathbf{e}_2 + \mathbf{n}\vee\mathbf{n}) + \left\langle(\Omega_3^2 - \Omega_1^2)\mathcal{E}(\mathbf{u})\right\rangle\mathbf{n}\vee\mathbf{n}\\
&= \frac{1}{2}\left\langle(1-\Omega_3^2)\mathcal{E}(\mathbf{u})\right\rangle\mathrm{Id} + \frac{1}{2}\left\langle(3\Omega_3^2 - 1)\mathcal{E}(\mathbf{u})\right\rangle\mathbf{n}\vee\mathbf{n},
\end{aligned} \tag{4.52}$$

where we have used the fact that $\mathbf{e}_1\vee\mathbf{e}_1 + \mathbf{e}_2\vee\mathbf{e}_2 + \mathbf{n}\vee\mathbf{n}$ is the identity. From the definition of χ_2, we conclude that

$$\Pi^{\mathrm{M}_1} = \frac{E}{2}\left[(1-\chi_2)\mathrm{Id} + (3\chi_2 - 1)(\mathbf{n}\vee\mathbf{n})\right]. \tag{4.53}$$

Similarly for $k = 3$,

$$\begin{aligned}
Q^{\mathrm{M}_1} &= 3\left\langle\Omega_1^2\Omega_3\mathcal{E}(\mathbf{u})\right\rangle(\mathbf{e}_1\vee\mathbf{e}_1 + \mathbf{e}_2\vee\mathbf{e}_2 + \mathbf{n}\vee\mathbf{n})\vee\mathbf{n} + \left\langle(\Omega_3^2 - 3\Omega_1^2)\Omega_3\mathcal{E}(\mathbf{u})\right\rangle\mathbf{n}^{\vee 3}\\
&= \frac{3}{2}\left\langle(1-\Omega_3^2)\Omega_3\mathcal{E}(\mathbf{u})\right\rangle\mathrm{Id}\vee\mathbf{n} + \frac{1}{2}\left\langle(5\Omega_3^2 - 3)\Omega_3\mathcal{E}(\mathbf{u})\right\rangle\mathbf{n}^{\vee 3}\\
&= \frac{3cE}{2}\left[(\chi_1 - \chi_3)(\mathrm{Id}\vee\mathbf{n}) + (5\chi_3 - 3\chi_1)\mathbf{n}^{\vee 3}\right].
\end{aligned} \tag{4.54}$$

□

Proposition 3. *The perturbed M_1 system is hyperbolic if* $\Pi^{\mathrm{D}} = 0$ *and* $cE \neq |F|$.

Proof. Without loss of generality, we consider $c = 1$ and prove that the eigenvalues of the Jacobian associated with the convective flux in (4.77) are real. To do so, the following definitions are introduced:

$$\alpha := \frac{\partial}{\partial E}\left(\xi E + r_a a T^4 + \frac{S}{\sigma_{\mathrm{t}}}\right), \qquad \beta := \frac{\partial}{\partial F}\left(\xi E + r_a a T^4 + \frac{S}{\sigma_{\mathrm{t}}}\right),$$
$$\xi(f) := \chi(f) + r_s \eta(f), \qquad f := F/E.$$

We show that the radical $\alpha + \beta^2/4$ in the formula for the eigenvalues is positive for all $f \neq 1$. Note that (4.44b) implies $\eta = 1/3 + \chi' f - \chi$ and hence,

$$\xi = r_s\left(\frac{1}{3} - \chi + \chi' f\right) + \chi. \tag{4.55}$$

The prime notation always refers to the derivative with respect to f. With this, we conclude

$$\beta^2 + 4\alpha = \xi'^2 - 4f\xi' + 4\xi = \xi'^2 - 4f\xi' + 4r_s\left(\frac{1}{3} - \chi + \chi' f\right) + 4\chi \tag{4.56}$$

$$= (\xi' - 2f)^2 + 4r_s\left(\frac{1}{3} - \chi + \chi' f\right) + 4(\chi - f^2). \tag{4.57}$$

Using (4.64), straight-forward calculations imply

$$\chi - f^2 > 0 \quad \text{for all } f \neq 1 \quad \text{and} \quad \frac{1}{3} - \chi + \chi' f \geq 0. \tag{4.58}$$

Applying (4.58) on (4.57) completes the proof.

□

Lemma 10. *For the M_1 model, the multiplier $\hat{\alpha}_1$ is co-linear with F, that is*

$$\frac{\hat{\alpha}_1}{|\hat{\alpha}_1|} = \frac{F}{|F|} \tag{4.59}$$

Proof. If $\mathcal{E}(\mathbf{u}) = \eta_*'\left(-\frac{h\nu c}{k}(\hat{\alpha}_0 + \hat{\alpha}_1 \mathbf{m}_1)\right)$ solves the optimization problem (4.13), then by definition

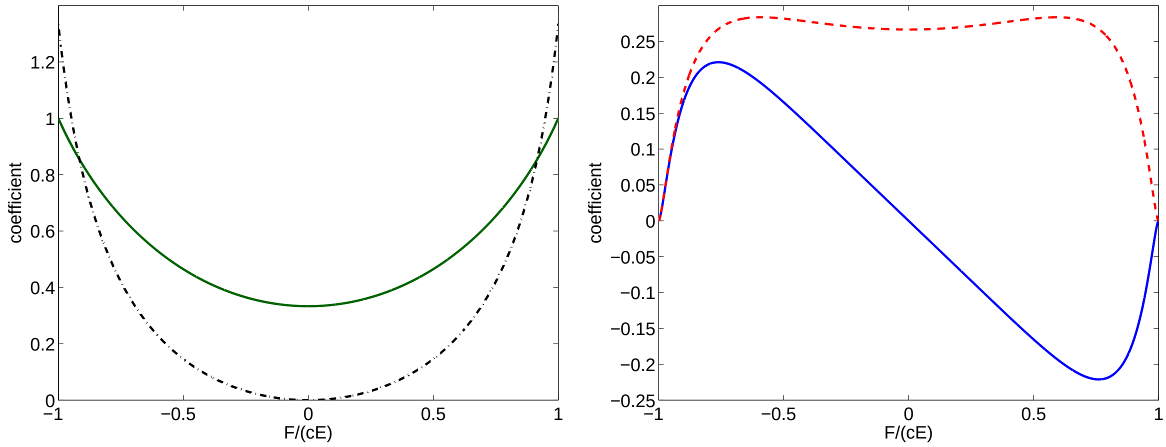
$$F = \left\langle \Omega\, \eta_*'\left(-\frac{h\nu c}{k}(\hat{\alpha}_0 + \hat{\alpha}_1 \mathbf{m}_1)\right)\right\rangle. \tag{4.60}$$

Let R be any orthogonal 3×3 matrix which preserves F. Then multiplying (4.60) by R gives

$$F = \left\langle R\Omega\, \eta_*'\left(-\frac{h\nu c}{k}(\hat{\alpha}_0 + \hat{\alpha}_1 \mathbf{m}_1)\right)\right\rangle = \left\langle \Omega\, \eta_*'\left(-\frac{h\nu c}{k}(\hat{\alpha}_0 + R\hat{\alpha}_1 \mathbf{m}_1)\right)\right\rangle, \tag{4.61}$$

where we have used the fact that the measure $d\Omega$ is invariant under the action of R. Because the solution of the optimization is unique, we conclude that $R\hat{\alpha}_1 = \hat{\alpha}_1$ and therefore, since R is arbitrary, $\hat{\alpha}_1$ and F must be co-linear.

□



(a) Convection coefficients.

(b) Diffusion coefficients.

Figure 4.1: Perturbed M1 model coefficients. Left: χ (dark green solid line) and η (black dash-dot line). Right: D_E (blue solid line) and D_F (red dashed line).

Finally, we end up with the following expressions for the components of the pressure term $\Pi = \Pi^{\mathrm{M_1}} + \Pi^{\mathrm{C}} + \Pi^{\mathrm{D}}$ which can be computed from Lemma 8 and Lemma 9:

$$\Pi^{\mathrm{M_1}} = \chi(E,F)E\,, \quad \Pi^{\mathrm{C}} = \eta(E,F)\left(r_{\mathrm{s}}E + r_{\mathrm{a}}aT^4 + \frac{S}{c\Sigma_t}\right)\,, \tag{4.62}$$

$$\Pi^{\mathrm{D}} = -\frac{1}{c\sigma_{\mathrm{t}}}\left[D_E(E,F)\partial_x E + D_F(E,F)\partial_x F\right] =: \mathbf{D}(\mathbf{u})\partial_x \mathbf{u}\,, \tag{4.63}$$

where the convection and diffusion coefficients are given by

$$\chi(E,F) = \frac{1+3\gamma^2}{3+\gamma^2}, \qquad \eta = \frac{8\gamma^2}{3(3-\gamma^2)}, \tag{4.64}$$

$$D_E(E,F) = \frac{3(\gamma^2+5)\,(\gamma^2-1)^2}{2\gamma^4(\gamma^2-3)^2}\left[(\gamma^2-3)\ln\left(\frac{1-\gamma}{1+\gamma}\right) - 6\gamma\right], \tag{4.65}$$

$$D_F(E,F) = \frac{9(\gamma^2+1)(\gamma^2-1)^2}{2\gamma^5(\gamma^2-3)^2}\left[(\gamma^2-3)\ln\left(\frac{1-\gamma}{1+\gamma}\right) - 6\gamma\right], \tag{4.66}$$

and

$$\gamma = \frac{-3F}{2cE + \sqrt{4(cE)^2 - 3F^2}}. \tag{4.67}$$

These coefficients are displayed in Figure 4.1. Note that χ, η, and D_F are even functions of the ratio $F/(cE)$, while D_E is odd.

4.3.4 Controlling the Perturbations

While the entropy-based ansatz in (4.17) is positive for all Ω, the addition of the perturbation in (4.30) may lead to an ansatz which is not. As a consequence, the moments of the perturbed ansatz may not satisfy the realizability condition (4.1). To correct for this defect, we introduce a modification and approximate ψ as

$$\mathcal{E}(\mathbf{u}) = \mathcal{B}(\hat{\boldsymbol{\alpha}}) + \delta\tilde{\psi}, \tag{4.68}$$

where $\delta(x,t)$ is a scalar control parameter. Several different choices are possible. For example, one could select δ to ensure that $\mathcal{E}(\mathbf{u})$ is positive everywhere. However, this choice requires pointwise evaluations with respect to Ω—a task we would like to avoid. Instead, we select δ in such a way as to preserve (4.1) in the numerical computation. While the exact form of δ depends on the details of the numerical method, the general framework relies on the realizability conditions for the moments $(cE, F, c\Pi)$.

Definition 2. *An array* $(\Psi_0, \Psi_1, \ldots, \Psi_N)$ *is called realizable with respect to* $(1, \Omega, \ldots, \Omega^{\otimes N})$ *if there exists a non-negative measure on* $d\Omega d\nu$ *with density* $\Psi(\Omega, \nu)$ *such that* $\Psi_k = \langle \Omega^{\otimes k} \Psi \rangle$ *for* $k = 1, \ldots, N$. *The set* $\mathcal{R}_N$ *of all such vectors is called the realizable set.*

With the ansatz (4.68), the pressure term in the M_1 model becomes

$$\Pi_\delta = \Pi^{\mathrm{M}_1} + \delta(\Pi^{\mathrm{C}} + \Pi^{\mathrm{D}}). \tag{4.69}$$

Roughly speaking, we select δ to ensure the $(cE, F, c\Pi_\delta) \in \mathcal{R}_3$. Details are given in Section 4.4.3. Note that such a δ always exists: When $\delta = 0$, $\Pi_\delta = \Pi^{\mathrm{M}_1}$, and since the M_1 ansatz is always positive, $(cE, F, c\Pi^{\mathrm{M}_1}) \in \mathcal{R}_3$.

4.4 Numerical Simulation using Discontinuous Galerkin

In slab geometries, the diffusion-corrected M_1 model and the material energy (4.6) reduce to

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, \partial_x \mathbf{u}) = \mathcal{S}(\mathbf{u}), \quad (x,t) \in (x_L, x_R) \times (0, t_{\mathrm{final}}), \tag{4.70a}$$

$$\partial_t T = \frac{c\sigma_{\mathrm{a}}}{C_v}(E - aT^4), \tag{4.70b}$$

where

$$\mathbf{u} = \begin{bmatrix} cE \\ F \end{bmatrix}, \quad \mathcal{S}(\mathbf{u}) = \begin{bmatrix} -c^2\sigma_{\mathrm{a}}(E - aT^4) + S \\ -c\sigma_{\mathrm{t}} F \end{bmatrix}, \quad \mathbf{f}(\mathbf{u}, \partial_x \mathbf{u}) = \begin{bmatrix} cF \\ c^2 \Pi_\delta \end{bmatrix}, \tag{4.70c}$$

$\Pi_\delta = \Pi^{\mathrm{M}_1} + \delta(\Pi^{\mathrm{C}} + \Pi^{\mathrm{D}})$ and $C_v = \frac{\partial e}{\partial T}$ is the specific heat at constant volume.

We simulate the system (4.70) using a Runge-Kutta discontinuous Galerkin (RKDG) method. The RKDG method is a method of lines: the DG discretization is only applied to spatial variables while time discretization is achieved by explicit Runge-Kutta time integrators. The presentation here is rather brief and relies on details found in Chapter 3, where the method was applied to the M_1 model. A general description of the RKDG method can be found, for example, in [33, 34].

4.4.1 Spatial Discretization

We divide the computational domain $[x_L, x_R]$ into J cells with edges

$$x_L = x_{1/2} < x_{3/2} < \ldots < x_{J+1/2} = x_R,$$

and let x_j denote the center of each cell $I_j = (x_{j-1/2}, x_{j+1/2})$. Let $h_j := x_{j+1/2} - x_{j-1/2}$ be the length of the interval I_j and $h := \max_j h_j$. Moreover, we denote the finite-dimensional approximation space by

$$V_h^k = \{v \in L^1(x_L, x_R) : v_{|I_j} \in \mathcal{P}^k(I_j),\ j = 1, \ldots, J\},$$

where $\mathcal{P}^k(I_j)$ is the space of polynomials of degree at most k on the interval I_j.

The semidiscrete DG scheme is derived from a weak formulation of (4.70). However, following [35] we first reduce the convection-diffusion equations (4.70) to a system of first order equations by introducing the auxiliary variable $\mathbf{v}$:

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, \mathbf{v}) = \mathcal{S}(\mathbf{u}), \tag{4.71a}$$

$$\partial_x \mathbf{u} = \mathbf{v}, \tag{4.71b}$$

$$\partial_t T = \frac{c\sigma_\mathrm{a}}{C_v}(E - aT^4). \tag{4.71c}$$

Then the exact solutions $\mathbf{u}(\cdot, t)$, $\mathbf{v}(\cdot, t)$ and $T(\cdot, t)$ are replaced by approximations $\mathbf{u}_h(\cdot, t)$, $\mathbf{v}_h(\cdot, t) \in V_h^k \times V_h^k$ and $T_h(\cdot, t) \in V_h^k$. The resulting set of equations is then required to hold for all test functions $\varphi_h \in V_h^k$:

$$\int_{I_j} \varphi_h(x)\partial_t \mathbf{u}_h(x,t)dx - \int_{I_j} \mathbf{f}(\mathbf{u}_h(x,t), \mathbf{v}_h(x,t))\partial_x \varphi_h(x)dx \tag{4.72a}$$

$$+ [\![\mathbf{f}\varphi_h(x)]\!]_j = \int_{I_j} \mathcal{S}(\mathbf{u}_h(x,t))\varphi_h(x)dx$$

$$\int_{I_j} \varphi_h(x)\mathbf{v}_h(x,t)dx + \int_{I_j} \mathbf{u}_h(x,t)\partial_x \varphi_h(x)dx - [\![\mathbf{g}\varphi_h(x)]\!]_j = 0 \tag{4.72b}$$

$$\int_{I_j} \varphi_h(x)\partial_t T_h(x,t)dx = \int_{I_j} \varphi_h(x)\frac{c\sigma_\mathrm{a}(x)}{C_v}(E_h(x,t) - aT_h^4(x,t))dx. \tag{4.72c}$$

Here we use the bracket notation:

$$[\![\mathbf{f}\varphi_h(x)]\!]_j = \mathbf{f}_{j+1/2}\varphi_h(x_{j+1/2}^-) - \mathbf{f}_{j-1/2}\varphi_h(x_{j-1/2}^+) \tag{4.73}$$

where

$$\mathbf{f}_{j\pm1/2}(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{u}(x_{j\pm1/2}, t), \mathbf{v}(x_{j\pm1/2}, t)) \quad \text{and} \quad \mathbf{g}_{j\pm1/2}(\mathbf{u}) = \mathbf{u}(x_{j\pm1/2}, t) \tag{4.74}$$

and

$$\varphi_h(x_{j+1/2}^-) = \lim_{\varepsilon\to0^+} \varphi_h(x_{j+1/2} - \varepsilon), \quad \varphi_h(x_{j-1/2}^+) = \lim_{\varepsilon\to0^+} \varphi_h(x_{j-1/2} + \varepsilon) \tag{4.75}$$

are the right and left limits of φ_h at the cell interfaces $x_{j\pm1/2}$. The term $[\![\hat{\mathbf{g}}\varphi_h(x)]\!]_j$ is defined in an analogous fashion.

Since the components of $\mathbf{u}_h(.,t)$ and $\mathbf{v}_h(.,t)$ are piecewise polynomials, the edge values of $\mathbf{u}$ and $\mathbf{v}$ in (4.74) are not strictly defined. Thus, the nonlinear flux function $\mathbf{f}$ is replaced by a numerical flux $\hat{\mathbf{f}}$ which depends on the pointwise limits of $\mathbf{u}_h$, $\mathbf{v}_h$ on either side of the edge at $x_{j\pm1/2}$:

$$\hat{\mathbf{f}}_{j\pm1/2} = \hat{\mathbf{f}}(\mathbf{u}_h(x_{j\pm1/2}^-, t), \mathbf{u}_h(x_{j\pm1/2}^+, t), \mathbf{v}_h(x_{j\pm1/2}^-, t), \mathbf{v}_h(x_{j\pm1/2}^+, t)). \tag{4.76}$$

The notations for $\hat{\mathbf{g}}$ carry over analogously.

It remains to choose suitable numerical fluxes $\hat{\mathbf{f}}$ and $\hat{\mathbf{g}}$. Since (4.70) has both a convective flux

$$\mathbf{f}^\mathrm{C}(\mathbf{u}) := \begin{bmatrix} cF \\ c^2\,\Pi^{\mathrm{M}_1} + \frac{c}{\sigma_\mathrm{t}}\eta(E, F)\left(c\sigma_\mathrm{s}E + c\sigma_\mathrm{a}aT^4 + S\right) \end{bmatrix} \tag{4.77}$$

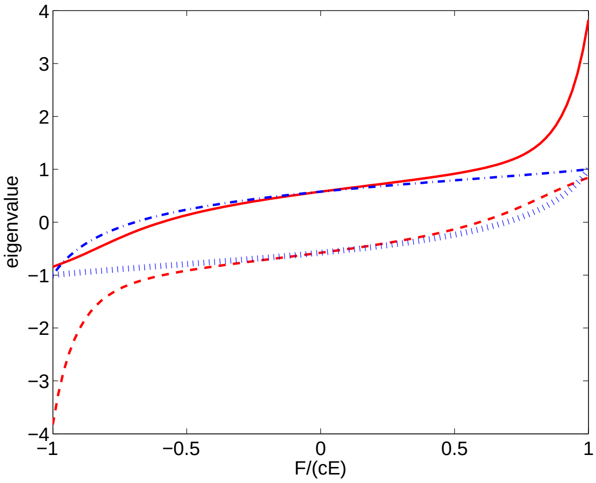


Figure 4.2: Eigenvalues of the hyperbolic flux Jacobian: M1 model (blue lines), perturbed M1 model (red lines).

and a diffusive flux

$$\mathbf{f}^{\mathrm{D}}(\mathbf{u},\mathbf{v}) = \begin{bmatrix} 0 \\ c^2\,\mathbf{D}(\mathbf{u})\cdot\mathbf{v} \end{bmatrix} \tag{4.78}$$

the choice is not obvious. Several approaches have been presented in literature [11, 115, 134, 167]. In [11], the prescription for the diffusive term is given by

$$\hat{\mathbf{f}}^{\mathrm{D}}_{j\pm1/2} = \frac{1}{2}\left[\mathbf{f}^{\mathrm{D}}(\mathbf{u}^-_{j\pm1/2},\mathbf{v}^-_{j\pm1/2}) + \mathbf{f}^{\mathrm{D}}(\mathbf{u}^+_{j\pm1/2},\mathbf{v}^+_{j\pm1/2})\right], \tag{4.79}$$

$$\hat{\mathbf{g}}_{j\pm1/2} = \frac{1}{2}\left[\mathbf{u}^-_{j\pm1/2} + \mathbf{u}^+_{j\pm1/2}\right]. \tag{4.80}$$

Combining this term with the Lax-Friedrichs flux for $\mathbf{f}^{\mathrm{C}}(\mathbf{u})$ gives the following total numerical flux:

$$\hat{\mathbf{f}}_{j\pm1/2} = \frac{1}{2}\left[\mathbf{f}(\mathbf{u}^-_{j\pm1/2},\mathbf{v}^-_{j\pm1/2}) + \mathbf{f}(\mathbf{u}^+_{j\pm1/2},\mathbf{v}^+_{j\pm1/2}) - \lambda(\mathbf{u}^+_{j\pm1/2} - \mathbf{u}^-_{j\pm1/2})\right], \tag{4.81}$$

where λ is the largest magnitude of any eigenvalue of the Jacobian associated with $\mathbf{f}^{\mathrm{C}}$. These eigenvalues, in general, depend on material properties, the temperature T and the source term S. In contrast to the M1 model, they are not bounded by the speed of light c. For example, neglecting the temperature and source the maximum value is approximately $9.12\,c$. However, we instead use the smaller value of $\lambda = c$, which is the particle speed in the transport equation and is consistent with the application of the control parameter to enforce realizability (see Section 4.3.4). Figure 4.2 shows the comparison between M1 and perturbed M1 for $c = 1$, $\sigma_{\mathrm{s}} = 1$, $\sigma_{\mathrm{t}} = 3$, $T = 0 = S$.

Remark 16. *The above formulation is formally very similar to a DG discretization of a purely hyperbolic system. However, we stress that it is a mixed convection-diffusion problem which is suitably rewritten into a larger system with first order derivatives. This first order form and the special choice for the numerical fluxes will be useful for the proof of Lemma 14 in Section 4.4.3.*

The DG solutions $\mathbf{u}_h$, $\mathbf{v}_h$ and T_h are expanded in terms of local basis functions $\{\phi_l^j\}_{l=0}^k$ for $\mathcal{P}^k(I_j)$ in each cell I_j:

$$\mathbf{u}_h^j(x,t) = \sum_{l=0}^k \mathbf{u}_l^j(t)\phi_l^j(x), \quad \mathbf{v}_h^j(x,t) = \sum_{l=0}^k \mathbf{v}_l^j(t)\phi_l^j(x), \quad T_h^j(x,t) = \sum_{l=0}^k T_l^j(t)\phi_l^j(x).$$

The standard choice of basis for $\mathcal{P}(I_j)$ is generated by Legendre polynomials P_l, defined on the reference cell $[-1,1]$ for $j \in \{1,\dots,J\}$ and $l \in \{0,\dots,k\}$:

$$\phi_l^j(x) = P_l\left(\frac{2(x-x_j)}{h_j}\right), \tag{4.82}$$

which satisfy the orthoganility condition

$$\int_{-1}^1 P_l(y)P_m(y)dy = \frac{2}{2m+1}\delta_{l,m}. \tag{4.83}$$

With $\xi_j(y) := x_j + yh_j/2$, this gives a formulation defined on the reference cell:

$$\begin{aligned}
&\frac{h_j}{2m+1}\partial_t \mathbf{u}_m^j(t) - \int_{-1}^1 \mathbf{f}(\mathbf{u}_h^j(\xi_j(y),t), \mathbf{v}_h^j(\xi_j(y),t))\partial_y P_m(y)dy && \text{(4.84a)}\\
&\qquad + \hat{\mathbf{f}}_{j+1/2} - (-1)^m \hat{\mathbf{f}}_{j-1/2} = \frac{h_j}{2}\int_{-1}^1 \mathcal{S}(\mathbf{u}_h^j(\xi_j(y),t))P_m(y)dy, \\
&\frac{h_j}{2m+1}\mathbf{v}_m^j(t) + \sum_{l=0}^k \mathbf{u}_l^j(t)\,\mathcal{C}_{l,m} - \hat{\mathbf{g}}_{j+1/2} + (-1)^m\hat{\mathbf{g}}_{j-1/2} = 0, && \text{(4.84b)}\\
&\frac{h_j}{2m+1}\partial_t T_m^j(t) = \frac{ch_j}{2C_v}\sum_{l=0}^k E_l^j(t)\left(\int_{-1}^1 P_m(y)P_l(y)\sigma_{\mathrm{a}}(\xi_j(y))\right)dy && \text{(4.84c)}\\
&\qquad - \frac{ach_j}{2C_v}\int_{-1}^1 P_m(y){T_h^j}^4(\xi_j(y),t)\sigma_{\mathrm{a}}(y)dy,
\end{aligned}$$

where

$$\mathcal{C}_{l,m} = \int_{-1}^1 P_l(y)\partial_y P_m(y)dy. \tag{4.85}$$

The remaining integrals are calculated by a quadrature rule. Note that (4.84b) can be solved locally for $\mathbf{v}_m^j(t)$ in each cell I_j. Each $\mathbf{v}_m^j(t)$ depends on coefficients $\mathbf{u}_m^{j-1}$, $\mathbf{u}_m^j$, $\mathbf{u}_m^{j+1}$ and can be plugged in (4.84a).

Equations (4.84) form a system of ODEs for the coefficients $\mathbf{u}_m^j(t)$ and $T_m^j(t)$. For all $j \in \{1,\dots,J\}$ and $m \in \{0,\dots,k\}$, we write this system in the abstract form:

$$\partial_t \mathbf{u}_m^j(t) = \mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_0^{j-1},\dots,\mathbf{u}_k^{j-1},\mathbf{u}_0^j,\dots,\mathbf{u}_k^j,\mathbf{u}_0^{j+1},\dots,\mathbf{u}_k^{j+1}), \tag{4.86}$$

$$\partial_t T_m^j(t) = \mathcal{L}_{T,m}^j(E_0^j,\dots,E_k^j,T_0^j,\dots,T_k^j). \tag{4.87}$$

Here, $\mathcal{L}_{\mathbf{u},m}^j$ and $\mathcal{L}_{T,m}^j$ are the respective right-hand sides of the ODEs.

4.4.2 Time Discretization: Explicit SSP Runge-Kutta Schemes

The purpose of high-order, strong stability preserving (SSP) Runge-Kutta time integration methods is to achieve high-order accuracy in time while preserving desirable properties of the forward Euler method (for a review, see [66]). In our implementaion, we only use explicit schemes, which compute values of the unknowns at several intermediate stages. Each stage is a convex combination of forward Euler operators and this usually leads to modified CFL restrictions.

Let $\{t^n\}_{n=0}^N$ be an equidistant partition of $[0, t_{\text{final}}]$ and set $\Delta t := t_{\text{final}}/N$. Let Λ denote the application of a generic slope limiter. The algorithm for the optimal third-order SSP Runge-Kutta (SSPRK(3,3)) method [166] reads as follows:

- For all $j \in \{1, \dots, J\}$ and $m \in \{0, \dots, k\}$, set $\mathbf{u}_m^{j,0} = \Lambda\{\pi_{V_h^m}(\mathbf{u}_0)\}$.
- For all $n \in \{0, \dots, N-1\}$, $j \in \{1, \dots, J\}$, and $m \in \{0, \dots, k\}$,
    (1) Compute the intermediate stages

$$\begin{aligned} \mathbf{u}_m^{j,(1)} &= \Lambda\left\{\mathbf{u}_m^{j,n} + \Delta t \mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_h^{j,(0)})\right\} \\ \mathbf{u}_m^{j,(2)} &= \Lambda\left\{\frac{3}{4}\mathbf{u}_m^{j,n} + \frac{1}{4}\mathbf{u}_m^{j,(1)} + \frac{1}{4}\Delta t \mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_h^{j,(1)})\right\} \\ \mathbf{u}_m^{j,(3)} &= \Lambda\left\{\frac{1}{3}\mathbf{u}_m^{j,n} + \frac{2}{3}\mathbf{u}_m^{j,(2)} + \frac{2}{3}\Delta t \mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_h^{j,(2)})\right\} \end{aligned} \tag{4.88}$$

    (2) Set $\mathbf{u}_m^{j,n+1} = \mathbf{u}_m^{j,(3)}$.

In the initial step, $\pi_{V_h^m}(\mathbf{u}_0)$ is the projection of the initial condition $\mathbf{u}_0$ onto the finite dimensional space V_h^m. Note that Λ is applied at every Runge-Kutta stage.
For the sake of completeness, we also state the optimal fourth order scheme SSPRK(5,4) [66]:

$$\begin{aligned} \mathbf{u}_m^{j,(1)} &= \Lambda\{\mathbf{u}_h^{(n)} + 0.391752226571890\,\Delta t\,\mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_m^{j,(n)})\} \\ \mathbf{u}_m^{j,(2)} &= \Lambda\{0.444370493651235\,\mathbf{u}_m^{j,(n)} + 0.555629506348765\,\mathbf{u}_m^{j,(1)} \\ &\quad + 0.368410593050371\,\Delta t\,\mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_m^{j,(1)})\} \\ \mathbf{u}_m^{j,(3)} &= \Lambda\{0.620101851488403\,\mathbf{u}_m^{j,(n)} + 0.379898148511597\,\mathbf{u}_m^{j,(2)} \\ &\quad + 0.251891774271694\,\Delta t\,\mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_m^{j,(2)})\} \\ \mathbf{u}_m^{j,(4)} &= \Lambda\{0.178079954393132\,\mathbf{u}_m^{j,(n)} + 0.821920045606868\,\mathbf{u}_m^{j,(3)} \\ &\quad + 0.544974750228521\,\Delta t\,\mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_m^{j,(3)})\} \\ \mathbf{u}_h^{(n+1)} &= \Lambda\{0.517231671970585\,\mathbf{u}_m^{j,(2)} + 0.096059710526147\,\mathbf{u}_m^{j,(3)} \\ &\quad + 0.386708617503269\,\mathbf{u}_m^{j,(4)} + 0.063692468666290\,\Delta t\,\mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_m^{j,(3)}) \\ &\quad + 0.226007483236906\,\Delta t\,\mathcal{L}_{\mathbf{u},m}^j(\mathbf{u}_m^{j,(4)})\}. \end{aligned} \tag{4.89}$$

Note that SSPRK(3,3) permits a timestep of the same size as forward Euler, while the SSPRK(5,4) method is less restrictive, allowing for a time step that is 1.508 times larger the forward Euler scheme.

4.4.3 Limiters

As in Chapter 3, two types of limiters are used. The first is standard; it is used to suppress spurious oscillations and maintain stability. There are many such limiters available. Here, we apply the moment limiter from [25], which is a modification of the original limiter in [15]. This limiter is applied to the variables $\mathbf{u}$, but not the auxiliary variables $\mathbf{v}$ or the temperature T. Additional details can be found in Chapter 3.

Realizability-Preserving Limiter

The second limiter is a *realizability-preserving limiter* that is needed to ensure that the cell averages of E and F satisfy the condition (4.1) at each stage of the numerical computation. The application of the limiter is very similar to what was done in Chapter 3 for the M_1 model. The major difference here is the addition of the control parameter δ.

An essential ingredient of the realizability limiter is the Gauss-Lobatto quadrature set

$$\{x_{j-1/2} = \hat{x}_j^1, \hat{x}_j^2, \dots, \hat{x}_j^{M-1}, \hat{x}_j^M = x_{j+1/2}\} \subset I_j, \tag{4.90}$$

where, for a spatial reconstruction of order k, M is the smallest integer such that $2M-3 \geq 2k+1$. This condition on M ensures accuracy of the scheme [32]. The weaker condition $2M-3 \geq k$ ensures that the quadrature integrates elements of the approximation space V_h^k exactly.

The realizability limiter is defined in order to ensure that $\mathbf{u}_h(\hat{x}_j^\ell) \in \mathcal{R}_2$ at each point $\hat{x}_j^\ell$ in the quadrature set. However, we enforce the convexity condition indirectly by requiring the positivity of the intermediate quantities[2]

$$Q := \frac{cE+F}{2} \quad \text{and} \quad R := \frac{cE-F}{2}\,. \tag{4.91}$$

The inverse transformation that maps $(Q,R) \mapsto (cE,F)$ is given by

$$E = \frac{Q+R}{c} \quad \text{and} \quad F = Q-R\,. \tag{4.92}$$

An additional limiter is also used to enforce the positivity of the temperature reconstruction at each quadrature point.

We now proceed to define the limiters. Let $\mathbf{u}_h^{j,n} = (cE_h^{j,n}, F_h^{j,n})$ and $T_h^{j,n}$ be the approximations of $\mathbf{u}$ and T in cell I_j at time t^n, and let $\hat{\mathbf{u}}_h^{j,n}$ and $\hat{T}_h^{j,n}$ denote the modifications of $\mathbf{u}_h^{j,n}$ and $T_h^{j,n}$ that are generated by the limiting. We assume that the cell average of $\mathbf{u}_h^{j,n}$, which we denote by $\overline{\mathbf{u}}_h^{j,n}$, is realizable, i.e., $\overline{\mathbf{u}}_h^{j,n} \in \mathcal{R}_2$. We also assume that the cell average of $T_h^{j,n}$, which we denote by $\overline{T}_h^{j,n}$, is positive. Let $Q_h^{j,n}(x)$ and $R_h^{j,n}(x)$ be the approximations of Q and R, respectively, and define limited variables by

$$\hat{Q}_h^{j,n}(x) = \theta_Q^{j,n} Q_h^{j,n}(x) + (1-\theta_Q^{j,n})\overline{Q}_h^{j,n}\,, \tag{4.93a}$$

$$\hat{R}_h^{j,n}(x) = \theta_R^{j,n} R_h^{j,n}(x) + (1-\theta_R^{j,n})\overline{R}_h^{j,n}\,, \tag{4.93b}$$

$$\hat{T}_h^{j,n}(x) = \theta_T^{j,n} T_h^{j,n}(x) + (1-\theta_T^{j,n})\overline{T}_h^{j,n}\,, \tag{4.93c}$$

---

[2] The meaning of all subsequent subscripts, superscripts and adornments of Q and R will be inherited from analogous definitions for E and F.

where

$$\theta_Q^{j,n} := \min\left\{\frac{\overline{Q}_h^{j,n} - \varepsilon/2}{\overline{Q}_h^{j,n} - Q_{\min}^{j,n}}, 1\right\}, \quad Q_{\min}^{j,n} := \min_{\ell=1,\dots,M} Q_h^{j,n}(\hat{x}_j^\ell)\,, \tag{4.93d}$$

$$\theta_R^{j,n} := \min\left\{\frac{\overline{R}_h^{j,n} - \varepsilon/2}{\overline{R}_h^{j,n} - R_{\min}^{j,n}}, 1\right\}, \quad R_{\min}^{j,n} := \min_{\ell=1,\dots,M} R_h^{j,n}(\hat{x}_j^\ell)\,, \tag{4.93e}$$

$$\theta_T^{j,n} := \min\left\{\frac{\overline{T}_h^{j,n}}{\overline{T}_h^{j,n} - T_{\min}^{j,n}}, 1\right\}, \quad T_{\min}^{j,n} := \min_{\ell=1,\dots,M} T_h^{j,n}(\hat{x}_j^\ell)\,. \tag{4.93f}$$

The parameter $\varepsilon > 0$ is chosen to maintain numerical stability with finite precision arithmetic; its value should be small relative to the magnitude of the variables in a given problem. The components of $\hat{\mathbf{u}}_h^{j,n}$ are then defined using (4.92). They satisfy the following property which is a key ingredient for maintaining realizability in the RKDG scheme.

Lemma 11 (Lemma 6, Chapter 3)**.** *If* $\overline{\mathbf{u}}_h^{j,n} \in \mathcal{R}_2$ *(respectively:* $\overline{T}_h^{j,n} \geq 0$*), then* $\hat{\mathbf{u}}_h^{j,n}(\hat{x}_j^\ell) \in \mathcal{R}_2^\varepsilon := \mathcal{R}_2 + [\varepsilon, 0]^T$ *(respectively:* $\hat{T}_h^{j,n}(\hat{x}_j^\ell) \geq 0$*) for* $\ell = 1, \dots, M$*.*

Setting the Control Parameter

We now proceed to define the control parameter δ, discussed in Section 4.3.4, using the following result.

Lemma 12 (Lemma 3, Chapter 3)**.** *In the one dimensional setting, a necessary condition for* $(cE, F, c\Pi_\delta) \in \mathcal{R}_3$ *is that*

(C1) $\Pi_\delta \leq E$*,*

(C2) $|F \pm c\Pi_\delta| \leq cE \pm F$*.*

Rather than to require $(cE, F, c\Pi_\delta) \in \mathcal{R}_3$, we choose $\delta \in [0, 1]$ to ensure the weaker conditions (C1) and (C2). More specifically, for any $(cE, F) \in \mathcal{R}_2$, we set

$$\delta(E, F) = \begin{cases} \delta_0(E, F), & \Pi^{\mathrm{C}}(E, F) + \Pi^{\mathrm{D}}(E, F) > 0, \\ \delta_1(E, F), & \Pi^{\mathrm{C}}(E, F) + \Pi^{\mathrm{D}}(E, F) < 0, \end{cases} \tag{4.94a}$$

where

$$\delta_0 = \min\left\{\frac{E - \Pi^{\mathrm{M}_1}}{\Pi^{\mathrm{C}} + \Pi^{\mathrm{D}}}, 1\right\}, \tag{4.94b}$$

$$\delta_1 = \min\left\{\frac{-2F + cE + c\Pi^{\mathrm{M}_1}}{c|\Pi^{\mathrm{C}} + \Pi^{\mathrm{D}}|}, \frac{2F + cE + c\Pi^{\mathrm{M}_1}}{c|\Pi^{\mathrm{C}} + \Pi^{\mathrm{D}}|}, 1\right\}. \tag{4.94c}$$

Lemma 13. *For all* $(cE, F) \in \mathcal{R}_2$*,* $\Pi_\delta := \Pi^{\mathrm{M}_1} + \delta[\Pi^{\mathrm{C}} + \Pi^{\mathrm{D}}]$ *satisfies (C1) – (C2).*

Proof. The assertion $(cE, F) \in \mathcal{R}_2$ implies $(cE, F, c\Pi^{\mathrm{M}_1}) \in \mathcal{R}_3$. It follows then that for $\Pi^{\mathrm{D}} = 0$, conditions (C1) – (C2) are trivially satisfied. It remains only to show the following inequalities:

$$c\Pi_\delta \leq cE \quad \text{and} \quad c\Pi_\delta \geq 2F - cE \quad \text{and} \quad c\Pi_\delta \geq -2F - cE. \tag{4.95}$$

These relations are easily verified by applying the definition of δ and using the fact that $(cE, F, c\Pi^{\mathrm{M}_1}) \in \mathcal{R}_3$. $\square$

With δ given by (4.94), one can show that cell averages of $\mathbf{u}_h$ remain realizable and that the cell average of T_h remains positive in a forward Euler step. Let

$$\mathbf{u}_\ell^{j,n} := \mathbf{u}_h^{j,n}(\hat{x}_j^\ell)\,, \quad T_\ell^{j,n} := T_h^{j,n}(\hat{x}_j^\ell)\,, \quad \Pi_{\delta,\ell}^{j,n} := \Pi_\delta(\mathbf{u}_\ell^{j,n})\,, \quad \sigma_{\mathrm{t},\ell} := \sigma_{\mathrm{t}}(\hat{x}_j^\ell). \tag{4.96}$$

Lemma 14. *Assume that* $2M - 3 \geq k$ *and for each* $\ell = 1, \ldots, M$,

$$\mathbf{u}_\ell^{j,n} \in \mathcal{R}_2, \quad T_\ell^{j,n} \geq 0, \tag{4.97}$$

and $\Pi_{\delta,\ell}^{j,n}$ *satisfies (C1) and (C2). Assume further that* Δt *satisfies the following conditions:*

(A1) $\Delta t < \min\limits_{\ell=1,\ldots,M} \left\{ \dfrac{1}{c\sigma_{\mathrm{t},\ell}} \right\}$,

(A2) $\Delta t < \min\limits_{\ell=1,\ldots,M} \left\{ \dfrac{w_\ell h}{c(1 + w_\ell \sigma_{\mathrm{t},\ell} h)} \right\}$,

(A3) $\Delta t \leq \min\limits_{\ell=1,\ldots,M} \left\{ \dfrac{C_v}{\sigma_{\mathrm{a},\ell} ac (T_\ell^{j,n})^3} \right\}$.

where $h := \min_j h_j$. *Then after a forward Euler time step,*

$$\overline{\mathbf{u}}_h^{j,n+1} \in \mathcal{R}_2 \quad \textit{and} \quad \overline{T}_h^{j,n+1} \geq 0. \tag{4.98}$$

Proof. We refer the reader to the proof of Theorem 3 in Chapter 3 for the M_1 model, which relies exactly on the conditions (A1)–(A3) and (C1)–(C2). The only difference is that (C1) and (C2) are assumed in Lemma 14, while in Chapter 3 they are naturally satisfied by the M_1 model. $\square$

Theorem 5. *The Runge-Kutta discontinuous Galerkin scheme which combines*

(1) the space discretization in (4.84),

(2) the limiters in (4.93),

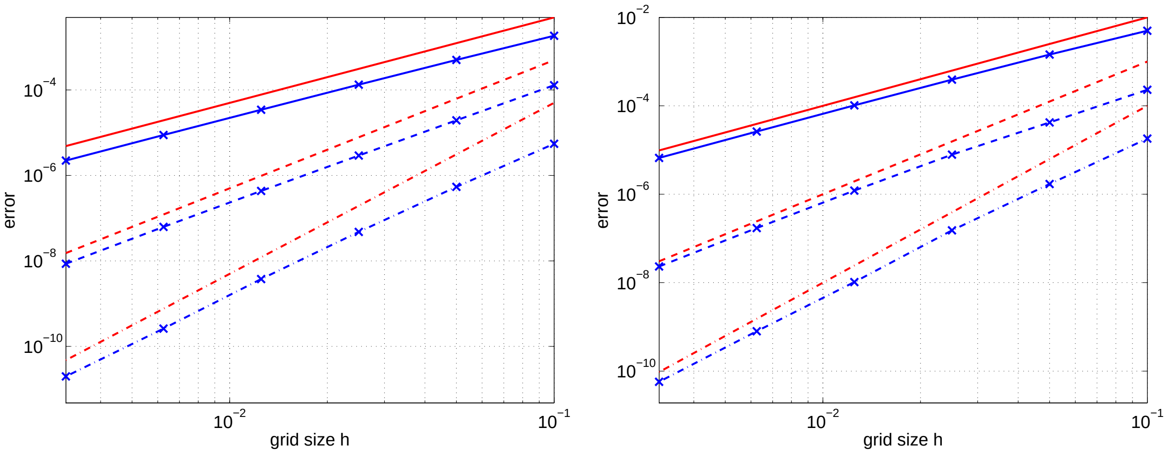
(3) the modified pressure Π_δ *in* (4.69) *with control parameter* δ *given by* (4.94),

(4) a strong-stability-preserving Runge Kutta time integrator, and

(5) a sufficiently accurate Gauss-Lobatto quadrature

preserves the realizability of the moments in the sense of cell averages. In particular, if the time step conditions (A1)-(A2) in the statement of Lemma 14 hold and if $\overline{\mathbf{u}}_h^{j,n} \in \mathcal{R}_2$, *then* $\overline{\mathbf{u}}_h^{j,n+1} \in \mathcal{R}_2$.

Proof. Application of the limiters in (4.93) ensures that the conditions of Lemma 14 hold at each stage in the SSP-RK scheme. Each successive stage is an application of the forward Euler operator to the current stage with an appropriately modified time step. Thus, the conclusions of Lemma 14 apply at the next stage, including the final stage, which gives $\mathbf{u}_h^{j,n+1}$. $\square$



(a) L^1-error

(b) L^∞-error

Figure 4.3: Accuracy test at $t = 0.1$, realizability limited: $k = 1$ (blue cross solid line), $k = 2$ (blue cross dashed line), $k = 3$ (blue cross dash-dot line), slope 2 (red solid line), slope 3 (red dashed line), slope 4 (red dash-dot line).

4.5 Numerical Results

We perform numerical computations for a choice of test cases which are common for the M1 model. The goal is to compare and contrast the perturbed M1 model with the M1 model and to point out benefits and drawbacks. Benchmark solutions are generated by the discrete ordinates method, high-order spherical harmonics or semi-analytic expressions. As in Chapter 3 our algorithm is implemented in MATLAB and Legendre Gauss-Lobatto quadrature on [-1,1] is used. Additionally, the Runge-Kutta time integration methods as well as parameters for the admissibility limiter are applied in the same way. Following Lemma 14 and to guarantee stability the time step is set to

$$\Delta t < \min\left\{c_1, c_2, c_3, c_4\right\},$$

$$c_1 = \frac{1}{c\sigma_{\mathrm{t,max}}}, \quad c_2 = \frac{h\, w_{\min}}{c(1 + w_{\max}\sigma_{\mathrm{t,max}}\, h)}, \quad c_3 = \frac{C_v}{ac\tau_{\max}} \quad \text{and} \quad c_4 = \frac{h^2}{2(2k+1)},$$

where k is the polynomial degree, $h = \min_j h_j$, $w_{\min}$ and $w_{\max}$ are the minimum and maximum quadrature weights, respectively. The quantities $\sigma_{t,\max}$ and $\tau_{\max}$ are the maximum values of $\sigma_{\mathrm{t},\ell}$ and $\sigma_{\mathrm{a},\ell}(T_\ell^{j,n})^3$.

The constant c_4 is needed to keep the numerical scheme stable. As we discretize a system of time-dependent equations with diffusion terms one might expect a parabolic CFL condition. Indeed, our numerical calculations indicate that this condition is needed. Otherwise, for $\Delta t > c_4$ we observe instabilities in the solutions. Obviously, the parabolic CFL restriction leads to small time steps. However, it should be emphasized that this restriction is not necessary to preserve realizability of the moments.

The stability parameter for the realizability limiter of E and F is set to $\varepsilon = 10^{-10}$. The same value is also used to enforce conditions (C1) and (C2), i.e., the control parameter in (4.94) is chosen such that

$$c\Pi_\delta \leq \Psi_0 - \varepsilon \quad \text{and} \quad |\Psi_1 \pm c\Pi_\delta| \leq \Psi_0 \pm \Psi_1 \pm \varepsilon.$$

	h	$\mathcal{P}^1$ order	$\mathcal{P}^2$ order	$\mathcal{P}^3$ order
unlimited	1/10	–	–	–
	1/20	1.88	2.74	3.35
	1/40	1.92	2.74	3.50
	1/80	1.95	2.75	3.67
	1/160	1.97	2.80	3.85
	1/320	1.98	2.86	3.72
realiz. limited	1/10	–	–	–
	1/20	1.88	2.74	3.35
	1/40	1.92	2.74	3.50
	1/80	1.95	2.75	3.67
	1/160	1.97	2.80	3.85
	1/320	1.98	2.86	3.72
slope limited	1/10	–	–	–
	1/20	1.40	2.62	9.33
	1/40	2.22	2.42	3.83
	1/80	2.09	2.70	3.76
	1/160	2.08	3.02	3.92
	1/320	2.07	3.00	3.78
slope + realiz. limited	1/10	–	–	–
	1/20	1.40	2.62	9.33
	1/40	2.22	2.42	3.83
	1/80	2.09	2.70	3.76
	1/160	2.08	3.02	3.92
	1/320	2.07	3.00	3.79

Table 4.1: L^1-order of accuracy for E.

	h	$\mathcal{P}^1$ order	$\mathcal{P}^2$ order	$\mathcal{P}^3$ order
unlimited	1/10	–	–	–
	1/20	1.79	2.45	3.42
	1/40	1.88	2.42	3.48
	1/80	1.94	2.70	3.89
	1/160	1.97	2.82	3.69
	1/320	1.98	2.89	3.80
realiz. limited	1/10	–	–	–
	1/20	1.79	2.45	3.42
	1/40	1.88	2.42	3.48
	1/80	1.94	2.70	3.89
	1/160	1.97	2.82	3.69
	1/320	1.98	2.89	3.80
slope limited	1/10	–	–	–
	1/20	0.85	2.22	9.98
	1/40	1.58	1.26	3.29
	1/80	1.74	2.33	3.12
	1/160	1.83	2.84	3.71
	1/320	1.91	2.95	4.88
slope + realiz. limited	1/10	–	–	–
	1/20	0.85	2.22	9.98
	1/40	1.58	1.26	3.29
	1/80	1.74	2.33	3.12
	1/160	1.83	2.84	3.71
	1/320	1.91	2.95	4.89

Table 4.2: L^∞-order of accuracy for E.

In Sections 4.5.1-4.5.4, we study simulations with $c = 1$ and neglect the energy equation which is included in the last two cases from Section 4.5.5. Unless otherwise stated, slope and realizability limiters are always turned on for all DG calculations. If transformation to characteristic variables for the slope limiter is used, it will be explicitly stated.

Note that the flux $\mathbf{f}$ in eqs. (4.70) depends explicitly on a space-dependent source S, the temperature T as well as material coefficients σ_a and σ_s. To account for discontinuous fluxes the numerical scheme needs to be treated with special care. In [181], e.g., auxiliary equations are introduced for discontinuous variables to reduce the system to a standard form. Although in Section 4.5.3 and the benchmark problem of *Su and Olson* (Section 4.5.5) discontinuous material properties are defined we set the discontinuities at cell edges and do not modify our numerical scheme.

4.5.1 Accuracy Test: Manufactured Solutions

Our code is verified by the method of manufactured solutions [153]. By adding or modifying the source terms in the original equations in an appropriate manner, this technique enables to choose simple analytical functions which solve the respective equations. Enforcing periodic boundary conditions we keep the nonlinear coefficients as simple as possible and set

$$\alpha = \cos(\pi x)/6 + 1/4, \quad x \in [-1, 1],$$
$$E(x,t) = (\cos(\pi x) + 2)\, t, \quad x \in [-1, 1], \quad t \geq 0.$$

The first moment can then be computed to

$$F(x,t) = \frac{-48\,(2\,\cos(\pi x)+3)\,(\cos(\pi x)+2)}{4\,(\cos(\pi x))^2 + 12\,\cos(\pi x) + 441}\,t \quad x \in [0,1], \quad t \geq 0. \tag{4.100}$$

In particular, α is chosen such that E and F remain inside the realizability set $\mathcal{R}_2$ for all $t > 0$. Moreover, both moments fulfill conditions (C1) and (C2) and material constants are set to $\sigma_\mathrm{a} = 0.5$ and $\sigma_\mathrm{s} = 1$. For every polynomial degree, L^∞- and L^1-error are calculated for different numbers of spatial discretization points ranging from $2^0 \cdot 10 = 10$ to $2^5 \cdot 10 = 320$.

Figure 4.3 displays the convergence of approximate solutions with the realizability limiter turned on. We obtain optimal convergence orders of $k+1$ for polynomials of degree at most k. The slope limited results also show optimal convergence rates in both the L^1- and L^∞-norm.

Table 4.1 and Table 4.2 show the convergence orders in detail. At $h = 1/320$ small degradations in the L^1-error for third order polynomials occur. The reason is that the generated source terms can only be evaluated with a precision of roughly 10^{-10}. They include complicated expressions leading to numerical instabilities.

4.5.2 Two Beam Instability

We impose two incoming beams at the boundaries of the domain $[-0.5, 0, 5]$ and set $c = 1$, $S = 0$. Particles stream from both boundaries in a purely absorbing material with $\sigma_a = 4 = \sigma_\mathrm{t}$ until they meet at $x = 0$. In our moment model, this is realized by the boundary conditions

$$\mathbf{u}(0,t) = [1, 0.9999]^T, \quad \mathbf{u}(1,t) = [1, -0.9999]^T, \quad t > 0$$

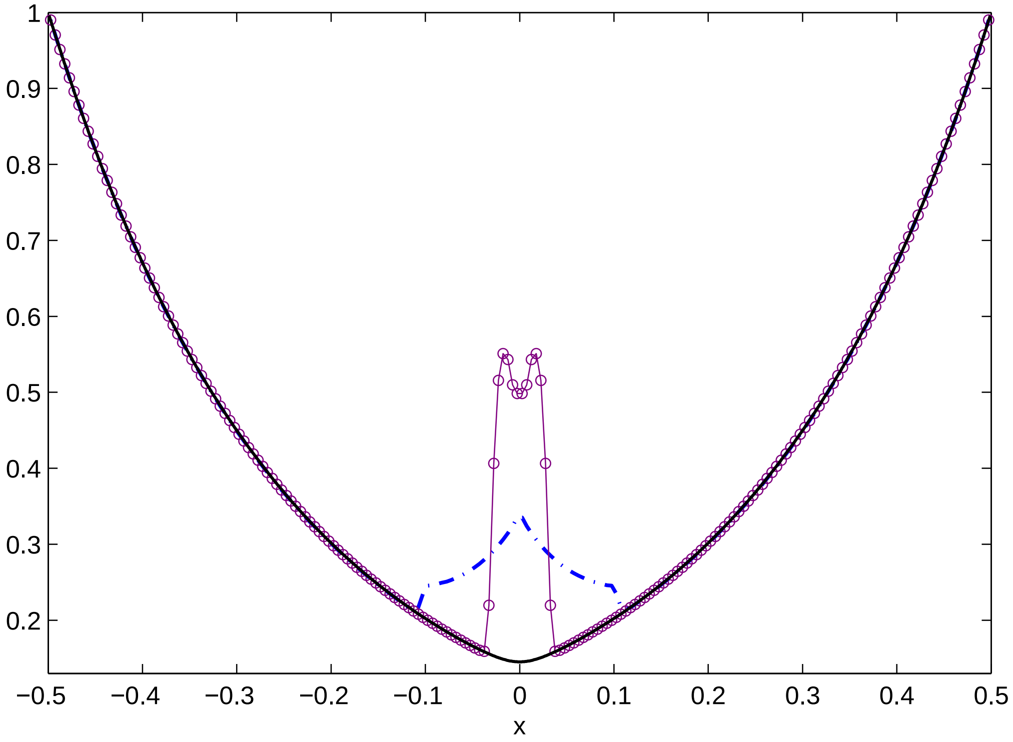
and initial conditions

$$\mathbf{u}_0(x) = [1, 0.9999]^T, \quad x \in (-0.5, 0, 5).$$

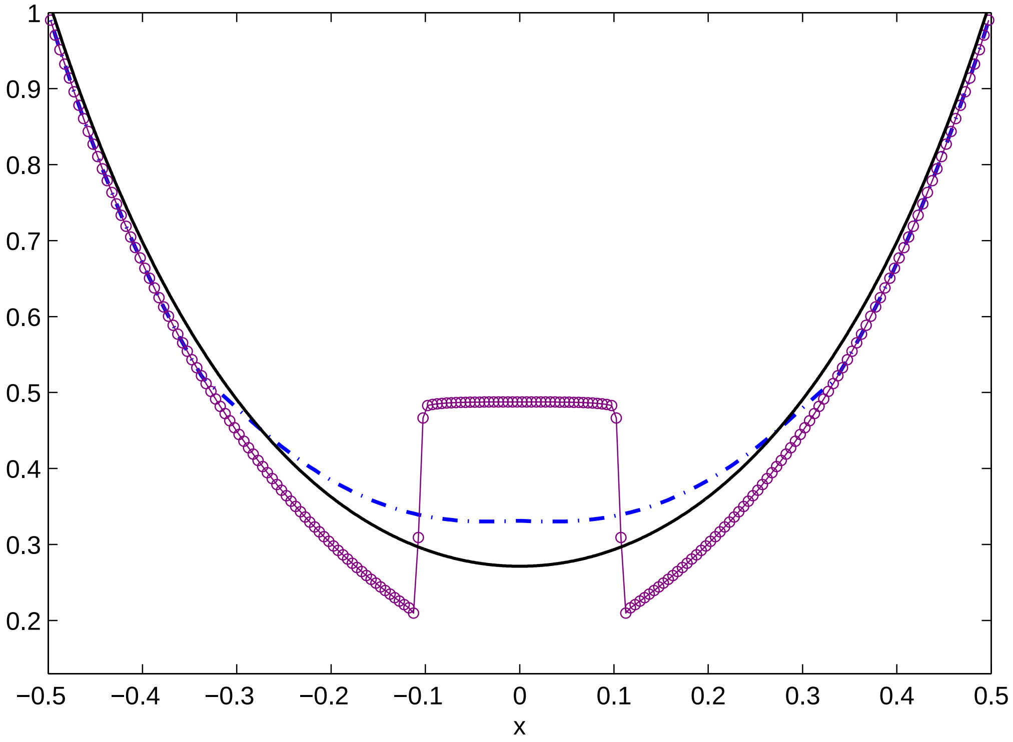
Again, $T = 0$ and the additional energy ODE from (4.6) is not included.

Figure 4.4 shows the shock of the M_1 solution in the middle of the domain. The perturbed M_1 model also develops an unphysical increase of the particle number. However, the shock is much smaller and becomes hardly visible in the steady state at $t = 4$. In the shock region, particles are distributed towards the boundaries which leads to kinks in the solution at $x \approx \pm 0.3$. For comparison, discrete ordinates solutions are plotted for which 256 discretization points in angle and 1000 points in space are used. The perturbed M_1 is throughout closer to the transport solution.

Remark 17. *Precise explanations for the occurence of shocks and kinks in the steady-state perturbed* M_1 *solution require an additional analysis. For example, one might use the continuity criterion from Brunner and Holloway in [21] to show why shocks are formed in the lowest-order moment. However, this goes beyond the purpose of this work and must be postponed to future work.*

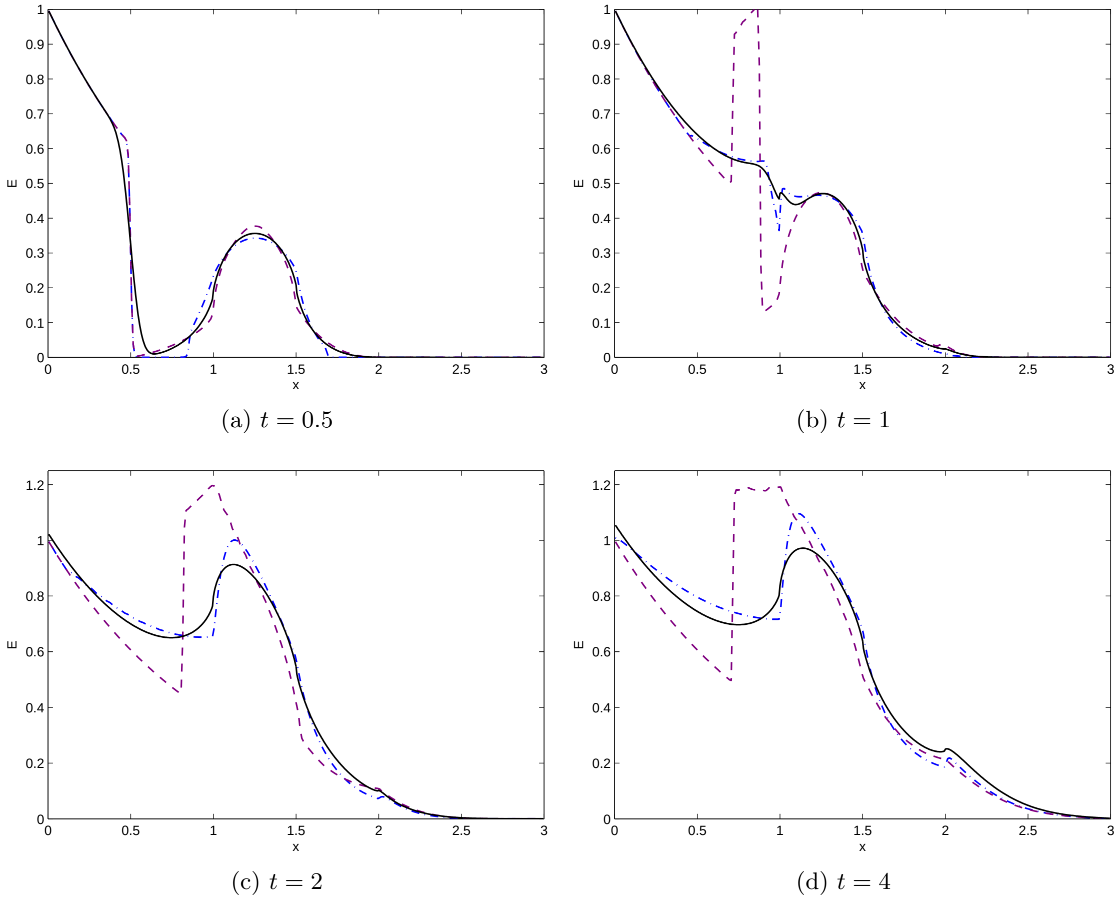


(a) $t = 0.6$



(b) $t = 3$

Figure 4.4: Two beams instability. $J = 200$, $k = 2$: M_1 (purple circle line), perturbed M_1 (blue dash-dot line), transport (black solid line).



(a) $t = 0.5$ (b) $t = 1$

(c) $t = 2$ (d) $t = 4$

Figure 4.5: Source-beam problem. $J = 300$, $k = 2$: M_1 (purple dashed line), perturbed M_1 (blue dash-dot line), transport solution (black solid line)

4.5.3 Source-Beam Problem

An incoming beam

$$\mathbf{u}(0,t) = [1, 0.9999]^T, \quad t > 0,$$

on the left boundary of the domain $[0, 3]$ hits an isotropic source $S = 1/2$ generating particles on the interval $1 \leq x < 1.5$. On the right boundary, particles are absorbed and zero Dirichlet conditions

$$\mathbf{u}(3,t) = [\varepsilon, 0]^T, \quad t > 0,$$

are set. Initially, there are no particles in the system, i.e.,

$$\mathbf{u}_0(x) = [\varepsilon, 0]^T, \quad x \in (0, 3).$$

The value of c is again set to one. We design the material properties with discontinuous cross sections:

$$\sigma_\mathrm{a} = \begin{cases} 1, & 0 \leq x < 2 \\ 0, & \text{else} \end{cases}, \quad \text{and} \quad \sigma_\mathrm{s} = \begin{cases} 2, & 1 \leq x < 2 \\ 10, & 2 \leq x < 3 \\ 0, & \text{else.} \end{cases}$$

M_1 and perturbed M_1 results are compared to transport solutions in the time evolution. Classic M_1 calculations are slope limited in the characteristic variables. The transport solution is computed with 600 spatial and 256 angular discretization points.

At the beginning, particles penetrating the medium from the left are simply absorbed in the region $0 \leq x < 1$ where $\sigma_\mathrm{a} = 1 = \sigma_\mathrm{t}$. As particles are emitted into the system simultaneously, Figure 4.5a shows the scalar flux at $t = 0.5$ which strongly decrease between $0 \leq x \lesssim 0.5$ and forms a bulge between $0.5 \lesssim x \leq 2$. Both models give results close to the reference although the M_1 solution is slightly more accurate.

For increasing time, the ingoing beam encounters the source and M_1 profiles diverge from the transport solution more and more (Figure 4.5b-d). Even at steady state $t = 4$ there is a large difference for $x \leq 1$. However, the perturbed M_1 calculations capture the reference behavior much better. Although they overshoot the discrete ordinates results in the source region for larger times, significant benefits can be observed especially at $t = 1$.

4.5.4 Gaussian Source

The next test case simulates particles emerging from an initial Gaussian distribution for the scalar flux:

$$\mathbf{u}_0(x) = \left[\frac{1}{\xi\sqrt{2\pi}}\,e^{-\frac{x^2}{2\xi^2}}, 0\right]^T, \quad \xi = 0.1, \quad x \in (-L, L).$$

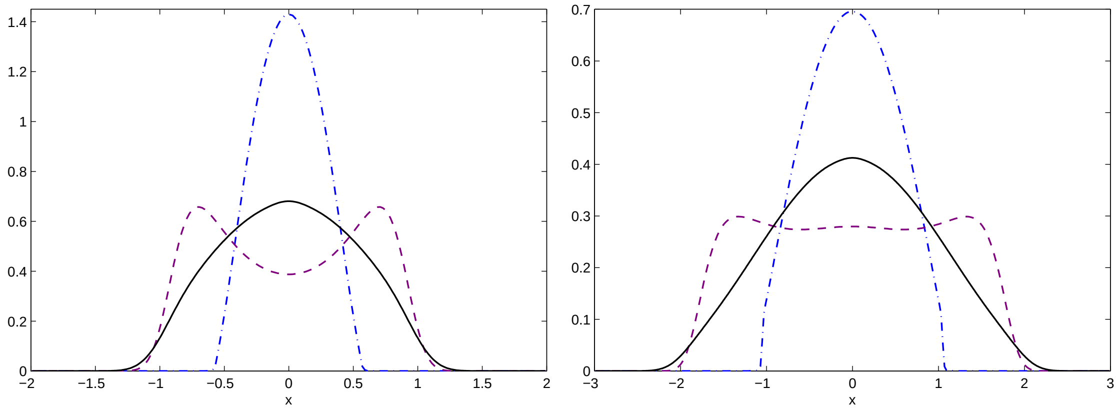
Periodic boundary conditions on $[-L, L]$ are prescribed where $L = t + 1$. The computational domain is always chosen large enough to ensure that a negligible amount of particles reaches the boundaries. No internal source $S = 0$ is assumed and the medium is purely scattering with $\sigma_\mathrm{s} = 1 = \sigma_\mathrm{t}$. The velocity is set to one and (4.6) is neglected. All DG results are computed with $h = 0.01$ and polynomial degree $k = 2$. Discrete ordinates solutions are obtained with $h = 0.005$ and 128 angular points.

Figure 4.6 displays the solutions at $t = 1, 2, 3, 10$. The M_1 model gives the expected wave effects which are washed out at larger times. These effects do not occur in the perturbed M_1 results. However, they form Gaussian bells which are higher and more narrow than the benchmark solution. At lower times, their maximum propagation speed is roughly half the correct velocity. Nevertheless, at $t = 10$ the perturbed M_1 catches up with the reference and is in a good agreement.

Since the perturbation $\tilde{\psi}$ from Section 4.3.2 can be interpreted as the difference between the M_1 and transport solution Figure 4.6 indicates that this quantity is highly time-dependent. Additionally, the spatial gradient of $\tilde{\psi}$ is large at lower times. Hence, this numerical example violates the assumptions made in the derivation of the perturbed M_1 model in Section 4.3.2 such that more accurate solutions cannot be expected.

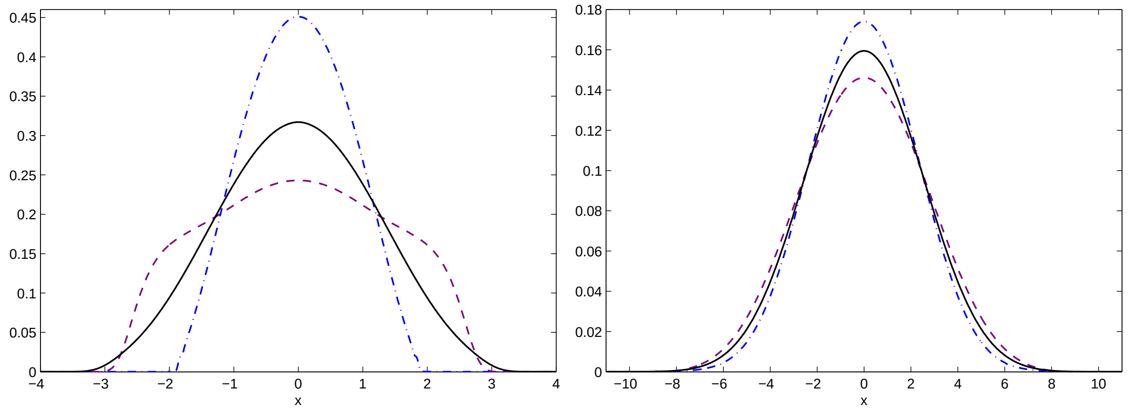
4.5.5 Including the Material Energy Equation

Our results compared moment models without the material energy equation so far. The next two examples are chosen to approximate the transport equation (4.2) coupled to the energy equation (4.6). The linearized Marshak wave problem from [173] is analyzed first and a thin Marshak wave from Chapter 3 is studied last.



(a) $t = 1$

(b) $t = 2$



(c) $t = 3$

(d) $t = 10$

Figure 4.6: Gaussian source. $J = 100\,(t + 1)$, $k = 2$: M_1 (purple circle line), perturbed M_1 (blue dash-dot line), transport (black solid line).

Su-Olson's Benchmark Problem

Su and Olson provide tabulated data of analytic solutions to a linearized Marshak wave problem which serves as a validation of numerical algorithms in the radiative transfer community [173]. In particular, this semi-analytic benchmark is also compared to diffusion-corrected P_N approximations in [156]. It is therefore of interest to study solutions of the perturbed M_1 model to this problem.

Precisely, we compute approximations to (4.2) and (4.6) in slab geometry with the following physical data

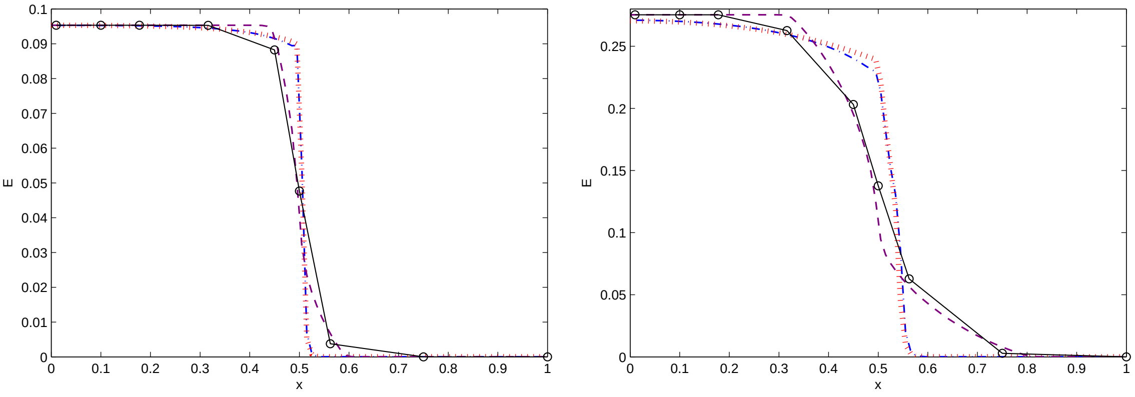
$$C_v = T^3, \quad c = 1, \quad \sigma_\mathrm{a} = 1 = \sigma_\mathrm{t}, \quad S(x,t) = \begin{cases} 1, & 0 \le t \le 10, \quad \text{and} \quad -0.5 \le x \le 0.5 \\ 0, & \text{else.} \end{cases}$$

Initially, the medium is cold and there is no radiation:

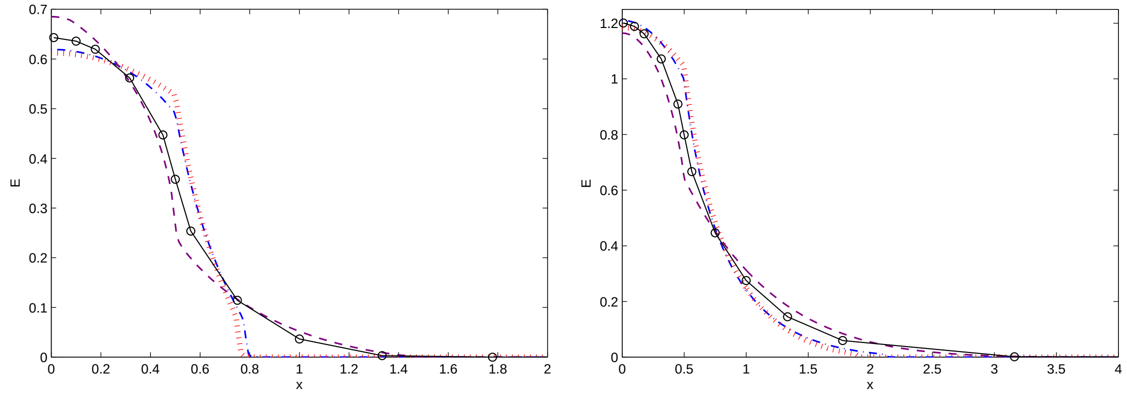
$$\psi(x, \mu, \nu, t = 0) = 0 \quad \text{and} \quad T(x, t = 0) = 0.$$

Additionally, zero boundary conditions are enforced on an infinite domain:

$$\lim_{x\to\pm\infty} \psi(x, \mu, \nu, t) = 0 \quad \text{and} \quad \lim_{x\to\pm\infty} T(x, t) = 0.$$



(a) $t = 0.1$ (b) $t = 0.31623$



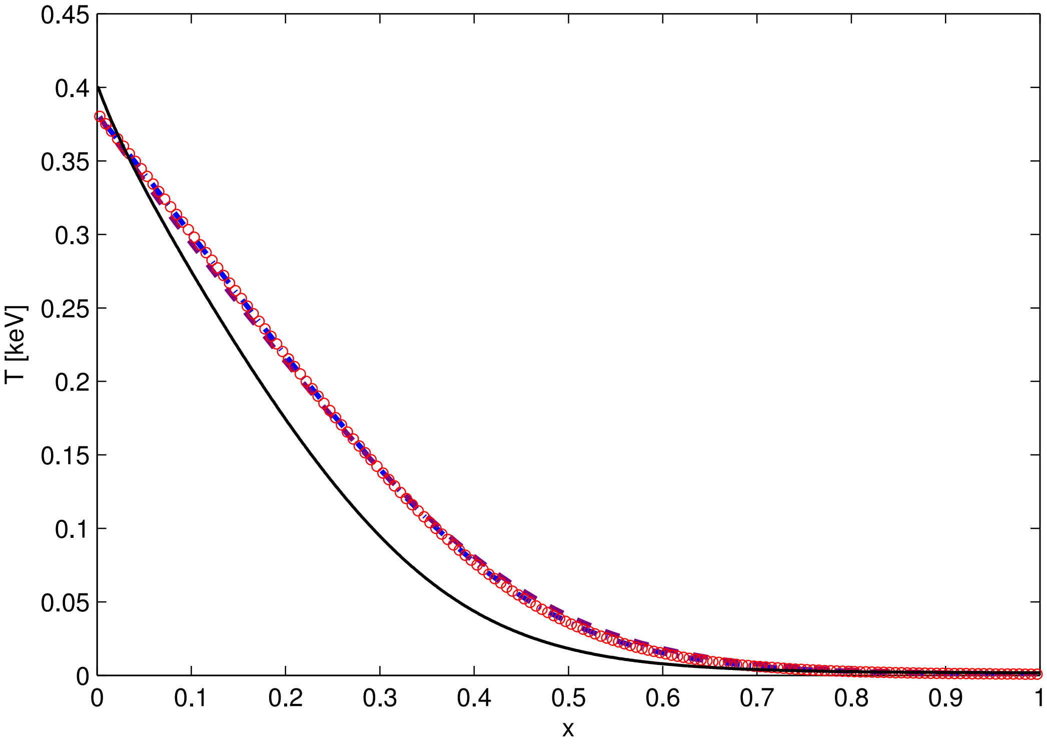
(c) $t = 1$ (d) $t = 3.16228$

Figure 4.7: Su-Olson's Problem. $J = 200\,L$, $k = 2$: M_1 (purple dashed line), perturbed M_1 (blue dash-dot line), perturbed M_1 with $\eta = 0$ (red dotted line), semi-analytic (black solid circle line).

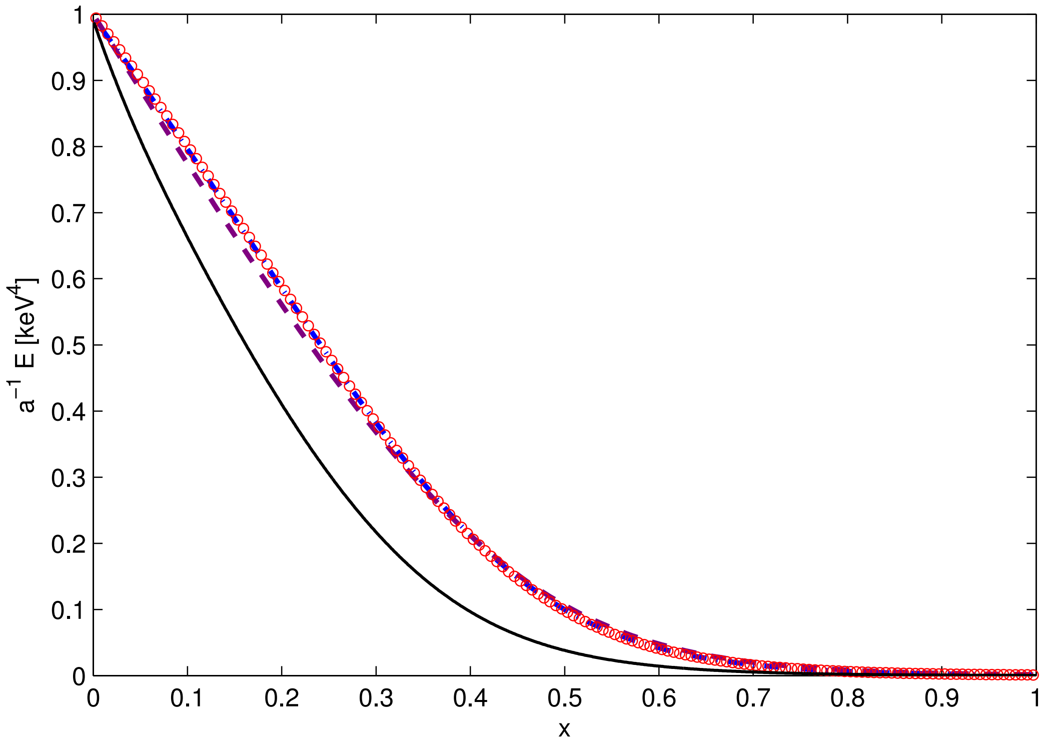
In practice, we impose periodic boundary conditions on a large domain $[-L, L]$ where $L = \lfloor t \rfloor + 1$.

Solutions at different times are provided in Figure 4.7 for the half plane $x \geq 0$. A grid size of $h = 0.01$ and polynomial degree of $k = 2$ are chosen for all DG solutions. Classic M_1 computations are slope limited in the characteristic variables. They are throughout close to the semi-analytic results. However, the perturbed M_1 solutions form larger slopes at the source discontinuity $x \approx 0.5$ at lower times. There, the perturbed M_1 model yields larger deviations from the reference. Only at $t = 3.16228$ solutions from both models are close to each other as well as to the semi-analytic points.

Additionally, we included solutions neglecting the drift term η. Small differences to the full perturbed M_1 result can be observed. However, in comparison to the reference, $\eta = 0$ does not yield any advantage.



(a) Temperature at $t = 0.1$ns.



(b) Scaled scalar flux $a^{-1}E$ at $t = 0.1$ns.

Figure 4.8: Thin Marshak wave. $J = 160$, $k = 2$: M_1 (purple dashed line), perturbed M_1 (blue dash-dot line), perturbed M_1 with $\eta = 0$ (red circles), P_{99} (black solid circle line).

Thin Marshak Wave

Ingoing radiation is prescribed on the left boundary by well-posed boundary conditions,

$$\begin{aligned}
&T(0,t) = 1, \quad T(1,t) = 0, \quad t > 0, \\
&\mathbf{u}(0,t) = [T(0,t)^4 a, 0.8 \cdot T(0,t)^4 ac]^T, \quad \mathbf{u}(1,t) = [2\varepsilon, 0]^T, \quad t > 0,
\end{aligned}$$

and the material is assumed to be purely absorbing:

$$\sigma_{\mathrm{a}}(T) = \frac{1}{(T+0.5)^3}\frac{\mathrm{keV}^3}{\mathrm{cm}}, \quad \sigma_s = 0, \quad S = 0.$$

The physical constants are given by

$$\begin{aligned}
c &= 3 \cdot 10^{10}\,\mathrm{cm/s} && \text{speed of light,} \\
a &= 1.372 \cdot 10^{14}\,\mathrm{erg/(cm^3keV^4)} && \text{radiation constant,} \\
C_v &= 3 \cdot 10^{15}\,\mathrm{erg/(cm^3keV)} && \text{heat capacity,}
\end{aligned}$$

which implies units of cm^{-1} for cross sections and keV for temperature T. Initially, the material is cold

$$\begin{aligned}
T_0(x) &= 5 \cdot 10^{-4}, \quad x \in (0,1), \\
\mathbf{u}_0(x) &= [T_0(x)^4 a, 0]^T, \quad x \in (0,1).
\end{aligned}$$

Due to above incoming radiation on the left boundary, radiation propagates through the medium from left to the right. This is why the material temperature T strictly diminishes near the left boundary down to zero on the right boundary (Figure 4.8a). Figure 4.8b also shows a smooth decrease of the scalar flux E. It is a consequence of a starting wavefront which moves to the right and is smoothed out with increasing time. Both models give solutions with similar results which hardly deviate. No differences to computations with $\eta = 0$ can be observed. The transport solution is more curved and decreases faster. It was computed with a P_{99} semi-implicit time integration scheme from [122] and 800 spatial points.

Chapter 5

Discussion and Conclusions

In this work, several approximate models to the linear Boltzmann equation have been studied. In the general framework of moment methods, these models can all be derived from certain closures required to make the associated underdetermined system of equations solvable. The central goals we pursued have always been

- to derive models based on physical concepts or on desired quantities which are supposed to be maintained,
- to keep the dimension of the resulting system of equations small and as simple as possible and
- to preserve important physical as well as model properties in the numerical scheme.

Thus, our approaches contained both mathematical models based on certain approximations as well as numerical methods for their discretization. In this way, it was possible to approach the results from two perspectives: first, deriving theoretical predictions and second, observing numerical experiments including common and challenging settings. Only the combination of both revealed final conclusions about benefits and drawbacks of the underlying model. As all investigations in this work focused on simplified equations in 1D or 2D, applications to and practicability for real-world problems require further research and should be the subject of future work. In particular, we want to highlight detailed conclusions of the previous chapters and explain open problems:

5.1 Electron, Photon and Proton Scattering Processes

Cross sections for photons, electrons and protons are investigated in Chapter 1. In the transport equation, they describe physical interactions of particles with matter. In principle, analytic formulas and databases are available for the most part. However, some state-of-the-art databases are not easily accessible. Moreover, the quantities and notations referred to in these databases are designed for either Monte Carlo codes or physical applications; thus, transformations to mathematical quantities in the Boltzmann equation are required. We therefore adapt the quantities of the most relevant effects in such a way that they can be included in deterministic codes.

Electron motion is characterized by small energy loss in collision events as well as multiple, small-angle elastic scattering with target nuclei. This is different for protons. The largest energy is transferred in proton-electron collisions whereas little energy is lost in proton-nucleus events. On the other hand, photons are neutral particles and behave in a very different way. Their mean free path is much larger than that of charged particles, they can be absorbed in collision events, and they experience large angle scattering. The intensity of the photon beam decreases exponentially with the travelled path.

As we are primarily interested in applications to radiotherapy our main goal is an accurate simulation of the deposited dose in tissue. For photon and electron dose calculations, stochastic Monte Carlo codes have been benchmarked against physical experiments and show a good agreement [162, 178]. It is therefore valuable to compare deterministic dose computations with Monte Carlo solutions. This has only been done for electron transport in Section 1.2. Therein, the accuracy of presented electron cross sections is demonstrated for dose calculations in water. However, a coupling to photons is still necessary because photon-electron interactions play an important role. Evaluations of cross sections for photons and protons are missing in Chapter 1 and should be performed in the future. In particular, protons and ions show great promise for radiotherapy applications because of their local energy deposition in tissue. Since the development of accurate models for Monte Carlo proton dose calculations is the subject of current research [163], improvements are expected to come.

A further future topic is the inclusion of the cross sections studied in Chapter 1 in the approximate models investigated in Chapters 2–4 and their application to simulations in radiotherapy.

5.2 Time-Dependent Simplified P_N Equations

Giving a theoretical foundation for the time-dependent SP_N equations and confirming the asymptotic analysis from Section 2.3 is the main purpose of Chapter 2. The main result consists of eqs. (2.26), a first-order system of PDEs. No special methods, or variations of known methods, are needed. Even more, it turns out that the developed time-dependent equations are equivalent to those which are derived by simply adding the partial derivative with respect to time to the steady-state SP_N equations in first order form. By a semi-discretization in time it is therefore possible to extend existent steady-state SP_N codes without too much effort.

One feature of eqs. (2.26) is that in 1-D planar geometry, they reduce *exactly* to the time-dependent 1-D planar-geometry P_3 equations. Hence, *in planar geometry*, numerical solutions of the time-dependent SP_3 and P_3 equations are indistinguishable – they are equivalent. Numerical calculations are therefore performed in two space dimensions in Section 2.4. Moreover, computations in Section 2.4.3 demonstrate that there are even cases outside the asymptotic limit where P_N and SP_N are equivalent in 2D.

Problems in heterogeneous media with cross sections of small scattering ratios $c = \frac{\Sigma_s}{\Sigma_t}$ show the following behavior: For small scaling parameters ε, SP_N computations are closer to transport solutions whereas large ε lead to larger errors (Section 2.4.3). Solutions to problems which do not satisfy the homogeneity assumption, made in the asymptotic analysis, still show accurate solutions (Section 2.4.3). Furthermore, all comparisons to diffusion solutions confirm that SP_3 results are throughout superior and especially in cases where large gradients are formed.

Finally, a few possible future tasks are discussed in the following:

1. The asymptotic analysis presented in this chapter implies time-dependent SP_3 equations which differ from those developed in [56]. From the theoretical point of view, the main reason is the different scaling which especially yields an additional ε^2 in front of the time-derivative in [56]. Amongst others, one consequence of the different choice is that the previous time-dependent SP_3 equations do not reduce *exactly* to the time-dependent planar-geometry P_3 equations. Another difference is that the time-dependent SP_3 theory in [56] contains an arbitrary constant α. Although all admissible values of α imply the same asymptotic order of accuracy it could be still valuable to compare numerical solutions of those equations with numerical solutions of equations proposed in this publication. Central challenges are to work out different behaviors of these equations for certain problems and find out whether computational results confirm the theoretical predictions.

2. Recently, *Larsen* presented *modified* diffusion and SP_N equations in [99] which are designed to be more accurate for deep penetration problems. The asymptotic analysis therein covers the steady-state, anisotropically scattering linear Boltzmann equation and is based on a different scaling. Can a similar procedure be applied to the time-dependent SP_N equations derived in Section 2.3? Answers to this question could give a mathematical foundation for the time-dependent SP_N equations in a much more general field of applications.

3. One possibility to investigate the accuracy of angular moment approximations to the Boltzmann equation is to perform a moment analysis for the regarding approximation model. It is difficult to calculate exact angular fluxes of the transport equation and obviously, even more difficult to compare them to approximations. Nevertheless, it is still possible to gain a different type of information about the analytic solution. The main idea of this method is to calculate angular and/or spatial moments of the angular flux and compare them to corresponding quantities of the regarding approximation method. Consequently, one can draw conclusions about the accuracy of the method. *Densmore and McClarren* [42], e.g., compared the previous time-dependent simplified P_N methods from [56] to other approximations. A similar analysis would provide additional information about moments which are preserved by solutions of the time-dependent SP_N equations studied above.

4. *McClarren* suggests one important aspect about the SP_N equations to be studied in the future [120]: What is the optimal order of SP_N equations that determines a limit where your benefit in accuracy is still larger than the additional computational costs? No investigations concerning this issue have been performed up to now. However, the publicly available MATLAB code [161] allows to perform P_N and SP_N calculations in 2D of an extremely high order. In Section 2.4.3 the error behavior between SP_N and transport solution is studied for increasing order and different scaling parameters. The largest benefit in the special problem appears to be from first to third and third to fifth order. This is purely heuristic and lacks mathematical foundations but could, however, be a motivation for future analysis.

5. As we only considered mono-energetic transport problems here the extension to multi-group calculations will allow for calculations of realistic applications and is left to future work.

5.3 A Realizability-Preserving DG Method: The M_1 Model

The guiding system of equations in Chapter 3 is the M_1 model of radiative transfer coupled to an ODE represented by the material energy equation in (3.3). Our main purpose is to show one possibility how cell averages of solutions calculated by a DG approach can be forced to remain realizable whenever they leave the realizability region. This is achieved by introducing a limiting procedure which, under a more restrictive CFL condition, is proved to keep the cell averages in every time step inside the realizability domain. The performed numerical simulations confirm that our proposed limiter enforces realizability and is, indeed, necessary for the computation of reasonable approximations which are consistent with the underlying model. In the following, we point out the most important observations and end our discussion with suggestions for future work:

Our examples show that neglecting the realizability limiter sometimes leads to numerical results which remain unrealizable throughout the whole time-marching process. Consequently, running computations without the realizability limiter one is basically faced with three difficulties:

- As soon as $|F| > cE$, the M_1 model becomes ill-posed so that the whole theoretical background is not valid any more.
- At $|F| = cE$, both eigenvalues of the Jacobian of the corresponding system of PDEs are equal and hence, the hyperbolicity of the PDE is lost at these points.
- For the Eddington factor to make sense, one has to ensure that in the course of numerical calculations, E and F remain in the set $\mathcal{D}_\chi := \{(E,F) : |F| \leq (2/\sqrt{3})cE\}$. On this set $1/3 \leq \chi \leq 5/3$.

Without changing the PDEs, the first two items cannot be avoided in the numerical computations. To get rid of the last mentioned issue we implement the cutoff in (3.104) and try to take the well-posedness of the model into account by enforcing $\chi \leq 1$. This is sometimes used for simulations with the M_1 model. Nevertheless this cutoff can not handle all test cases and sometimes gives completely false results (Sections 3.5.3–3.5.4). Instead, setting the cutoff at 5/3 may yield better results, as it did in Section 3.5.3.

The algorithm for the realizability limiter is less involved and computationally costly than for the slope limiter described in Section 3.3.2. Turning on merely the realizability limiter can sometimes be enough to obtain accurate results (Section 4.5.5) and so, one could save computational effort in this case. In general, there is, however, no guarantee that this is sufficient. Especially in cases where very steep slopes occur, the realizability limiter still produces spurious oscillations (Section 3.5.3).

One drawback of the constructed realizability limiter is the more restrictive CFL conditions from Theorem 2 which are additionally problem-dependent. As this leads to larger computational costs, it is therefore of interest to apply this limiter to temporal implicit schemes and enhance the performance of the algorithm.

Although in 2-D and 3-D problems it is already difficult to formulate realizability conditions, generalizing the results derived in Section 3.4 to multiple dimensions should be subject of future work. Moreover, our results motivate to construct admissibility limiters for similar models or moment models with different closures.

5.4 Perturbed, Entropy-Based Closure

Moment equations are used to obtain approximations to the radiative transfer equation. Since an approximation of the highest moment is required, entropy-based closures have been derived. In Chapter 4, we study perturbations of entropy closures and basically pursue two major intentions:

- We introduce perturbations to standard entropy closures and present rigorous derivations of the regarding moment equations.
- Applying the new perturbed closure to the M_1 model we additionally compare numerical simulations of both the standard and the perturbed M_1 model.

Our derivations of the perturbative model reveal final equations containing an additional convective and diffusive term which are added to the flux term of the standard closure. This is different to perturbations to standard P_N closures [156] which only gain a diffusive component.

Explicit equations and their parameters are presented for the perturbed M_1 model which is the first member in the moment hierarchy. The resulting system of equations is a convection-diffusion system which is discretized by using a Runge-Kutta discontinuous Galerkin method. By introducing an additional control parameter we modify the pressure term of the perturbed M_1 equations and ensure that cell averages of the moments remain realizable.

Improvements to the standard M_1 model are observed in cases where particles move in opposing directions. Whereas the classic M_1 model generates large shocks the perturbed M_1 model significantly suppresses this unphysical behavior and shows results which are much closer to the transport solution (Sections 4.5.2-4.5.3). However, in cases of discontinuous sources (Sections 4.5.4-4.5.5) the new model is not superior to the standard M_1 and only gives comparably accurate results for larger time evolutions.

Finally, we discuss some open problems in this framework which might be addressed in future:

- Moment systems from entropy-based closures are proven to be hyperbolic and their entropy is locally decreasing. Both properties have not been proven for the perturbed closures. Only the partial result in Proposition 2 confirms that the diffusive term dissipates the entropy. Moreover, neglecting the diffusion contribution the system indeed becomes hyperbolic for the special case of the M_1 model.
- The derived perturbed M_1 model yields convection-diffusion equations with a semi-positive definite diffusion matrix. It still has to be investigated whether the model is well-posed and has a unique solution.
- In Section 4.4.3 the control parameter is chosen to guarantee conditions (C1)-(C2). However, this ansatz is a crude modification of the original perturbative

model. Hence, additional errors are possible and could distort numerical solutions. The question should be answered if it is possible to come up with a more subtle limiter which modifies $\mathbf{u}_h^{j,n}$ similar to (4.93) in such a way that the pressure term Π^{D} additionally fulfills the necessary conditions.

- An undesirable issue of the RKDG method is the explicit time integrator which entails CFL conditions. Especially in this case of a mixed type of convection-diffusion equations this time step restriction is very harsh. To lower the computational effort, implicit time descritizations are therefore necessary.
- Another issue is the formation of unphysical shocks in the (perturbed) M_1 solution. Further analysis could help to determine where discontinuities appear and it gives an explanation why they occur.

Appendix A

Transport Coefficients: Different Polynomial Kernels

Descriptions in Chapter 1 focused on the quantities

$$\xi_n(\varepsilon) = 2\pi \int_{-1}^{1} (1-\mu)^n \sigma(\varepsilon,\mu) d\mu, \quad \mu = \cos(\theta), \quad n \in \mathbb{N}_0,$$

which are needed to compute approximative solutions derived in the generalized Fokker-Planck theory [107]. However, there are approximation theories to the Boltzmann equation which require different transport coefficients and in general look like

$$2\pi \int_{-1}^{1} p_n(\mu) \sigma_{\mathrm{el}}(\varepsilon,\mu) d\mu,$$

where p_n is a polynomial of at most degree n. Many approximations are based on moment analysis and often use either monomials

$$\{M_0, M_1, M_2, M_3, ...\}, \quad M_n(\mu) := \mu^n,$$

or Legendre polynomials

$$\{P_0, P_1, P_2, P_3, ...\},$$

where $P_n(\mu)$ is the n-th Legendre polynomial. If we define

$$\zeta_n(\varepsilon) = 2\pi \int_{-1}^{1} M_n(\mu) \sigma(\varepsilon,\mu) d\mu, \quad n \in \mathbb{N}_0, \tag{A.1}$$

$$\eta_n(\varepsilon) = 2\pi \int_{-1}^{1} P_n(\mu) \sigma(\varepsilon,\mu) d\mu, \quad n \in \mathbb{N}_0, \tag{A.2}$$

then one can use $\xi_n(\varepsilon)$ to compute above quantities:

$$\begin{aligned}
\zeta_0(\varepsilon) &= \xi_0(\varepsilon),\\
\zeta_1(\varepsilon) &= \xi_0(\varepsilon) - \xi_1(\varepsilon),\\
\zeta_2(\varepsilon) &= \xi_0(\varepsilon) - 2\xi_1(\varepsilon) + \xi_2(\varepsilon),\\
\zeta_3(\varepsilon) &= \xi_0(\varepsilon) - 3\xi_1(\varepsilon) + 3\xi_2(\varepsilon) - \xi_3(\varepsilon),\\
\zeta_4(\varepsilon) &= \xi_0(\varepsilon) - 4\xi_1(\varepsilon) + 6\xi_2(\varepsilon) - 4\xi_3(\varepsilon) + \xi_4(\varepsilon),\\
\zeta_5(\varepsilon) &= \xi_0(\varepsilon) - 5\xi_1(\varepsilon) + 10\xi_2(\varepsilon) - 10\xi_3(\varepsilon) + 5\xi_4(\varepsilon) - \xi_5(\varepsilon),\\
\zeta_6(\varepsilon) &= \xi_0(\varepsilon) - 6\xi_1(\varepsilon) + 15\xi_2(\varepsilon) - 20\xi_3(\varepsilon) + 15\xi_4(\varepsilon) - 6\xi_5(\varepsilon) + \xi_6(\varepsilon),
\end{aligned}$$

and

$$
\begin{aligned}
&\eta_0(\varepsilon) = \xi_0(\varepsilon),\\
&\eta_1(\varepsilon) = \xi_0(\varepsilon) - \xi_1(\varepsilon),\\
&\eta_2(\varepsilon) = \xi_0(\varepsilon) - 3\xi_1(\varepsilon) + \frac{3\xi_2(\varepsilon)}{2},\\
&\eta_3(\varepsilon) = \xi_0(\varepsilon) - 6\xi_1(\varepsilon) + \frac{15\xi_2(\varepsilon)}{2} - \frac{5\xi_3(\varepsilon)}{2},\\
&\eta_4(\varepsilon) = \xi_0(\varepsilon) - 10\xi_1(\varepsilon) + \frac{45\xi_2(\varepsilon)}{2} - \frac{35\xi_3(\varepsilon)}{2} + \frac{35\xi_4(\varepsilon)}{8},\\
&\eta_5(\varepsilon) = \xi_0(\varepsilon) - 15\xi_1(\varepsilon) + \frac{105\xi_2(\varepsilon)}{2} - 70\xi_3(\varepsilon) + \frac{315\xi_4(\varepsilon)}{8} - \frac{63\xi_5(\varepsilon)}{8},\\
&\eta_6(\varepsilon) = \xi_0(\varepsilon) - 21\xi_1(\varepsilon) + 105\xi_2(\varepsilon) - 210\xi_3(\varepsilon) + \frac{1575\xi_4(\varepsilon)}{8} - \frac{693\xi_5(\varepsilon)}{8} + \frac{231\xi_6(\varepsilon)}{16}.
\end{aligned}
$$

Appendix B

Time-Dependent SP_N Equations

B.1 Classic Derivation: Time-Dependent SP_N equations of Arbitrary Order

Although a mathematical foundation is only given for the SP_3 equations in Chapter 2, we explain how time-dependent SP_N equations for arbitrary $N \geq 1$ can be derived.

First, we recall the 1-D time-dependent P_N equations in planar geometry:

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(\frac{n+1}{2n+1}\phi_{n+1}(x,t) + \frac{n}{2n+1}\phi_{n-1}(x,t)\right) + \Sigma_t(x,t)\phi_n(x,t) \tag{B.1a}$$
$$= \delta_{n,0}\left(\Sigma_s(x,t)\phi_0(x,t) + Q(x,t)\right)$$

$$\phi_{N+1} = 0, \tag{B.1b}$$

for $n = 0, 1, \ldots, N$. Let $i = 0, 1, 2, \ldots$, then odd and even moments can be written as

$$\mathcal{T}\phi_{2i-1}(x,t) + \mathcal{X}\left(\frac{2i}{4i-1}\phi_{2i}(x,t) + \frac{2i-1}{4i-1}\phi_{2i-2}(x,t)\right) + \phi_{2i}(x,t) = 0 \tag{B.2a}$$

$$\mathcal{T}\phi_{2i}(x,t) + \mathcal{X}\left(\frac{2i+1}{4i+1}\phi_{2i+1}(x,t) + \frac{2i}{4i+1}\phi_{2i-1}(x,t)\right) + \phi_{2i}(x,t) \tag{B.2b}$$
$$= \frac{\delta_{i,0}}{\Sigma_t(x,t)}(\Sigma_s\phi_0(x,t) + Q(x,t)),$$

where

$$\phi_{-1} = 0 = \phi_{-2}, \quad \mathcal{T} = \frac{1}{v\Sigma_t(x,t)}\frac{\partial}{\partial t}, \quad \text{and} \quad \mathcal{X} = \frac{1}{\Sigma_t(x,t)}\frac{\partial}{\partial x}. \tag{B.3}$$

Second, from eqs. (B.2), we can solve for odd moments

$$\phi_{2i-1}(x,t) = -(I+\mathcal{T})^{-1}\mathcal{X}\left(\frac{2i}{4i-1}\phi_{2i}(x,t) + \frac{2i-1}{4i-1}\phi_{2i-2}(x,t)\right) \tag{B.4}$$

and introduce (B.4) into (B.2b):

$$\mathcal{T}\phi_{2i}(x,t) + \mathcal{X}\left\{\frac{2i+1}{4i+1}\phi_{2i+1}(x,t) - \frac{2i}{4i+1}(I+\mathcal{T})^{-1}\mathcal{X}\left(\frac{2i}{4i-1}\phi_{2i}(x,t)\right.\right. \tag{B.5}$$
$$\left.\left. + \frac{2i-1}{4i-1}\phi_{2i-2}(x,t)\right)\right\} + \phi_{2i}(x,t) = \frac{\delta_{i,0}}{\Sigma_t(x,t)}(\Sigma_s\phi_0(x,t) + Q(x,t)).$$

Now, if $\phi_{2i+1} = 0$ then we end up with

$$\mathcal{T}\phi_{2i}(x,t) - \mathcal{X}(I+\mathcal{T})^{-1}\mathcal{X}\left\{\frac{4i^2}{(16i^2-1)}\phi_{2i}(x,t) + \frac{2i(2i-1)}{(16i^2-1)}\phi_{2i-2}(x,t)\right\} \tag{B.6}$$
$$+ \phi_{2i}(x,t) = \frac{\delta_{i,0}}{\Sigma_t(x,t)}(\Sigma_s\phi_0(x,t) + Q(x,t)).$$

Otherwise, if $\phi_{2i+1} \neq 0$ we use (B.4) and get

$$\mathcal{T}\phi_{2i}(x,t) - \mathcal{X}(I+\mathcal{T})^{-1}\mathcal{X}\left\{\frac{2(2i^2+3i+1)}{(4i+1)(4i+3)}\phi_{2i+2}(x,t) + \left(\frac{32i^3+24i^2-1}{(16i^2-1)(4i+3)}\right)\phi_{2i}(x,t)\right.$$
$$\left. + \frac{2i(2i-1)}{(16i^2-1)}\phi_{2i-2}(x,t)\right\} + \phi_{2i}(x,t) = \frac{\delta_{i,0}}{\Sigma_t(x,t)}(\Sigma_s\phi_0(x,t) + Q(x,t)). \tag{B.7}$$

Altogether, the set of SP_N equations can be written as

$$\mathcal{T}\phi_{2i}(x,t) - \mathcal{X}(I+\mathcal{T})^{-1}\mathcal{X}\left\{k_i\phi_{2i+2}(x,t) + l_i\phi_{2i}(x,t) + m_i\phi_{2i-2}(x,t)\right\} + \phi_{2i}(x,t)$$
$$= \frac{\delta_{i,0}}{\Sigma_t(x,t)}(\Sigma_s\phi_0(x,t) + Q(x,t)), \tag{B.8a}$$

for all $i = 0, 1, \dots, \lfloor \frac{N}{2} \rfloor$, where $\lfloor\ \rfloor$ is the floor function and the coefficients are defined as

$$k_i = \frac{2i(2i-1)}{(16i^2-1)} \tag{B.9a}$$

$$l_i = \begin{cases} \dfrac{32i^3+24i^2-1}{(16i^2-1)(4i+3)}, & N \text{ odd}, \\ \dfrac{4i^2}{(16i^2-1)}, & N \text{ even}, \end{cases} \tag{B.9b}$$

$$m_i = \begin{cases} \dfrac{2(2i^2+3i+1)}{(4i+1)(4i+3)}, & 2(i+1) \leq N, \\ 0, & \text{else}. \end{cases} \tag{B.9c}$$

Again, the 1-D operator is replaced by its 3-D extension

$$\mathcal{X}(I+\mathcal{T})^{-1}\mathcal{X} \quad \longrightarrow \quad \underline{\mathcal{X}} \cdot (I+\mathcal{T})^{-1}\underline{\mathcal{X}} := \frac{1}{\Sigma_t(\underline{x},t)}\underline{\nabla} \cdot (I+\mathcal{T})^{-1}\frac{1}{\Sigma_t(\underline{x},t)}\underline{\nabla}$$

and the same is also done for the spatial argument. Introducing the vector-valued variables

$$\underline{J}_{2i}(\underline{x},t) := -(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\left\{k_i\phi_{2i+2}(\underline{x},t) + l_i\phi_{2i}(\underline{x},t) + m_i\phi_{2i-2}(\underline{x},t)\right\}, \tag{B.10}$$

we can conclude the following hyperbolic system of SP_N equations:

$$\frac{1}{v}\frac{\partial\phi_{2i}}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot\underline{J}_{2i}(\underline{x},t) + \Sigma_t(\underline{x},t)\phi_{2i}(\underline{x},t) = \delta_{i,0}(\Sigma_s\phi_0(\underline{x},t) + Q(\underline{x},t)), \tag{B.11}$$
$$\frac{1}{v}\frac{\partial\underline{J}_{2i}}{\partial t}(\underline{x},t) + \underline{\nabla}\left\{k_i\phi_{2i+2}(\underline{x},t) + l_i\phi_{2i}(\underline{x},t) + m_i\phi_{2i-2}(\underline{x},t)\right\} + \Sigma_t(x,t)\underline{J}_{2i}(\underline{x},t). \tag{B.12}$$

Finally, some examples of simplified spherical harmonics equations up to $N = 5$ are listed below:

B.1.1 SP_1 Equations

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot \underline{J}_0(\underline{x},t) + \Sigma_a(\underline{x},t)\phi_0(\underline{x},t) = Q(\underline{x},t), \tag{B.13a}$$

$$\frac{1}{v}\frac{\partial \underline{J}_0}{\partial t}(\underline{x},t) + \frac{1}{3}\underline{\nabla}\phi_0(\underline{x},t) + \Sigma_t(\underline{x},t)\underline{J}_0 = 0. \tag{B.13b}$$

B.1.2 SP_2 Equations

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot \underline{J}_0(\underline{x},t) + \Sigma_a(\underline{x},t)\phi_0(\underline{x},t) = Q(\underline{x},t), \tag{B.14a}$$

$$\frac{1}{v}\frac{\partial \phi_2}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot \underline{J}_2(\underline{x},t) + \Sigma_t(\underline{x},t)\phi_2(\underline{x},t) = 0, \tag{B.14b}$$

$$\frac{1}{v}\frac{\partial \underline{J}_0}{\partial t}(\underline{x},t) + \frac{1}{3}\underline{\nabla}\Big(\phi_0(\underline{x},t) + 2\phi_2(\underline{x},t)\Big) + \Sigma_t(\underline{x},t)\underline{J}_0(\underline{x},t) = 0, \tag{B.14c}$$

$$\frac{1}{v}\frac{\partial \underline{J}_2}{\partial t}(\underline{x},t) + \frac{2}{15}\underline{\nabla}\Big(\phi_0(\underline{x},t) + 2\phi_2(\underline{x},t)\Big) + \Sigma_t(\underline{x},t)\underline{J}_2(\underline{x},t) = 0. \tag{B.14d}$$

B.1.3 SP_3 Equations

Here, eqs. (B.14a)–(B.14c) remain the same and (B.14d) is replaced by

$$\frac{1}{v}\frac{\partial \underline{J}_2}{\partial t}(\underline{x},t) + \frac{1}{3}\underline{\nabla}\left(\frac{2}{5}\phi_0(\underline{x},t) + \frac{11}{7}\phi_2(\underline{x},t)\right) + \Sigma_t(\underline{x},t)\underline{J}_2(\underline{x},t) = 0. \tag{B.15}$$

B.1.4 SP_4 Equations

$$\frac{1}{v}\frac{\partial \phi_0}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot \underline{J}_0(\underline{x},t) + \Sigma_a(\underline{x},t)\phi_0(\underline{x},t) = Q(\underline{x},t), \tag{B.16a}$$

$$\frac{1}{v}\frac{\partial \phi_2}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot \underline{J}_2(\underline{x},t) + \Sigma_t(\underline{x},t)\phi_2(\underline{x},t) = 0, \tag{B.16b}$$

$$\frac{1}{v}\frac{\partial \phi_4}{\partial t}(\underline{x},t) + \underline{\nabla}\cdot \underline{J}_4(\underline{x},t) + \Sigma_t(\underline{x},t)\phi_4(\underline{x},t) = 0, \tag{B.16c}$$

$$\frac{1}{v}\frac{\partial \underline{J}_0}{\partial t}(\underline{x},t) + \frac{1}{3}\underline{\nabla}(\phi_0(\underline{x},t) + 2\phi_2(\underline{x},t)) + \Sigma_t(\underline{x},t)\underline{J}_0(\underline{x},t) = 0, \tag{B.16d}$$

$$\begin{aligned}\frac{1}{v}\frac{\partial \underline{J}_2}{\partial t}(\underline{x},t) + \underline{\nabla}\left(\frac{2}{15}\phi_0(\underline{x},t) + \frac{11}{21}\phi_2(\underline{x},t) + \frac{12}{35}\phi_4(\underline{x},t)\right)& \\ + \Sigma_t(\underline{x},t)\underline{J}_2(\underline{x},t) = 0,&\end{aligned} \tag{B.16e}$$

$$\frac{1}{v}\frac{\partial \underline{J}_4}{\partial t}(\underline{x},t) + \frac{4}{21}\underline{\nabla}\left(\phi_2(\underline{x},t) + \frac{4}{3}\phi_4(\underline{x},t)\right) + \Sigma_t(\underline{x},t)\underline{J}_4(\underline{x},t) = 0. \tag{B.16f}$$

B.1.5 SP_5 Equations

Eq. (B.16f) is changed by

$$\frac{1}{v}\frac{\partial \underline{J}_4}{\partial t}(\underline{x},t) + \frac{1}{7}\underline{\nabla}\left(\frac{4}{3}\phi_2(\underline{x},t) + \frac{39}{11}\phi_4(\underline{x},t)\right) + \Sigma_t(\underline{x},t)\underline{J}_4(\underline{x},t) = 0 \tag{B.17}$$

and eqs. (B.16a)–(B.16e) are the same as for SP_4.

B.2 Expanding Operators

Here, we provide detailed algebraic manipulations for the simplification of the expansion operators

$$\mathcal{L}_n = \frac{1}{4\pi} \int_{S^2} \underline{\Omega} \cdot \underline{\nabla} (\mathcal{M}\, \underline{\Omega} \cdot \underline{\mathcal{X}})^n d\Omega, \tag{B.18}$$

where

$$\mathcal{T} = \frac{1}{v\Sigma_t} \frac{\partial}{\partial t}, \quad \underline{\mathcal{X}} = \frac{1}{\Sigma_t} \underline{\nabla}, \quad \mathcal{P} = \frac{1}{4\pi} \int_{S^2} (\cdot)\, d\Omega, \quad \mathcal{M} = (I + \mathcal{T})^{-1} (I - \mathcal{P}). \tag{B.19}$$

If we denote the cosine of the polar angle by $-1 \leq \mu \leq 1$ and the azimuthal angle by $0 \leq \phi < 2\pi$ then the unit vector $\underline{\Omega} = (\Omega_1, \Omega_2, \Omega_3)$ on the unit sphere is given by its scalar components

$$\Omega_1 = \mu, \tag{B.20}$$

$$\Omega_2 = \sqrt{1 - \mu^2} \cos(\phi), \tag{B.21}$$

$$\Omega_3 = \sqrt{1 - \mu^2} \sin(\phi). \tag{B.22}$$

Much of the following analysis is based on results of the integral

$$\int_{S^2} \Omega_{i_1} \Omega_{i_2} \dots \Omega_{i_n} d\Omega$$

for all possible combinations of $1 \leq i_1, \dots, i_n \leq 3$. Without a proof, we explicitly state some solutions which are needed later on:

$$\int_{S^2} \Omega_{i_1} \Omega_{i_2}\, d\Omega = \frac{4\pi}{3} \delta_{i_1,i_2}, \tag{B.23a}$$

$$\int_{S^2} \Omega_{i_1} \Omega_{i_2} \Omega_{i_3} \Omega_{i_4}\, d\Omega = \frac{4\pi}{15} \left(\delta_{i_1,i_2} \delta_{i_3,i_4} + \delta_{i_1,i_3} \delta_{i_2,i_4} + \delta_{i_1,i_4} \delta_{i_2,i_3} \right), \tag{B.23b}$$

$$\int_{S^2} \Omega_{i_1} \Omega_{i_2} \dots \Omega_{i_n}\, d\Omega = 0 \quad \text{for odd } n. \tag{B.23c}$$

We first focus on calculating $\mathcal{L}_n$ for even n because above integral is zero for odd n:

$$\mathcal{L}_0 = \frac{1}{4\pi} \int_{S^2} \underline{\Omega} \cdot \underline{\nabla}\, d\Omega = 0. \tag{B.24}$$

$$\mathcal{L}_2 = \frac{1}{4\pi} \int_{S^2} \underline{\Omega} \cdot \underline{\nabla} (\mathcal{M}\, \underline{\Omega} \cdot \underline{\mathcal{X}})^2\, d\Omega \tag{B.25}$$

$$= \frac{1}{4\pi} \int_{S^2} \underline{\Omega} \cdot \underline{\nabla} (\mathcal{M}\, \underline{\Omega} \cdot \underline{\mathcal{X}}) (\mathcal{M}\, \underline{\Omega} \cdot \underline{\mathcal{X}})\, d\Omega \tag{B.26}$$

$$= \frac{1}{4\pi} \int_{S^2} \underline{\Omega} \cdot \underline{\nabla} (\mathcal{M}\, \underline{\Omega} \cdot \underline{\mathcal{X}}) (I + \mathcal{T})^{-1} (I - \mathcal{P})\, (\underline{\Omega} \cdot \underline{\mathcal{X}})\, d\Omega. \tag{B.27}$$

Since

$$\mathcal{P}(\underline{\Omega}\cdot X)\phi_0(\underline{x},t) = \frac{1}{\Sigma_t}\sum_{i=1}^{3}\partial_i\phi_0(\underline{x},t)\int_{S^2}\Omega_i\, d\Omega = 0, \tag{B.28}$$

it follows

$$\begin{aligned}
\mathcal{L}_2 &= \frac{1}{4\pi}\int_{S^2}\underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})(I+\mathcal{T})^{-1}d\Omega && \text{(B.29)}\\
&= \frac{1}{4\pi}\int_{S^2}\underline{\Omega}\cdot\underline{\nabla}(I+\mathcal{T})^{-1}(I-\mathcal{P})\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,(I+\mathcal{T})^{-1}\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,d\Omega && \text{(B.30)}\\
&= \frac{1}{4\pi}\int_{S^2}\underline{\Omega}\cdot\underline{\nabla}(I+\mathcal{T})^{-1}\{\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,(I+\mathcal{T})^{-1}\,(\underline{\Omega}\cdot\underline{\mathcal{X}}) \\
&\qquad\qquad - \mathcal{P}\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,(I+\mathcal{T})^{-1}\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,\}\,d\Omega && \text{(B.31)}\\
&= \frac{1}{4\pi}\int_{S^2}\underline{\Omega}\cdot\underline{\nabla}(I+\mathcal{T})^{-1}\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,(I+\mathcal{T})^{-1}\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,d\Omega && \text{(B.32)}\\
&= 0, && \text{(B.33)}
\end{aligned}$$

where last two equalities follow from the fact that

$$\left[\mathcal{P}\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,(I+\mathcal{T})^{-1}\,(\underline{\Omega}\cdot\underline{\mathcal{X}})\,\right]\phi_0(\underline{x},t) \tag{B.34}$$

is independent of $\underline{\Omega}$. Hence, only integrals of an odd number of $\underline{\Omega}$ are left which all vanish according to (B.23c). Similarly, we obtain

$$\begin{aligned}
\mathcal{L}_4 &= 0, && \text{(B.35)}\\
\mathcal{L}_6 &= 0, && \text{(B.36)}
\end{aligned}$$

and attend to odd-numbered operators $\mathcal{L}_n$:

Note that (B.28) yields

$$\begin{aligned}
\underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}}) &= \underline{\Omega}\cdot\underline{\nabla}(I+\mathcal{T})^{-1}(I-\mathcal{P})\frac{1}{\Sigma_t}\underline{\Omega}\cdot\underline{\nabla}\\
&= \underline{\Omega}\cdot\underline{\nabla}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\underline{\Omega}\cdot\underline{\nabla} && \text{(B.37)}
\end{aligned}$$

Using (B.23a), we obtain

$$\begin{aligned}
\mathcal{L}_1 = \frac{1}{4\pi}\int_{S^2}\underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})d\Omega &= \frac{1}{4\pi}\sum_{i_1,i_2=1}^{3}\int_{S^2}\Omega_{i_1}\partial_{i_1}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\Omega_{i_2}\partial_{i_2}\,d\Omega && \text{(B.38)}\\
&= \frac{1}{4\pi}\sum_{i_1,i_2=1}^{3}\frac{4\pi}{3}\delta_{i_1,i_2}\left(\partial_{i_1}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_2}\right) && \text{(B.39)}\\
&= \frac{1}{3}\sum_{i_1=1}^{3}\partial_{i_1}(I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_1} && \text{(B.40)}\\
&= \frac{1}{3}\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}. && \text{(B.41)}
\end{aligned}$$

Next, making use of (B.23c) we calculate

$$\mathcal{L}_3 = \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})d\Omega \tag{B.42}$$

$$= \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})[(I+\mathcal{T})^{-1}(I-\mathcal{P})\underline{\Omega}\cdot\underline{\mathcal{X}}]d\Omega \tag{B.43}$$

$$= \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})[(I+\mathcal{T})^{-1}(I-\mathcal{P})\underline{\Omega}\cdot\underline{\mathcal{X}}][(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}]d\Omega \tag{B.44}$$

$$\begin{aligned} &= \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})[(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}][(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}]d\Omega \\ &\quad - \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})(I+\mathcal{T})^{-1}\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}\right\}d\Omega. \end{aligned} \tag{B.45}$$

According to (B.23c), last integral simplifies to

$$- \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}(\mathcal{M}\,\underline{\Omega}\cdot\underline{\mathcal{X}})(I+\mathcal{T})^{-1}\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}\right\}d\Omega \tag{B.46}$$

$$\begin{aligned} =& - \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}[(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}](I+\mathcal{T})^{-1}\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}\right\}d\Omega \\ &+ \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}(I+\mathcal{T})^{-1}\underbrace{\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}\right\}\right\}}_{=0}d\Omega \end{aligned} \tag{B.47}$$

$$= - \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}[(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}](I+\mathcal{T})^{-1}\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}\right\}d\Omega. \tag{B.48}$$

Similarly, we can get rid of the operator $\mathcal{P}$ in the first term of (B.45) and obtain

$$\begin{aligned} \mathcal{L}_3 =& \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}[(I+\mathcal{T})^{-1}(I-\mathcal{P})\underline{\Omega}\cdot\underline{\mathcal{X}}][(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}][(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}]d\Omega \\ &- \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}[(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}](I+\mathcal{T})^{-1}\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}\right\}d\Omega \end{aligned} \tag{B.49}$$

$$\begin{aligned} =& \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}[(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}][(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}][(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}]d\Omega \\ &- \frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}[(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}](I+\mathcal{T})^{-1}\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}\right\}d\Omega \end{aligned} \tag{B.50}$$

We consider last two integral terms separately and expand the dot products therein:

$$\frac{1}{4\pi}\int_{S^2} \underline{\Omega}\cdot\underline{\nabla}[(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}][(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}][(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}]d\Omega \tag{B.51}$$

$$\begin{aligned} =&\frac{1}{4\pi}\sum_{i_1,\dots,i_4=1}^{3}\int_{S^2}\Omega_{i_1}\partial_{i_1}\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\Omega_{i_2}\partial_{i_2}\right) \\ &\cdot\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\Omega_{i_3}\partial_{i_3}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\Omega_{i_4}\partial_{i_4}\right)d\Omega \end{aligned} \tag{B.52}$$

$$
\begin{aligned}
&=\frac{1}{4\pi}\sum_{i_1,\dots,i_4=1}^{3}\partial_{i_1}\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_2}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_3}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_4}\right) \qquad (\mathrm{B.53})\\
&\qquad\qquad\qquad\qquad\qquad\qquad \cdot\int_{S^2}\Omega_{i_1}\Omega_{i_2}\Omega_{i_3}\Omega_{i_4}\,d\Omega\\
&=\sum_{i_1,\dots,i_4=1}^{3}\partial_{i_1}\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_2}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_3}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_4}\right) \qquad (\mathrm{B.54})\\
&\qquad\qquad\qquad\qquad\qquad\qquad \cdot\frac{1}{15}\left(\delta_{i_1,i_2}\delta_{i_3,i_4}+\delta_{i_1,i_3}\delta_{i_2,i_4}+\delta_{i_1,i_4}\delta_{i_2,i_3}\right),
\end{aligned}
$$

where last equality follows from (B.23b). Following the same arguments as in eqs. (B.38)-(B.39), the second integral term in (B.50) can be rewritten to

$$
\begin{aligned}
&-\frac{1}{4\pi}\int_{S^2}\underline{\Omega}\cdot\underline{\nabla}[(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}](I+\mathcal{T})^{-1}\mathcal{P}\left\{\underline{\Omega}\cdot\underline{\mathcal{X}}(I+\mathcal{T})^{-1}\underline{\Omega}\cdot\underline{\mathcal{X}}\right\}d\Omega \qquad (\mathrm{B.55})\\
&=-\sum_{i_1,\dots,i_4=1}^{3}\partial_{i_1}\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_2}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_3}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_4}\right)\cdot\frac{1}{9}\delta_{i_1,i_2}\delta_{i_3,i_4}.
\end{aligned}
$$

Altogether, we get

$$
\begin{aligned}
\mathcal{L}_3=&\sum_{i_1,\dots,i_4=1}^{3}\partial_{i_1}\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_2}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_3}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_4}\right)\\
&\qquad\qquad\qquad \cdot\frac{1}{45}\left(-2\delta_{i_1,i_2}\delta_{i_3,i_4}+3\delta_{i_1,i_3}\delta_{i_2,i_4}+3\delta_{i_1,i_4}\delta_{i_2,i_3}\right).
\end{aligned}
\tag{B.56}
$$

Remark 18. *Eq. (B.56) is exact and holds for arbitrary geometries. In principle, (B.56) can be used to derive some kind of equations which are asymptotically valid in heterogeneous media. However, due to the complexity of (B.56) one would probably lose the simple structure of the SP_N equations as presented here. To keep this main advantage of the SP_N method further simplifications are needed and one possibility is shown in the following.*

Assuming that the medium is either

- one dimensional (where the sum of 81 terms reduces to one single term) or
- homogeneous (where Σ_t is independent of the spatial variable $\underline{x}$),

we can rearrange the derivative operators in (B.56) and conclude

$$
\begin{aligned}
\mathcal{L}_3&=\sum_{i_1,\dots,i_4=1}^{3}\frac{4\delta_{i_1,i_2}\delta_{i_3,i_4}}{45}\,\partial_{i_1}\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_2}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_3}\right)\left((I+\mathcal{T})^{-1}\frac{1}{\Sigma_t}\partial_{i_4}\right)\\
&=\frac{4}{45}\left[\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right](I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right].
\end{aligned}
\tag{B.57}
$$

A similar analysis can also be performed for the next operator in the hierarchy. Since these expressions become even more involved and lengthy, we do not present them here. A road map to this analysis is given in [99] where a detailed derivation is also presented for $\mathcal{L}_3$. Making the same assertions as above the operator simplifies to

$$
\mathcal{L}_5=\frac{44}{945}\left[\underline{\nabla}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right](I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right](I+\mathcal{T})^{-1}\left[\underline{\mathcal{X}}\cdot(I+\mathcal{T})^{-1}\underline{\mathcal{X}}\right].
$$

B.3 Accuracy Tests: Manufactured Solutions

We investigate the convergence properties of our algorithms for the diffusion equation, P_N and SP_N equations for $N = 1, 3$ in this section. Our code is verified by the method of manufactured solutions [153]. By adding or modifying the source terms in the original equations in an appropriate manner, this technique enables to choose simple analytical functions which solve the respective equations. For every forthcoming numerical method, L^∞- and L^1-error are calculated for different numbers of discretization points ranging from $2^3 = 8$ to $2^9 = 512$ in each spatial dimension. Moreover, our computational domain is a square of side length one and the scalar flux is always chosen as

$$\phi_0(x, y, t) = e^{-t}(1 + \cos(2\pi(x + y))) \quad (x, y) \in [0, 1] \times [0, 1]. \tag{B.58}$$

We solve the PDEs on a torus and enforce periodic boundary conditions. The convergence analysis is performed for $\phi_0(x, y, t = 1)$ which is displayed in Figure B.1a. Our material properties are set to

$$\Sigma_a(x, y, t) = 1 + \cos(2\pi(x + y))t, \quad \Sigma_s = 1 \tag{B.59}$$

and we begin with the

B.3.1 Diffusion Equation

For the validation of our code gained by a finite difference Crank-Nicolson discretization of (2.77), we choose

$$\begin{aligned} Q(x, y, t) = \frac{\mathrm{e}^{-t}}{6(2 + \cos(2\pi(x + y))\,t)} & \left\{96 \cos^3(2\pi(x + y))\,\pi^2 t^2 - 11 - 96\pi^2 t \right. \\ & \left. - \left(6t - 288\pi^2 t\right) \cos(2\pi(x + y)) + \left(-6t - 11 - 48\pi^2 t^2 + 192\pi^2\right) \cos^2(2\pi(x + y))\right\}, \end{aligned} \tag{B.60}$$

as well as

$$\Sigma_a(x, y, t) = \frac{1}{6(2 + \cos(2\pi(x + y))\,t)} = \Sigma_s(x, y, t). \tag{B.61}$$

Reflection boundary conditions are set on the computational domain $[0, 1]^2$. The second order convergence is shown in Figure B.1b.

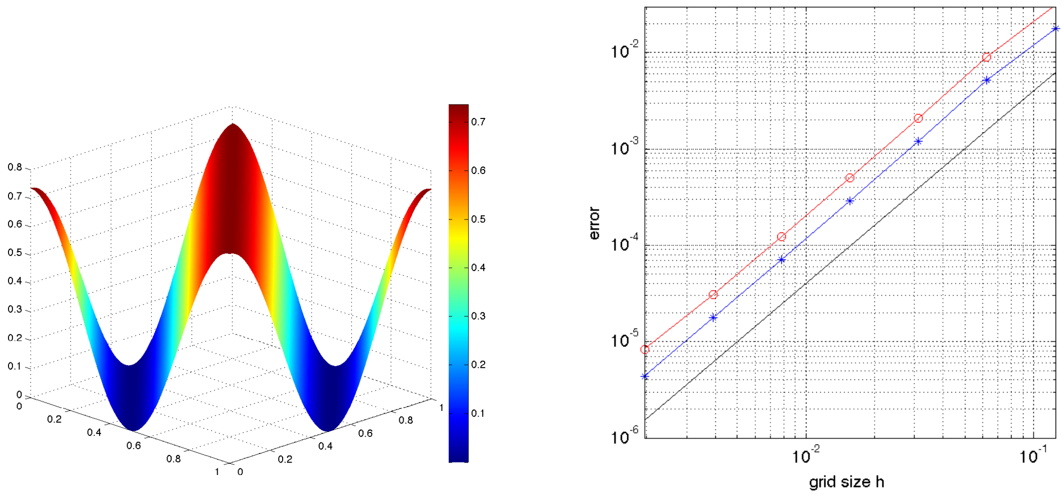
B.3.2 SP_N Equations

A. SP_1 Equations

Analytical expressions are set up for solutions of the three PDEs

$$\frac{\partial \phi_0}{\partial t}(x, y, t) + \underline{\nabla} \cdot \underline{J}_0(x, y, t) + \Sigma_a \phi_0(x, y, t) = Q_0^\phi(x, y, t), \tag{B.62a}$$

$$\frac{\partial \underline{J}_0}{\partial t}(x, y, t) + \frac{1}{3}\underline{\nabla}\phi_0(x, y, t) + \Sigma_t \underline{J}_0(x, y, t) = \underline{Q}_0^J(x, y, t). \tag{B.62b}$$



(a)

(b)

Figure B.1: Manufactured solutions: (a) Analytical solution $e^{-1}(1+\cos(2\pi(x+y)))$ (b) Error convergence for the diffusion approximation to ϕ_0 at $t_{\text{final}} = 1$: L^∞-error (red circle line), L^1-error (blue asterisk line), slope two (black solid line)

In addition to the scalar flux in (B.58) our desired solution for $\underline{J}_0$ reads:

$$J_{0,x}(x,y,t) = e^{-t}\left(1 - \frac{2\pi}{3}\sin(2\pi(x+y))\right), \tag{B.63}$$

$$J_{0,y}(x,y,t) = \frac{2\pi}{3}e^{-t}\sin(2\pi(x+y)). \tag{B.64}$$

which implies the following source terms:

$$Q_0^\phi(x,y,t) = \mathrm{e}^{-t}\cos\left(2\pi\left(x+y\right)\right)t[1+\cos\left(2\pi\left(x+y\right)\right)], \tag{B.65a}$$

$$Q_{0,x}^J(x,y,t) = \mathrm{e}^{-t}\cos\left(2\pi\left(x+y\right)\right)t\left(1-\frac{2\pi}{3}\sin\left(2\pi\left(x+y\right)\right)\right) \\ + \mathrm{e}^{-t}\left(1-\frac{4\pi}{3}\sin\left(2\pi\left(x+y\right)\right)\right), \tag{B.65b}$$

$$Q_{0,y}^J(x,y,t) = \frac{2\pi}{3}\mathrm{e}^{-t}\sin\left(2\pi\left(x+y\right)\right)\cos\left(2\pi\left(x+y\right)\right)t. \tag{B.65c}$$

B. SP_3 Equations
Apart from the the scalar flux in (B.58) the remaining variables are determined by

$$\phi_2(x,y,t) = -\frac{1}{2}e^{-t}\cos(2\pi(x+y)), \tag{B.66}$$

$$J_{0,x}(x,y,t) = e^{-t}\sin(2\pi(x+y)), \tag{B.67}$$

$$J_{0,y}(x,y,t) = e^{-t}(1-\sin(2\pi(x+y))), \tag{B.68}$$

$$J_{2,x}(x,y,t) = \frac{22}{21}e^{-t}(1+\sin(2\pi(x+y))), \tag{B.69}$$

$$J_{2,y}(x,y,t) = -\frac{22}{21}e^{-t}\sin(2\pi(x+y)). \tag{B.70}$$

Setting the source terms to

$$Q_0^{\phi}(x,y,t) = \mathrm{e}^{-t}\left(1+\cos\left(2\pi\left(x+y\right)\right)\right)\cos\left(2\pi\left(x+y\right)\right)t, \tag{B.71a}$$

$$Q_2^{\phi}(x,y,t) = -1/2\mathrm{e}^{-t}\cos\left(2\pi\left(x+y\right)\right)\left(1+\cos\left(2\pi\left(x+y\right)\right)t\right), \tag{B.71b}$$

$$Q_{0,x}^{J}(x,y,t) = \mathrm{e}^{-t}\sin\left(2\pi\left(x+y\right)\right)\left(1+\cos\left(2\pi\left(x+y\right)\right)t\right), \tag{B.71c}$$

$$\begin{aligned} Q_{0,y}^{J}(x,y,t) = {} & \mathrm{e}^{-t}\left(t-\sin\left(2\pi\left(x+y\right)\right)t\right)\cos\left(2\pi\left(x+y\right)\right) \\ & + \mathrm{e}^{-t}\left(1-\sin\left(2\pi\left(x+y\right)\right)\right), \end{aligned} \tag{B.71d}$$

$$\begin{aligned} Q_{2,x}^{J}(x,y,t) = {} & \frac{1}{21}\mathrm{e}^{-t}\left(22+\frac{27\pi}{5}+22\cos\left(2\pi\left(x+y\right)\right)t\right)\sin\left(2\pi\left(x+y\right)\right) \\ & + \frac{1}{21}\mathrm{e}^{-t}\left(22+22\cos\left(2\pi\left(x+y\right)\right)t\right), \end{aligned} \tag{B.71e}$$

$$Q_{2,y}^{J}(x,y,t) = \frac{1}{21}\mathrm{e}^{-t}\sin\left(2\pi\left(x+y\right)\right)\left(\frac{27\pi}{5}-22-22\cos\left(2\pi\left(x+y\right)\right)t\right) \tag{B.71f}$$

all functions in eqs. (2.77) solve the six PDEs

$$\frac{\partial\phi_0}{\partial t}(x,y,t) + \underline{\nabla}\cdot\underline{J}_0(x,y,t) + \Sigma_a\phi_0(x,yt) = Q_0^{\phi}(x,y,t), \tag{B.72a}$$

$$\frac{\partial\phi_2}{\partial t}(x,y,t) + \underline{\nabla}\cdot\underline{J}_2(x,y,t) + \Sigma_t\phi_2(x,y,t) = Q_2^{\phi}(x,y,t), \tag{B.72b}$$

$$\frac{\partial\underline{J}_0}{\partial t}(x,y,t) + \frac{1}{3}\underline{\nabla}(\phi_0(x,y,t)+2\phi_2(x,y,t)) + \Sigma_t\underline{J}_0(x,y,t) = \underline{Q}_0^{J}(x,y,t), \tag{B.72c}$$

$$\begin{aligned} \frac{\partial\underline{J}_2}{\partial t}(x,y,t) + \frac{1}{3}\underline{\nabla}\left(\frac{2}{5}\phi_0(x,y,t)+\frac{11}{7}\phi_2(x,y,t)\right) + \Sigma_t\underline{J}_2(x,y,t) \\ = \underline{Q}_2^{J}(x,y,t). \end{aligned} \tag{B.72d}$$

After implementing both approximations we compare numerical to expected solutions and achieve a second order convergence (Figure B.3).

B.3.3 P_N Equations

A. P_1 Equations

The P_1 equations in [161] with additional sources in the last two equations read

$$\frac{\partial\phi_0}{\partial t}(x,y,t) + \frac{1}{\sqrt{3}}\left(\frac{\partial\xi_2}{\partial y}(x,y,t) - \frac{\partial\xi_1}{\partial x}(x,y,t)\right) + \Sigma_a\phi_0(x,y,t) = Q_0(x,y,t), \tag{B.73a}$$

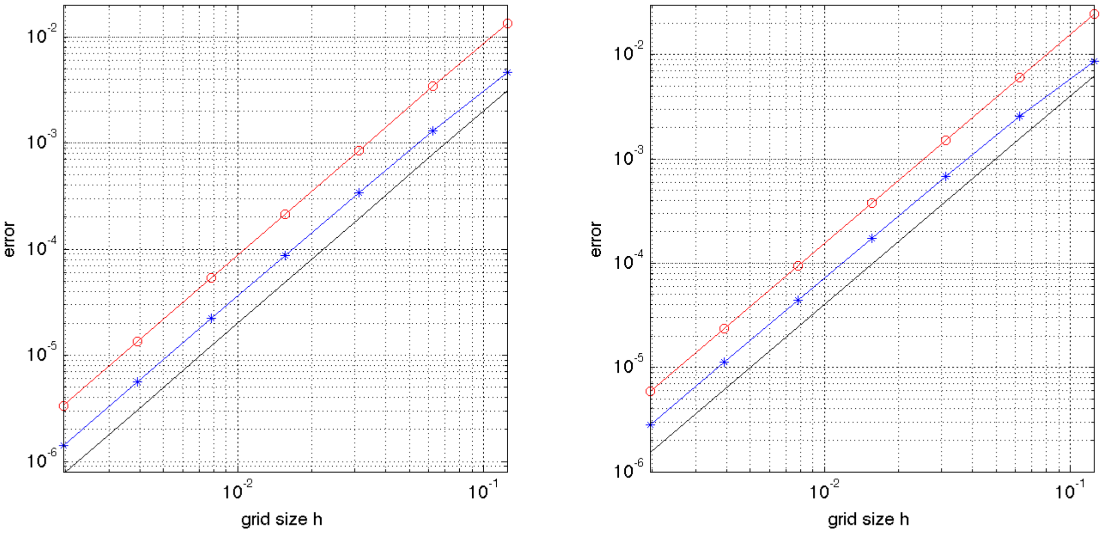
$$\frac{\partial\xi_1}{\partial t}(x,y,t) - \frac{1}{\sqrt{3}}\frac{\partial\phi_0}{\partial x}(x,y,t) + \Sigma_t\xi_1(x,y,t) = Q_1(x,y,t), \tag{B.73b}$$

$$\frac{\partial\xi_2}{\partial t}(x,y,t) + \frac{1}{\sqrt{3}}\frac{\partial\phi_0}{\partial y}(x,y,t) + \Sigma_t\xi_2(x,y,t) = Q_2(x,y,t). \tag{B.73c}$$

The choice for our solutions

$$\xi_1(x,y,t) = \frac{2\pi}{\sqrt{3}}e^{-t}\sin(2\pi(x+y)), \tag{B.74a}$$

$$\xi_2(x,y,t) = 1 + \frac{2\pi}{\sqrt{3}}e^{-t}\sin(2\pi(x+y)). \tag{B.74b}$$



(a) P_1 approximation

(b) P_3 approximation

Figure B.2: Error convergence for approximations to ϕ_0 at $t_{\text{final}} = 1$: L^∞-error (red circle line), L^1-error (blue asterisk line), slope two (black solid line)

implies the right-hand side

$$Q_0(x,y,t) = \mathrm{e}^{-t}\left(1+\cos\left(2\,\pi\;(x+y)\right)\right)\cos\left(2\,\pi\;(x+y)\right)t, \tag{B.75a}$$

$$Q_1(x,y,t) = \frac{2\pi}{\sqrt{3}}\left(2+\cos\left(2\,\pi\;(x+y)\right)t\right)\mathrm{e}^{-t}\sin\left(2\,\pi\;(x+y)\right), \tag{B.75b}$$

$$Q_2(x,y,t) = \mathrm{e}^{-t}\left(t+\frac{2\pi}{\sqrt{3}}t\sin\left(2\,\pi\;(x+y)\right)\right)\cos\left(2\,\pi\;(x+y)\right)+\mathrm{e}^{-t}. \tag{B.75c}$$

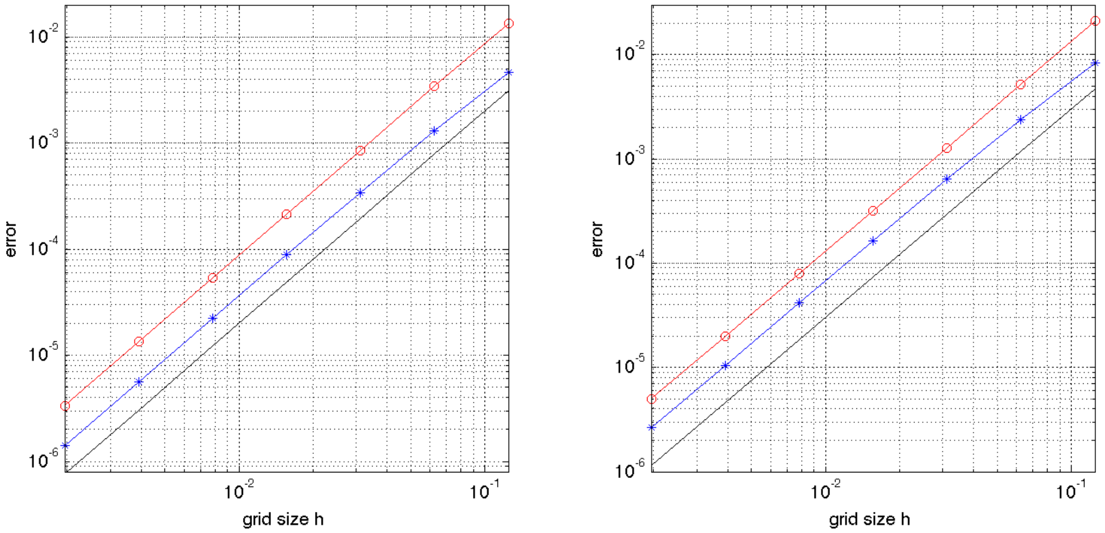
B. P_3 Equations

We go two steps further in the hierarchy and design manufactured solutions for the system of ten P_3 equations. The procedure is similar to the P_1 approximation. However, the regarding equations and their variables become lengthy and even more complicated. To simplify our notation all space- and time-arguments are dropped. We design analytical solutions to perform a convergence test for the following ten equations:

$$\frac{\partial \xi_0}{\partial t} + \frac{1}{\sqrt{3}}\left(-\frac{\partial \xi_1}{\partial x} + \frac{\partial \xi_2}{\partial y}\right) + \Sigma_a \xi_0 = Q_0 \tag{B.76}$$

$$\frac{\partial \xi_1}{\partial t} + \frac{\partial}{\partial x}\left(-\frac{1}{\sqrt{3}}\xi_0 - \frac{1}{\sqrt{5}}\xi_3 + \frac{1}{\sqrt{15}}\xi_5\right) + \frac{1}{\sqrt{5}}\frac{\partial \xi_4}{\partial y} + \Sigma_t \xi_1 = Q_1, \tag{B.77}$$

$$\frac{\partial \xi_2}{\partial t} - \frac{1}{\sqrt{5}}\frac{\partial \xi_4}{\partial x} + \frac{\partial}{\partial y}\left(\frac{1}{\sqrt{3}}\xi_0 - \frac{1}{\sqrt{5}}\xi_3 - \frac{1}{\sqrt{15}}\xi_5\right) + \Sigma_t \xi_2 = Q_2, \tag{B.78}$$



(a) SP_1 approximation (b) SP_3 approximation

Figure B.3: Error convergence for ϕ_0 at $t_{\text{final}} = 1$: L^∞-error (red circle line), L^1-error (blue asterisk line), slope two (black solid line)

$$\frac{\partial \xi_3}{\partial t} + \frac{\partial}{\partial x}\left(-\frac{1}{\sqrt{5}}\xi_1 - \sqrt{\frac{3}{14}}\xi_6 + \frac{1}{\sqrt{70}}\xi_8\right) + \frac{\partial}{\partial y}\left(-\frac{1}{\sqrt{5}}\xi_2 + \sqrt{\frac{3}{14}}\xi_7 + \frac{1}{\sqrt{70}}\xi_9\right) + \Sigma_t \xi_3 = Q_3, \tag{B.79}$$

$$\frac{\partial \xi_4}{\partial t} + \frac{\partial}{\partial x}\left(-\frac{1}{\sqrt{5}}\xi_2 - \sqrt{\frac{3}{14}}\xi_7 + \frac{1}{\sqrt{70}}\xi_9\right) + \frac{\partial}{\partial y}\left(\frac{1}{\sqrt{5}}\xi_1 - \sqrt{\frac{3}{14}}\xi_6 - \frac{1}{\sqrt{70}}\xi_8\right) + \Sigma_t \xi_4 = Q_4, \tag{B.80}$$

$$\frac{\partial \xi_5}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{\sqrt{15}}\xi_1 - \sqrt{\frac{6}{35}}\xi_8\right) + \frac{\partial}{\partial y}\left(-\frac{1}{\sqrt{15}}\xi_2 + \sqrt{\frac{6}{35}}\xi_9\right) + \Sigma_t \xi_5 = Q_5, \tag{B.81}$$

$$\frac{\partial \xi_6}{\partial t} - \sqrt{\frac{3}{14}}\left(\frac{\partial \xi_3}{\partial x} + \frac{\partial \xi_4}{\partial y}\right) + \Sigma_t \xi_6 = Q_6, \tag{B.82}$$

$$\frac{\partial \xi_7}{\partial t} - \sqrt{\frac{3}{14}}\left(\frac{\partial \xi_4}{\partial x} - \frac{\partial \xi_3}{\partial y}\right) + \Sigma_t \xi_7 = Q_7, \tag{B.83}$$

$$\frac{\partial \xi_8}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{\sqrt{70}}\xi_3 - \sqrt{\frac{6}{35}}\xi_5\right) - \frac{1}{\sqrt{70}}\frac{\partial \xi_4}{\partial y} + \Sigma_t \xi_8 = Q_8, \tag{B.84}$$

$$\frac{\partial \xi_9}{\partial t} + \frac{1}{\sqrt{70}}\frac{\partial \xi_4}{\partial x} + \frac{\partial}{\partial y}\left(\frac{1}{\sqrt{70}}\xi_3 + \sqrt{\frac{6}{35}}\xi_5\right) + \Sigma_t \xi_9 = Q_9. \tag{B.85}$$

If the sources are set to

$$Q_0 = \mathrm{e}^{-t} t \cos\left(2\pi\left(x+y\right)\right)\left[1+\cos\left(2\pi\left(x+y\right)\right)\right], \tag{B.86}$$

$$Q_1 = \mathrm{e}^{-t} \sin\left(2\pi\left(x+y\right)\right)\left[1+\cos\left(2\pi\left(x+y\right)\right)t\right], \tag{B.87}$$

$$Q_2 = \left(\cos\left(2\pi\left(x+y\right)\right)t + 1 + \frac{4\pi}{\sqrt{5}}\right)\mathrm{e}^{-t}\sin\left(2\pi\left(x+y\right)\right) + \mathrm{e}^{-t}\left(1+\cos\left(2\pi\left(x+y\right)\right)t\right), \tag{B.88}$$

$$Q_3 = \mathrm{e}^{-t}\cos\left(2\pi\left(x+y\right)\right)\left(\cos\left(2\pi\left(x+y\right)\right)t + 1 - \frac{4\pi}{\sqrt{5}}\right) - \frac{2\pi\sqrt{2}}{\sqrt{35}}\mathrm{e}^{-t}\sin\left(2\pi\left(x+y\right)\right), \tag{B.89}$$

$$Q_4 = \mathrm{e}^{-t}\cos\left(2\pi\left(x+y\right)\right)\left(t\cos\left(2\pi\left(x+y\right)\right) + 1 - \frac{2\pi\sqrt{6}}{\sqrt{7}} + t\right) + \mathrm{e}^{-t}, \tag{B.90}$$

$$Q_5 = \mathrm{e}^{-t}\cos\left(2\pi\left(x+y\right)\right)\left(\sqrt{5}\cos\left(2\pi\left(x+y\right)\right)t + 2t + \sqrt{5}\right) + 2\mathrm{e}^{-t}, \tag{B.91}$$

$$Q_6 = \left(\cos\left(2\pi\left(x+y\right)\right)t + 1 + \frac{2\pi\sqrt{6}}{\sqrt{7}}\right)\mathrm{e}^{-t}\sin\left(2\pi\left(x+y\right)\right) + \left[1+\cos\left(2\pi\left(x+y\right)\right)t\right]\mathrm{e}^{-t}, \tag{B.92}$$

$$Q_7 = \left(1+\cos\left(2\pi\left(x+y\right)\right)t\right)\mathrm{e}^{-t}\sin\left(2\pi\left(x+y\right)\right), \tag{B.93}$$

$$Q_8 = \mathrm{e}^{-t}\cos\left(2\pi\left(x+y\right)\right)\left[\cos\left(2\pi\left(x+y\right)\right)t + 1 + t\right] + \mathrm{e}^{-t}\left(\frac{2\pi\sqrt{6}}{\sqrt{7}}\sin\left(2\pi\left(x+y\right)\right)\pi + 1\right), \tag{B.94}$$

$$Q_9 = -\mathrm{e}^{-t}\frac{2\pi\sqrt{2}}{\sqrt{35}}\left(5+\sqrt{15}\right)\sin\left(2\pi\left(x+y\right)\right) + \mathrm{e}^{-t}\cos\left(2\pi\left(x+y\right)\right)\left[\cos\left(2\pi\left(x+y\right)\right)t + 1\right], \tag{B.95}$$

then

$$\xi_1 = e^{-t}\sin(2\pi(x+y)), \tag{B.96}$$

$$\xi_2 = e^{-t}(1+\sin(2\pi(x+y))), \tag{B.97}$$

$$\xi_3 = e^{-t}\cos(2\pi(x+y)), \tag{B.98}$$

$$\xi_4 = e^{-t}(1+\cos(2\pi(x+y))), \tag{B.99}$$

$$\xi_5 = e^{-t}(2+\sqrt{5}\cos(2\pi(x+y))), \tag{B.100}$$

$$\xi_6 = e^{-t}(1+\sin(2\pi(x+y))), \tag{B.101}$$

$$\xi_7 = e^{-t}\sin(2\pi(x+y)), \tag{B.102}$$

$$\xi_8 = e^{-t}(1+\cos(2\pi(x+y))), \tag{B.103}$$

$$\xi_9 = e^{-t}\cos(2\pi(x+y)), \tag{B.104}$$

solve the ten PDEs stated above. All our calculations were performed by means of the commercial software MAPLE.
A second order converge can be observed in Figure B.2 for both P_1 and P_3 methods.

Appendix C

Discontinuous Galerkin Method: Slope Limiting with Characteristic Variables

In this appendix, we give details of the limiting procedure for the characteristic variables of the M_1 system in Chapter 3. Let

$$\mathbf{u}_l^j = \begin{bmatrix} u_l^{j,(1)} \\ u_l^{j,(2)} \end{bmatrix} \tag{C.1}$$

be the updated coefficients of the state variable $\mathbf{u}$ from the previous stage of the limiting process. Here l is the moment index, j is the cell index, and the additional superscripts denote the components of the vector. Let $\mathcal{J}$ be the set of all j such that cell j requires limiting at the next stage and let l_* be the moment that will be limited in that stage.

Moreover, let $S(\mathbf{u}) \in \mathbb{R}^{2\times 2}$ be the matrix which transforms state to characterstic variables and let $S_j = S(\bar{\mathbf{u}}^j)$ where $\bar{\mathbf{u}}^j$ is the average of $\mathbf{u}$ on cell j. Note that for fixed j, S_j is the same throughtout the limiting process.

For all $j \in \mathcal{J}$, compute the characterstic variables

$$\mathbf{w}_{l_*}^{j,k} = \begin{bmatrix} w_{l_*}^{j,k,(1)} \\ w_{l_*}^{j,k,(2)} \end{bmatrix} = S_j \mathbf{u}_{l_*}^{j+k} \tag{C.2}$$

and

$$\mathbf{w}_{l_*-1}^{j,k} = \begin{bmatrix} w_{l_*-1}^{j,k,(1)} \\ w_{l_*-1}^{j,k,(2)} \end{bmatrix} = S_j \mathbf{u}_{l_*-1}^{j+k} \tag{C.3}$$

where $k \in \{-1, 0, 1\}$ is the index for the surrounding stencil. From this, one can compute

$$\Delta_{l_*-1,\pm}^{j,(\beta)} = \pm \left(w_{l_*-1}^{j,\pm 1,(\beta)} - w_{l_*-1}^{j,0,(\beta)} \right), \quad \beta = 1, 2. \tag{C.4}$$

Although it might be more efficient to use the fact that

$$\Delta_{l_*-1,\pm}^{j,(\beta)} = \pm S_j \left(u_{l_*-1}^{j\pm 1,(\beta)} - u_{l_*-1}^{j,(\beta)} \right). \tag{C.5}$$

Additionally, we also need to calculate

$$H_{*,\pm}^{j,(\beta)} = \Delta_{l_*-1,\pm}^{j,(\beta)} - (2l_* - 1) w_{l_*}^{j,\pm 1,(\beta)}. \tag{C.6}$$

Note that one only needs to send $H_{*,\pm}^{(\beta),j}$, $\Delta_{l_*-1,\pm}^{(\beta),j}$ and $w_{l_*}^{j,0,(\beta)}$ to the limiter function.

Now apply the limiter function componentwise: Let $\mathcal{J}'$ be the set of all j such that $w_{l_*}^{j,0,(\beta)}$ is updated to $\tilde{w}_{l_*}^{j,0,(\beta)}$ by the limiter for $\beta = 1$ or $\beta = 2$. The last step is then to transform the updated characteristic variables to state variables via

$$\tilde{\mathbf{u}}_{l_*}^{j} = S_j^{-1}\tilde{\mathbf{w}}_{l_*}^{j,0} \tag{C.7}$$

and change $\mathbf{u}_{l_*}^{j}$ to $\tilde{\mathbf{u}}_{l_*}^{j}$ only for $j \in \mathcal{J}'$. Then update $\mathcal{J}$, the index set to be checked for limiting in the next step and l_* the moment to be limited in the next step.

Bibliography

[1] *Directory of Radiotherapy Centres.* http://www-naweb.iaea.org/nahu/dirac/query3.asp, accessed on 2011-11-03.

[2] *PTCOG: Particle Therapy Co-Operative Group.* http://ptcog.web.psi.ch/, accessed on 2011-10-11.

[3] S. Agostinelli et al., *Geant4–a simulation toolkit*, Nucl. Instrum. Methods B, 506 (2003), pp. 250–303.

[4] G. 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*, submitted, (2012).

[5] J. Allison et al., *Geant4 developments and applications*, IEEE Transactions on Nuclear Science, 53 (2006), pp. 270–278.

[6] A. Anile, W. Allegretto, and C. Ringhofer, *Mathematical Problems in Semiconductor Physics*, vol. 1823 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 2003. Lectures given at the C.I.M.E. Summer School held in Cetraro, Italy on July 15-22, 1998.

[7] A. M. Anile and O. Muscato, *Improved hydrodynamical model for carrier transport in semiconductors*, Phys. Rev. B, 51 (1995), pp. 16728–16740.

[8] A. M. Anile and S. Pennisi, *Thermodynamic derivation of the hydrodynamical model for charge transport in semiconductors*, Phys. Rev. B, 46 (1992), pp. 187–193.

[9] A. M. Anile and V. Romano, *Hydrodynamical modeling of charge carrier transport in semiconductors*, Meccanica, 35 (2000), pp. 249–296.

[10] R. Backofen, T. Bilz, A. Ribalta, and A. Voigt, *SP_N-approximations of internal radiation in crystal growth of optical materials*, Journal of Crystal Growth, 266 (2004), pp. 264–270.

[11] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[12] C. Beckert and U. Grundmann, *Development and verification of a nodal approach for solving the multigroup SP_3 equations*, Ann. Nucl. Energy, 35 (2008), pp. 75–86.

[13] C. Berthon, P. Charrier, and B. Dubroca, *An HLLC scheme to solve the M1 model of radiative transfer in two space dimensions*, J. Sci. Comput., 31 (2007), pp. 347–389.

[14] H. A. Bethe, *Zur Theorie des Durchgangs schneller Korpuskularstrahlen durch Materie*, Ann. d. Physik, 397 (1930), pp. 325–400.

[15] R. Biswas, K. D. Devine, and J. E. Flaherty, *Parallel, adaptive finite element methods for conservation laws*, Applied Numerical Mathematics, 14 (1994), pp. 255 – 283.

[16] S. A. Bludman and J. Cernohorsky, *Stationary neutrino radiation transport by maximum entropy closure*, Phys. Rep., 256 (1995), pp. 37 – 51.

[17] B. H. Bransden and C. J. Joachain, *Physics of atoms and molecules*, Prentice Hall, 2nd ed., 2003.

[18] P. Brantley and E. W. Larsen, *The simplified P_3 approximation*, Nucl. Sci. Eng., 134 (2000), pp. 1–21.

[19] T. A. Brunner, *Riemann Solvers for Time-Dependent Transport Based on the Maximum Entropy and Spherical Harmonics Closures*, PhD thesis, University of Michigan, 2000.

[20] ——, *Forms of approximate radiation transport*, Tech. Rep. SAND2002-1778, Sandia National Laboratories, 2002.

[21] T. A. Brunner and J. P. Holloway, *One-dimensional Riemann solvers and the maximum entropy closure*, J. Quant Spect. and Radiative Trans, 69 (2001), pp. 543 – 566.

[22] ——, *Two-dimensional time-dependent Riemann solvers for neutron transport*, J. Comp Phys., 210 (2005), pp. 386–399.

[23] C. Buet and S. Cordier, *An asymptotic preserving scheme for hydrodynamics radiative transfer models: numerics for radiative transfer*, Numer. Math., 108 (2007), pp. 199–221.

[24] C. Buet and B. Després, *Asymptotic preserving and positive schemes for radiation hydrodynamics*, J. Comput. Phys., 215 (2006), pp. 717–740.

[25] A. Burbeau, P. Sagaut, and C.-H. Bruneau, *A problem-independent limiter for high-order runge-kutta discontinuous galerkin methods*, J. Comput. Phys., 169 (2001), pp. 111–150.

[26] N. J. Carron, *An Introduction to the Passage of Energetic Particles through Matter*, Taylor & Francis Group, 2007.

[27] J. Cernohorsky and S. A. Bludman, *Stationary neutrino radiation transport by maximum entropy closure*, Tech. Rep. LBL–36135, Lawrence Berkely National Laboratory, 1994.

[28] J. Cernohorsky, L. J. van den Horn, and J. Cooperstein, *Maximum entropy Eddington factors in flux-limited neutrino diffusion*, Journal of Quantitative Spectroscopy and Radiative Transfer, 42 (1989), pp. 603 – 613.

[29] M. Chadwick, P. G. Young, G. M. Hale, et al., *LA150 documentation of cross sections, heating and damage: Part A (incident neutrons) and part B (incident protons)*, Tech. Rep. LA-UR-99-1222, Los Alamos National Laboratory, 1999.

[30] S. Chandrasekhar, *Radiative transfer*, Oxford University Press, Dover, 1950.

[31] G. Chiba, *Application of the hierarchical domain decomposition boundary element method to the simplified P_3 equation*, Ann. Nucl. Energy, 38 (2011), pp. 1033–1038.

[32] B. Cockburn, S. Hou, and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[33] B. Cockburn, G. Karniadakis, and C. Shu, *Discontinuous Galerkin methods: theory, computation, and applications*, Lecture notes in computational science and engineering, Springer, 2000.

[34] B. Cockburn, S. Lin, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems.*, J. Comput. Phys., 84 (1989), pp. 90–113.

[35] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservation laws V*, J. Comput. Phys., 141 (1998), pp. 199–224.

[36] J.-F. Coulombel and T. Goudon, *Entropy-based moment closure for kinetic equations: Riemann problem and invariant regions*, JHDE, 3 (2006), pp. 649–671.

[37] D. E. Cullen, J. H. Hubbel, and L. Kissel, *EPDL97, the evaluated photon data library, '97 version*, Tech. Rep. UCRL-50400, Lawrence Livermore National Laboratory, 1997.

[38] R. Dautray and J. L. Lions, *Mathematical Analysis And Numerical Methods For Science And Technology, Volume 6: Evolution Problems II*, Spinger-Verlag, Berlin, 2000.

[39] B. Davison, *Neutron Transport Theory*, Oxford Clarendon Press, 1957.

[40] C. M. Davisson and R. D. Evans, *Gamma-ray absorption coefficients*, Rev. Mod. Phys., 24 (1952), pp. 79–107.

[41] P. Degond and C. Ringhofer, *Quantum moment hydrodynamics and the entropy principle*, J. Stat. Phys., 112 (2003), pp. 587–627.

[42] J. D. Densmore and R. G. McClarren, *Moment analysis of angular approximation methods for time-dependent radiation transport*, Transport Theor. Stat. Phys., 39 (2011), pp. 192–233.

[43] T. Downar, D. Lee, Y. Xu, and T. Kozlowski, *PARCS v2.6, U.S. NRC Core Neutronics Simulator*, School of Nuclear Engineering, Purdue University, draft ed., 2004. https://engineering.purdue.edu/PARCS.

[44] W. Dreyer, *Maximisation of the entropy in non-equilibrium*, Journal of Physics A Mathematical General, 20 (1987), pp. 6505–6517.

[45] W. Dreyer, M. Herrmann, and M. Kunik, *Kinetic solutions of the Boltzmann-Peierls equation and its moment systems*, Continuum Mechanics and Thermodynamics, 16 (2004), pp. 453–469.

[46] B. Dubroca and J. L. Feugeas, *Entropic moment closure hierarchy for the radiative transfer equation*, C. R. Acad. Sci. Paris Ser. I, 329 (1999), pp. 915–920.

[47] ——, *Étude théorique et numérique d'une hiérarchie de modèles aus moments pour le transfert radiatif*, C.R. Acad. Sci. Paris, I. 329 (1999), pp. 915–920.

[48] B. Dubroca and A. Klar, *Half-moment closure for radiative transfer equations*, J. Comput. Phys., 180 (2002), pp. 584–596.

[49] C. Dunford, *ENDF Utility Codes Release 6.13*, National Nuclear Data Center, Brookhaven National Laboratory, 2002.

[50] E. Everhart, G. Stone, and R. J. Carbone, *Classical calculation of differential cross section for scattering from a Coulomb potential with exponential screening*, Phys. Rev., 99 (1955), pp. 1287–1290.

[51] J. Ferlay et al., *GLOBOCAN 2008 v1.2, cancer incidence and mortality worldwide: IARC CancerBase no. 10.* http://globocan.iarc.fr, accessed on 2011-11-03.

[52] M. Frank, *Approximate models for radiative transfer*, Bulletin of the Institute of Mathematics Academia Sinica (New Series), 2 (2007), pp. 409–432.

[53] ——. Private communication, 2010.

[54] M. Frank, B. Dubroca, and A. Klar, *Partial moment entropy approximation to radiative heat transfer*, J. Comput. Phys., 218 (2006), pp. 1–18.

[55] M. Frank, C. D. Hauck, and E. Olbrant, *Perturbed, entropy-based closure for radiative transfer*, in preparation, (2012).

[56] M. Frank, A. Klar, E. W. Larsen, and S. Yasuda, *Time-dependent simplified P_N approximation to the equations of radiative transfer*, J. Comput. Phys., 226 (2007), pp. 2289–2305.

[57] M. Frank, A. Klar, and R. Pinnau, *Optimal control of glass cooling using simplified pn theory*, Transport Theor. Stat. Phys., 39 (2011), pp. 282–311.

[58] M. Frank, J. Lang, and M. Schäfer, *Adaptive finite element simulation of the time-dependent simplified P_N equations*, J. Sci. Comput., 49 (2011), pp. 332–350.

[59] M. Frank and B. Seibold, *Optimal prediction for radiative transfer: A new perspective on moment closure*, Kinet. Relat. Models., 4 (2011), pp. 717–733.

[60] B. D. Ganapol, *Homogeneous infinite media time-dependent analytic benchmarks for X-TM transport methods development*, tech. rep., Los Alamos National Laboratory, March 1999.

[61] B. D. Ganapol, C. T. Kelley, and G. C. Pomraning, *Asymptotically exact boundary conditions for the P_N equations*, Nucl. Sci. Eng., 114 (1993), pp. 12–19.

[62] J. C. Garth, *Electron/photon transport and its applications*, The Monte Carlo Method: Versatility Unbounded in a Dynamic Computing World, (2005).

[63] E. M. Gelbard, *Applications of spherical harmonics method to reactor problems*, Tech. Rep. WAPD-BT-20, Bettis Atomic Power Laboratory, 1960.

[64] ——, *Simplified spherical harmonics equations and their use in shielding problems*, Tech. Rep. WAPD-T-1182, Bettis Atomic Power Laboratory, 1961.

[65] ——, *Applications of the simplified spherical harmonics equations in spherical geometry*, Tech. Rep. WAPD-TM-294, Bettis Atomic Power Laboratory, 1962.

[66] S. Gottlieb, D. Ketcheson, and C.-W. Shu, *High order strong stability preserving time discretizations*, J. Sci. Comput., 38 (2009), pp. 251–289.

[67] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge-Kutta schemes*, Math. Comput., 67 (1998), pp. 73–85.

[68] S. Gottlieb, C.-W. Shu, and E. Tadmore, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[69] B. Grosswendt and E. Waibel, *Transport of low energy electrons in nitrogen and air*, Nucl. Instrum. Methods, 155 (1978), pp. 145–156.

[70] C. Groth and J. McDonald, *Towards physically realizable and hyperbolic moment closures for kinetic theory*, Continuum Mechanics and Thermodynamics, (2009), pp. 467—493.

[71] U. Grundmann and U. Rohde, *DYN3D - A 3-dimensional core model for steady state and transient analysis in thermal reactors*, in PHYSOR 96 - Breakthrough of Nuclear Energy by Reactor Physics, 1996, pp. 70–79.

[72] D. E. H. Nikjoo, S. Uehara and A. Brahme, *Heavy charged particles in radiation biology and biophysics*, New J. Phys., 10 (2008), pp. 1–28.

[73] G. Habetler and B. Matkowsky, *Uniform asymptotic expansions in transport theory with small mean free paths*, J. Math. Phys., 16 (1975), pp. 846–854.

[74] G. M. Hale, *Use of R-matrix methods for light element evaluations*, in Proceedings of the Conference on Nuclear Data Evaluation Methods and Procedure, held at Brookhaven National Laboratory, Upton, New York 1197, Los Alamos Scientific Laboratory, University of California, 1980, pp. 509–521.

[75] C. Hauck and R. G. McClarren, *Positive P_N closures*, Tech. Rep. LA-UR 09-03865, Los Alamos Report, 2009.

[76] C. D. Hauck, *Entropy-Based Moment Closures in Semiconductor Models*, PhD thesis, University of Maryland, College Park, 2006.

[77] ——, *High-order entropy-based closures for linear transport in slab geometry*, Communications in Mathematical Sciences, 9 (2011), pp. 187–205.

[78] 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 (2008), pp. 1977–2015.

[79] C. D. Hauck and R. G. McClarren, *Positive P_N closures*, SIAM Journal on Scientific Computing, 32 (2010), pp. 2603–2626.

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

[81] W. Huang, *Heuristic solutions to polynomial moment problems with some convex entropic objectives*, Numerical Algorithms, 12 (1996), pp. 297–308.

[82] J. Hubbell et al., *Atomic form factors, incoherent scattering functions, and photon scattering cross sections*, J. Phys. Chem. Ref. Data, 4 (1975), pp. 471–538.

[83] ICRU, *Stopping Power and Ranges for Protons and Alpha Particles*, no. 49, Oxford University Press, 1993.

[84] ——, *Nuclear Data for Neutron and Proton Radiotherapy and for Radiation Protection*, vol. 63, Oxford Journals, 1999.

[85] ——, *Elastic scattering of electrons and photons*, vol. 7, Oxford University Press, 2007.

[86] Y. Itikawa and N. Mason, *Cross sections for electron collisions with water molecules*, J. Phys. Chem. Ref. Data, 34 (2005), pp. 1–22.

[87] J. H. Jeans, *The equations of radiative transfer of energy*, Monthly Notices Royal Astronomical Society, 78 (1917), pp. 28–36.

[88] B. Jones and N. Burnet, *Radiotherapy for the future, protons and ions hold much promise*, British Medical Journal, 330 (2005), pp. 979–980.

[89] A. Jüngel, S. Krause, and P. Pietra, *A hierarchy of diffusive higher-order moment equations for semiconductors*, SIAM Journal on Applied Mathematics, 68 (2007), pp. 171–198.

[90] M. Junk, *Domain of definition of Levermore's five moment system*, J. Stat. Phys., 93 (1998), pp. 1143–1167.

[91] ——, *Maximum entropy for reduced moment problems*, Math. Mod. Meth. Appl. S., 10 (2000), pp. 1001–1025.

[92] M. Junk and V. Romano, *Maximum entropy moment system of the semiconductor Boltzmann equation using Kane's dispersion relation*, Continuum Mechanics and Thermodynamics, 17 (2005), pp. 247–267.

[93] A. D. Klose and E. W. Larsen, *Light transport in biological tissue based on the simplified spherical harmonics equations*, J. Comput. Phys., 220 (2006), pp. 441–470.

[94] A. D. Klose and T. Pöschinger, *Excitation-resolved fluorescence tomography with simplified spherical harmonics equations*, Phys. Med. Biol., 56 (2011), pp. 1443–1469.

[95] P. Kotiluoto, *Adaptive tree multigrids and simplified spherical harmonics approximation in deterministic neutral and charged particle transport*, PhD thesis, University of Helsinki, VTT Technical Research Centre of Finland, 2007.

[96] H. Krieger, *Grundlagen der Strahlungsphysik und des Strahlenschutzes*, Teubner Verlag, Wiesbaden, 2. ed., 2007.

[97] E. Larsen, *The asymptotic diffusion limit of discretized transport problems*, Nucl. Sci. Eng., 112 (1992), pp. 336–346.

[98] E. W. Larsen, *Diffusion theory as an asymptotic limit of transport theory for nearly critical systems with small mean free paths*, Ann. Nucl. Energy, 7 (1980), pp. 249–255.

[99] ——, *Asymptotic diffusion and simplified P_N approximations for diffusive and deep penetration problems. part 1: Theory*, Transport Theor. Stat. Phys., 39 (2011), pp. 110–163.

[100] E. W. Larsen et al., *Electron dose calculations using the Method of Moments*, Med. Phys., 24 (1997), pp. 111–125.

[101] E. W. Larsen and J. Keller, *Asymptotic solution of neutron transport problems for small mean free paths*, J. Math. Phys., 15 (1974), pp. 75–81.

[102] E. W. Larsen, J. Morel, and J. McGhee, *Asymptotic derivation of the simplified P_N equations*, Proc. ANS Topical Meeting, Mathematical Methods and Supercomputing in Nuclear Applications, 1 (1993), pp. 718–730.

[103] 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 (1991), pp. 76–85.

[104] ——, *The P_N theory as an asymptotic limit of transport theory in planar geometry I: Analysis*, Nucl. Sci. Eng, 109 (1991), pp. 49–75.

[105] E. W. Larsen, G. Thömmes, A. Klar, M. Seaïd, and T. Götz, *Simplified P_N approximations to the equations of radiative heat transfer and applications*, J. Comput. Phys., 183 (2002), pp. 652–675.

[106] J. A. LaVerne and S. M. Pimblott, *Effect of elastic collisions on energy deposition by electrons in water*, J. Phys. Chem., 101 (1997), pp. 4504–4510.

[107] C. L. Leakeas and E. W. Larsen, *Generalized Fokker-Planck approximations of particle transport with highly forward-peaked scattering*, Nucl. Sci. Eng., 137 (2001), pp. 236–250.

[108] C. D. Levermore, *Boundary conditions for moment closures.* Presented at Institute for Pure and Applied Mathematics University of California, Los Angeles, CA on May 27, 2009.

[109] ——, *Moment closure hierarchies for kinetic theories*, J. Stat. Phys., 83 (1996), pp. 1021–1065.

[110] ——, *Moment closure hierarchies for the Boltzmann-Poisson equation*, VLSI Design, 6 (1998), pp. 97–101.

[111] ——, *Boundary conditions for moment closures.* Presentation at The Annual Kinetic FRG Meeting, 2009.

[112] C. D. Levermore, W. J. Morokoff, and B. T. Nadiga, *Moment realizability and the validity of the Navier-Stokes equations for rarefied gas dynamics*, Phys. Fluids, 10 (1998), pp. 3214–3226.

[113] E. E. Lewis and J. W. F. Miller, *Computational Methods in Neutron Transport*, John Wiley and Sons, New York, 1984.

[114] C. K. Li and R. D. Petrasso, *Stopping of directed energetic electrons in high-temperature hydrogenic plasmas*, Physical Review E, 70 (2004), pp. 1–4.

[115] H. Liu and J. Yan, *The Direct Discontinuous Galerkin (DDG) methods for diffusion problems*, SIAM J. Numer. Anal., 47 (2009), pp. 475–698.

[116] R. B. Lowrie and J. E. Morel, *Methods for hyperbolic systems with stiff relaxation*, Int. J. Numer. Meth. Fluids, 40 (2002), pp. 413–423.

[117] T. M. MacRobert, *Spherical Harmonics: An Elementary Treatise on Harmonic Functions with Applications*, Dover Publications, Inc., New York, 1973.

[118] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, New York, 1990.

[119] R. E. Marshak, *Effect of radiation on shock wave behaviour*, Phys. Fluids 1, 1 (1958), pp. 24–29.

[120] R. G. McClarren, *Theoretical aspects of the simplified P_N equations*, Transport Theor. Stat. Phys., 39 (2011), pp. 73–109.

[121] R. G. McClarren and R. P. Drake, *Anti-diffusive radiation flow in the cooling layer of a radiating shock*, J. Quant. Spectrosc. Radiat. Transfer, 111 (2010), pp. 2095–2105.

[122] R. G. McClarren et al., *Semi-implicit time integration for P_N thermal radiative transfer*, J. Comp Phys., 227 (2008), pp. 7561–7586.

[123] R. G. McClarren and C. D. Hauck, *Robust and accurate filtered spherical harmonics expansions for radiative transfer*, J. Comput. Phys., 229 (2010), pp. 5597–5614.

[124] R. G. McClarren, J. P. Holloway, and T. A. Brunner, *On solutions to the p_n equations for thermal radiative transfer*, J. Comput. Phys., 227 (2008), pp. 2864–2885.

[125] S. A. McKee, W. A. Wulf, and T. C. Landon, *Bounds on memory bandwidth in streamed computations*, in Euro-Par '95: Proceedings of the First International Euro-Par Conference on Parallel Processing, London, UK, 1995, Springer-Verlag, pp. 83–99.

[126] Members of the Cross Sections Evaluation Working Group, *ENDF-6 Formats Manual ENDF-6 Formats Manual: Data Formats and Procedures for the Evaluated Nuclear Data File ENDF/B-VI and ENDF/B-VII*, National Nuclear Data Center, Brookhaven National Laboratory, 2009.

[127] G. N. Minerbo, *Maximum entropy Eddington factors*, J. Quant. Spectrosc. Radiat. Transfer, 20 (1978), pp. 541—545.

[128] C. Møller, *Zur Theorie des Durchgangs schneller Elektronen durch Materie*, Ann. d. Physik, 14 (1932), pp. 531–585.

[129] P. Monreal and M. Frank, *Higher order minimum entropy approximations in radiative transfer*. http://arxiv.org/abs/0812.3063.

[130] J. Morel, J. McGhee, and E. W. Larsen, *A three-dimensional time-dependent unstructured tetrahedral-mesh SP_N method*, Nucl. Sci. Eng., 123 (1996), pp. 319–327.

[131] J. E. Morel, *An improved Fokker-Planck angular differencing scheme*, Nucl. Sci. Eng., 89 (1985), pp. 131–136.

[132] I. Müller and T. Ruggeri, *Rational Extended Thermodynamics*, vol. 37 of Springer Tracts in Natural Philosophy, Springer-Verlag, New York, second ed., 1993.

[133] National Institute of Standards and Technology, *PSTAR: Stopping power and range tables for protons*. http://physics.nist.gov/PhysRefData/Star/Text/PSTAR.html, accessed on 2011-09-30.

[134] N. C. Nguyen, J. Peraire, and B. Cockburn, *An implicit high-order hybridizable Discontinuous Galerkin method for nonlinear convection–diffusion equations*, J. Comput. Phys., 228 (2009), pp. 8841–8855.

[135] K. S. Oh, *A study of low order spherical harmonic closures for rapid transients in radiation transport*, PhD thesis, University of Michigan, 2009.

[136] K. S. Oh and J. P. Holloway, *A quasi-static closure for 3rd order spherical harmonics time- dependent radiation transport in 2-D*, in Proceedings of the 2009 International Conference on Mathematics and Computational Methods and Reactor Physics, American Nuclear Society, May 2008.

[137] E. Olbrant and M. Frank, *Generalized Fokker-Planck approximations of particle transport with highly forward-peaked scattering*, Comp. Math. Methods in Medicine, 11 (2010), pp. 313–339.

[138] E. Olbrant, C. D. Hauck, and M. Frank, *A realizability-preserving discontinuous Galerkin method for the M_1 model of radiative transfer*, submitted to J. Comp Phys., (2012).

[139] E. Olbrant, E. W. Larsen, M. Frank, and B. Seibold, *Asymptotic derivation and numerical investigation of time-dependent simplified P_N equations*, submitted to J. Comput. Phys., (2012).

[140] G. L. Olson, *Second-order time evolution of P_N equations for radiation transport*, J. Comput. Phys., 228 (2009), pp. 3072–3083.

[141] A. Ore, *Entropy of radiation*, Phys. Rev., 98 (1955), p. 887.

[142] S. Osher, *Riemann solvers, the entropy condition, and difference*, SIAM J. Numer. Anal., 21 (1984), pp. 217–235.

[143] B. Perthame, *Boltzmann type schemes for gas dynamics and the entropy property*, SIAM J. on Numer. Anal., 27 (1990), pp. 1405–1421.

[144] ——, *Second-order Boltzmann schemes for compressible euler equations in one and two space dimensions*, SIAM J. Numer. Anal., 29 (1992), pp. 1–19.

[145] G. C. Pomraning, *Radiation Hydrodynamics*, Pergamon Press, New York, 1973.

[146] ——, *The non-equilibrium Marshak wave problem*, J. Quant. Spectrosc. Radiat. Transfer, 21 (1979), pp. 249–261.

[147] ——, *The Fokker-Planck operator as an asymptotic limit*, Math. Models Methods Appl. Sci., 1 (1992), pp. 21–36.

[148] ——, *Asymptotic and variational derivations of the simplified P_N equations*, Ann. Nucl. Energy, 20 (1993), pp. 623–637.

[149] B. Povh, K. Rith, C. Scholz, and F. Zetsche, *Teilchen und Kerne, Eine Einführung in die physikalischen Konzepte*, Springer, 7 ed., 2006.

[150] J. Qiu and C.-W. Shu, *A comparison of troubled-cell indicators for Runge-Kutta discontinuous Galerkin methods using weighted essentially nonoscillatory limiters*, SIAM J. Sci. Comput., 27 (2005), pp. 995–1013.

[151] S. L. Rosa, G. Mascali, and V. Romano, *Exact maximum entropy closure of the hydrodynamical model for Si semiconductors: The 8-moment case*, SIAM Journal on Applied Mathematics, 70 (2009), pp. 710–734.

[152] P. Rosen, *Entropy of radiation*, Phys. Rev., 96 (1954), p. 555.

[153] K. Salari and P. Knupp, *Code verification by the method of manufactured solutions*, Tech. Rep. SAND2000-1444, Sandia National Laboratories, 2000.

[154] F. Salvat and J. M. Fernández-Varea, *Overview of physical interaction models for photon and electron transport used in Monte Carlo codes*, Metrologia, 46 (2009), pp. 112–138.

[155] F. Salvat, J. M. Fernández-Varea, and J. Sempau, *PENELOPE-20011: A Code System for Monte Carlo Simulation of Electron and Photon Transport*, OECD, 2011.

[156] M. Schäfer, M. Frank, and C. D. Levermore, *Diffusive corrections to P_N approximations*, Multiscale Model. Simul., 9 (2011), pp. 1–28.

[157] E. Schneider, M. Seaïd, J. Janicka, and A. Klar, *Validation of simplified P_N models for radiative transfer in combustion systems*, Commun. Numer. Meth. Engng., 24 (2008), pp. 85–96.

[158] J. Schneider, *Entropic approximation in kinetic theory*, Math. Model. Numer. Anal., 38 (2004), pp. 541–561.

[159] M. Seaïd, M. Frank, A. Klar, R. Pinnau, and G. Thömmes, *Efficient numerical methods for radiation in gas turbines*, J. Comp. Applied Math., 170 (2004), pp. 217–239.

[160] B. Seibold and M. Frank, *Optimal prediction for moment models: Crescendo diffusion and reordered equations*, Contin. Mech. Thermodyn., 21 (2009), pp. 511–527.

[161] ——, *`StaRMAp` – A Staggered grid Radiation Moment Approximation solver*, in preparation, (2012).

[162] J. Sempau, J. M. Fernández-Varea, E. Acosta, and F. Salvat, *Experimental benchmarks of the monte carlo code penelope*, Nucl. Instrum. Methods B, 207 (2003), pp. 107–123.

[163] D. R. Shipley, H. Palmans, A. Kacperek, and C. Baker, *Geant4 simulation of an ocular proton beam and benchmark against other Monte Carlo codes*, Proceedings of the Monte Carlo Method: Versatility Unbounded in a Dynamic Computing World, Monte Carlo Topical Meeting, Chattanooga, USA, (17 - 21 April 2005).

[164] C.-W. Shu, *TVB uniformly high order schemes for conservation laws*, Math. Comp., 49 (1987), pp. 105–121.

[165] C.-W. Shu and S. Osher, *Effcient implementation of essentially non-oscillatory shock capturing schemes*, J. Comp. Phys., 77 (1988), pp. 439–471.

[166] ——, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1989), pp. 439–471.

[167] Y. Shu and C.-W. Shu, *Local Discontinuous Galerkin methods for high-order time-dependent partial differential equations*, Commun. Comput. Phys., 7 (2010), pp. 1–46.

[168] J. M. Smit, J. Cernohorsky, and C.-P. Dullemond, *Hyperbolicity and critical points in two-moment approximate radiative transfer.*, Astrophys. J., 325 (1997), pp. 203–211.

[169] J. M. Smit, L. J. van den Horn, and S. A. Bludman, *Closure in flux-limited neutrino diffusion and two-moment transport*, Astron. Astrophys., 356 (2000), pp. 559–569.

[170] R. Sternheimer, *Density effect for the ionization loss of charged particles*, Phys. Rev., 145 (1965), pp. 247–250.

[171] H. Struchtrup, *Kinetic schemes and boundary conditions for moment equations*, Z. Angew. Math. Phys., 51 (2000), pp. 346–365.

[172] ——, *Linear kinetic heat transfer: Moment equations, boundary conditions, and Knudsen layers*, Physica A Statistical Mechanics and its Applications, 387 (2008), pp. 1750–1766.

[173] B. Su and G. L. Olson, *An analytical benchmark for non-equilibrium radiative transfer in an isotropically scattering medium*, Ann. Nucl. Energy, 24 (1997), pp. 1035–1055.

[174] G. Thomas and K. Stamnes, *Radiative Transfer in the Atmosphere and Ocean*, Cambridge University Press, Cambridge, 1999.

[175] D. I. Tomasevic and E. W. Larsen, *The simplified P_2 approximation*, Nucl. Sci. Eng., 122 (1996), pp. 309–325.

[176] R. Turpault, *A consistent multigroup model for radiative transfer and its underlying mean opacities*, J. Quant. Spectrosc. Radiat. Transfer, 94 (2005), pp. 357–371.

[177] J. Venselaar, H. Welleweerd, and B. Mijnheer, *Tolerances for the accuracy of photon beam dose calculations of treatment planning systems*, Radiother. Oncol., 60 (2001), pp. 191–201.

[178] F. Verhaegen and J. Seuntjens, *Monte carlo modelling of external radiotherpay photon beams*, Phys. Med. Biol., 48 (2003), pp. R107–R164.

[179] D. Wright, M. Frank, and A. Klar, *The minimum entropy approximation to the radiative transfer equation*, Proc. Symp. Appl. Math, 29 (2009), pp. 987–996.

[180] W. A. Wulf and S. A. McKee, *Hitting the memory wall: implications of the obvious*, SIGARCH Comput. Archit. News, 23 (1995), pp. 20–24.

[181] P. Zhang, S. Wong, and C.-W. Shu, *A weighted essentially non-oscillatory numerical scheme for a multi-class traffic flow model on an inhomogeneous highway*, J. Comput. Phys., 212 (2006), pp. 739–756.

[182] X. Zhang and C.-W. Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, J. Comput. Phys., 229 (2010), pp. 3091–3120.

[183] ——, *On positivity-preserving high order discontinuous Galerkin schemes for compressible euler equations on rectangular meshes*, J. Comput. Phys., 229 (2010), pp. 8918–8934.

[184] H. Zheng and W. Han, *On simplified spherical harmonics equations for the radiative transfer equation*, J. Math. Chem., 49 (2011), pp. 1785–1797.

[185] J. F. Ziegler, *The stopping of energetic light ions in elemental matter*, J. Appl. Phys. / Rev. Appl. Phys., 85 (1999), pp. 1249–1272.

LEBENSLAUF

Persönliche Daten

Name:	Olbrant
Vorname:	Edgar
Geburtstag:	14.04.1984
Geburtsort:	Maikain (Kasachstan)
Staatsangehörigkeit:	deutsch

Qualifikationen

2003	Abitur
2004 – 2008	Studium der Technomathematik *Anwendungsfach Physik*
Abschluss:	Diplom (Studienprogramm Mathematics International)
2008 – 2009	Studium für das Lehramt an Gymnasien *Fächer Mathematik und Physik*
Abschluss:	Erstes Staatsexamen
ab 2008	Beginn der Promotion in Mathematik