Models, Numerical Methods, and Uncertainty Quantification for Radiation Therapy

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

vorgelegt von
M. Sc. Kerstin Küpper
aus Herdecke

Berichter: Univ.-Prof. Dr. rer. nat. Martin Frank
Univ.-Prof. Dr. rer. nat. Michael Herty
Prof. Ph. D. Cory D. Hauck

Tag der mündlichen Prüfung: 11.11.2016

Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.
Acknowledgements

First of all I would like to thank my Ph. D. supervisor Martin Frank for the continuous support of my scientific career during the last years. His ability to explain things clearly and simply, his motivation, his new ideas, and his sound advice helped me in all time of research and writing of papers and this thesis. I am grateful that he also supported my personal development and introduced me to many people in the scientific community.

Besides my advisor, I would like to thank Michael Herty and Cory Hauck for reviewing this thesis.

I also wish to thank Cory Hauck for the fruitful cooperation in the past years and for providing me an opportunity to join his team for a research stay in Oak Ridge, Tennessee. It was extremely instructive and a great experience to work with him.

I thank Shi Jin for sharing his thoughts and ideas on numerical methods based on his expertise on asymptotic preserving schemes. I also want to acknowledge him for the support of the research network KI-Net, which supported my research stay in Oak Ridge and allowed me to participate in inspiring conferences.

Many thanks go to my fellows at the MathCCES who contributed to a pleasant working atmosphere. In particular, I would like to thank Graham Alldredge, Thomas Camminady, Philipp Otte, Teddy Pichard, Pascal Richter, Birte Schmidtmann, and Torsten Trimborn for their support, helpful discussions, and all the good times we shared.

Lastly, I would like to thank my friends, above all Christian Lax, and my family for their indulgence and patience in stressful times and for creating a sense of trust and well-being.
Abstract

The evolution of particles traveling through a background medium can be described by the linear Boltzmann transport equation. For this equation, we study several numerical methods with the focus on radiation therapy. The main challenges to solve the linear transport equation numerically are the large dimensionality of the problem, the integro-differential form of the equation, and the highly varying material parameters in the scattering operator. In this work, we consider four main topics. First, we introduce a numerical method for electron transport based on the continuous slowing down approximation. We adapt a second order staggered grid method for spherical harmonic moment equations of radiative transfer to our application. Secondly, we investigate the filtered spherical harmonic (FP$_N$) equations, which have the potential to reduce the Gibbs phenomena around discontinuities. In particular, we prove global $L^2$ convergence properties of the FP$_N$ equations and show how the convergence rates depend on the smoothness of the solution and the order of the filter. Thirdly, we extend a one-dimensional staggered grid method to two spatial dimensions. Due to the staggering, the method requires less unknowns than one might have expected and leads to a compact stencil in the discrete diffusion limit. In addition, we rigorously analyze the underlying one-dimensional scheme and show that the scheme is asymptotic preserving. Lastly, we derive a two-level sampling strategy based on a model hierarchy to improve a stochastic collocation sparse grid method for uncertainty quantification. The method combines a reduced model for the hyperbolic relaxation system with a correction in a two-level framework to balance the different error terms and to minimize the computational cost. For all major topics, we present several numerical experiments to confirm the theoretical predictions.
# Contents

**Preface** ix  
**List of Figures** xiii  
**List of Tables** xxi  

## 1 Electron Transport 1  
1.1 Introduction .................................................. 1  
1.1.1 Radiation Transport ........................................ 5  
1.2 The CSD Approximation ........................................ 6  
1.2.1 Simplifications ............................................. 8  
1.2.2 Transformations ............................................ 9  
1.3 Numerical Method ............................................. 11  
1.3.1 Spherical Harmonic ($P_N$) Equations ..................... 11  
1.3.2 StaRMAP in Three Spatial Dimensions .................... 14  
1.3.3 StaRMAP with CSD ........................................... 17  
1.3.3.1 Implementation of the Dose ......................... 18  
1.3.3.2 Implementation of Physical Parameters ............... 19  
1.4 Numerical Results ........................................... 19  
1.4.1 Phantoms .................................................. 21  
1.4.1.1 Water Phantom ......................................... 21  
1.4.1.2 Water Phantom with Air Cavity ....................... 22  
1.4.2 CT Scans .................................................. 23  
1.4.2.1 Two-Dimensional Test Case ......................... 24  
1.4.2.2 Three-Dimensional Test Case ....................... 26
CONENTS

2 Convergence of $\text{FP}_N$ Equations
  2.1 Introduction ........................................... 31
    2.1.1 Filtered Spherical Harmonic (FP$_N$) Equations .... 33
  2.2 Error Estimate ....................................... 36
    2.2.1 Preliminaries ..................................... 36
    2.2.2 Main Result ....................................... 39
    2.2.3 A Sharper Estimate ............................... 44
  2.3 Numerical Results .................................... 45
    2.3.1 Comparison of FP$_N$ Methods .................... 46
      2.3.1.1 Line Source Test .......................... 46
      2.3.1.2 Manufactured Solutions ................... 47
    2.3.2 Numerical Rates of Convergence .................. 49
      2.3.2.1 Gaussian Test ................................ 54
      2.3.2.2 Lattice Test ................................ 54
      2.3.2.3 Hemisphere Test ............................. 59
    2.3.3 FP$_N$ for Electron Transport .................... 60
      2.3.3.1 Water Phantom ............................... 60
      2.3.3.2 CT Scan ..................................... 62

3 AP Staggered Grid Method ............................... 65
  3.1 Introduction .......................................... 65
    3.1.1 The Diffusion Limit ............................. 68
  3.2 The Numerical Method ................................ 71
    3.2.1 Spatial Discretization .......................... 72
    3.2.2 Time Discretization ............................. 75
  3.3 The AP Property ...................................... 77
    3.3.1 The Discrete Diffusion Limit ................... 77
    3.3.2 Stability ........................................ 78
  3.4 Numerical Results .................................... 83
    3.4.1 Convergence Order ................................ 84
      3.4.1.1 Method of Manufactured Solutions ........... 85
      3.4.1.2 Gauss Test .................................. 85
    3.4.2 Variable Scattering ............................. 87
Preface

The evolution of particles traveling through a background medium can be described by the linear Boltzmann transport equation

\[ \partial_t \psi(t, x, \Omega) + \Omega \cdot \nabla_x \psi(t, x, \Omega) = \mathcal{C}(\psi)(t, x, \Omega), \]  

where \( \psi \) is the density of particles at time \( t \in [0, \infty) \), position \( x \in \mathbb{R}^3 \), and moving in direction \( \Omega \in S^2 \). While the left-hand side describes particles moving in straight lines, the collision operator \( \mathcal{C} \) incorporates the interactions with the background medium. The linear transport equation can be used for the simulation of radiation therapy, which is one of the main tools in the therapy of cancer.

The main challenges to solve the linear transport equation numerically are the large dimensionality of the problem (one in time, three in space, two in angle), the integro-differential form of the equation (\( \mathcal{C} \) is a linear integral operator), and the highly varying material parameters in the scattering operator. In view of these challenges, we study several numerical methods for the linear transport equation with the focus on radiation therapy in this work.

In the first chapter, we introduce a numerical method for electron transport. The method relies on the continuous slowing down equation, which is a simplified model for the physical interactions. For this model, we use the spherical harmonic (\( P_N \)) approximation to discretize the angular variable. Since the spherical harmonic basis functions diagonalize the scattering operator, this leads to an efficient implementation of the equation. In addition, we use physical parameters extracted from the ICRU database [101] to model electrons traveling through tissue with high energy. Finally, we compare the method with a benchmark (Monte Carlo) code for several numerical radiation therapy problems.
As a spectral method, the above method suffers from the Gibbs phenomena around discontinuities, which can lead to spurious oscillations and negative solutions. However, it has been shown in several numerical experiments that the oscillations can be dampened by the use of a filter function [46, 98]. The filtered spherical harmonic (FP$_N$) equations capture the rotational invariance and the simplicity of the method, but a detailed analysis is needed to choose a suitable filter. In the second chapter, we analyze the convergence of the FP$_N$ equations and prove global $L^2$ convergence properties. In particular, we show explicitly how the convergence rate of the spectral method to the solution of the transport equation depends on the smoothness of the solution and on the order of the filter.

Besides the structure of the equation and the large dimensionality of the problem, the highly varying material parameters are a major challenge for numerical methods, as mentioned above. In the context of radiation therapy, the material parameters correlate with the density inside the human body, which has a wide range varying from a kinetic to a diffusive regime. To capture the correct physical behavior in all regimes is the objective of asymptotic preserving (AP) schemes. In Chapter 3, we propose a two-dimensional AP scheme for linear transport equations with diffusive scalings based on [60]. We propose a staggering in two dimensions that requires fewer unknowns than one could have naively expected. In addition, we rigorously analyze the one-dimensional scheme of Jin, Pareschi, and Toscani [60] and show that the scheme is asymptotic preserving.

Finally, we investigate another issue in radiation therapy, which is related to uncertainty quantification. In radiation therapy, the simulation relies on measurement data from a computed tomography scan of the patient and on material parameters, which introduce uncertainties into the model. In Chapter 4, we develop a new sampling strategy for hyperbolic relaxation systems to quantify the effect of uncertain input parameters on the solution. Our method is a two-level sparse grid stochastic collocation method, but instead of using different discretizations of the underlying system, we use a model hierarchy. Similar to Chapter 3, we consider a reduced model for the hyperbolic relaxation system, which is a limiting equation. Usually a reduced model has good accuracy in a collisional regime, while it loses accuracy if we approach the rarefied regime. We show that
in a transition regime the combination of the reduced model and a correction in
a two-level framework can be advantageous.

Part of this work relies on papers, which were published with coauthors. The
content of Chapter 2 is based on [34], which was published with Martin Frank
and Cory Hauck in Communications in Mathematical Sciences. While a sketch of
the proof of the main Theorem 2.8 was developed by my coauthors, I was working
out the details of the proof and was able to sharpen the results under additional
mild conditions (cf. Theorem 2.10). Moreover, I designed and also preformed all
numerical test cases presented in section 2.3. To a large extend Chapter 3 relies
on [63], which was published with Martin Frank and Shi Jin in SIAM Journal on
Numerical Analysis. The idea of extending the numerical scheme, developed by
Jin, Pareschi, and Toscani in [60], to two dimensions and a basic sketch was given
by my coauthors. I realized the idea and worked out a stability proof to show
the asymptotic preserving property. Furthermore, I implemented the numerical
scheme and designed test cases to confirm the predicted results. I also came up
with possible extensions of the method, which are presented in section 3.5.
List of Figures

1.1 Segment of three-dimensional staggered grids: grids with filled markers are odd grids; grids with empty markers are even grids. 15

1.2 Water phantom with initial beam entering from the left with energy (a) $E_{\text{max}} = 5$ MeV or (b) $E_{\text{max}} = 10$ MeV: Normalized dose and $\gamma$-index on a selection of the computational domain. 22

1.3 Water phantom with air cavity and initial beam entering from the left with $E_{\text{max}} = 5$ MeV: Normalized dose and $\gamma$-index shown on the interval $[0, 3.5]$. 23

1.4 Two-dimensional segmented CT scan: The locations and the directions of the two initial beams are marked in black. 25

1.5 Two-dimensional CT scan with $E_{\text{max}} = 21$ MeV: (a) and (b) show isodose curves computed with $P_{17}$ and MC; (c) and (d) show the $\gamma$-index limited to $[0, 2]$; regions with $\gamma(x) > 1$ are marked by red borders. 26

1.6 Geometry of the three-dimensional test case: the CTV and the PTV are marked in green; the red box indicates the computational domain; the arrows indicate the position and the direction of the initial condition. 27

1.7 Three-dimensional CT scan with $E_{\text{max}} = 21$ MeV: Cut along $z = 6.45$: (a) and (b) show isodose curves computed with $P_{17}$ and MC; (c) shows the $\gamma$-index limited to $[0, 2]$; regions with $\gamma(x) > 1$ are marked by red borders. 28
LIST OF FIGURES

1.8 Three-dimensional CT scan with $E_{\text{max}} = 21\, \text{MeV}$: the first row is a view from the back right; the second row is a top view; and the third row is an enlarged selection of the top view. (a), (b), (d), and (e) show 5%, 10%, 25%, 50%, 70%, 80%, 90%, 95%, 98% isodose surfaces. (g) and (h) show 50%, 70%, 80%, 90%, 95%, 98% isodose surfaces. (c), (f), and (i) show $\gamma$-values greater than 1. The data is cut in the x-y plane. 

2.1 Exponential filter of order $\alpha = 2, 4, 8, 16$. 

2.2 Linesource test. (a)-(c) Scalar flux $\phi = \langle \psi \rangle$ at $t = 0.5$: for $P_{99}$ (a); for $P_{13}$ (b); and FP$_{13}$ in the “modified” case (c). Negative values are shown in white. (d) Absolute error of the scalar flux along $y = 0$ computed with $P_{13}$ and FP$_{13}$ with different filters. 

2.3 Manufactured solution test: $L^1$, $L^2$, and $L^\infty$ error (in space and angle) for different resolutions and implementations of the filter method: (a) “modified”, (b) “full step”, and (c) “substep”. 

2.4 Gauss test computed with $P129$: (a) Log-log plot of the sequences $B_\ell$ and $D_\ell$ vs. the order $\ell$; (b) Log-log plot of the error terms $E_N$ as a function of the moment order approximation $N$. 

2.5 Lattice test: (a) Material coefficients: isotropic source (white square) $S = 1$; purely scattering $\sigma_t = \sigma_s = 1$ (orange and white squares); purely absorbing $\sigma_t = \sigma_a = 10$ (black squares). (b) Scalar flux $\phi = \langle \psi \rangle$ at $t = 2.8$ for $P_{129}$, computed with $500 \times 500$ grid points. The values are plotted in a logarithmic scale and limited to seven orders of magnitude. 

2.6 Lattice test: Log-log plot of the sequences $B_\ell$ and $D_\ell$ vs. the order $\ell$. The dashed lines indicate until which point the relative difference between the sequences computed with $P_{128}$ and $P_{129}$ differs by no more than 3%. 

2.7 Lattice test: Log-log plot of the error terms $E_N$ and $R_N$ as a function of the moment order approximation $N$. 

xv
2.8 Hemisphere test: Log-log plot of the sequences $B_\ell$ and $D_\ell$ against the order $\ell$. The dashed lines indicate until which point the sequences computed with P$_{98}$ and P$_{99}$ coincide up to machine precision. .................................................. 60

2.9 Hemisphere test: Log-log plot of the error terms $E_N$ and $R_N$ as a function of the moment order approximation $N$. ................................. 61

2.10 Water phantom with initial energy entering from the left with energy $E_{\text{max}} = 5\text{MeV}$. Solution computed with MC (as a reference), P$_5$, and FP$_5$ with different filter orders $\alpha = 2, 4$ and filter effective opacities $f_{\text{eff}} = 2, 4$. .................................................. 62

2.11 Two dimensional CT scan with $E_{\text{max}} = 21\text{MeV}$: (a) and (b) show isodose curves computed with P$_{11}$ and FP$_{11}$ with negative values (less than $-3\%$ of the maximal dose) marked by black borders; (c) and (d) show the $\gamma$-index with respect to the MC solution ($\gamma(x) > 1$ is marked by red borders). .................................................. 63

3.1 Staggered grids and control volumes: red circles (vertices and cell centers), $r^{(1)}$, $r^{(2)}$, $\rho$, $q$; blue diamonds (face centers), $j^{(1)}$, $j^{(2)}$. ........................................ 72

3.2 Convergence order (MMS): $\ell^2$-error as a function of the spatial resolution. Hyperbolic (filled markers) or parabolic (empty markers) CFL condition. ........................................ 86

3.3 Convergence order (Gauss test): $\ell^2$-error as a function of the spatial resolution. Hyperbolic (filled markers) or parabolic (empty markers) CFL condition. ........................................ 87

3.4 Variable scattering: Density $\rho$ at different end times $\ell = 0.1, 0.5, 1.0$, computed on a $32 \times 32$ grid (first column) or a $512 \times 512$ grid (second column). ........................................ 89

3.5 Variable scattering: Density $\rho$ and error at different end times $\ell = 0.1, 0.5, 1.0$ on a cut along $y = 0$, computed with $n \times n$ grid points. ........................................ 90
# LIST OF FIGURES

3.6 Two material test: (a) geometry – source (orange), purely scattering $\sigma_t = \sigma_s = 1$ (white and orange), purely absorbing $\sigma_t = \sigma_a = 100$ (black); (b) and (c) density $\rho$ at $t = 1.7$, computed on a $64 \times 64$ grid or $512 \times 512$ grid; (d) absolute error in the coarse grid compared to the fine grid. Logarithmic scaling, values are limited to seven orders of magnitude. 92

3.7 Stability: Density on a cut along $y = 0$, computed on a $300 \times 300$ grid up to time $t = 0.36$ using different values of $\phi$. 93

3.8 Black solid line denotes the boundary of the domain $\partial X$; filled markers denote inner points; empty markers denote boundary and ghost points. 94

4.1 Two-dimensional sparse grid and tensor product grid constructed with the one-dimensional Gauss-Patterson formulas with maximal seven points. 102

4.2 Cost ratio as a function of the convergence order $\mu$ and the single computational cost ratio $\kappa$ (values are limited to $[0, 1]$). Results are computed for the reduced models of order $p = 1, 2$ with different relaxation parameters $\varepsilon = 0.1$ and $\varepsilon = 0.5$. 108

4.3 DVM in 1D with 10 velocities: Error in the full model, in the reduced model of order $p = 1, 2$, and in the difference between the two as a function of the number of points of the quadrature formula for $\varepsilon = 0.1$. 117

4.4 DVM in 1D with ten velocities: Reduced models of order $p = 1, 2$ with relaxation parameter $\varepsilon = 0.1$. Cost ratio and error as a function of the error tolerance Error_{tol}. The maximal accuracy of the reduced model $\alpha_R$, the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical and the estimated approach. 119
LIST OF FIGURES

4.5  DVM in 1D with ten velocities: First-order reduced model \((p = 1)\) with various relaxation parameters \(\varepsilon = 0.5, 0.1, 0.02\). Cost ratio and error as a function of the error tolerance \(\text{Error}_{\text{tol}}\). The maximal accuracy of the reduced model \(\alpha_R\), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical, the iterative, and the greedy approach. ............................................................... 120

4.6  DVM in 1D with ten velocities: Second-order reduced model \((p = 2)\) with various relaxation parameters \(\varepsilon = 0.5, 0.1, 0.02\). Cost ratio and error as a function of the error tolerance \(\text{Error}_{\text{tol}}\). The maximal accuracy of the reduced model \(\alpha_R\), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical, the iterative, and the greedy approach. ............................................................... 121

4.7  DVM in 2D with eight velocities: Error in the full model, in the reduced model of order \(p = 1, 2\), and in the difference between the two as a function of the number of points of the quadrature formula for \(\varepsilon = 0.1\). ............................................................... 123

4.8  DVM in 2D with eight velocities: Reduced models of order \(p = 1, 2\) with relaxation parameter \(\varepsilon = 0.1\). Cost ratio and error as a function of the error tolerance \(\text{Error}_{\text{tol}}\). The maximal accuracy of the reduced model \(\alpha_R\), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical and the estimated approach. .......... 125

4.9  DVM in 2D with eight velocities: First-order reduced model \((p = 1)\) with various relaxation parameters \(\varepsilon = 0.5, 0.1, 0.02\). Cost ratio and error as a function of the error tolerance \(\text{Error}_{\text{tol}}\). The maximal accuracy of the reduced model \(\alpha_R\), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical, the iterative, and the greedy approach. ............................................................... 126
LIST OF FIGURES

4.10 DVM in 2D with eight velocities: Second-order reduced model
\((p = 2)\) with various relaxation parameters \(\varepsilon = 0.5, 0.1, 0.02\). Cost ratio and error as a function of the error tolerance \(\text{Error}_{\text{tol}}\). The maximal accuracy of the reduced model \(\alpha_R\), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical, the iterative, and the greedy approach. ................................. 127

4.11 Hemisphere Test: (a) samples of the scattering kernel; (b) moments of the scattering kernel. The mean is shown as a black solid line and 20 random samples are shown as colored dotted lines. Black dashed lines indicate the position of the interpolation points and the order of the moments. ................................. 130

4.12 Hemisphere Test: (a) expected value of the zeroth order moment of the \(P_{11}\) solution computed with a high level \((\ell = 11)\) SC method; (b) \(L^2\) error in the full model \((P_{11})\), the reduced model \((P_5)\), and the difference between the two as a function of the number of points of the quadrature. ................................. 131

4.13 Hemisphere Test with \(\varepsilon = 0.1\): (a) cost ratio between the two-level sampling approach (using \(P_5\) and \(P_{11}\)) and the standard SC approach (using \(P_{11}\)); (b) \(L^2\) error in the two-level sampling approach compared to the high level standard SC method. The maximal accuracy of the reduced model \(\alpha_R\), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. ................................. 132

4.14 Hemisphere Test with \(\varepsilon = 0.1\): (a) cost ratio between the two-level sampling approach (using \(P_3\) and \(P_{11}\)) and the standard SC approach (using \(P_{11}\)); (b) \(L^2\) error in the two-level sampling approach compared a high level standard SC method. The maximal accuracy of the reduced model \(\alpha_R\), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. ................................. 132
4.15 Water phantom: (a) 10 random samples of the uncertain input parameter $\sigma(\xi, E)$; (b) mean and variance of the depth dose profile computed with SC ($P_{21}$ and level 11 quadrature formula). The $\gamma$-index is computed for $E \pm \text{Var}$ compared to $E$ on a 3%, 3 mm level. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 134

4.16 Water phantom: $L^2$ error (4.67) in the full model ($P_{21}$), the reduced model ($P_{11}$), and the difference between the two as a function of the number of points of the quadrature for different spatial resolutions: (a) $n = 150$ grid points; (b) $n = 300$ grid points. . . . . . . . . . . . . . . 135

4.17 Water phantom: (a) cost ratio between the two-level sampling approach (using $P_{11}$ and $P_{21}$) and the standard SC approach (using $P_{21}$); (b) $L^2$ error (4.67) in the two-level sampling approach compared to a high level standard SC method. The maximal accuracy of the reduced model $\alpha_R$, the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. . . . . . . . . 136
LIST OF FIGURES

xx
List of Tables

2.1 Manufactured solution test: Approximate rate of convergence in the $L^2$ norm for different implementations of the FP$_N$ equations: “modified”, “full step”, and “substep”. The term $\tilde{E}^n_{n_2}$ is the convergence rate when going from $n_1$ to $n_2$ grid points. The degree defines the considered solution components. 51

2.2 Approximate order of convergence for different filter orders and test cases. Computed with different spatial resolutions and implementations of the filter: (a) spatial resolution $150 \times 150$ (Gaussian, hemisphere), $250 \times 250$ (lattice); (b) and (d) spatial resolution $300 \times 300$ (Gaussian, hemisphere), $500 \times 500$ (lattice); (a) and (c) filter implementation “modified”; (b) filter implementation “substep”. Filter order of $\infty$ means that no filter is applied. 53

2.3 Gauss test: Filter order of $\infty$ means that no filter is applied. The term $E^N_{N_2}$ is the convergence rate when going from $N_1$ to $N_2$. 55

2.4 Lattice test: Approximate decay rates of the sequence of the moments $B_\ell$ (and the moments of the differentials $D_\ell$). (a) Even order moments $N_2 = 2^{k+1}$ vs. $N_1 = 2^k$ and (b) odd order moments $N_2 = 2^{k+1} + 1$ vs. $N_1 = 2^k + 1$ with $k = 1, \ldots, 5$ (Computed with $P_{129}$). 58

2.5 Lattice Test: Filter order of $\infty$ means that no filter is applied. The term $E^N_{N_2}$ is the convergence rate when going from $N_1$ to $N_2$. 58
2.6 Hemisphere test: Approximated decay rates of the sequence of the moments $B_\ell$ (and the moments of the differentials $D_\ell$). (a) Even order moments $N_2 = 2^{k+1}$ vs. $N_1 = 2^k$ and (b) odd order moments $N_2 = 2^{k+1} + 1$ vs. $N_1 = 2^k + 1$ with $k = 1, \ldots, 5$ (Computed with $P_{99}$). 61

2.7 Hemisphere test: Filter order of $\infty$ means that no filter is applied. The term $E^{N_2}_{N_1}$ is the convergence rate when going from $N_1$ to $N_2$. 61

3.1 Convergence order (MMS): The term $E^{N_2}_{N_1}$ is the convergence rate when going from $N_1 \times N_1$ to $N_2 \times N_2$ grid points for a fixed mean free path $\varepsilon_k = 10^{-k}, k = 0, 1, 2, 3$. The dashed line indicates the switch from the hyperbolic to the parabolic CFL condition. 86

3.2 Convergence order (Gauss test): The term $E^{N_2}_{N_1}$ is the convergence rate when going from $N_1 \times N_1$ to $N_2 \times N_2$ grid points for a fixed mean free path $\varepsilon_k = 10^{-k}, k = 0, 1, 2$. The dashed line indicates the switch from the hyperbolic to the parabolic CFL condition. 87
Chapter 1

Electron Transport

1.1 Introduction

The use of ionization radiation is one of the main tools in the therapy of cancer. In practice, photons are commonly used for the radiation, but radiation with electrons, protons, and heavily charged ions gain in importance. In all cases, the aim is to shrink the tumor by delivering a certain dose to it. However, since all cells are damaged by radiation, a careful treatment planning is necessary.

The treatment planning starts with a simulation, which relies on a computed tomography (CT) scan of the patient. In particular, this includes the computation of the dose deposition in the body. We focus on the transport of electrons in the human body. Electrons and other charged particles have a different behavior than uncharged particles such as photons. But photons trigger electrons as secondary particles, so that a model for electron transport is also needed for the simulation of high-energy photon radiation.

When electrons are traveling through matter, they can interact in various ways. The most important interactions in the high energy range (1 – 30 MeV), which is relevant for radiation therapy, are elastic scattering, soft inelastic scattering, hard inelastic scattering, and bremsstrahlung [92].

A common assumption to simplify the scattering process is to assume that electrons lose their energy continuously, when they are traveling through the domain. It is known as the continuous-slowing-down (CSD) approximation [67]. This is a nonphysical assumption as electrons deflect and lose energy at discrete
1. ELECTRON TRANSPORT

collisions. However, it is a good approximation for multiple collisions with small energy deviation and frequently used for electrons with energies above 1 keV [36]. For instance, the Monte-Carlo code Penelope [100] uses this approximation for soft inelastic collisions. The decay is then described by the so called stopping power. If in addition secondary electrons and bremsstrahlung are neglected, electron transport can be described by the Boltzmann continuous-slowing-down (CSD) equation [67]

\[- \partial_E \left( S(E, x) \psi(E, x, \Omega) \right) + \Omega \cdot \nabla_x \psi(E, x, \Omega) + \Sigma_t(E, x) \psi(E, x, \Omega) \]

\[= \int_{S^2} \Sigma_s(E, x, \Omega \cdot \Omega') \psi(E, x, \Omega') \, d\Omega' + q(E, x, \Omega) \quad (1.1a)\]

with initial condition

\[\psi(E_{\text{max}}, x, \Omega) = \psi_{\text{max}}(x, \Omega) \quad (1.1b)\]

and inflow boundary conditions. Here, the particle density \(\psi(E, x, \Omega)\) describes electrons with a certain energy \(E\) at a position \(x\) and traveling into a direction \(\Omega \in S^2\). The electrons interact with a background medium, which is described by the cross sections \(\Sigma_t\) and \(\Sigma_s\) and the stopping power \(S\). The initial condition defines the incoming radiation at energy \(E_{\text{max}}\), which is assumed to be the maximal energy of the system. Further, a source inside the domain is defined by \(q\). The CSD equation is an evolution equation of the particle density. Under suitable conditions on the input parameters, such as cross sections, stopping power, and source, the CSD equation has a unique solution, which is \(C^1\) in energy and \(L^2\) in space and angle [113]. The solution of the CSD equation also enables the computation of the absorbed dose

\[D(x) = \frac{1}{\rho(x)} \int_0^{\infty} \int_{S^2} S(E, x) \psi(E, x, \Omega) \, d\Omega dE , \quad (1.2)\]

which is of main interest in applications.

To solve the Boltzmann CSD equation is a challenging problem, since it is high dimensional (one dimension in energy, three dimensions in space, and two dimensions in angle) and the physical parameters are highly varying. However, under certain assumptions, the CSD equation can be transformed into a radiation
transport equation with a space dependent flux term (see Lemma 1.4). The radiation transport equation is widely studied and there exists a variety of solution methods. As an overview, we briefly present a selection of the methods, which is by far not complete (see for example [33, 70, 97] and references therein).

**Monte Carlo Methods** use repeated random sampling of the pathways of particles to compute a density. Since the methods rely on the law of large numbers, many individual simulations are necessary to obtain reliable results and the convergence rate is typically very slow. Therefore, Monte Carlo methods are computationally intense compared to other methods. However, the convergence rate is independent of the dimension and they can achieve a very high accuracy. In particular, for photon and electron transport, there are code systems like Geant4 [3, 4] or Penelope [100], which have a very good agreement with physical experiments. Moreover, there are multiple approaches to speed up the computations. For instance, Quasi-Monte Carlo methods [18, 86, 107], which use a careful selection of the samples, Multilevel Monte Carlo methods [38, 47, 48], which balance statistical and deterministic errors, and non-analog Monte Carlo methods [23], which use weighting to reduce the variance.

**Entropy-based Methods** are moment methods, which rely on a minimum entropy principle [30, 80]. The radiation transport equation is transformed into a moment system, which is an unclosed system of equations. The system is closed by solving a constrained, convex optimization problem coming from the entropy principle. Usually, this leads to a complicated relationship between the coefficients and the moments of the expansion [1, 2]. Therefore, the implementation of high-order expansions is computationally expensive. Moreover, limiters, which push the solution into the realizable set, are usually required to ensure solvability, but realizability conditions are only fully known in one-dimension. However, entropy-based methods yield positive approximations for low-order expansions and have produced promising results in several applications [11, 31, 41, 73, 85, 115].
1. ELECTRON TRANSPORT

**Discrete-Ordinates Methods** evaluate the transport equation for discrete angular directions with the help of quadrature formulas [66, 70]. The derivation of a discrete ordinates (S\textsubscript{N}) method is very simple and it can be implemented efficiently. However, in more than one spatial dimension, the method may exhibit severe artifacts, which are called ray effects [15, 68, 70, 79]. As ray effects are most likely to arise in problems with very different characteristics and we have beams and varying densities in our application, the method might not be well suited for electron transport.

**Spherical harmonic Methods** are moment methods, which use spherical harmonic functions as basis functions. The well-known spherical harmonic (P\textsubscript{N}) system [21, 70, 96] can be derived via a global spectral approximation in angle of the solution to the radiation transport equation. Like any spectral approximation, the P\textsubscript{N} system may exhibit oscillations around discontinuities and large gradients, in some cases causing negative particle concentrations [15]. This fact is considered one of the major drawbacks of the P\textsubscript{N} method (we address this issue in Chapter 2). On the other hand, advantages of the P\textsubscript{N} method are the rotational symmetric solutions and the simplicity of the scattering operator\textsuperscript{1}. The simplicity and the accuracy of the P\textsubscript{N} system lead to efficient solution methods [103].

In this chapter, we introduce a numerical method for electron transport. Our method relies on the spherical harmonic (P\textsubscript{N}) approximation of the CSD equation (1.1) and uses an existing code for the P\textsubscript{N} system of the radiation transport equation. Thus, before we discuss the CSD equation, we introduce the radiation transport equation. At the end of the chapter, we present numerical experiments to show the validity of the method.

\textsuperscript{1}The spherical harmonics diagonalize the scattering operator.
1.1 Introduction

1.1.1 Radiation Transport

We consider the Cauchy problem

\[\frac{\partial}{\partial t}\psi(t, x, \Omega) + \Omega \cdot \nabla_x \psi(t, x, \Omega) + \sigma_a(x)\psi(t, x, \Omega) + (\Omega \psi)(t, x, \Omega) = q(t, x, \Omega) , \tag{1.3a}\]

\[\psi(0, x, \Omega) = \psi_0(x, \Omega) , \tag{1.3b}\]

where \(\sigma_a \geq 0\) is the absorption cross-section; the source \(q \geq 0\) is given; and the unknown \(\psi(t, x, \Omega)\) is the density of particles, with respect to the measure \(d\Omega d\mathbf{x}\), that at time \(t \in \mathbb{R}\) are located at position \(\mathbf{x} = (x, y, z) \in \mathbb{R}^3\) and moving in the direction \(\Omega \in \mathbb{S}^2\). The scattering operator \(\mathbb{Q}\) describes the change in particle direction due to collisions with a fixed background medium

\[(\mathbb{Q}\psi)(t, x, \Omega) = \sigma_s(x) \left(\psi(t, x, \Omega) - \int_{\mathbb{S}^2} g(x, \Omega \cdot \Omega') \psi(t, x, \Omega') \, d\Omega'\right) , \tag{1.4}\]

where \(\sigma_s\) is the non-negative scattering cross-section and for each \(\mathbf{x}\), \(g(\mathbf{x}, \cdot)\) is a non-negative probability density on the interval \([-1, 1]\). Thus \(Q\) is an integral operator in \(\Omega\), but local in \(x\) and \(t\). Note that the cross sections \(\sigma_a\) and \(\sigma_s\) and the kernel \(g\) can also depend on time \(t\).

For each fixed \(x\) and \(t\), \(\mathbb{Q}\) is a self-adjoint bounded linear operator from \(L^2(\mathbb{S}^2)\) to itself. It has a nontrivial null space comprised of functions that are constant with respect to \(\Omega\). Orthogonal to the null space, \(\mathbb{Q}\) is coercive—that is, there exists a constant \(c > 0\) such that

\[\int_{\mathbb{S}^2} h(\mathbb{Q}h) \, d\Omega \geq c\|h\|^2_{L^2(\mathbb{S}^2)} \quad \text{for all } h \in [\text{Null}(\mathbb{Q})]^\perp . \tag{1.5}\]

These properties are standard results in kinetic transport theory; their proofs can be found e.g. in [25, Chapter XXI] or in [24].

For (1.3a), we also use the abstract notation \(\mathcal{T}\psi = q\), where \(\mathcal{T}\) denotes the linear integro-differential operator on the left-hand side. We also use angle brackets as a short-hand for angular integration over \(\mathbb{S}^2\)

\[\langle \cdot \rangle = \int_{\mathbb{S}^2} \langle \cdot \rangle \, d\Omega . \tag{1.6}\]
1. ELECTRON TRANSPORT

1.2 The CSD Approximation

To motivate the Boltzmann CSD equation for electron transport, we first introduce the linear Boltzmann equation

\[
\frac{1}{v} \partial_t \psi(t, E, x, \Omega) + \Omega \cdot \nabla_x \psi(t, E, x, \Omega) + \Sigma_t(E, x) \psi(t, E, x, \Omega) = \int_0^{\infty} \int_{S^2} \Sigma(E', E, \Omega \cdot \Omega') \psi(t, E', x, \Omega') d\Omega' dE' + q(E, x, \Omega), \tag{1.7}
\]

where \( \psi(t, E, x, \Omega) \) is the density of particles at time \( t \) in a phase volume \( dE dx d\Omega \), \( q \) represents a source of particles, and \( v \) is the particle speed. The linear Boltzmann equation describes the evolution of particles, which are interacting with a background medium. The interactions are described via the absorption coefficient \( \Sigma_a \) and the scattering kernel \( \Sigma \). The scattering kernel \( \Sigma \) is proportional to the probability of a particle having energy \( E' \) and direction \( \Omega' \) before the collision and energy \( E \) and direction \( \Omega \) after the collision. Out-scattering and absorption add up to the total cross-section \( \Sigma_t \) given by

\[
\Sigma_t = \Sigma_a + \Sigma_{s,0} \quad \text{with} \quad \Sigma_{s,0}(x, E) = 2\pi \int_0^{\infty} \int_{-1}^{1} \Sigma(E', E, x, \mu) dE' d\mu. \tag{1.8}
\]

The linear Boltzmann equation can be used to model the propagation of electrons in a human body. Since the speed of electrons is very high compared to the irradiation period, the time dependence can be neglected

\[
\psi(t, E, x, \Omega) = \psi(E, x, \Omega). \tag{1.9}
\]

Moreover, in the CSD approximation, we assume electrons lose energy continuously. This means the energy does not change instantaneously in a collision and the in-scattering term becomes

\[
\int_0^{\infty} \int_{S^2} \Sigma(E', E, x, \Omega \cdot \Omega') dE' \psi(E, x, \Omega') d\Omega' = \int_{S^2} \Sigma_s(E, x, \Omega \cdot \Omega') \psi(E, x, \Omega') d\Omega', \tag{1.10}
\]

where

\[
\Sigma_s(E, x, \Omega \cdot \Omega') = \int_{S^2} \Sigma(E', E, x, \Omega \cdot \Omega') dE'. \tag{1.11}
\]
In addition, the mean energy loss is given by the stopping power
\[
S(E, x) = \int_{0}^{\infty} E' \int_{S^2} \Sigma(E', x, \mu) \, d\mu \, dE'
\] (1.12)
and the continuous decay introduces an additional differential operator. Altogether, this yields the Boltzmann CSD equation
\[
- \partial_{E} (S(E, x) \psi(E, x, \Omega)) + \Omega \cdot \nabla_{x} \psi(E, x, \Omega) + \Sigma_{t}(E, x, \Omega) \psi(E, x, \Omega) = \int_{S^2} \Sigma_{s}(E, x, \Omega \cdot \Omega') \psi(E, x, \Omega') \, d\Omega' + q(E, x, \Omega), \tag{1.13}
\]
where the total cross-section is given by the out-scattering
\[
\Sigma_{t}(E, x) = \Sigma_{s,0}(E, x) = 2\pi \int_{-1}^{1} \Sigma_{s}(E, x, \mu) \, d\mu, \tag{1.14}
\]
since for electrons there is no absorption.

**Remark 1.1.** Using the theory of evolution equations, it can be shown that (1.13) has a unique solution [113]. To apply the theory, we have to make various assumptions on the parameters of the system. As before, we assume that there exists a maximal energy of the system \(E_{\text{max}}\) and \(\psi(E_{\text{max}}, x, \Omega) = \psi_{\text{max}}(x, \Omega)\). However, the initial condition can be set to zero by transforming \(\psi_{\text{max}}\) into a source term. Further, let the computational domain be given by \(I \times X \times S^2\) with \(I = [0, E_{\text{max}}]\) and \(X \subset \mathbb{R}^3\). Then [113, Theorem 3.19] can be applied and we obtain a unique solution \(\psi \in C^{1}(I, L^{2}(X \times S^2))\) if the following assumptions are satisfied:

1. The stopping power satisfies
\[
S \in C^{2}(I, L^{\infty}(X)) , \quad \nabla_{x} S \in L^{\infty}(I \times X) ,
\]
and
\[
\inf_{(E, x) \in I \times X} S(E, x) > 0. \tag{1.15}
\]

2. The scattering cross-section satisfies
\[
\Sigma_{s,0} \in L^{\infty}(I \times X) , \quad \Sigma_{s} \in C^{1}(I, L^{\infty}(X \times [-1, 1])) , \quad \Sigma_{s} \geq 0 , \quad \text{and} \quad \int_{S^2} \Sigma_{s}(E, x, \Omega \cdot \Omega') \, d\Omega' < M \tag{1.16}
\]
for some constant \(M\).
3. The source term satisfies $q \in C^1(I, L^2(X \times S^2))$.

4. The inflow boundary condition is zero.

Note that (1.15) and (1.16) are usually fulfilled by physical parameters. Moreover, the last assumption can be generalized, but we do not consider boundary conditions.

Remark 1.2. The CSD approximation is similar to the Fokker-Planck approximation, as it also introduces an additional differential operator of the form $\frac{\partial}{\partial E}(S\psi)$. The Fokker-Planck operator can be derived by a formal asymptotic analysis for small energy losses, small angle deviations, and a small mean free path. Originally, it neglects large angle and large energy-loss scattering [95], but there exist improvements for large angle scattering [92].

In the following, we use the CSD approximation for electron transport and present further simplifications and transformations to obtain a kinetic equation, which is similar to the radiation transport equation in section 1.1.1.

### 1.2.1 Simplifications

The stopping power and the transport coefficient depend on energy, the density of the matter, and the structure of the matter. For simplicity, we neglect the dependence on the structure of the matter and assume all substances are water with different densities. Note that this assumption is often, but not always, used in applications.

**Assumption 1.3.** Assume the material coefficients satisfy

$$
S(E, \mathbf{x}) = S_{H_2O}^E(\rho(\mathbf{x})) ,
$$

$$
\Sigma_t(E, \mathbf{x}) = \Sigma_{t_{H_2O}}^E(\rho(\mathbf{x})) ,
$$

$$
\Sigma_{s}(E, \mathbf{x}, \Omega \cdot \Omega') = \Sigma_{s_{H_2O}}^E(E, \Omega \cdot \Omega')\rho(\mathbf{x}) ,
$$

where $\rho(\mathbf{x})$ is the density of the substance at position $\mathbf{x}$ and the superscript $H_2O$ indicates the material coefficients of water.
1.2 The CSD Approximation

Applying Assumption 1.3 to the above transport equation (1.1a) and dropping the superscript $H_2O$, we obtain

$$
- \partial_x E \psi(E, x, \Omega) - \partial_x \psi(E, x, \Omega) + \rho(x) \Sigma_x(E) \psi(E, x, \Omega) = \rho(x) \int_{S^2} \Sigma_s(E, \Omega') \psi(E, x, \Omega') d\Omega' + q(E, x, \Omega). \quad (1.18)
$$

Analogously, the dose (1.2) becomes

$$
D(x) = \int_{0}^{\infty} \int_{S^2} S(E) \psi(E, x, \Omega) d\Omega dE. \quad (1.19)
$$

1.2.2 Transformations

We want to adapt an existing radiation transport code to solve the CSD equation (1.18). The main idea is to use a time-dependent code and consider the energy as time, i.e., we start with the highest energy and propagate the solution backward in energy (cf. [11]). To this end, we first transform the equation into a form, which is similar to the radiation transport equation (1.3a).

In the first step, we make a transformation to remove the stopping power from the energy derivative. We multiply (1.18) by $S(E)$ and define

$$
\tilde{\psi}(E, x, \Omega) := S(E) \rho(x) \psi(E, x, \Omega) \quad (1.20)
$$

to obtain

$$
- S(E) \partial_x \tilde{\psi}(E, x, \Omega) - \partial_x \tilde{\psi}(E, x, \Omega) + \rho(x) \Sigma_x(E) \tilde{\psi}(E, x, \Omega) = \rho(x) \int_{S^2} \Sigma_s(E, \Omega') \tilde{\psi}(E, x, \Omega') d\Omega' + S(E) q(E, x, \Omega). \quad (1.21)
$$

In the second step, we transform the energy variable by

$$
\tilde{E} := \int_{0}^{E} \frac{1}{S(E')} dE' \quad (1.22)
$$

and denote the evaluation of the functions at the transformed energy with a tilde, e.g.,

$$
\tilde{\psi}(\tilde{E}, x, \Omega) := \tilde{\psi}(E, x, \Omega) = S(E) \rho(x) \psi(E, x, \Omega). \quad (1.23)
$$
1. ELECTRON TRANSPORT

Then we obtain

\[-\partial_{\tilde{E}} \tilde{\psi}(\tilde{E}, \mathbf{x}, \Omega) + \Omega \cdot \nabla_{\mathbf{x}} \tilde{\psi}(\tilde{E}, \mathbf{x}, \Omega) + \tilde{\Sigma}_s(\tilde{E}) \tilde{\psi}(\tilde{E}, \mathbf{x}, \Omega) + \tilde{\Sigma}_t(\tilde{E}) \tilde{\psi}(\tilde{E}, \mathbf{x}, \Omega) + \rho(\mathbf{x}) + \tilde{\Sigma}_t(\tilde{E}) \tilde{\psi}(\tilde{E}, \mathbf{x}, \Omega) = \int_{S^2} \tilde{\Sigma}_s(\tilde{E}, \Omega', \Omega) \tilde{\psi}(\tilde{E}, \mathbf{x}, \Omega') d\Omega' + \tilde{S}(\tilde{E}) \tilde{q}(\tilde{E}, \mathbf{x}, \Omega).\]  

(1.24)

In the last step, we change the direction of the energy

\[\tilde{E} = \tilde{E}_{\text{max}} - \tilde{E}\]  

(1.25)

and define again functions for the transformed energy. This changes (1.1b) to an initial condition and changes the sign in front of the time derivative. In particular, if \(E_{\text{max}}\) is the maximal energy of the system, this step makes the transformed energy to a pseudo time.

**Lemma 1.4.** Under Assumption 1.3, the CSD approximation (1.18) can be transformed into the following form

\[\partial_E \psi(E, \mathbf{x}, \Omega) + \Omega \cdot \nabla_{\mathbf{x}} \psi(E, \mathbf{x}, \Omega) + \Sigma(E) \psi(E, \mathbf{x}, \Omega) = \int_{S^2} \Sigma_s(E, \Omega, \Omega') \psi(E, \mathbf{x}, \Omega') d\Omega' + S(E) q(E, \mathbf{x}, \Omega),\]  

(1.26a)

\[\psi(0, \mathbf{x}, \Omega) = \psi_{\text{max}}(\mathbf{x}, \Omega).\]  

(1.26b)

This is a standard radiation transport equation except for the density term in the spatial gradient.

**Corollary 1.5.** Applying the same transformations to the dose (1.19) yields

\[D(\mathbf{x}) = \frac{1}{\rho(\mathbf{x})} \int_0^\infty S(E) \int_{S^2} \psi(E, \mathbf{x}, \Omega) d\Omega dE,\]  

(1.27)

where \(E\) and \(\psi\) are the transformed variables from (1.26).

**Remark 1.6.** Solving (1.26) with a time-marching scheme in the transformed variables, requires to evaluate the material coefficients at the original energy. This means we need the transformation

\[\tilde{E}(E) = \tilde{E}_{\text{max}} - \tilde{E} = \int_0^{E_{\text{max}}} \frac{1}{S(E')} dE' - \int_0^E \frac{1}{S(E')} dE'\]  

(1.28)

and its inverse. Also, remember that we are actually solving for

\[S(E) \rho(\mathbf{x}) \psi(E, \mathbf{x}, \Omega).\]  

(1.29)
Remark 1.7. The density term in the spatial derivative in (1.26) has to be handled by changing the radiation transport code. If we change the transformation, the term can be pulled outside the derivative. However, in three spatial dimensions, it cannot be removed from the equation (in one dimension this can be done by rescaling space).

1.3 Numerical Method

In the previous section, we have described the transformation from the CSD approximation to a radiation transport equation. This can be used to adapt an existing transport code to electron transport. Our method is based on a two-dimensional staggered grid method for spherical harmonics moment equations of radiative transfer, which is called StaRMAP [103] (the code is available online at [104]). The main idea of the method is to use the structure of the transport equation to construct staggered grids in space and time. In this section, we describe the adaptations to the method for the use in electron transport problems. First, we introduce the spherical harmonic equations of radiative transfer. Second, we generalize the StaRMAP method from two to three spatial dimensions. Finally, we derive the \( P_N \) equations for the CSD equations and describe the implementation of physical transport coefficients.

1.3.1 Spherical Harmonic (\( P_N \)) Equations

The spherical harmonic (\( P_N \)) equations are derived from a spectral Galerkin discretization of the transport equation, using spherical harmonic functions as a basis. These functions and their properties are classical (see, for example, [6, 84]), but for completeness we briefly describe them here. We write

\[
\Omega = (\Omega_x, \Omega_y, \Omega_z)^T = (\sqrt{1 - \mu^2 \cos(\varphi)}, \sqrt{1 - \mu^2 \sin(\varphi)}, \mu)^T, \tag{1.30}
\]

where \( \mu := \Omega_z \in [-1, 1] \) and \( \varphi \in [0, 2\pi) \) is the angle between the \( x \)-axis and the projection of \( \Omega \) onto the \( x-y \) plane. Given integers \( \ell \geq 0 \) and \( k \in [-\ell, \ell] \), the complex-valued spherical harmonic of degree \( \ell \) and order \( k \) is expressed in terms
1. ELECTRON TRANSPORT

of $\mu$ and $\varphi$ as

$$Y_\ell^k(\Omega) = \sqrt{\frac{2\ell + 1 (\ell - k)!}{4\pi (\ell + k)!}} e^{ik\varphi} P_\ell^k(\mu), \quad (1.31)$$

where $P_\ell^k$ is an associated Legendre function. With this scaling, the spherical harmonics form an orthonormal system on the sphere

$$\int_{S^2} Y_\ell^k(\Omega) \overline{Y_{\ell'}^{k'}(\Omega')} d\Omega = \delta_{\ell,\ell'} \delta_{k,k'} . \quad (1.32)$$

The primary motivation for using the spherical harmonics is that they form a complete set of eigenfunctions for the scattering operator $Q$ given in (1.4). The eigenvalue relation can be derived by first expanding the kernel $g$ in Legendre polynomials

$$g(x, \Omega \cdot \Omega') = \sum_{\ell=0}^{\infty} \frac{2\ell + 1}{4\pi} g_\ell(x) P_\ell(\Omega \cdot \Omega') \quad (1.33)$$

with $g_\ell(x) = 2\pi \int_{-1}^{1} g(x, \mu) P_\ell(\mu) d\mu$ and $P_\ell$ is the degree $\ell$ Legendre polynomial with normalization $\int_{-1}^{1} P_\ell^2 d\mu = \frac{2}{2\ell + 1}$. Second, we apply the addition formula for spherical harmonics (see, for example, [70, Appendix A] or [67])

$$P_\ell(\Omega \cdot \Omega') = \frac{4\pi}{2\ell + 1} \sum_{k=-\ell}^{\ell} \overline{Y_{\ell'}^{m}(\Omega')} Y_\ell^k(\Omega) \quad (1.34)$$

with $\overline{Y_\ell^k}$ being the complex conjugate. This yields

$$\int_{S^2} g(x, \Omega \cdot \Omega') Y_\ell^k(\Omega') d\Omega' = \int_{S^2} \sum_{\ell'=0}^{\infty} \frac{2\ell' + 1}{4\pi} g_{\ell'}(x) P_{\ell'}(\Omega \cdot \Omega') Y_\ell^k(\Omega') d\Omega' \quad (1.35)$$

Therefore, the eigenvalue relation of the scattering operator is given by

$$Q Y_\ell^k = \sigma_s (1 - g_\ell) Y_\ell^k , \quad \ell = 0, 1, 2, \ldots \text{ and } k = -\ell, \ldots, 0, \ldots, \ell , \quad (1.36)$$

and it reduces the approximation of $Q$ in the Galerkin method from an $O(N^2)$ to an $O(N)$ operation.
1.3 Numerical Method

For convenience, we use the real-valued spherical harmonics which, up to a normalization factor, are the real and imaginary parts of each $Y^k_\ell$:

$$m^k_\ell = \begin{cases} \frac{(-1)^k}{\sqrt{2}} (Y^k_\ell + (-1)^k Y^{-k}_\ell), & k > 0, \\ Y^0_\ell, & k = 0, \\ -\frac{(-1)^k}{\sqrt{2}} (Y^{-k}_\ell - (-1)^k Y^k_\ell), & k < 0. \end{cases}$$ (1.37)

We collect the $n_\ell := 2\ell + 1$ real-valued harmonics of degree $\ell$ together into a vector-valued function $m^\ell$ and then for given $N$, set

$$m = (m_0, m_1^T, \ldots, m_N^T)^T = (m_0^0, m_1^{-1}, m_1^0, m_1^1, \ldots, m_N^N)^T.$$ (1.38)

In all, $m$ has $n := \sum_{\ell=0}^N n_\ell = (N+1)^2$ components which form an orthonormal basis for the space

$$P_N = \left\{ \sum_{\ell=0}^N \sum_{k=-\ell}^\ell c^k_\ell m^k_\ell : c^k_\ell \in \mathbb{R} \text{ for } 0 \leq \ell \leq N, |k| \leq \ell \right\}.$$ (1.39)

Finally, the spherical harmonics fulfill a recursion relation of the form

$$\Omega_i m^\ell = \left( a^{(i)}_\ell \right)^T m^{\ell-1} + a^{(i)}_{\ell+1} m^{\ell+1} , \text{ where } a^{(i)}_\ell \in \mathbb{R}^{(2\ell-1) \times (2\ell+1)}.$$ (1.40)

The $P_N$ equations are derived by approximating $\psi$ by a function $\psi_P \in P_N$

$$\psi \approx \psi_P \equiv m^T u_P,$$ (1.41)

where $u_P : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^n, (t, x) \mapsto u_P(t, x)$ solves

$$\left\langle m^T (m^T u_P(t, x)) = q(t, x), \quad (t, x) \in (0, \infty) \times \mathbb{R}^3, \right.$$ (1.42a)

$$u_P(0, x) = \left\langle m^T \psi(x, \cdot) \right\rangle, \quad x \in \mathbb{R}^3, \quad (1.42b)$$

and $q := \langle mq \rangle$. Using (1.3a), the system (1.42a) can be written more explicitly as

$$\partial_t u_P + A \cdot \nabla_x u_P + \sigma_a u_P + \sigma_s G u_P = q,$$ (1.43)

where $G$ is diagonal with components $G_{(\ell,k),(\ell,k)} = 1 - g_\ell$, and we have used the eigenvalue relation (1.36) and the fact that $\langle mm^T \rangle = I$. The inner product between $A$ and the gradient is understood as

$$A \cdot \nabla_x \equiv A_x \partial_x + A_y \partial_y + A_z \partial_z ,$$ (1.44)

\[1\] Note that the dependence of $m$ on $N$ has been suppressed.
1. ELECTRON TRANSPORT

where $A_i = \langle \mathbf{m} \mathbf{m}^T \Omega_i \rangle$ with $i \in \{x,y,z\}$. Due to the recursion relation (1.40), the structure of the matrices $A_i$ is very specific:

$$A_i = \begin{bmatrix}
0 & a^{(i)}_1 \\
(a^{(i)}_1)^T & 0 & a^{(i)}_2 \\
& \ddots & \ddots & \ddots \\
& & (a^{(i)}_N)^T & 0
\end{bmatrix}.$$  

(1.45)

More details, including exact expressions for the matrices $a^{(i)}_\ell$, can be found in Appendix A.

1.3.2 StaRMAP in Three Spatial Dimensions

The $P_N$ equations are introduced in three dimensions in section 1.3.1. Compared to the two dimensional system, the system size increases from $(N+1)(N+2)/2$ to $(N+1)^2$. The size of the matrices $A_x$ and $A_y$ changes accordingly and the matrix $A_z$ is created (see Appendix A). To discretize the system, we transfer the methods used in the two-dimensional StaRMAP code to three dimensions.

For the spatial discretization, we need to define three-dimensional staggered grids. Note that in one dimension, there are two staggered regular grids (the cell centers and the cell edges), whereas in two dimensions, there are four staggered grids (cell centers, vertical and horizontal cell edges, and cell corners). Here, in three dimensions, there are eight staggered grids given by (see Figure 1.1)

$$G_{111} = \{(i\Delta x, j\Delta y, k\Delta z) : i, j, k \in \mathbb{Z}\},$$  

$$G_{112} = \{(i\Delta x, j\Delta y, (k - \frac{1}{2})\Delta z) : i, j, k \in \mathbb{Z}\},$$  

$$G_{222} = \{(i - \frac{1}{2})\Delta x, (j - \frac{1}{2})\Delta y, (k - \frac{1}{2})\Delta z) : i, j, k \in \mathbb{Z}\},$$  

(1.46)

where $(\Delta x, \Delta y, \Delta z)$ is the grid size. Here, we describe the method for infinite meshes, but the boundary conditions, used in [103], can easily be adapted.

We call the grids with an even digit sum index, e.g. $G_{222}$, “even” grids and the ones with an odd digit sum index, e.g. $G_{111}$, “odd” grids. Then, in every axis direction the even and odd grids are staggered to each other. Using half-grid central finite differences, the derivative of a variable lying on an even grid
1.3 Numerical Method

Figure 1.1: Segment of three-dimensional staggered grids: grids with filled markers are odd grids; grids with empty markers are even grids.

are lying on an odd grid and vice versa. This can be used to place the solution components in an efficient way on the grids. This means, we start by placing the zeroth degree component on the center grid $G_{111}$. Then, we distribute all other components according to the pattern of the matrices $A$, such that all derivatives can be approximated by half-grid central differences. In particular, even solution components ($u^k_\ell$ with $\ell$ even) are mapped to even grids and odd components ($u^k_\ell$ with $\ell$ odd) are mapped to odd grids. Note that this mapping is possible due to the structure in the transport equation and the recursion relation of the spherical harmonics.

The time discretization of the StaRMAP code only relies on the splitting into even and odd components and can remain unchanged. Thus, we re-sort the solution components in (1.43) into even and odd components and obtain

$$\partial_t \begin{bmatrix} u^e_{P_N} \\ u^o_{P_N} \end{bmatrix} + \begin{bmatrix} 0 & A^{eo} \\ A^{oe} & 0 \end{bmatrix} \cdot \nabla_x \begin{bmatrix} u^e_{P_N} \\ u^o_{P_N} \end{bmatrix} + \begin{bmatrix} C^e & 0 \\ 0 & C^o \end{bmatrix} \begin{bmatrix} u^e_{P_N} \\ u^o_{P_N} \end{bmatrix} = \begin{bmatrix} q^e \\ q^o \end{bmatrix} \quad (1.47)$$

with $C := \sigma_a + \sigma_s G$. To introduce a second order time-stepping scheme for this system, we first eliminate the time dependence of the coefficients by evaluating
the coefficients at the mid-time step \( t + \frac{\Delta t}{2} \) and assuming they are constant over a time step. Then, we split the system into two parts.

\[
\begin{align*}
\frac{\partial}{\partial t} \begin{bmatrix} u_{PN}^c \\ u_{PN}^o \end{bmatrix} &+ \begin{bmatrix} 0 & A^c \\ 0 & 0 \end{bmatrix} \cdot \nabla_x \begin{bmatrix} u_{PN}^c \\ u_{PN}^o \end{bmatrix} + \begin{bmatrix} C^c & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{PN}^c \\ u_{PN}^o \end{bmatrix} = \begin{bmatrix} q^c \\ 0 \end{bmatrix} \quad (1.48a) \\
\frac{\partial}{\partial t} \begin{bmatrix} u_{PN}^o \\ u_{PN}^c \end{bmatrix} &+ \begin{bmatrix} 0 & 0 \\ A^{o,0} & 0 \end{bmatrix} \cdot \nabla_x \begin{bmatrix} u_{PN}^o \\ u_{PN}^c \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & C^o \end{bmatrix} \begin{bmatrix} u_{PN}^o \\ u_{PN}^c \end{bmatrix} = \begin{bmatrix} 0 \\ q^o \end{bmatrix} \quad (1.48b)
\end{align*}
\]

and solve both parts analytically for one time step. In (1.48a) the odd components \( u_{PN}^o \) are constant over time and the even components satisfy

\[
\frac{\partial}{\partial t} u_{PN}^e + C^e u_{PN}^e = r^e \quad \text{with} \quad r^e = q^e - A^{e,0} \cdot \nabla_x u_{PN}^o.
\] (1.49)

Since \( C^e \) and \( r^e \) are independent of time and \( C^e \) is diagonal, it remains to solve scalar differential equations of the form

\[
\frac{\partial}{\partial \tau} u_k(\tau, x) + c_k(x) u_k(\tau, x) = r_k(x)
\] (1.50)

from \( \tau = t \) to \( \tau = t + \Delta t \), which can be solved analytically by

\[
\begin{align*}
u_k(t + \Delta t, x) &= \exp(-c_k(x) \Delta t) u_k(t, x) - \frac{1}{c_k(x)} (1 - \exp(-c_k(x) \Delta t)) r_k(x) \\
&= u_k(t, x) + \Delta t (r_k(x) - c_k(x) u_k(t, x)) E(-c_k(x) \Delta t),
\end{align*}
\] (1.51)

where \( E(c) = \frac{\exp(c) - 1}{c} \). For the implementation, we use the latter part in (1.51) together with a robust implementation of the function \( E \), so that the update rule can also be applied for \( c_k(x) = 0 \). Altogether, an analytic solution of (1.48a) is defined by the following evolution operator

\[
S^c_{t+\Delta t, t} \begin{bmatrix} u_{PN}^c \\ u_{PN}^o \end{bmatrix} = \begin{bmatrix} u_{PN}^o \\ u_{PN}^c \end{bmatrix} + \Delta t \begin{bmatrix} (r^e - C^e u_{PN}^e) \\ 0 \end{bmatrix} E(-C^e \Delta t).
\] (1.52)

Similarly, we can define an evolution operator for (1.48b)

\[
S^o_{t+\Delta t, t} \begin{bmatrix} u_{PN}^c \\ u_{PN}^o \end{bmatrix} = \begin{bmatrix} u_{PN}^o \\ u_{PN}^c \end{bmatrix} + \Delta t \begin{bmatrix} 0 \\ (r^o - C^o u_{PN}^o) \end{bmatrix} E(-C^o \Delta t).
\] (1.53)

Finally, a full solution step from \( t \) to \( t + \Delta t \) is given by

\[
S_{t+\Delta t} = S^o_{t+\Delta t, t} \circ S^c_{t+\Delta t, t} \circ S^o_{t, \frac{\Delta t}{2}} \circ S^c_{t, \frac{\Delta t}{2}} \circ S^o_{t, \frac{\Delta t}{2}} \circ S^c_{t, \frac{\Delta t}{2}}.
\] (1.54)
Remark 1.8. Since the evolution operator \( S_{t+\Delta t}^e \) yields an analytic solution of (1.48a), we have
\[
S_{t+\Delta t,t}^e = S_{t,\frac{\Delta t}{2}}^e \circ S_{t+\frac{\Delta t}{2},t}^e,
\]
which can also be seen from the solution formula (1.51) if we compute
\[
\begin{align*}
\psi_k(t + \Delta t, x) &= \exp \left( -c_k(x) \frac{\Delta t}{2} \right) \psi_k(t, x) - \frac{r_k(x)}{c_k(x)} \left( 1 - \exp \left( -c_k(x) \frac{\Delta t}{2} \right) \right) \\
&= \psi_k(t, x) - \frac{r_k(x)}{c_k(x)} \left( 1 - \exp \left( -c_k(x) \frac{\Delta t}{2} \right) \right) \right).
\end{align*}
\]
Therefore, (1.54) is a Strang splitting [111] of (1.47) with constant coefficients. Together with the evaluation of the time-dependent coefficients at the mid-time step, this leads to a second order time discretization.

Remark 1.9. The CFL condition of the scheme needs to be adapted to the dimension. In the simplest one-dimensional case, the stability analysis leads to \( \Delta t < \Delta x \). In two dimensions, the condition becomes \( \Delta t < \frac{\Delta x}{\sqrt{2}} \) and in three dimensions \( \Delta t < \frac{\Delta x}{\sqrt{3}} \). This accounts for the growth in each spatial direction.

1.3.3 StaRMAP with CSD

To implement (1.26) into the StaRAMAP code, we first derive the \( P_N \) equations for the CSD equation. We use the same notation as in section 1.3.1, where the \( P_N \) equations for the radiation transport equations are discussed. Thus, let \( \mathbf{m} \) be a vector containing the spherical harmonics basis functions up to degree \( N \) and \( \mathbf{m}^T \mathbf{u}_{P_N} \) denote the \( P_N \) approximation of \( \psi \). Then, the \( P_N \) equations of (1.26) are given by
\[
\partial_E \mathbf{u}_{P_N} + \mathbf{A} \cdot \nabla_x \frac{\mathbf{u}_{P_N}}{\rho} + \mathbf{C} \mathbf{u}_{P_N} = \mathbf{S} \mathbf{q}
\]
with \( \mathbf{A} := \langle \mathbf{mm}^T \Omega \rangle \) (see Appendix A). In particular, the scattering cross-sections are encoded in a diagonal matrix \( \mathbf{C} \) with entries \( \Sigma_s - \Sigma_{s\ell} \), where \( \Sigma_{s\ell} \) are the Legendre expansion coefficients of the scattering cross-section
\[
\Sigma_{s\ell}(E) = 2\pi \int_{-1}^{1} P_{\ell}(\mu) \Sigma_s(E, \mu) \, d\mu.
\]
1. ELECTRON TRANSPORT

The density introduces an explicit space dependence in the flux term and may be strongly varying. This is a major difficulty for numerical approximations, which has been addressed by several authors and is still a current topic of research (see e.g. [119] for an overview). Without going into further theoretical details, we implement the space depend flux by evaluating the density function \( \rho(x) \) on each grid and divide the solution components at the conforming positions. Since this changes the characteristics of the system, the selection of the time step needs to be adapted. To ensure stability, we multiply the time step with the minimal density of the system. Note that this slows down the computations substantially if the minimal density is small. For instance, the density of air is nearly one-thousandth times smaller than the density of water. Although one can manipulate the minimal density and use for instance \( \rho_{\text{min}} = 0.01 \) or even \( \rho_{\text{min}} = 0.05 \) for the numerics, the computational cost still rises by a factor of 100 or 20.

The material parameters, i.e. stopping power and transport coefficient, are usually given as tabulated data. In our method, the stopping power is, in particular, needed for the transformation of variables. Note that the forward transformation is only needed to transform the initial condition, but the backward transformation is always needed when material coefficients are evaluated. Therefore, the transformation should be implemented efficiently. To compute the forward transformation

\[
\tilde{E}(E) = \tilde{E}_{\text{max}} - \tilde{E} = \int_0^{E_{\text{max}}} \frac{1}{S(E')} \, dE' - \int_0^E \frac{1}{S(E')} \, dE',
\]

we can use the tabulated data for the stopping power and the trapezoidal rule for the integral. With this we can precompute tabulated data for the transformed energy. Then, we use linear interpolation to evaluate the forward and the backward transformation.

1.3.3.1 Implementation of the Dose

So far, we considered the implementation of the CSD equation, but we are in particular interested in the dose. To compute the dose, we need to integrate the
first moment of the solution against the stopping power

\[
D(x) = \frac{1}{\rho(x)} \int_0^\infty \int_{S^2} S(E, x) \psi(E, x, \Omega) \, d\Omega \, dE = \frac{1}{\rho(x)} \int_0^\infty \psi_0^0(E, x) \, dE \ , \quad (1.61)
\]

where

\[
\psi_0^0(E, x) = \int_{S^2} \psi(E, x, \Omega) \, d\Omega \quad (1.62)
\]

is approximated by the zeroth order solution component. Since the method is second order in time, we use the trapezoidal rule to approximate the integral. This can be done in a memory efficient way if part of the integral is computed after each time-step.

1.3.3.2 Implementation of Physical Parameters

In our model, the input parameters are the density of the medium, the stopping power and the transport coefficient. These parameters have to be extracted from measurement data or databases. In radiotherapy, the computations usually rely on a CT scan of the patient. The measurement values of the CT scan (in Hounsfield units) lead to a good approximation of the density. For the stopping power and the transport coefficient, we use tabulated data extracted from a state of the art Monte Carlo code (Penelope [99, 100]) and the ICRU77 database [101]. More details can be found in [91, 92].

1.4 Numerical Results

In this section, we validate the numerical method by comparison with a well-established Monte Carlo (MC) code, which was evaluated against experiments. To compute the reference solution we use penEasy2012 [105], which is based on Penelope2011 [100] and can handle voxelized geometries. Note that the MC code includes all relevant physical effects, whereas our method relies on a simplified physical model. Since in applications the absorbed dose is of main interest, we compare the dose computed with the \( P_N \) and the MC method.

To evaluate our dose distribution, we use the \( \gamma \) index [72]. The \( \gamma \)-index is a space-dependent quantitative measure for the accuracy of the computed dose \( D \).
1. ELECTRON TRANSPORT

against a reference solution $\bar{D}$. It is given by

$$\gamma(x) = \min_x \sqrt{\left(\frac{\|x - \bar{x}\|}{r}\right)^2 + \left(\frac{D(x) - \bar{D}(\bar{x})}{R}\right)^2},$$

(1.63)

where $r$ is a tolerance for the spatial error and $R$ is a tolerance for the total error. At a given tolerance the $\gamma$-index yields an acceptance criterion:

$$\gamma(x) \leq 1, \quad \text{result passes},$$

$$\gamma(x) > 1, \quad \text{result fails.}$$

(1.64)

This means the acceptance criterion can only be met if there exists a point $x_0$ in the ball with radius $r$ around $x$, which satisfies the dose-difference criterion $|D(x) - \bar{D}(x_0)| < R$. Here, we test the $P_N$ solution against a Monte Carlo solution at a tolerance $r = 3$ mm and $R = 3\%$ of the maximal dose.

Remark 1.10. In the continuous case, the distance-to-agreement (DTA)

$$d(x) = \min \{\|x - \bar{x}\| : D(x) = \bar{D}(\bar{x})\} \leq r$$

(1.65)

and the dose-difference criterion [44]

$$\delta(x) = |D(x) - \bar{D}(x)| \leq R$$

(1.66)

are easily related to the $\gamma$-criterion. If the DTA holds, there exists a point $x_0$ such that $\|x - x_0\| \leq r$ and $D(x) = \bar{D}(x_0)$. This implies $\gamma(x) \leq \frac{\|x - x_0\|}{r} \leq 1$. On the other hand, if the dose-difference criterion holds, we have $|D(x) - \bar{D}(x)| \leq R$. This also implies $\gamma(x) \leq \frac{|D(x) - \bar{D}(x)|}{R} \leq 1$. In particular, if one of the criteria holds, the $\gamma$-criterion is met, but the contrary is not necessarily true.

Remark 1.11. Sometimes a threshold cutoff of e.g. 10\% of the maximum dose is used in addition to the $\gamma$-criterion [7]. Although this can improve the results in regions with a lower radiation dose, which are less relevant in applications, we do not consider this here.

For all test cases, we compute a $P_N$ solution with the method described in section 1.3 and a MC solution as a reference. As far as possible, the solutions are computed with the same problem parameters, e.g., size of the domain, number of grid points, initial condition. The main differences, which are specific to the
respective method (besides physical parameters), are the boundary conditions, the number of random samples in the MC simulation, and the order $N$ of the moment approximation in the $P_N$ method. Since we do not consider boundary conditions, we always choose the size of the domain, such that for the considered times the solution is compactly supported inside the domain. Furthermore, we reduce the statistical error in the MC solution by using a lot of random samples, and we reduce the Gibbs phenomena in the $P_N$ method by using a high order approximation. Note that in all our test cases, this yields in essence non negative $P_N$ solutions, meaning that the solution does not fall below $-3\%$ of the maximal dose. In particular, we do not filter out any negative values of the solution.

In the first test cases, we investigate the model error of the $P_N$ method using simple geometries, for instance we model a phantom containing water. Afterwards we consider complex geometries, given by a CT scan of a human torso, to test the full potential of the method in two and three spatial dimensions.

### 1.4.1 Phantoms

In the following test cases, we investigate the impact of the model on the dose profile for varying initial energies and densities. First, we model a one-dimensional water phantom with a beam entering from the left with a certain energy. Second, we model a cavity inside the water phantom, which is filled with an air-like substance.

#### 1.4.1.1 Water Phantom

We consider a water phantom with density

$$\rho(x) = \rho_{H_2O} = 1 \ \text{g/cm}^3$$

and a beam located at zero and moving to the right. Here, we model a beam as a Gaussian with variance $s = 0.1$ in space and a delta function in angle, i.e., the initial condition is given by

$$\psi(E_{\text{max}}, x, \Omega) = \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{x^2}{2s^2}\right) \delta(\Omega - \Omega_0)$$
1. ELECTRON TRANSPORT

Figure 1.2: Water phantom with initial beam entering from the left with energy (a) \( E_{\text{max}} = 5 \text{ MeV} \) or (b) \( E_{\text{max}} = 10 \text{ MeV} \): Normalized dose and \( \gamma \)-index on a selection of the computational domain.

With \( \Omega_0 \) pointing in x-direction. The computational domain is \([-2.5, 7.5]\) and the grid size is \( \Delta x = 0.05 \text{ cm} \). We compute the \( P_N \) solution for \( N = 21 \) and compare the normalized dose to a MC solution with the help of the \( \gamma \)-index\(^1\).

The results are shown for \( E_{\text{max}} = 5 \text{ MeV} \) and \( E_{\text{max}} = 10 \text{ MeV} \) in Figure 1.2. Compared to the MC solution, the dose profile is similar, but the dose falls off faster. In the case of \( E_{\text{max}} = 5 \text{ MeV} \) this behavior is still acceptable (\( \gamma(x) \leq 1 \) for all \( x \)), whereas for \( E_{\text{max}} = 10 \text{ MeV} \) the gap between the two solutions becomes larger and the \( \gamma \)-index exceeds the critical bound (\( \gamma(x) > 1 \) for \( 4.325 < x < 5.025 \)).

1.4.1.2 Water Phantom with Air Cavity

Let the density of the phantom be given by

\[
\rho(x) = \begin{cases} 
0.01 \frac{\text{g}}{\text{cm}^3}, & 1.5 < x < 2, \\
1 \frac{\text{g}}{\text{cm}^3}, & \text{otherwise}. 
\end{cases}
\]  

(1.69)

Note that the density of air is smaller than the density of the cavity, but a smaller density increases the computational cost and taking a smaller density does not

---

\(^1\)The \( \gamma \)-index is computed by first interpolating the data on a 100 times finer grid and then using (1.63), where we minimize over all grid points in a surrounding of \( \bar{x} \) with radius \( 2R \).
change the result substantially. Similar to the previous test case, we consider a beam at zero moving to the right, which we model by a Gaussian in space and a delta function in angle. Thus, let the initial condition be given by (1.68) with \( s = 0.1 \) and \( E_{\text{max}} = 5 \text{ MeV} \), the computational domain be given by \([-2.5, 7.5]\) and the grid size be \( \Delta x = 0.05 \text{ cm} \).

![Figure 1.3](image)

**Figure 1.3:** Water phantom with air cavity and initial beam entering from the left with \( E_{\text{max}} = 5 \text{ MeV} \): Normalized dose and \( \gamma \)-index shown on the interval \([0, 3.5]\).

Figure 1.3 shows the comparison of the normalized dose profile between the \( P_N \) solution with \( N = 21 \) and the MC solution. As in the previous test case of a pure water phantom with \( E_{\text{max}} = 5 \text{ MeV} \), the \( \gamma \)-index satisfies \( \gamma(x) \leq 1 \) on the hole computational domain, and with this the result has a very good accuracy.

### 1.4.2 CT Scans

In this test case, we validate the numerical method with real data of a patient, who has a tumor in the lungs \(^1\). The data is given as a three-dimensional CT scan of the torso with a resolution of \(0.977\text{mm} \times 0.977\text{mm} \times 3\text{mm}\). In addition, the clinical target volume (CTV) and the planning target volume (PTV) of the tumor are marked and certain organs are segmented. We use this information to select a computational domain and to set up the test cases, whereas we only use the CT scan for the computations. First, we consider a two-dimensional test case using a

\(^1\)The data is provided by Nuria Escobar-Corral, Klinik für Radioonkologie und Strahlentherapie, Uniklinik RWTH Aachen.
1. ELECTRON TRANSPORT

horizontal slice of the CT scan. Here, we explore the method in two-dimensions with a high resolution. Then the method is tested on a three-dimensional test case. For both test cases, we compute the density from the Hounsfield units in the CT scan using

\[ \rho(x) \approx \frac{\text{HU}(x) \cdot \text{slope} + \text{intercept}}{1000} + \rho_{\text{H}_2\text{O}} \]

with \( \text{slope} = 1 \), \( \text{intercept} = -1024 \), and \( \rho_{\text{H}_2\text{O}} = 1 \text{g/cm}^3 \) being the density of water. We set the minimal density to \( 0.05 \text{g/cm}^3 \), which is about forty times the density of air. This speeds up the computations, while it does not change the results substantially. In addition, we set the density outside the body to \( \rho_{\text{H}_2\text{O}} \), since in the following test cases we are not interested in the dose outside the body. This helps to obtain accurate \( P_N \) solutions, since it softens the discontinuity at the boundary of the body and shields the radiation from the outer boundary of the computational domain.

As before, we compute a \( P_N \) and a MC solution and compare them using the \( \gamma \)-index (1.63) (with \( r = 3\text{mm} \) and \( R = 3\% \) of the maximal dose). Note again that we compute highly resolved \( P_N \) solutions, so that the Gibbs phenomena is hardly visible, the dose remains positive in the greater part of the computational domain, and the dose does not drop below \( -R \).

1.4.2.1 Two-Dimensional Test Case

We use a slice of the CT scan to validate the method in two spatial dimensions. The slice shows a cut through a human torso with a tumor in the lungs. The clinical target volume, the planning target volume, and some organs are segmented. With this, we define an initial condition corresponding to a maximal energy \( E_{\text{max}} = 21 \text{MeV} \) (see Figure 1.4).

The computational domain is about \( 43.0 \times 28.5 \text{cm}^2 \) with \( 440 \times 292 \) data points and the grid size is \( \Delta x = \Delta y = 0.977 \text{mm} \). We model the initial condition by delta functions in angle, Gaussian distributions with variance \( s = 0.5 \text{cm} \) in the moving directions, and characteristic functions with width \( 2 \text{cm} \) in the perpendicular direction. Note that we compute one solution for the \( P_N \) solution, whereas for the MC solution, we split the initial condition into two parts and compute two
Figure 1.4: Two-dimensional segmented CT scan: The locations and the directions of the two initial beams are marked in black.

separate solutions. Due to the linearity of the model and the linearity of the MC method, adding the solutions yields the same result as considering the complete initial condition at once.

The solution inside the body is shown in Figure 1.5: in (a) and (b), isodose curves of the normalized dose computed with $P_{17}$ and MC are shown; in (c) and (d), the $\gamma$-index is shown and regions where the $\gamma$-criterion is violated are marked by red borders. Taking all pixels with a significant dose (higher than 3% of the maximal dose in the reference solution), there are 4.8% not passing data values.
1. ELECTRON TRANSPORT

![Images of dose distributions](image_url)

**Figure 1.5:** Two-dimensional CT scan with $E_{\text{max}} = 21$ MeV: (a) and (b) show isodose curves computed with $P_{17}$ and MC; (c) and (d) show the $\gamma$-index limited to $[0, 2]$; regions with $\gamma(x) > 1$ are marked by red borders.

### 1.4.2.2 Three-Dimensional Test Case

We take a three-dimensional selection of the CT scan, which includes the tumor and some surrounding tissue. In the same way as for the two-dimensional test case, we use the PTV and the CTV to define an initial condition, which
corresponds to a maximal energy $E_{\text{max}} = 21$ MeV (see Figure 1.6).

This time, the computational domain is about $20.5 \times 14.6 \times 13.2$ cm$^3$. We use the full resolution with $210 \times 150 \times 44$ grid points for the MC solution, while we reduce the resolution in the $x$-$y$-plane to $105 \times 75$ data points for the computation of the $P_N$ solution. However, since the $P_N$ method uses staggered grids, we can still use the finer density map. This reduces the memory and the computational time of the method, but incorporates the full resolution. The initial condition is again modeled by a delta function in angle, Gaussian distributions with variance $s = 0.5$ cm in the moving directions, and now characteristic functions on a rectangle of $2 \times 1.8$ cm$^2$ in the perpendicular plane.

We compute the electron dose deposition with $P_{13}$ and Monte Carlo, and compare the solutions with the $\gamma$-index, where we interpolate the $P_{13}$ solution on the finer grid. For the inside of the body, isodose surfaces and violations of the $\gamma$-threshold are shown in Figure 1.7 and 1.8. Figure 1.7 shows a two-dimensional slice of the solution cut along $z = 6.4$, which is located in the middle of the two beams. In addition, Figure 1.8 shows the three-dimensional data set for two different views and one enlarged selection. We observe that the regions, which
1. ELECTRON TRANSPORT

violate the $\gamma$-criterion, are mainly to the sides of the beams, while the greater part of the inner regions are passing the criterion. Further, about 2.8% of all voxels with a significant dose (higher than 3% of the maximal dose in the reference solution) are not passing the $\gamma$-criterion.

Figure 1.7: Three-dimensional CT scan with $E_{\text{max}} = 21$ MeV: Cut along $z = 6.45$: (a) and (b) show isodose curves computed with $P_{13}$ and MC; (c) shows the $\gamma$-index limited to $[0, 2]$; regions with $\gamma(x) > 1$ are marked by red borders.
1.4 Numerical Results

Figure 1.8: Three-dimensional CT scan with $E_{\text{max}} = 21$ MeV: the first row is a view from the back right; the second row is a top view; and the third row is an enlarged selection of the top view. (a), (b), (d), and (e) show 5%, 10%, 25%, 50%, 70%, 80%, 90%, 95%, 98% isodose surfaces. (g) and (h) show 50%, 70%, 80%, 90%, 95%, 98% isodose surfaces. (c), (f), and (i) show $\gamma$-values greater than 1. The data is cut in the x-y plane.
1. ELECTRON TRANSPORT
Chapter 2

Convergence of Filtered Spherical Harmonics Equations

2.1 Introduction

The purpose of this chapter is to analyze the global convergence properties of the filtered spherical harmonic (FP\(N\)) equations [75, 98], a system of hyperbolic balance laws that are used to model radiation transport. These equations are a modification of the well-known spherical harmonic (P\(N\)) system, which we introduced in 1.3.1. As mentioned above, the P\(N\) system may exhibit oscillations around discontinuities and large gradients, in some cases causing negative particle concentrations. This fact is considered one of the major drawbacks of the P\(N\) method. The natural way to address deficiencies in the P\(N\) equations is to modify the spectral approximation; indeed, the P\(N\) approximation is just a linear combination of spherical harmonics and is not guaranteed to be positive.

There are a variety of nonlinear approximations that ensure positivity. For example, entropy-based methods [30, 80] yield, among other things, positive approximations for low-order expansions and have produced promising results in several applications (cf. 1.1). However, the implementation of high-order expansions is computationally expensive because of the complicated relationship between the coefficients and the moments of the expansion [1, 2].\(^1\) Positivity can

\(^1\)For a standard spectral method like the P\(N\) equations, this relationship is linear and often diagonal.
also be enforced directly through inequality constraints [46] or by penalty methods [35]. However, these approaches are also computationally expensive when compared to the $P_N$ equations. In addition, all of these methods still suffer from Gibbs-like phenomena around discontinuities.

Another method that uses a positive approximation of the transport solution is the quadrature method of moments (QMOM) [77]. Although the theoretical properties of this method are not well-understood, the solution algorithm is simple and relatively fast. There are several variations of QMOM. (See, for example, [74] and references therein.) One such variation, known as extended QMOM (EQMOM) has been used to simulate thermal radiative transfer in one-dimensional slab geometries [117]. However, its fidelity for multi-dimensional radiative transfer problems has yet to be evaluated.

A very simple modification of the $P_N$ method, which does not significantly increase the computational cost, is to dampen, or filter, the coefficients in the expansion. Filtering has been widely used in conjunction with spectral methods to handle instabilities and oscillations that often arise when simulating linear and nonlinear advection. There are many papers on filtering in the literature. We refer the interested reader to [13, 42, 50] for analysis, further background, and a host of additional references.

Filters were first applied to the $P_N$ equations in [75, 76]. It is well known [49] that by suppressing oscillations, filters can improve the local accuracy of spectral approximations around unresolved gradients. This property was leveraged in [75, 76] to significantly improve simulation results for several challenging, multi-dimensional problems in radiative transfer using expansions of relatively moderate size. In its original form, the filter was applied after each stage of a time integration scheme; unfortunately, this approach is not consistent with any continuum equation in the limit of a vanishing time step. However in [98], the strength of the filter was made to depend on the time step in such a way as to give a modified system of equations in the continuum limit. This new system contains an additional artificial scattering operator that is analogous to the artificial viscosity induced by filtering methods for spatial discretizations of hyperbolic equations [13]. As with the original discrete approach, the filter strength is still an adjustable parameter. However, because of the consistent implementation,
2.1 Introduction

The parameter can be tuned once on a relatively coarse mesh and then held fixed under mesh refinement. In addition, the modified equations are more amenable to numerical analysis than the original filtering procedure is.

The filtering approach does have some drawbacks. For example, filtering moments may actually increase the numerical error, especially in smooth regions where the $P_N$ equations already perform quite well. Such effects are most noticeable for low-order expansions \cite{35}. In addition, the approximation is not guaranteed to be positive. Finally, there is no optimal value for the filter strength; rather it may require adjustments for different problems. The “best value” of the filter strength depends on the local solution: suppressing oscillations in some regions causes a loss of accuracy in others. In spite of these drawbacks, the $FP_N$ equations are a promising tool for simulating radiation transport due to their efficiency, overall accuracy, and simplicity.

In this chapter, we analyze the global convergence properties of the $FP_N$ equations derived in \cite{98}. In particular, we show explicitly how the global $L^2$ convergence rate depends on the smoothness of the solution and the order of the filter. The analysis results are confirmed by numerical experiments. Such results are a helpful guide for practitioners who will use the equations for scientific simulation.

In what follows, we first introduce the $FP_N$ equations. Then, in section 2.2, we state and prove our main theorem, which gives the convergence rate of the filtered $P_N$ method. Finally, several numerical tests are presented in section 2.3, to compare the $FP_N$ equations with the original filtering introduced in \cite{75}, to confirm the dependence of the convergence rate on the regularity of the solution and the order of the filter, and to investigate the impact on applications in radiation therapy.

2.1.1 Filtered Spherical Harmonic (FP$_N$) Equations

The filtered spherical harmonics (FP$_N$) method was originally introduced as a modification in a time integration scheme \cite{75, 76}. After each time step, the spherical harmonic expansion is filtered, using for example a spectral spline filter.
2. CONVERGENCE OF $\text{FP}_N$ EQUATIONS

It was later shown \cite{98} that if the filtering function is raised to some strength parameter depending on the time step $\Delta t$, then the filtering procedure is consistent with a set of modified equations.

Before presenting the $\text{FP}_N$-equations, we first introduce the definition of a filter.

**Definition 2.1.** A filter of order $\alpha$ is a real-valued, non-negative, and monotonically decreasing function $f \in C^\alpha(\mathbb{R})$ that satisfies

(i) $f(0) = 1$,  
(ii) $f^{(a)}(0) = 0$, for $a = 1, \ldots, \alpha - 1$,  
(iii) $f^{(\alpha)}(0) \neq 0$.  

(2.1)

**Remark 2.2.** There are several slightly different definitions of a filter in the literature \cite{13, 20, 49, 50, 98, 116}. In \cite{50} a filter of order $\alpha$ is a real-valued, even function $f \in C^{\alpha-1}(\mathbb{R})$ that, in addition to conditions (i) and (ii) above, satisfies

(iv) $f(\eta) = 0$ for $|\eta| \geq 1$ and  
(v) $f^{(a)}(1) = 0$ for $a = 0, 1, \ldots, \alpha - 1$.  

(2.2)

Conditions (i) and (ii) are essential to every filter, but the other requirements may vary slightly. Commonly used filters are additionally strictly monotone decreasing on the interval $[0, 1]$ and smoother than required. At the same time, some filters do not satisfy conditions (iv) or (v). For example, neither the fourth order spherical spline filter, $f(\eta) = \frac{1}{\eta^{4+1}}$, nor the exponential filter of order $\alpha$:

$$f(\eta) = \exp(c\eta^\alpha) \quad \text{with} \quad c = \log(\varepsilon_M),$$  

(2.3)

where $\varepsilon_M$ is the machine accuracy, satisfy these conditions. Since filtering functions like the exponential filter are suitable for our purposes, we neglect conditions (iv) and (v) in the above definition. Condition (iii) has been added so that the filter order becomes a unique property.

**Assumption 2.3.** In what follows, we make the additional technical assumption that the filter $f$ satisfies

(vi) $f(\eta) \geq C(1 - \eta)^k$,  

$\eta \in [\eta_0, 1]$  

(2.4)

for some $k \geq 0$, some constant $C$, and some $\eta_0 \in (0, 1)$. This condition will be used in the proof of Theorem 2.8.
2.1 Introduction

Figure 2.1: Exponential filter of order $\alpha = 2, 4, 8, 16$.

Remark 2.4. Filters in the sense of Definition 2.1 that satisfy Assumption 2.3 include the exponential filter (see Figure 2.1). We use the exponential filter in computations.

In [98], the truncated filtered spherical harmonic expansion of a function $\Phi$ with expansion coefficients $\Phi^k_\ell$ is given by

$$
\sum_{\ell=0}^{N} \sum_{k=-\ell}^{\ell} \left( f \left( \frac{\ell}{N+1} \right) \right)^{\sigma_\ell \Delta t} \Phi^k_\ell \eta^k_\ell,
$$

where $f$ is the filter and $\sigma_\ell \Delta t$ is a filter strength that is tuned by the selection of the filtering cross-section $\sigma_\ell$ or, equivalently, the filter effective opacity

$$
f_{\text{eff}} = -\sigma_\ell \log \left( f \left( \frac{N}{N+1} \right) \right).
$$

An update of the solution components with a filter step can be written as

$$
\Phi^k_\ell(t + \Delta t, x) = \left( f \left( \frac{\ell}{N+1} \right) \right)^{\sigma_\ell \Delta t} \Phi^k_\ell(t, x)
= \Phi^k_\ell(t, x) + \Delta t \exp \left( \sigma_\ell \log \left( f \left( \frac{\ell}{N+1} \right) \Delta t \right) \right) - 1 \frac{\Delta t}{\Phi^k_\ell(t, x)},
$$

and has the formal limit $\Delta t \to 0$

$$
\partial_t \Phi^k_\ell = \sigma_\ell \log \left( f \left( \frac{\ell}{N+1} \right) \right) \Phi^k_\ell.
$$
2. CONVERGENCE OF FP$_N$ EQUATIONS

Thus, the dependence of the filter strength on $\Delta t$ allows one to express, in the formal limit $\Delta t \to 0$, the filtered spherical harmonic (FP$_N$) equations (1.43) in the following continuum form $[98]$:

$$\partial_t u_{\text{FP}_N} + A \cdot \nabla_x u_{\text{FP}_N} + \sigma_a u_{\text{FP}_N} + \sigma_s G u_{\text{FP}_N} + \sigma_l G l u_{\text{FP}_N} = q,$$  \hspace{1cm} (2.9)

where $G_l$ is a diagonal matrix with components $(G_l)_{(\ell,k),(\ell,k)} = -\log\left(f\left(\frac{\ell}{N+1}\right)\right)$.

Back in the abstract notation (introduced in (1.6)), (2.9) can be written as

$$\langle m T(\psi_{\text{FP}_N}) \rangle + \langle m Q_f(\psi_{\text{FP}_N}) \rangle = q,$$  \hspace{1cm} (2.10)

where the operator $Q_l$ depends on $G_l$:

$$Q_l(\Phi) = \sigma_l m^T G_l (m \Phi).$$  \hspace{1cm} (2.11)

In a way, the FP$_N$ equations can be viewed as a Galerkin method for the transport equation (3.1), but with an additional scattering operator that depends on $N$. A benefit of the continuum formulation is that it allows a user to tune the filter strength with a fairly coarse space-time mesh and then refine the solution.

2.2 Error Estimate

In this section we analyze the $L^2$ convergence of the FP$_N$ method. This will require assumptions on the regularity of the transport solution.

2.2.1 Preliminaries

We begin by defining the spaces and operators that will be used in the analysis, using $\Phi$ and $u$ to denote generic scalar and vector-valued functions.

- For any nonnegative integer $s$, $H^s(\mathbb{S}^2)$ denotes the Sobolev space on the unit sphere with norm

$$\|\Phi\|_{H^s(\mathbb{S}^2)} := \sqrt{\sum_{|\alpha| \leq s} \int_{\mathbb{S}^2} |D^\alpha \Phi(\Omega)|^2 \, d\Omega},$$  \hspace{1cm} (2.12)
2.2 Error Estimate

where the sum is over integer multi-indices $\alpha$ and the case $q = 0$ recovers the regular $L^2$ norm on $S^2$. An equivalent weighted $L^2$ norm can be derived using the Beltrami (surface Laplacian) operator

$$\mathcal{L} = \frac{\partial}{\partial \mu} \left((1 - \mu^2) \frac{\partial}{\partial \mu}\right) + \frac{1}{1 - \mu^2} \frac{\partial^2}{\partial \varphi^2},$$

for which the spherical harmonics are also eigenfunctions:

$$\mathcal{L} m^k_{\ell} = -\lambda^k_{\ell} m^k_{\ell}, \quad \lambda^k_{\ell} = \ell (\ell + 1).$$

Therefore, the expansion coefficients $\Phi^k_{\ell} := \langle m^k_{\ell} \Phi \rangle$ of any function $\Phi \in H^2(S^2)$ satisfy

$$\Phi^k_{\ell} = \langle m^k_{\ell} \Phi \rangle = \frac{1}{(-\lambda^k_{\ell})^s} \langle (\mathcal{L}^s m^k_{\ell}) \Phi \rangle = \frac{1}{(-\lambda^k_{\ell})^s} \langle m^k_{\ell} \mathcal{L}^s \Phi \rangle,$$

where the last equality in (2.15) follows from the differentiability of $\Phi$ and the fact that the Beltrami operator is self-adjoint. Therefore

$$\Phi \mapsto \left(\sum_{\ell=0}^{\infty} \sum_{k=0}^{\ell} \ell^s (\ell + 1)^s |\Phi^k_{\ell}|^2\right)^{1/2} = \left(\sum_{\ell=0}^{\infty} \ell^s (\ell + 1)^s \sum_{k=0}^{\ell} |\Phi^k_{\ell}|^2\right)^{1/2}$$

defines an equivalent norm on $H^s(S^2)$ that can then be extended to non-integer values of $s$ [42, p. 317]. This norm will be used in the proofs of Theorems 2.8 and 2.10.

- For vectors $u \in \mathbb{R}^n$ we define the Euclidean norm in the usual way: $\|u\|_{\mathbb{R}^n} = \sqrt{u^T u}$. Since $\langle m m^T \rangle = I$, it follows that $\|m^T u\|_{L^2(S^2)} = \|u\|_{\mathbb{R}^n}$.

- For functions of space and angle, we define the space $L^2(\mathbb{R}^3; H^s(S^2))$ by the norm

$$\|\Phi\|_{L^2(\mathbb{R}^3; H^s(S^2))} := \left(\sum_{|\alpha| \leq s} \int_{\mathbb{R}^3} \int_{S^2} |D^\alpha \Phi(t, x, \Omega)|^2 d\Omega dx\right)^{1/2}.$$  

For vector-valued functions of space, we define $L^2(\mathbb{R}^3; \mathbb{R}^n)$ by

$$\|u\|_{L^2(\mathbb{R}^3; \mathbb{R}^n)} := \left(\int_{\mathbb{R}^3} u(x)^T u(x) dx\right)^{1/2}.$$
2. CONVERGENCE OF FP \(_N\) EQUATIONS

- We add time dependence and define the space \( C^0([0, T]; L^2(\mathbb{R}^3; H^s(\mathbb{S}^2))) \) by

\[
\| \Phi \|_{C^0([0, T]; L^2(\mathbb{R}^3; H^s(\mathbb{S}^2)))} := \sup_{t \in [0, T]} \left( \sum_{|\alpha| \leq s} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} |D^\alpha \Phi(t, x, \Omega)|^2 \, d\Omega \, dx \right)^{1/2}
\]

and \( C^0([0, T]; L^2(\mathbb{R}^3; \mathbb{R}^n)) \) by

\[
\| u \|_{C^0([0, T]; L^2(\mathbb{R}^3; \mathbb{R}^n))} := \sup_{t \in [0, T]} \left( \int_{\mathbb{R}^3} u(t, x)^T u(t, x) \, dx \right)^{1/2}.
\]

- The mapping

\[
P_N \Phi = \mathbf{m}^T \langle \mathbf{m} \mathbf{m}^T \rangle^{-1} \langle \mathbf{m} \Phi \rangle = \mathbf{m}^T \langle \mathbf{m} \Phi \rangle
\]

is the \( L^2 \)-orthogonal projection of a generic function \( \Phi \in L^2(\mathbb{S}^2) \) onto \( P_N \).

For any non-negative integer \( \ell \),

\[
(P_\ell - P_{\ell-1}) \Phi = \mathbf{m}_\ell^T \langle \mathbf{m}_\ell \mathbf{m}_\ell^T \rangle^{-1} \langle \mathbf{m}_\ell \Phi \rangle = \mathbf{m}_\ell^T \langle \mathbf{m}_\ell \Phi \rangle
\]

is the \( L^2 \)-orthogonal projection of \( \Phi \) onto the space of spherical harmonics of degree \( \ell \). It is easy to show that

\[
\| \langle \mathbf{m}_\ell \Phi \rangle \|_{\mathbb{R}^{n_\ell}} \equiv \| (P_\ell - P_{\ell-1}) \Phi \|_{L^2(\mathbb{S}^2)} \leq \| (I - P_{\ell-1}) \Phi \|_{L^2(\mathbb{S}^2)}
\]

and the equivalent \( H^s \) norm in (2.16) is equal to \( \sum_{\ell=0}^{\infty} \ell^s (\ell + 1)^s \| \langle \mathbf{m}_\ell \Phi \rangle \|_{\mathbb{R}^{n_\ell}}^2 \).

A standard existence and uniqueness result for the transport equation is

**Theorem 2.5** ([25, Theorem XXI.2.3]). Let

\[
\sigma_s, \sigma_a \in L^\infty(\mathbb{R}^3) \quad \text{with} \quad \sigma_s, \sigma_a \geq 0.
\]

Let the initial condition \( \psi_0 \) be such that

\[
\psi_0 \in L^2(\mathbb{R}^3; L^2(\mathbb{S}^2)) \quad \text{and} \quad \Omega \cdot \nabla_x \psi_0 \in L^2(\mathbb{R}^3; L^2(\mathbb{S}^2)).
\]

Furthermore, let the source \( q \) satisfy

\[
q \in C^1([0, T]; L^2(\mathbb{R}^3; L^2(\mathbb{S}^2))).
\]

Then there exists a unique solution that satisfies

\[
u \in C^1([0, T]; L^2(\mathbb{R}^3; L^2(\mathbb{S}^2))) \quad \text{and} \quad \Omega \cdot \nabla_x \psi \in C^0([0, T]; L^2(\mathbb{R}^3; L^2(\mathbb{S}^2))).
\]

(2.27)
Remark 2.6. The assumptions of the theorem can be weakened to allow for initial conditions that are not in the domain of the advection operator $\Omega \cdot \nabla_x$. Then there exists a $C^0$ solution. However, the convergence results below require estimates on the spatial gradient of $\psi$ and additional regularity in angle. As Theorem 2.5 is a semigroup result, this could be achieved by additional regularity in the initial condition, see e.g. [14, Theorem 7.5]. In all our numerical examples, the initial condition is either zero or smooth.

The FP$_N$ equations are a symmetric hyperbolic system. Thus if the initial value $u_{FP_N}(0, \cdot) \in L^2(\mathbb{R}^3; \mathbb{R}^n)$, then a straight-forward Fourier analysis (see [106, Chapter 3]) shows that there is a unique solution $u_{FP_N} \in C^0([0,T]; L^2(\mathbb{R}^3; \mathbb{R}^n))$ and that, in the absence of a source, $\|u_{FP_N}\|_{L^2(\mathbb{R}^3; \mathbb{R}^n)}$ is bounded uniformly in time by the initial data. However, with the assumptions of Theorem 2.5, more can be said.

Theorem 2.7. Let $u_{FP_N}(0, x) = \langle m_\psi_0 \rangle$ and $q = \langle m q \rangle$, where $\psi_0$ and $q$ satisfy (2.25) and (2.26), respectively. Then there exists a unique solution that satisfies

$$u_{FP_N} \in C^1([0,T]; L^2(\mathbb{R}^3; \mathbb{R}^n)) \quad \text{and} \quad A \cdot \nabla_x u_{FP_N} \in C_0([0,T]; L^2(\mathbb{R}^3; \mathbb{R}^n)).$$

(2.28)

This result follows from standard semigroup theory (see, e.g., [14, Theorem 7.4]). The regularity provided by (2.28) is sufficient for the analysis that follows.

2.2.2 Main Result

We now state and prove the main convergence result.

Theorem 2.8. Assume the transport solution $\psi$ satisfies the additional regularity conditions

$$\psi \in C^0([0,T]; L^2(\mathbb{R}^3; H^s(\mathbb{S}^2))) \quad \text{and} \quad \partial_x^i \psi \in C^0([0,T]; L^2(\mathbb{R}^3; H^r(\mathbb{S}^2))), \quad (2.29)$$

for each $i \in \{1, 2, 3\}$, where $r$ and $q$ are positive constants. Let $\psi_{FP_N} = m^T u_{FP_N}$ be the reconstructed solution to (2.10). Then for any $t \in [0,T]$

$$\|\psi(t, \cdot, \cdot) - \psi_{FP_N}(t, \cdot, \cdot)\|_{L^2(\mathbb{R}^3; L^2(\mathbb{S}^2))} \leq \|\psi(t, \cdot, \cdot) - \mathcal{P}_N \psi(t, \cdot, \cdot)\|_{L^2(\mathbb{R}^3; L^2(\mathbb{S}^2))} + t \left\{ \|a_{N+1} \cdot \nabla_x \langle m_{N+1} \psi \rangle\|_{C^0([0,T]; L^2(\mathbb{R}^3; L^{2N+1}))} + \sigma I \|G_t \langle m \psi \rangle\|_{C^0([0,T]; L^2(\mathbb{R}^3; \mathbb{R}^n))} \right\}$$

(2.30)
and as $N \to \infty$, we have the following rates:\footnote{Throughout the paper, we use $C$ as a generic positive constant.}

\[ \| \psi(t, \cdot, \cdot) - \mathcal{P}_N \psi(t, \cdot, \cdot) \|_{L^2(\mathbb{R}^3; L^2(\mathbb{S}^2)))} \leq C N^{-s}, \quad (2.31a) \]
\[ \| a_{N+1} \cdot \nabla_x \langle m_{N+1} \psi \rangle \|_{C^0([0,T]; L^2(\mathbb{R}^3; L^2(\mathbb{R}^{2N+1})))} \leq C N^{-r}, \quad (2.31b) \]
\[ \| G_f \langle m \psi \rangle \|_{C^0([0,T]; L^2(\mathbb{R}^3; L^2(\mathbb{R}^n)))} \leq \begin{cases} C N^{-s+1/2}, & \alpha > s - \frac{1}{2}, \\ C N^{-\alpha + \epsilon} \forall \epsilon > 0, & \alpha \leq s - \frac{1}{2}. \end{cases} \quad (2.31c) \]

Theorem 2.8 allows one to predict the order of convergence of the $\mathcal{FP}_N$ solution as $N \to \infty$, depending on the order of the filter $\alpha$ and the smoothness of $\psi$. The term in (2.31a) is the projection error. We refer to the term in (2.31b) as the closure error, and the term in (2.31c) as the filter error.

The remainder of this section is dedicated to proving Theorem 2.8. The strategy is a Galerkin-type estimate similar to [102]. As is standard, we split the total error into the projection error and a remainder that is an element of $\mathbb{P}_N$

\[ \psi - \psi_{\mathcal{FP}_N} = (\psi - \mathcal{P}_N \psi) + m^T r, \quad (2.32) \]

where $r = \langle m(\mathcal{P}_N \psi - \psi_{\mathcal{FP}_N}) \rangle = \langle m^T \psi \rangle - u_{\mathcal{FP}_N}$ inherits the regularity in (2.28). The first step is to control $r$.

**Lemma 2.9.** Let $\psi$ be the exact solution to (3.1) and $\psi_{\mathcal{FP}_N} = m^T u_{\mathcal{FP}_N}$ be the solution to (2.10). Then for any $t \in [0,T]$, the residual vector $r$ in (2.32) satisfies the estimate

\[ \| r(t, \cdot) \|_{L^2(\mathbb{R}^3; \mathbb{R}^n)} \leq t \left\{ \| a_{N+1} \cdot \nabla_x \langle m_{N+1} \psi(t, \cdot, \cdot) \rangle \|_{C^0([0,T]; L^2(\mathbb{R}^3; L^{2N+1}))} + \sigma_t \| G_f \langle m \psi(t, \cdot, \cdot) \rangle \|_{C^0([0,T]; L^2(\mathbb{R}^3; \mathbb{R}^n))} \right\}. \quad (2.33) \]

**Proof.** The proof is essentially a calculation. We apply $\mathcal{T}$ to (2.32), then multiply by $m^T r$ and integrate in angle. Since $\psi$ is the exact transport solution, $\langle m \mathcal{T}(\psi) \rangle = q$. Combined with (2.10), this gives for the left-hand side of (2.32)

\[ \langle m^T r \mathcal{T}(\psi - \psi_{\mathcal{FP}_N}) \rangle = \langle m^T r \mathcal{Q}_t(\psi_{\mathcal{FP}_N}) \rangle = \langle m^T r \mathcal{Q}_t(\mathcal{P}_N \psi) \rangle - \langle m^T r \mathcal{Q}_t(m^T r) \rangle. \quad (2.34) \]
2.2 Error Estimate

Equating this to the result for the right-hand side, we find

\[ \langle m^T r J(m^T r) \rangle + \langle m^T r Q_t(m^T r) \rangle = -\langle m^T r J(\psi - P_N \psi) \rangle + \langle m^T r Q_t(P_N \psi) \rangle . \]  

(2.35)

The individual terms in (2.35) can be explicitly computed:

\[ \langle m^T r T(m^T r) \rangle = \frac{1}{2} \partial_t |r|^2 + \frac{1}{2} \nabla_x \cdot (\Omega |m^T r|^2) + \sigma_s |r|^2 + \sigma_r r^T G r \]  

(2.36a)

\[ \langle m^T r Q_f(m^T r) \rangle = \sigma_f r^T G_f r \]  

(2.36b)

\[ \langle m^T r T(\psi - P_N \psi) \rangle = \langle m^T r \Omega \cdot \nabla_x (\psi - P_N \psi) \rangle = r^T_N a_{N+1} \cdot \nabla_x \langle m_{N+1} \psi \rangle \]  

(2.36c)

\[ \langle m^T r Q_f(P_N \psi) \rangle = \sigma_f r^T G_f \langle m \psi \rangle \]  

(2.36d)

In (2.36c), we have used the notation defined in (1.44) and the recursion relation of the spherical harmonics in (1.40). After integration with respect to \( x \) we obtain

\[ \frac{1}{2} \partial_t \int_{R^3} |r|^2 \, dx = -\int_{R^3} r^T N a_{N+1} \cdot \nabla_x \langle m_{N+1} \psi \rangle \, dx \]

\[ + \sigma_f \int_{R^3} r^T G_f \langle m \psi \rangle \, dx - \int_{R^3} r^T M r \, dx , \]  

(2.37)

where \( M := \sigma_s I + \sigma_r G + \sigma_f G_f \) is positive definite. This implies that

\[ \partial_t \|r\|_{L^2(R^3;R^n)}^2 \leq 2 \|r\|_{L^2(R^3;R^n)} \left( \|a_{N+1} \cdot \nabla_x \langle m_{N+1} \psi \rangle \|_{L^2(R^3;R^{2N+1})} + \sigma_f \|G_f \langle m \psi \rangle \|_{L^2(R^3;R^n)} \right) , \]  

(2.38)

which can also be written as an integral inequality using

\[ r(0) = \langle m^T \psi(0, \cdot, \cdot) \rangle - u_{FP_N}(0, \cdot) = 0 . \]  

(2.39)

Then, the inequality of Ou-Yang (see for instance [5, Theorem 3.4.1]) leads to

\[ \|r\|_{L^2(R^3;R^n)} \leq \int_0^t \left( \|a_{N+1} \cdot \nabla_x \langle m_{N+1} \psi \rangle \|_{L^2(R^3;R^{2N+1})} + \sigma_f \|G_f \langle m \psi \rangle \|_{L^2(R^3;R^n)} \right) \, ds . \]  

(2.40)

and the result in (2.33) follows immediately. \( \square \)

The next step is to prove the convergence rates in (2.31). The projection error rate in (2.31a) is well known (see, for example, [42, 43]) and the result in (2.31b) follows a similar argument. We rederive these rates for completeness. Our convergence proof for the filter error follows the approach used in [49]. For all three cases, we utilize the equivalent \( H^q \) norm in (2.16) to simplify the presentation.
2. CONVERGENCE OF FP\textsubscript{N} EQUATIONS

Projection error  Using equation (2.15) and Parseval’s identity, the projection error satisfies

\[ \| \psi(t, \cdot, \cdot) - P_N \psi(t, \cdot, \cdot) \|^2_{L^2(\mathbb{R}^3; L^2(\mathbb{S}^2))} = \int_{\mathbb{R}^3} \sum_{\ell=N+1}^{\infty} \sum_{k=-\ell}^{\ell} |\psi^\ell_k(t, x)|^2 \, dx \]
\[ \leq \frac{1}{(N + 1)^{2s}} \int_{\mathbb{R}^3} \sum_{\ell=N+1}^{\infty} \sum_{k=-\ell}^{\ell} \ell^{2s} |\psi^\ell_k(t, x)|^2 \, dx \]
\[ \leq \frac{C}{N^{2r}} \| \psi(t, \cdot, \cdot) \|^2_{L^2(\mathbb{R}^3; H^s(\mathbb{S}^2))}. \]  

(2.41)

Closure error  We note first that for each \( i \in \{1, 2, 3\} \), the scalar elements \( a^{(i)}_{N+1} \) are all bounded independently of \( N \). Moreover, the number of nonzero components in any row or column is also bounded independently of \( N \) (see Appendix A for more details). Thus \( \| a^{(i)}_{N+1} \|_2^2 \leq \| a^{(i)}_{N+1} \|_1 \| a^{(i)}_{N+1} \|_\infty \) is uniformly bounded in \( N \), and under the conditions of Theorem 2.8,

\[ \| a_{N+1} \cdot \nabla_x \langle m_{N+1} \psi(t, \cdot, \cdot) \rangle \|^2_{L^2(\mathbb{R}^3; L^2)} \leq C \sum_{i=1}^{3} \| \langle m_{N+1} \partial_x \psi(t, \cdot, \cdot) \rangle \|^2_{L^2(\mathbb{R}^3; L^2)} \]
\[ = C \sum_{i=1}^{3} \| (P_{N+1} - P_N)(\partial_x \psi(t, \cdot, \cdot)) \|^2_{L^2(\mathbb{R}^3; L^2)} \]
\[ \leq C \sum_{i=1}^{3} \| (\mathcal{J} - P_N)(\partial_x \psi(t, \cdot, \cdot)) \|^2_{L^2(\mathbb{R}^3; L^2)} \]
\[ \leq \frac{C}{N^{2r}} \sum_{i=1}^{3} \| \partial_x \psi(t, \cdot, \cdot) \|^2_{L^2(\mathbb{R}^3; H^r(\mathbb{S}^2))}. \]  

(2.42)

where we have used (2.23) and the estimate of the projection error in (2.41), replacing \( s \) by \( r \) and \( \psi \) by \( \partial_x \psi \). Taking the supremum over all \( t \in [0, T] \) on both sides yields the desired rate.
2.2 Error Estimate

**Filter error** The filtering error satisfies

\[
\| \mathbf{G}_f \langle \mathbf{m}_f \psi(t, \cdot) \rangle \|_{L^2(\mathbb{R}^3; \mathbb{R}^n)}^2 = \sum_{\ell=0}^{N} \log^2 \left( f\left( \frac{\ell}{N+1} \right) \right) \| \mathbf{m}_f \psi(t, \cdot) \|_{L^2(\mathbb{R}^3; \mathbb{R}^n)}^2 \\
= \sum_{\ell=1}^{N} \log^2 \left( f\left( \frac{\ell}{N+1} \right) \right) \| \mathcal{P}_\ell - \mathcal{P}_{\ell-1} \psi(t, \cdot) \|_{L^2(\mathbb{R}^3; L^2(S^2))}^2 \\
= C \sum_{\ell=1}^{N} \log^2 \left( f\left( \frac{\ell}{N+1} \right) \right) \| (\mathcal{I} - \mathcal{P}_{\ell-1}) \psi(t, \cdot) \|_{L^2(\mathbb{R}^3; L^2(S^2))}^2 \\
\leq C \sum_{\ell=1}^{N} \log^2 \left( f\left( \frac{\ell}{N+1} \right) \right) \frac{1}{\ell^2} \| \psi(t, \cdot) \|_{L^2(\mathbb{R}^3; H^s(S^2))}^2 ,
\]

(2.43)

where we have again used (2.23) and the estimate of the projection error in (2.41).

It remains to find an estimate for the sum in the last term of (2.43). We follow the strategy in [49], approximating this sum with a Riemann integral and then determining conditions under which the integral is bounded. For any $\theta \leq 2s$,

\[
\sum_{\ell=1}^{N} \log^2 \left( f\left( \frac{\ell}{N+1} \right) \right) \frac{1}{\ell^2s} \leq \frac{1}{(N+1)^{\theta-1}} \frac{1}{N+1} \sum_{\ell=1}^{N} \log^2 \left( f\left( \frac{\ell}{N+1} \right) \right) \left( \frac{N+1}{\ell} \right)^{\theta} . \tag{2.44}
\]

The quantity $\Sigma$ is a Riemann sum corresponding to the integral

\[
\int_{0}^{1} \log^2 (f(\eta)) \eta^{-\theta} d\eta , \tag{2.45}
\]

where the integrand is singular at $\eta = 0$ and $\eta = 1$. The singularity at $\eta = 1$ is because of the logarithm and is integrable under Assumption 2.3. The singularity at $\eta = 0$ is polynomial; for it to be integrable, one must impose additional conditions relating $\theta$ and the filter order $\alpha$. A Taylor expansion of $f$ around $\eta = 0$ yields

\[
\log f(\eta) = \log \left( f(0) + \eta f'(0) + \cdots + \eta^\alpha \frac{f^{(\alpha)}(\xi)}{\alpha!} \right) = \log \left( 1 + \eta^\alpha \frac{f^{(\alpha)}(\xi)}{\alpha!} \right)
\]

(2.46)

for some $\xi \in [0, \eta]$. Thus $\log f(\eta) \leq C \eta^\alpha$ for $\eta$ positive, but sufficiently small. As a consequence, the singularity at $\eta = 0$ will be integrable if

\[
\theta < 2\alpha + 1 . \tag{2.47}
\]
2. CONVERGENCE OF FP\textsubscript{N} EQUATIONS

There are two cases:

Case 1: \( \alpha > s - \frac{1}{2} \). In this case, convergence is limited by the regularity of \( \psi \), and (2.47) is valid for all \( \theta \leq 2s \). In particular, for \( \theta = 2s \), we obtain from (2.44) the estimate \( \| G_f \langle m\psi \rangle \|_{C^0([0,T];L^2(\mathbb{R}^3;\mathbb{R}^n))} \leq CN^{-s+1/2} \).

Case 2: \( \alpha \leq s - \frac{1}{2} \). In this case, convergence is limited by the filter order, and (2.47) is valid only for \( \theta = 2\alpha + 1 - \delta \), where \( \delta > 0 \) is arbitrary. We obtain from (2.44) the estimate \( \| G_f \langle m\psi \rangle \|_{C^0([0,T];L^2(\mathbb{R}^3;\mathbb{R}^n))} \leq CN^{-\alpha+\epsilon} \), where \( \epsilon = \delta/2 \).

This completes the proof of Theorem 2.8.

2.2.3 A Sharper Estimate

The estimates of the closure filter errors in the previous section rely on the projection error estimate and the conservative bound of the projection \( \mathcal{P}_\ell - \mathcal{P}_{\ell-1} \) that is expressed in (2.23). However, in the numerical results below, we observe faster decay rates that lead to sharper overall estimates.

Theorem 2.10. In addition to the assumptions of Theorem 2.8, suppose that for all \( \ell > 0 \),

\[
\| \langle m\psi \rangle \|_{C^0([0,T];L^2(\mathbb{R}^3;\mathbb{R}^n)))} \leq \frac{C}{\ell^{s+1/2}} \quad \text{and} \\
\| m_\ell \partial_i \psi \|_{C^0([0,T];L^2(\mathbb{R}^3;\mathbb{R}^n)))} \leq \frac{C}{\ell^r+1/2}, \quad i \in \{1,2,3\}.
\]  

(2.48)

Then the rates in (2.31b) and (2.31c) of Theorem 2.8 can be sharpened to

\[
\| a_{N+1} \cdot \nabla_x \langle m_{N+1}\psi \rangle \|_{C^0([0,T];L^2(\mathbb{R}^3;\mathbb{R}^{2N+1})))} \leq CN^{-r-\frac{1}{2}},
\]  

(2.49)

\[
\| G_f \langle m\psi \rangle \|_{C^0([0,T];L^2(\mathbb{R}^3;\mathbb{R}^n)))} \leq \begin{cases} 
CN^{-s}, & \alpha > s, \\
CN^{-\alpha+\epsilon} & \forall \epsilon > 0, \quad \alpha \leq s.
\end{cases}
\]  

(2.50)

Proof. The proof is a trivial modification of the Theorem 2.8 proof. One simply needs to insert the bounds assumed in (2.10) into the appropriate place.

Remark 2.11. The decay rates in (2.10) cannot be deduced from the Sobolev index alone. However, given sufficient smoothness, a subsequence of the expansion
coefficients will always satisfy (2.10).\footnote{For example, let the sequence \( \{a_\ell\}_{\ell=0}^\infty \) be non-negative. Then the convergence of the series \( \sum_{\ell=0}^\infty \ell^{2s}a_\ell^2 \) does not imply \( |a_\ell| \leq C\ell^{-(s+1/2)} \); consider for instance the counterexample \( a_\ell = \begin{cases} \ell^{-(s+1/4)} & \text{for } \ell = 4^j, \quad j \in \{1,2,\ldots\}, \\ \frac{\ell^{-j}}{2^j} & \text{otherwise}. \end{cases} \) In fact, \( a_\ell \) does not necessarily need to be bounded by \( C\ell^{-(s+\gamma)} \) for any \( \gamma > 0 \). However, a subsequence of \( a_\ell \) will always decay faster than \( \ell^{-(s+1/2)} \).} On the other hand, the decay rates in (2.10) imply that

\[
\psi \in C^0([0,T]; L^2(\mathbb{R}^3; H^{\tilde{s}}(S^2))) \quad \text{and} \quad \partial_{x_i}\psi \in C^0([0,T]; L^2(\mathbb{R}^3; H^{\tilde{r}}(S^2))) ,
\]

(2.51)

respectively, for any \( \tilde{s} < s \) and \( \tilde{r} < r \).

## 2.3 Numerical Results

In this section, we compute \( \text{FP}_N \) solutions for several test cases in two spatial dimensions (five dimensions total, including time). Here it is assumed that \( \psi \) is constant in \( z \). Thus we fix \( z \) and, in an abuse of notation, set \( x = (x,y) \) and adapt the relevant definitions in Section 2.2.1 from \( \mathbb{R}^3 \) to \( \mathbb{R}^2 \).

For the computations we use the code StaRMAP [103]. To modify the code for the \( \text{FP}_N \) equations, we can either use the continuum form (2.9) and implement a discretization of these modified equations or we can use the filtered expansion (2.5) and apply the filter in each time step. We consider the following three cases:

- **“modified”**: We use (2.9) and compute the \( \text{FP}_N \) solution by adding the filtering term \( \sigma_f G_f \) to the scattering term \( \sigma_s G_s \), i.e., in this case the modified equation are discretized directly.

- **“full step”**: We use (2.5) and compute the \( \text{FP}_N \) solution by applying the filter after each time step.

- **“substep”**: We use (2.5) and compute the \( \text{FP}_N \) solution by applying the filter in each substep to the components that are updated in that substep.

To make the methods comparable, the filter effective opacity is divided by 2.
2. CONVERGENCE OF FP$_N$ EQUATIONS

Note that the “fullstep” and the “substep” are also discretizations of (2.9), but they introduce an additional first order splitting into the method. Therefore, compared to the first implementation, these methods lose the second order convergence of the StaRMAP method. However, these methods are closer to the original FP$_N$ method.

First, we compare the different implementations of the FP$_N$ equations with the original FP$_N$ method introduced in [75]. Second, we compute the numerical rate of convergence for several test cases to confirm the results of Theorem 2.10. Third, we apply the filter method to the electron transport problem, which is introduced in Chapter 1.

2.3.1 Comparison of FP$_N$ Methods

We consider two test cases. The first one is the line source test, which shows the potential of filtering methods. The second one is based on the method of manufactured solutions and investigates the space/time convergence rate of the method.

2.3.1.1 Line Source Test

In this test case, we compare the unfiltered solution with the different implementations of the filter and the original filtered spherical harmonics method introduced in [75], which we refer to as spherical spline filter. The spherical spline filter relies on a similar expansion as in (2.5)

$$\sum_{\ell=0}^{N} \sum_{k=-\ell}^{\ell} \frac{1}{1 + \alpha \ell^2 (\ell+1)^2} \phi^k_{\ell} m^k_{\ell} \text{ with } \alpha = \frac{\omega}{N^2 (\sigma L + N)^2} ,$$

where $L = 1$ is a characteristic length scale and $\omega = \frac{\Delta t}{\Delta x}$ is a constant. To implement this filter, we apply the filter after each time step, i.e., we multiply the solution components with the respective filter terms $\frac{1}{1 + \alpha \ell^2 (\ell+1)^2}$.

We investigate the behavior of the different methods for the line source test. The line source problem is a challenging problem for the P$_N$ method as the initial condition is discontinuous in space. However, we model the initial condition by a
narrow Gaussian, so that the method is still converging. All moments are initially zero, except the first:

\[ u^k_\ell = \begin{cases} 
  \frac{1}{4\pi(3.2 \times 10^{-4})} \exp \left( - \frac{x^2 + y^2}{4(3.2 \times 10^{-4})} \right), & k = \ell = 0, \\
  0, & \text{otherwise}.
\] (2.53)

The medium is purely scattering with \( \sigma_t = \sigma_s = 1 \) and the computational domain is \([-0.6, 0.6]^2\) with 150 \(\times\) 150 grid points. The solution is computed up to time \( t = 0.5 \) with \( P_{99} \) (as a reference solution), \( P_{13} \), and \( \text{FP}_{13} \) (“modified”, “full step”, “substep”, and spherical spline filter), where we use the exponential filter of order 4 with filter effective opacity \( \sigma_{\text{eff}} = 40 \). Then, we compute the absolute error in the scalar flux \( \phi = \langle \psi \rangle \) between the order \( N = 13 \) solutions and the \( P_{99} \) reference solution. As expected, Figure 2.2 shows that the \( P_{13} \) solution exhibits nonphysical oscillations (Gibbs phenomena), while the oscillations can be dampened by filtering. In particular, the solution of the \( \text{FP}_{13} \) solution remains positive, while the \( P_{13} \) oscillates into negative values. Moreover, the error of the \( \text{FP}_{13} \) solution is substantially lower away from the leading edge. In this example, the effect of the different implementations can hardly be seen and a similar result can be achieved with the spherical spline filter.

2.3.1.2 Manufactured Solutions

In this test case, the space/time convergence rate is investigated for different implementations of the \( \text{FP}_N \) equations using the method of manufactured solutions. This means we first define a solution, cross sections, and a filter, then we compute the corresponding source term using (2.9). Let the solution be given by

\[ u^k_\ell(t, x, y) = \begin{cases} 
  \exp(-t) \sin(2\pi x)^2, & \ell = 0, \\
  \frac{1}{\sqrt{3}} \exp(-t) \sin(2\pi x)^2, & \ell = 1, \\
  0, & \text{otherwise},
\] (2.54)

the scattering coefficients be given by

\[ \sigma_a(t, x, y) = \cos(2\pi y)t \quad \text{and} \quad \sigma_{sm}(t, x, y) = g^m \quad \text{with} \quad g = 0.9, \] (2.55)

and the used filter be the exponential filter of order 2 with filter effective opacity \( f_{\text{eff}} = 20 \). We compute the \( \text{FP}_N \) solution with \( N = 3 \) for the three different
2. CONVERGENCE OF FP\textsubscript{N} EQUATIONS

![Images of plots for P\textsubscript{99}, P\textsubscript{13}, FP\textsubscript{13}, and error plots.]

**Figure 2.2:** Linesource test. (a)-(c) Scalar flux $\phi = \langle \psi \rangle$ at $t = 0.5$: for $P\textsubscript{99}$ (a); for $P\textsubscript{13}$ (b); and FP\textsubscript{13} in the “modified” case (c). Negative values are shown in white. (d) Absolute error of the scalar flux along $y = 0$ computed with $P\textsubscript{13}$ and FP\textsubscript{13} with different filters.
implementations: “modified”, “full step”, and “substep”. To investigate the convergence, we use \( n^2 \) grid points with \( n = \{16, 32, 64, 128\} \) and compute the \( L^1 \), \( L^2 \), and \( L^\infty \) error by

\[
\hat{E}_n^1 = \sum_{\ell=0}^{\infty} \sum_{k=\ell}^{\infty} \frac{1}{n^2} \sum_{i,j=1}^{n} |(\psi_{FP_N})_\ell^k(x_i, y_j, t) - (\psi_F)_\ell^k(x_i, y_j, t)| ,
\]

\[
\hat{E}_n^2 = \sqrt{\sum_{\ell=0}^{\infty} \sum_{k=\ell}^{\infty} \frac{1}{n^2} \sum_{i,j=1}^{n} |(\psi_{FP_N})_\ell^k(x_i, y_j, t) - (\psi_F)_\ell^k(x_i, y_j, t)|^2 ,}
\]

\[
\hat{E}_n^\infty = \max_{i,j=1..n} \sum_{\ell=0}^{\infty} \sum_{k=\ell}^{\infty} |(\psi_{FP_N})_\ell^k(x_i, y_j, t) - (\psi_F)_\ell^k(x_i, y_j, t) | ,
\]

where \( \psi_F \) is the solution of the modified equation. Independent of the considered norm, the results indicate that the solutions converge to the modified equations (see Figure 2.3). As expected, the first implementation seems to be a second-order approximation, while the other two implementations seems to be first-order approximations. For the \( L^2 \) norm we consider the error in more detail, and compute also the error in the moments of degree \( \ell = 0, 1, \ldots, N \) (by omitting the sum over \( \ell \) in (2.56b)). Then, we compute convergence rates from two values by

\[
\hat{E}_n^{n_1} = -\frac{\log(\hat{E}_{n_1}) - \log(\hat{E}_{n_2})}{\log(n_1) - \log(n_2)}. \tag{2.57}
\]

The results, shown in Table 2.1, reconfirm second-order convergence to the modified equations for the “modified” case, and first-order convergence for the two other cases. Note that the spatial and the temporal resolution are coupled through a CFL condition, so that we are actually shrinking both at the same time.

### 2.3.2 Numerical Rates of Convergence

In this subsection, we compute the numerical rate of convergence for several test cases in two spatial dimensions. Beyond the adaptation of the relevant definitions in Section 2.2.1 from \( \mathbb{R}^3 \) to \( \mathbb{R}^2 \), the results of Theorems 2.8 and 2.10 are unchanged.

We consider three test cases, each of which is designed to reveal one of the rates in Theorem 2.10. The Gaussian test has a smooth solution, so we expect a
2. CONVERGENCE OF FP$_N$ EQUATIONS

Figure 2.3: Manufactured solution test: $L^1$, $L^2$, and $L^\infty$ error (in space and angle) for different resolutions and implementations of the filter method: (a) “modified”, (b) “full step”, and (c) “substep”.

(a) “modified”  
(b) “full step”  
(c) “substep”
2.3 Numerical Results

Table 2.1: Manufactured solution test: Approximate rate of convergence in the $L^2$ norm for different implementations of the FP$_N$ equations: “modified”, “full step”, and “substep”. The term $\tilde{E}^{n_2}_{n_1}$ is the convergence rate when going from $n_1$ to $n_2$ grid points. The degree defines the considered solution components.

| degree | “modified” $| \tilde{E}^{32}_{16} \tilde{E}^{64}_{32} \tilde{E}^{64}_{128}$ | “full step” $| \tilde{E}^{32}_{16} \tilde{E}^{64}_{32} \tilde{E}^{64}_{128}$ | “substep” $| \tilde{E}^{32}_{16} \tilde{E}^{64}_{32} \tilde{E}^{64}_{128}$ |
|--------|---------------------------------|---------------------------------|---------------------------------|
| 0      | 2.00 2.00 2.00                  | 1.97 1.85 1.55                  | 1.94 1.89 1.73                  |
| 1      | 1.99 2.00 2.00                  | 0.82 0.90 0.94                  | 1.37 1.12 1.01                  |
| 2      | 2.09 2.02 2.00                  | 1.74 1.44 1.16                  | 1.56 1.33 1.18                  |
| 3      | 2.10 2.03 2.01                  | 1.47 1.14 0.97                  | 1.50 1.31 1.17                  |
| all    | 2.00 2.00 2.00                  | 1.07 0.95 0.95                  | 1.64 1.33 1.10                  |

convergence rate determined by the filter order $\alpha$. In the lattice test, we numerically determine the Sobolev indexes $s$ and $r$ of the true solution and its derivative; the convergence of the FP$_N$ solution is determined by these indices. The hemisphere test has a solution that is smooth in space but discontinuous in angle; here the convergence order depends on the Sobolev index $s$.

For each test case, we compute both the P$_N$ and FP$_N$ solutions for certain (odd) values of $N$, e.g. $N = 2^k + 1$ for $k \in \{1, 2, \ldots\}$. The FP$_N$ solutions are computed using the exponential filter, cf. (2.3), using a fixed moderate value for the effective filter opacity $f_{\text{eff}} = 20$ and several different filter orders: $\alpha \in \{2, 4, 8, 16\}$. We use the direct implementation of the modified equations “modified”. The other implementations yield similar results, as can be seen in summary for the “substep” implementation in Table 2.2b. The final time for each problem is chosen so that the boundary does not affect the solution. Although Theorem 2.10 actually gives an estimate for the error in the $C^0([0, T])$-norm in time, we consider the error at a fixed final time. We however observe the expected rates.

We denote by $E_N$ and $R_N$ the norm of the total and projected error, respec-
2. CONVERGENCE OF FP\_N EQUATIONS

\[ E_N = \| \psi - \psi_N \|_{L^2(\mathbb{R}^2; L^2(S^2))} \]

\[ R_N = \| \mathcal{P} \psi - \psi_N \|_{L^2(\mathbb{R}^2; L^2(S^2))} \]

where \( \psi_N \) is either \( \psi_{FP_N} \) or \( \psi_{P_N} \). To estimate \( E_N \) and \( R_N \) we use the trapezoidal rule for the integrals and we approximate \( \psi \) by a \( P_{true} \) solution, with \( N_{true} \gg N \) sufficiently large. Thus the reference solution has a sharply higher angular resolution than \( \psi_{FP_N} \) and \( \psi_{P_N} \), but the same spatial resolution.

The error terms \( E_N \) and \( R_N \) are determined for different values of \( N \), and we estimate the rate of convergence from two values \( N_1 \) and \( N_2 \) by

\[ E_{N_2}^{N_1} = -\log(E_{N_1}) - \log(E_{N_2}) \left/ \log(N_1) - \log(N_2) \right. \]

\[ R_{N_2}^{N_1} = -\log(R_{N_1}) - \log(R_{N_2}) \left/ \log(N_1) - \log(N_2) \right. \]

(2.59)

The spatial resolution is chosen so that the space-time errors are negligibly small. To check this, we have performed a grid convergence study. A summary of the results is given in Table 2.2a and, with doubled spatial resolution, in Table 2.2c.

According to Theorem 2.10, the order of \( E_N \) and \( R_N \) are given by

\[ E_N \sim R_N = O(N^{-\min\{s,r+\frac{1}{2}\}}) \]

\[ E_N = O(N^{-\min\{s,r+\frac{1}{2}\}}), \quad R_N = O(N^{-(r+\frac{1}{2})}) \]

(2.60)

In particular, both depend on the regularity of the solution, which is given by the values \( s \) and \( r \). To obtain an estimate for these values, we estimate the order of decay for the moments of the solution and their differentials, cf. Lemma 2.9.

Thus, we approximate

\[ B_\ell := \| \langle m_\ell \psi \rangle \|_{L^2(\mathbb{R}^2, \mathbb{R}^n)} \quad \text{and} \quad D_\ell := \left\| \sum_{i=1}^{3} \langle m_\ell \partial_{x_i} \psi \rangle \right\|_{L^2(\mathbb{R}^2, \mathbb{R}^n)} \]

(2.61)

\footnote{It can be shown that since \( \psi \) is independent of \( z \), it is also invariant under the mapping \( \Omega_z \mapsto -\Omega_z \). As a consequence, moments with respect to \( m_\ell^{\pm} \) vanish whenever \( \ell + k \) is odd. The total number of nonzero moments that remain is \((N+1)(N+2)/2\).}
### 2.3 Numerical Results

<table>
<thead>
<tr>
<th>Filter order</th>
<th>Gaussian $\varepsilon^{65}_{33}$</th>
<th>Lattice $\varepsilon^{65}_{33}$</th>
<th>Hemisphere $\varepsilon^{33}_{17}$</th>
<th>Hemisphere $\varepsilon^{33}_{17}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.92</td>
<td>1.92</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>4</td>
<td>3.98</td>
<td>3.98</td>
<td>1.06</td>
<td>1.03</td>
</tr>
<tr>
<td>8</td>
<td>8.00</td>
<td>8.00</td>
<td>1.10</td>
<td>1.11</td>
</tr>
<tr>
<td>16</td>
<td>15.99</td>
<td>15.99</td>
<td>1.11</td>
<td>1.22</td>
</tr>
<tr>
<td>$\infty$</td>
<td>18.32</td>
<td>17.43</td>
<td>1.04</td>
<td>0.97</td>
</tr>
</tbody>
</table>

(a) FP$_N$ “modified”. Lower spatial resolution.

<table>
<thead>
<tr>
<th>Filter order</th>
<th>Gaussian $\varepsilon^{65}_{33}$</th>
<th>Lattice $\varepsilon^{65}_{33}$</th>
<th>Hemisphere $\varepsilon^{33}_{17}$</th>
<th>Hemisphere $\varepsilon^{33}_{17}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.92</td>
<td>1.92</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>4</td>
<td>3.98</td>
<td>3.98</td>
<td>1.05</td>
<td>1.07</td>
</tr>
<tr>
<td>8</td>
<td>8.00</td>
<td>8.00</td>
<td>1.04</td>
<td>1.13</td>
</tr>
<tr>
<td>16</td>
<td>15.99</td>
<td>15.99</td>
<td>1.03</td>
<td>1.20</td>
</tr>
<tr>
<td>$\infty$</td>
<td>17.87</td>
<td>16.94</td>
<td>0.99</td>
<td>0.96</td>
</tr>
</tbody>
</table>

(b) FP$_N$ “substep”. Higher spatial resolution.

<table>
<thead>
<tr>
<th>Filter order</th>
<th>Gaussian $\varepsilon^{65}_{33}$</th>
<th>Lattice $\varepsilon^{65}_{33}$</th>
<th>Hemisphere $\varepsilon^{33}_{17}$</th>
<th>Hemisphere $\varepsilon^{33}_{17}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.92</td>
<td>1.92</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>4</td>
<td>3.98</td>
<td>3.98</td>
<td>1.05</td>
<td>1.06</td>
</tr>
<tr>
<td>8</td>
<td>8.00</td>
<td>8.00</td>
<td>1.06</td>
<td>1.16</td>
</tr>
<tr>
<td>16</td>
<td>15.99</td>
<td>15.99</td>
<td>1.04</td>
<td>1.20</td>
</tr>
<tr>
<td>$\infty$</td>
<td>17.87</td>
<td>16.94</td>
<td>1.00</td>
<td>0.96</td>
</tr>
</tbody>
</table>

(c) FP$_N$ “modified”. Higher spatial resolution.

**Table 2.2:** Approximate order of convergence for different filter orders and test cases. Computed with different spatial resolutions and implementations of the filter: (a) spatial resolution $150 \times 150$ (Gaussian, hemisphere), $250 \times 250$ (lattice); (b) and (d) spatial resolution $300 \times 300$ (Gaussian, hemisphere), $500 \times 500$ (lattice); (a) and (c) filter implementation “modified”; (b) filter implementation “substep”. Filter order of $\infty$ means that no filter is applied.
2. CONVERGENCE OF FP\(_N\) EQUATIONS

by using again the reference solution P\(_{N_{\text{true}}}\) to estimate \(\psi\) and the trapezoidal rule to approximate the spatial integrals. As in (2.59), we use specific data points to define approximate decay rates

\[
B_{N_1}^{N_2} = -\frac{\log(B_{N_1}) - \log(B_{N_2})}{\log(N_1) - \log(N_2)} \quad \text{and} \quad D_{N_1}^{N_2} = -\frac{\log(D_{N_1}) - \log(D_{N_2})}{\log(N_1) - \log(N_2)}.
\]  

(2.62)

Using (2.10), we approximate

\[
s \approx B_{N_1}^{N_2} - 0.5 \quad \text{and} \quad r \approx D_{N_1}^{N_2} - 0.5.
\]

(2.63)

2.3.2.1 Gaussian Test

This first test case has smooth input data to show that the convergence order of the FP\(_N\) solution is bounded by the filter order \(\alpha\). All moments are initially zero, except the first

\[
u_{k}^\ell = \begin{cases} 
\frac{1}{4\pi(3.2\times10^{-3})} \exp \left( -\frac{x^2+y^2}{4(3.2\times10^{-3})} \right), & k = \ell = 0, \\
0, & \text{otherwise}.
\end{cases}
\]

(2.64)

The medium is a void, with \(\sigma_t = \sigma_s = 0\). The computational domain is \([-0.6, 0.6] \times [-0.6, 0.6]\) and the solution is computed on a 300 \(\times\) 300 spatial grid (or 150 \(\times\) 150 for Table 2.2) up to time \(t = 0.4\). Errors are computed using a P\(_{99}\) reference solution. Since the initial condition is smooth, we expect spectral convergence for the P\(_N\) solution and a convergence order equal to the filter order for both \(E\) and \(R\). This behavior is clearly confirmed in Figure 2.4 and Table 2.3, where we observe that the convergence order increases until it reaches the filter order or until the error reaches machine precision.

2.3.2.2 Lattice Test

The lattice test was first proposed in [16]. It contains source terms and material cross-sections that are discontinuous in space. Due to the coupling of the spatial and the angular variable, this leads to a loss of regularity in the angular variable as well. Thus it is expected that the convergence order for \(E\) and \(R\) is determined by \(s\) and \(r\).
2.3 Numerical Results

Figure 2.4: Gauss test computed with $P^{129}$: (a) Log-log plot of the sequences $B_\ell$ and $D_\ell$ vs. the order $\ell$; (b) Log-log plot of the error terms $E_N$ as a function of the moment order approximation $N$.

Table 2.3: Gauss test: Filter order of $\infty$ means that no filter is applied. The term $\xi_{N_1}^{N_2}$ is the convergence rate when going from $N_1$ to $N_2$.

<table>
<thead>
<tr>
<th>Filter order</th>
<th>$\xi_{3}^{N}$</th>
<th>$\xi_{5}^{N}$</th>
<th>$\xi_{9}^{17}$</th>
<th>$\xi_{17}^{33}$</th>
<th>$\xi_{33}^{65}$</th>
<th>$R_{3}^{5}$</th>
<th>$R_{5}^{9}$</th>
<th>$R_{9}^{17}$</th>
<th>$R_{17}^{33}$</th>
<th>$R_{33}^{65}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.47</td>
<td>0.84</td>
<td>1.33</td>
<td>1.73</td>
<td>1.92</td>
<td>0.03</td>
<td>0.52</td>
<td>1.28</td>
<td>1.73</td>
<td>1.92</td>
</tr>
<tr>
<td>4</td>
<td>0.76</td>
<td>1.57</td>
<td>2.89</td>
<td>3.79</td>
<td>3.98</td>
<td>0.36</td>
<td>1.05</td>
<td>2.68</td>
<td>3.79</td>
<td>3.98</td>
</tr>
<tr>
<td>8</td>
<td>1.01</td>
<td>2.13</td>
<td>4.94</td>
<td>7.62</td>
<td>8.00</td>
<td>1.14</td>
<td>1.68</td>
<td>4.31</td>
<td>7.54</td>
<td>8.00</td>
</tr>
<tr>
<td>16</td>
<td>1.06</td>
<td>2.40</td>
<td>6.15</td>
<td>13.63</td>
<td>15.99</td>
<td>1.30</td>
<td>2.63</td>
<td>5.53</td>
<td>12.96</td>
<td>15.99</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.02</td>
<td>2.41</td>
<td>6.48</td>
<td>18.22</td>
<td>17.87</td>
<td>1.10</td>
<td>2.55</td>
<td>6.71</td>
<td>18.55</td>
<td>16.94</td>
</tr>
</tbody>
</table>

For this problem, the computational domain is a $7 \times 7$ square that is divided into smaller squares of length one. There is an isotropic source in the middle of the domain and a mixture of purely scattering and purely absorbing squares surrounding it (c.f. Figure 2.5a). The $P_N$ and $FP_N$ solutions are computed on a $500 \times 500$ spatial grid (or $250 \times 250$ for Table 2.2) up to time $t = 2.8$. We use $N_{\text{true}} = 129$ (i.e. a system with more than $2.1 \times 10^{10}$ degrees of freedom) to compute a reference solution and estimate convergence rates.

We estimate the decay rates of $\{B_\ell\}_{\ell=0}^\infty$ and $\{D_\ell\}_{\ell=0}^\infty$ which, due to the lack of regularity in the solution, converge very slowly. As a consequence numerical estimates of these values are not always reliable. In fact, we observe that they
2. CONVERGENCE OF FP\textsubscript{N} EQUATIONS

Figure 2.5: Lattice test: (a) Material coefficients: isotropic source (white square) \( S = 1 \); purely scattering \( \sigma_t = \sigma_a = 1 \) (orange and white squares); purely absorbing \( \sigma_t = \sigma_a = 10 \) (black squares). (b) Scalar flux \( \phi = \langle \psi \rangle \) at \( t = 2.8 \) for \( P_{129} \), computed with 500 \times 500 grid points. The values are plotted in a logarithmic scale and limited to seven orders of magnitude.

...depend on the parity of both \( \ell \) and the value of \( N_{\text{true}} \). This behavior can be observed in Figure 2.6. To address it, we approximate \( B_\ell \) and \( D_\ell \) using both \( P_{128} \) and \( P_{129} \) numerical solutions and then determine a cutoff \( \ell_{\text{max}} \) so that the relative difference in the two resulting approximations is acceptable for all \( \ell \leq \ell_{\text{max}} \). For a relative difference of three percent, \( \ell_{\text{max}} = 62 \) for \( B_\ell \) and \( \ell_{\text{max}} = 38 \) for \( D_\ell \) are sufficient. In this range, the even and odd subsequences \( \{B_\ell\}_{\ell=0}^{\infty} \) and \( \{D_\ell\}_{\ell=0}^{\infty} \) decay monotonically at a fairly constant rate. We use the slightly slower rates given by the odd subsequences: \( B^{33}_{17} = 1.55 \) and \( D^{33}_{17} = 0.77 \) (see Table 2.4). According to (2.60), we expect \( E_N \) and \( R_N \) to both converge at a rate of \( r + \frac{1}{2} \) when the filter is on and to converge at rates \( s \) and \( r + \frac{1}{2} \), respectively, when the filter is off. Using (2.63), we approximate \( s \approx B^{33}_{17} - \frac{1}{2} = 1.05 \) and \( r + \frac{1}{2} \approx D^{33}_{17} = 0.77 \). However from Table 2.5 and Figure 2.7, the convergence rate is roughly one in all cases, meaning that the observed convergence is actually slightly better than any of the estimates that depend on \( r \).
2.3 Numerical Results

![Graphs](image)

**Figure 2.6:** Lattice test: Log-log plot of the sequences $B_\ell$ and $D_\ell$ vs. the order $\ell$. The dashed lines indicate until which point the relative difference between the sequences computed with $P_{128}$ and $P_{129}$ differs by no more than 3%.
2. CONVERGENCE OF FPₙ EQUATIONS

<table>
<thead>
<tr>
<th>(N₁, N₂)</th>
<th>Bₙ₁^{N₂}</th>
<th>Dₙ₁^{N₂}</th>
<th>(N₁, N₂)</th>
<th>Bₙ₁^{N₂}</th>
<th>Dₙ₁^{N₂}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 4)</td>
<td>1.32</td>
<td>0.62</td>
<td>(3, 5)</td>
<td>1.62</td>
<td>0.78</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>1.82</td>
<td>0.82</td>
<td>(5, 9)</td>
<td>1.84</td>
<td>0.82</td>
</tr>
<tr>
<td>(8, 16)</td>
<td>1.52</td>
<td>0.83</td>
<td>(9, 17)</td>
<td>1.49</td>
<td>0.80</td>
</tr>
<tr>
<td>(16, 32)</td>
<td>1.58</td>
<td>0.87</td>
<td>(17, 33)</td>
<td>1.55</td>
<td>0.77</td>
</tr>
</tbody>
</table>

(a) even order moments

(b) odd order moments

Table 2.4: Lattice test: Approximate decay rates of the sequence of the moments Bₙ (and the moments of the differentials Dₙ). (a) Even order moments N₂ = 2ᵏ+¹ vs. N₁ = 2ᵏ and (b) odd order moments N₂ = 2ᵏ+¹ + 1 vs. N₁ = 2ᵏ + 1 with k = 1, … , 5 (Computed with P₁₂⁹).

![Log-log plot of the error terms Eₙ and Rₙ as a function of the moment order approximation N.](image)

**Figure 2.7:** Lattice test: Log-log plot of the error terms Eₙ and Rₙ as a function of the moment order approximation N.

<table>
<thead>
<tr>
<th>Filter order</th>
<th>E₅³</th>
<th>E₉⁵</th>
<th>E₉¹⁷</th>
<th>E₃³¹⁷</th>
<th>R₅³</th>
<th>R₅⁹</th>
<th>R₁⁷</th>
<th>R₃³¹⁷</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.89</td>
<td>0.80</td>
<td>0.94</td>
<td>1.05</td>
<td>0.86</td>
<td>0.78</td>
<td>0.93</td>
<td>1.05</td>
</tr>
<tr>
<td>4</td>
<td>1.01</td>
<td>1.15</td>
<td>1.13</td>
<td>1.05</td>
<td>0.97</td>
<td>1.21</td>
<td>1.21</td>
<td>1.06</td>
</tr>
<tr>
<td>8</td>
<td>1.20</td>
<td>1.22</td>
<td>1.04</td>
<td>1.06</td>
<td>1.31</td>
<td>1.55</td>
<td>1.14</td>
<td>1.16</td>
</tr>
<tr>
<td>16</td>
<td>1.61</td>
<td>1.30</td>
<td>1.03</td>
<td>1.04</td>
<td>2.10</td>
<td>2.11</td>
<td>1.22</td>
<td>1.20</td>
</tr>
<tr>
<td>∞</td>
<td>1.10</td>
<td>0.95</td>
<td>0.98</td>
<td>1.00</td>
<td>1.10</td>
<td>0.85</td>
<td>0.95</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 2.5: Lattice Test: Filter order of ∞ means that no filter is applied. The term Eₙᵣᵢ is the convergence rate when going from Nᵣ to Nᵢ.
2.3 Numerical Results

2.3.2.3 Hemisphere Test

In our final test, we consider a problem with input data that is smooth with respect to the spatial variable but a source term that is discontinuous in the angle variable. As a consequence, we expect that \( r = s < \alpha \) so that the convergence order does not depend on \( \alpha \).

The domain is a \( 1.2 \times 1.2 \) square centered at the origin with a \( 300 \times 300 \) spatial grid (or \( 150 \times 150 \) for Table 2.2). The final time is \( t = 0.3 \). There is no material medium (i.e. \( \sigma_t = 0 \)) and the initial condition is zero everywhere. The source term \( q \) is

\[
q(t, x, \Omega) = W(x)\chi_{\mathbb{R}^+}(\Omega_x), \tag{2.65}
\]

where \( W(x) = \frac{1}{4\pi \times 10^{-3}} \exp \left( -\frac{x^2 + y^2}{4 \times 10^{-3}} \right) \) and \( \chi_{\mathbb{R}^+} \) is the characteristic function over \( \mathbb{R}^+ \). Since \( q \) only depends on \( x \) and \( \Omega_x = \Omega \cdot e_x \) with \( e_x = (1, 0, 0) \), its expansion in spherical harmonics is

\[
q(t, x, \Omega) = W(x) \sum_{\ell=0}^{\infty} \sum_{k=-\ell}^{\ell} s_{\ell} m^k_\ell(e_x) m^k_\ell(\Omega) \tag{2.66}
\]

with

\[
s_{\ell} = 2\pi \int_0^1 P_{\ell}(\eta) d\eta = 2\pi \frac{P_{\ell-1}(0) - P_{\ell+1}(0)}{2\ell + 1}. \tag{2.67}
\]

In particular, all the moments with \( \ell \) even are zero.

As before, we determine the smoothness of the exact solution numerically. To this end, we compute P_N solutions which are highly resolved in the angular variable. Again, we observe parity in \( B_\ell \) and \( D_\ell \) with respect to \( \ell \). However, unlike the lattice problem, the values do not depend on the parity of \( N_{\text{true}} \). This fact is confirmed in Figure 2.8, which shows values of \( B_\ell \) and \( D_\ell \) approximated with \( N_{\text{true}} = 98 \) and \( N_{\text{true}} = 99 \), respectively. For \( \ell = 0, \ldots, 75 \) the values of \( B_\ell \) (as well as \( D_\ell \)) with \( N_{\text{true}} = 98 \) and \( N_{\text{true}} = 99 \) coincide up to machine precision. Figure 2.8 also shows that the odd subsequences of \( B_\ell \) and \( D_\ell \) have larger values than the even ones. This can be explained by the form of the source, whose even order moments are identically zero; nonzero values are only generated by spatial gradients of the odd moments. Although the values of the even and odd subsequences differ, the order of the decay rates of the sub-sequences are almost the same in the range \( 16 \leq \ell \leq 90 \). In particular, (2.63) with \( N_1 = 17 \) and
2. CONVERGENCE OF \(\text{FP}_N\) EQUATIONS

\(N_2 = 33\) (cf. Table 2.6) yields the estimate \(s \approx r \approx \frac{1}{2}\). According to (2.60), the order of the error terms are given by \(\min\{s, r + \frac{1}{2}, \alpha\} = s \approx \frac{1}{2}\), except the order of the unfiltered \(P_N\) error term \(R_N\), which is approximately one. These predictions match the observed orders of convergence given in Table 2.7 and Figure 2.9.

\[\begin{align*}
B_{\ell} & (\ell \text{ odd}) \\
B_{\ell} & (\ell \text{ even}) \\
D_{\ell} & (\ell \text{ odd}) \\
D_{\ell} & (\ell \text{ even})
\end{align*}\]

Figure 2.8: Hemisphere test: Log-log plot of the sequences \(B_{\ell}\) and \(D_{\ell}\) against the order \(\ell\). The dashed lines indicate until which point the sequences computed with \(P_{98}\) and \(P_{99}\) coincide up to machine precision.

2.3.3 \(\text{FP}_N\) for Electron Transport

In this subsection, we apply the filter to the continuous slowing down approximation for electron transport. In section 1.4 highly resolved \(P_N\) solutions are compared to Monte Carlo solutions. Here, we investigate if we can decrease the moment order approximation \(N\) by using a filter to dampen the arising oscillations. Note that the \(\text{FP}_N\) equations of the continuous slowing down equation (1.26) can simply be obtained by adding a filter term \(\sigma_f G_f\) to (1.58).

2.3.3.1 Water Phantom

We consider again the water phantom with \(E_{\text{max}} = 5\text{MeV}\) from 1.4.1.1, where it has been shown that the \(P_{21}\) solution yields good results compared to a Monte Carlo solution. Now, we decrease the moment order approximation from \(N = 21\) to \(N = 5\) and compute various \(\text{FP}_N\) solutions. Figure 2.10 shows the \(\text{FP}_N\)
2.3 Numerical Results

<table>
<thead>
<tr>
<th>((N_1, N_2))</th>
<th>(B_{N_1}^{N_2})</th>
<th>(D_{N_1}^{N_2})</th>
<th>((N_1, N_2))</th>
<th>(B_{N_1}^{N_2})</th>
<th>(D_{N_1}^{N_2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 4)</td>
<td>1.02</td>
<td>0.71</td>
<td>(3, 5)</td>
<td>1.18</td>
<td>1.05</td>
</tr>
<tr>
<td>(4, 8)</td>
<td>1.58</td>
<td>1.67</td>
<td>(5, 9)</td>
<td>1.25</td>
<td>1.57</td>
</tr>
<tr>
<td>(8, 16)</td>
<td>1.15</td>
<td>1.35</td>
<td>(9, 17)</td>
<td>1.02</td>
<td>1.14</td>
</tr>
<tr>
<td>(16, 32)</td>
<td>1.01</td>
<td>1.03</td>
<td>(17, 33)</td>
<td>0.99</td>
<td>1.01</td>
</tr>
<tr>
<td>(32, 64)</td>
<td>1.00</td>
<td>1.00</td>
<td>(33, 65)</td>
<td>0.99</td>
<td>1.00</td>
</tr>
</tbody>
</table>

(a) even order moments (b) odd order moments

Table 2.6: Hemisphere test: Approximated decay rates of the sequence of the moments \(B_\ell\) (and the moments of the differentials \(D_\ell\)), (a) Even order moments \(N_2 = 2^k + 1\) vs. \(N_1 = 2^k\) and (b) odd order moments \(N_2 = 2^{k+1} + 1\) vs. \(N_1 = 2^k + 1\) with \(k = 1, \ldots, 5\) (Computed with \(P_{99}\)).

![Log-log plot of the error terms \(E_N\) and \(R_N\) as a function of the moment order approximation \(N\).](image)

Figure 2.9: Hemisphere test: Log-log plot of the error terms \(E_N\) and \(R_N\) as a function of the moment order approximation \(N\).

<table>
<thead>
<tr>
<th>Filter order</th>
<th>(\varepsilon_3^5)</th>
<th>(\varepsilon_5^9)</th>
<th>(\varepsilon_9^{17})</th>
<th>(\varepsilon_{17}^{33})</th>
<th>(R_3^5)</th>
<th>(R_5^9)</th>
<th>(R_9^{17})</th>
<th>(R_{17}^{33})</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.55</td>
<td>0.58</td>
<td>0.57</td>
<td>0.58</td>
<td>0.44</td>
<td>0.61</td>
<td>0.59</td>
<td>0.52</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
<td>0.60</td>
<td>0.55</td>
<td>0.61</td>
<td>0.70</td>
<td>0.70</td>
<td>0.57</td>
<td>0.52</td>
</tr>
<tr>
<td>8</td>
<td>0.75</td>
<td>0.61</td>
<td>0.56</td>
<td>0.63</td>
<td>1.05</td>
<td>0.83</td>
<td>0.61</td>
<td>0.56</td>
</tr>
<tr>
<td>16</td>
<td>0.77</td>
<td>0.64</td>
<td>0.57</td>
<td>0.64</td>
<td>1.13</td>
<td>1.03</td>
<td>0.79</td>
<td>0.64</td>
</tr>
<tr>
<td>(\infty)</td>
<td>0.71</td>
<td>0.59</td>
<td>0.56</td>
<td>0.65</td>
<td>1.33</td>
<td>1.26</td>
<td>0.99</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 2.7: Hemisphere test: Filter order of \(\infty\) means that no filter is applied. The term \(\varepsilon_1^{N_2}_{N_1}\) is the convergence rate when going from \(N_1\) to \(N_2\).
2. CONVERGENCE OF FPₙ EQUATIONS

![Graphs showing dose distribution for different filter orders and effective opacities.]

(a) $\alpha = 2$

(b) $\alpha = 4$

Figure 2.10: Water phantom with initial energy entering from the left with energy $E_{\text{max}} = 5\text{MeV}$. Solution computed with MC (as a reference), $P_5$, and FP₅ with different filter orders $\alpha = 2, 4$ and filter effective opacities $f_{\text{eff}} = 2, 4$.

solution with filter order $\alpha = 2, 4$ and filter effective opacity $f_{\text{eff}} = 2, 4$. Note that compared to the $P_5$ solution, the FP₅ solution gains accuracy in the build up region, while it loses accuracy at the moving front.

2.3.3.2 CT Scan

We consider the two-dimensional example from 1.4.2.1, where we have compared the $P_N$ solution with $N = 17$ to a Monte Carlo solution. Here, we compute the $P_N$ solution with $N = 11$ and apply a weak filter to dampen the arising oscillations. A comparison between the $P_{11}$ and the FP_{11} solution with filter order $\alpha = 4$ and filter effective opacity $f_{\text{eff}} = 1$ is shown in Figure 2.11. At the first glance, the FPₙ solution seems to be more reasonable as it is less oscillatory and the solution remains essentially positive (the dose does not drop below the threshold of $-3\%$ of the maximal dose). However, if we compute the $\gamma$-index with respect to the Monte Carlo solution, we observe that the $\gamma$-criterion is more often violated for the FPₙ solution. In particular, the $\gamma$-index increases in the greater part of the inner domain and only decreases in the area to the sides of the source and in more distant regions from the source. Note that the $\gamma$-index rises further for a stronger filter, while we lose the damping effect if we use an even weaker filter.
2.3 Numerical Results

Figure 2.11: Two dimensional CT scan with $E_{\text{max}} = 21\text{MeV}$: (a) and (b) show isodose curves computed with $P_{11}$ and $FP_{11}$ with negative values (less than $-3\%$ of the maximal dose) marked by black borders; (c) and (d) show the $\gamma$-index with respect to the MC solution ($\gamma(x) > 1$ is marked by red borders).
2. CONVERGENCE OF $FP_N$ EQUATIONS
Chapter 3

An Asymptotic Preserving
Two-Dimensional Staggered Grid
Method

3.1 Introduction

In many applications, the material parameters are strongly varying from a kinetic
to a diffusive regime. For instance, in radiation therapy (cf. Chapter 1), the tissue
is in some regions extremely dense and in other regions rarefied. In this case, a
good overall accuracy is a challenge for numerical schemes, which is addressed by
the asymptotic preserving schemes.

In this chapter, we propose a two-dimensional asymptotic preserving scheme
for linear transport equations with diffusive scalings. The diffuse scaling arises
from the linear transport equation when the scattering dominates the propagation
process. In general, the model in scaled variables can be written as [60]

\[ \varepsilon \partial_t \psi + \Omega \cdot \nabla_x \psi = \frac{1}{\varepsilon} \left[ \frac{\sigma_s}{2\pi} \int_{S^1} \psi d\Omega' - \sigma_t \psi \right] + \varepsilon q, \]  

(3.1)

where \( \psi(t, x, \Omega) \) denotes the probability density distribution depending on time
\( t \), position \( x = (x, y) \in \mathbb{R}^2 \), and direction of velocity \( \Omega = (\Omega_x, \Omega_y) = (\xi, \eta) \in S^1 = \{ (\xi, \eta) : -1 \leq \xi, \eta \leq 1, \xi^2 + \eta^2 = 1 \} \) (two-dimensional flatland model – the
extension to three dimensions is straightforward). Moreover, \( \sigma_t = \sigma_s + \varepsilon^2 \sigma_a \) is
3. AP STAGGERED GRID METHOD

the total transport coefficient, $\sigma_s$ is the scattering coefficient, $\sigma_a$ is the absorption coefficient, and $q$ is a $\Omega$-independent source term. It is well-known that the limiting equation ($\varepsilon \to 0$) of (3.1) is the diffusion equation:

$$\partial_t \rho = \frac{1}{2} \nabla_x \cdot \left( \frac{1}{\sigma_s} \nabla_x \rho \right) - \sigma_a \rho + q,$$

(3.2)

where $\rho(t, x) = \frac{1}{4\pi} \int_{S^1} \psi(t, x, \Omega) \, d\Omega$, but for completeness we briefly describe the derivation at the end of this section.

In many applications, the scaling parameter of the transport equation $\varepsilon$ (mean free path) may differ in several orders of magnitude, ranging from the rarefied kinetic regime to the hydrodynamic diffusive regime. When $\varepsilon$ is small, in the diffusive regime, the equation becomes numerically stiff, which leads to numerical challenges: straightforward explicit implementations lead to high computational costs in the diffusive regimes; fully implicit schemes could be difficult to implement [27]; multiscale multiphysics domain-decomposition approaches, which couple models at different scales, have difficulties in the transition zones, since they need to transfer data from one scale to another [54]. Thus, it is desirable to develop schemes which are suitable for all regimes (no domain-decomposition), but do not require a resolved grid in space and small time compared to the mean free path. This is the objective of asymptotic preserving (AP) schemes.

A scheme is called AP if it preserves the discrete analogue of the asymptotic transition from the microscopic scale to the macroscopic one [53, 54], namely, in the limit $\varepsilon \to 0$, the discretization of the above transport equation (3.1) should yield a discretization of the diffusion equation (3.2). Such schemes allow mesh sizes and time steps much bigger than the mean free path/time, yet still capture the correct physical behavior. The development of such schemes started with steady problems of linear transport equations by Larsen, Morel, and Miller [65] and for boundary value problems by Jin and Levermore [55, 57]. Uniform convergence with respect to the mean free path for an AP scheme was first established by Golse, Jin, and Levermore [40].

In [65], and its follow-up [64], several space discretizations for steady transport problems were investigated, among them diamond differencing and a linear discontinuous Galerkin (LD) method, both of which were identified to behave
well in the asymptotic regime. Furthermore, LD leads to a compact stencil for the diffusion equation.

For time-dependent problems, AP schemes were first designed for nonlinear hyperbolic systems with relaxation by Jin and Levermore [52, 56]. There one needs to design both the time and the spatial discretization carefully [56], in particular, to overcome the stiffness of the source term. AP schemes for time-dependent transport equations with diffusive scaling started by Jin, Pareschi, and Toscani [59, 60] and Klar [62]. Since then there have been many new developments in the construction of AP schemes for a large class of kinetic equations (cf. reviews by Degond [26] and Jin [54]). Time discretizations usually need an implicit-explicit (IMEX) approach [12, 19, 52, 93], exponential integration methods [29], BGK-type penalty methods [32], or micro-macro decomposition-based schemes [69, 71]. See also [78, 121]. One key idea of the schemes is to split the equation into a nonstiff part, which is treated explicitly, and a stiff part, which will be implicit but can be implemented explicitly. The splitting should be taken in a way such that the combination preserves the AP property. Another possibility to treat the time-dependent case, which has been used often in the neutron transport literature, is to use an implicit semidiscretization in time first (e.g., backward Euler). This reduces the problem to the successive solution of steady problems (with a modified absorption term). For these, one can use a variety of techniques, among them the second order self-adjoint form of the transport equation. For a recent example, see [81], where different spatial discretizations are investigated. The second order self-adjoint form, however, does not exist for the fully time-dependent, undiscretized transport equations.

In this chapter, we present a two-dimensional AP scheme for the time-dependent transport equation (3.1), where we do not use a semidiscretization in time. We combine a well-known scheme [60] which is based on the parity equations (which in turn are well-known [70]) with staggered grids, which in one dimension are fully understood [54]. The advantage of the staggered approach, compared with the regular grid approach in [60], is that in the diffusion limit we approach a compact stencil, as pointed out by Jin [54]. To be more precise, in one space dimension, using a regular Cartesian grid, the discrete diffusion limit of the scheme of [60] approximates the diffusion operator in (3.2) by \( (\rho_{i+2} - 2\rho_i + \rho_{i-2})/(2\Delta x)^2 \)
3. AP STAGGERED GRID METHOD

(for the case $\sigma_t \equiv 1$), while the current scheme gives the compact discretization $(\rho_{i+1} - 2\rho_i + \rho_{i-1})/(\Delta x)^2$ which offers a better spatial resolution.

However, it is not trivial to extend the scheme from [60] to staggered grids in two dimensions. The first novelty of this paper is that we propose a staggering in two dimensions that requires fewer unknowns than one could have naively expected. The scheme shares similarities with diamond differencing but turns out to be different. The second contribution of this paper is that we present a rigorous stability analysis of the scheme [60] in one dimension and show uniform stability. Previously, only a stability argument was available. Similar to [69, 71], we obtain an explicit CFL condition, which couples a hyperbolic and a parabolic condition and guarantees uniform stability. The scheme contains a splitting parameter that distributes terms between the explicit and implicit parts. For the choice of this parameter, our stability analysis yields an upper bound that is more restrictive than the one that has previously been used [60]. We discuss the choices of the methods we combine in the text. The advantages and disadvantages of these have been described in the original papers. For the sake of completeness, we repeat them in the text.

The remainder of this chapter is organized as follows. In section 3.2, we first derive the parity equations for the linear transport equation. Then, we describe the numerical method. In section 3.3, we show the AP property. We consider the asymptotic limit of the scheme and we state and prove a stability result, which gives a CFL condition and an upper bound on the relaxation parameter. Finally, several numerical tests are presented in section 3.4 to confirm the AP property of the scheme. Finally, we comment on several possible extensions in section 3.5.

3.1.1 The Diffusion Limit

For completeness, we briefly describe the derivation of the diffusion limit (see for instance [25] for more details). We derive the diffusion limit from the radiation transport equation, which is introduced in section 1.1.1

\[
\frac{\partial \psi(t, x, \Omega)}{\partial t} + \Omega \cdot \nabla_x \psi(t, x, \Omega) + \sigma_a \psi(t, x, \Omega) + \sigma_s \left( \psi(t, x, \Omega) - \int_{S^2} \psi(t, x, \Omega') \ d\Omega' \right) = q(t, x, \Omega). \tag{3.3}
\]
For simplicity, we use space and time independent cross-sections $\sigma_a$ and $\sigma_s$ and set the scattering kernel to $g(x, \Omega \cdot \Omega') = 1$. Note that we assume that the velocity is already normalized and dimensionless. For all other variables, we define dimensionless quantities

$$t^* = \frac{t}{t_0}, \quad x^* = \frac{x}{x_0}, \quad \sigma^*_a = \frac{\sigma_a}{\sigma_a^0}, \quad \text{and} \quad \sigma^*_s = \frac{\sigma_s}{\sigma_s^0},$$

(3.4)

where $t_0$ is a characteristic time, $x_0$ is the macroscopic length scale, for example the size of the container, and so on. Further, we define the dimensionless kinetic density and the dimensionless source term

$$\psi^*(t^*, x^*, \Omega^*) = \frac{\psi(t, x, \Omega)}{\psi_0} \quad \text{and} \quad q^*(t^*, x^*, \Omega^*) = \frac{q(t, x, \Omega)}{q_0}. \quad (3.5)$$

After plugging the quantities into the kinetic equation and multiplying with $\frac{x_0}{\psi_0}$, we get (after immediately dropping the asterisks)

$$\frac{x_0}{t_0} \partial_t \psi(t, x, \Omega) + \Omega \cdot \nabla_x \psi(t, x, \Omega) + x_0 \sigma_0^0 \sigma_a \psi(t, x, \Omega)
$$

$$+ x_0 \sigma_s^0 \sigma_s \left( \psi(t, x, \Omega) - \int_{S^2} \psi(t, x, \Omega') \, d\Omega' \right) = \frac{x_0 q_0}{\psi_0} q(t, x, \Omega). \quad (3.6)$$

From this, we obtain the diffusion scaling of the linear transport equation

$$\varepsilon \partial_t \psi(t, x, \Omega) + \Omega \cdot \nabla_x \psi(t, x, \Omega) + \varepsilon \sigma_a \psi(t, x, \Omega)
$$

$$+ \frac{1}{\varepsilon \sigma_s} \left( \psi(t, x, \Omega) - \int_{S^2} \psi(t, x, \Omega') \, d\Omega' \right) = \varepsilon q(t, x, \Omega) \quad \text{(3.7)}$$

if the dimensionless numbers satisfy

$$\varepsilon = \frac{x_0}{t_0} = x_0 \sigma_0^0 = \frac{1}{x_0 \sigma_s^0} = \frac{x_0 q_0}{\psi_0}. \quad (3.8)$$

Thus, the diffusion scale assumes that compared to the length scale we are interested in a long time scale; there are few absorptions and lots of collisions happening; and the source is scaled appropriately.

In what follows, we consider the transport equation in scaled variables for a vanishing scaling parameter $\varepsilon > 0$, i.e. we want to compute $\lim_{\varepsilon \to 0} \psi_\varepsilon$ with $\psi_\varepsilon$
3. AP STAGGERED GRID METHOD

being the solution of (3.7). The first step is to expand the dependence of $\psi_\varepsilon$ on $\varepsilon$ as

$$
\psi_\varepsilon(t, x, \Omega) = \psi_0(t, x, \Omega) + \varepsilon\psi_1(t, x, \Omega) + \varepsilon^2\psi_2(t, x, \Omega) + \mathcal{R}_\varepsilon(t, x, \Omega),
$$

(3.9)

where neither $\psi_0$, $\psi_1$, nor $\psi_2$ depend on $\varepsilon$. We insert this expansion into (3.7), multiply by $\varepsilon$, and match powers of $\varepsilon$. Neglecting $\mathcal{R}_\varepsilon$, this gives in compact notation (cf. (1.4)):

$$
\mathcal{O}(\varepsilon^0) : 0 = -\Omega(\psi_0)(t, x, \Omega),
$$

(3.10a)

$$
\mathcal{O}(\varepsilon^1) : \Omega \cdot \nabla_x \psi_0(t, x, \Omega) = -\Omega(\psi_1)(t, x, \Omega),
$$

(3.10b)

$$
\mathcal{O}(\varepsilon^2) : \partial_t \psi_0(t, x, \Omega) + \Omega \cdot \nabla_x \psi_1(t, x, \Omega) + \sigma_s \psi_0(t, x, \Omega) - q(t, x, \Omega) = -\Omega(\psi_2)(t, x, \Omega).
$$

(3.10c)

The next step is to choose $\psi_0$, $\psi_1$, and $\psi_2$ consistent with these equations with the help of the properties of $\Omega$ (see section 1.1.1). First, (3.10a) implies that $\psi_0$ is in the null space of $\Omega$ and with this $\psi_0$ is constant in $\Omega$. Second, (3.10b) can be rearranged to

$$
\psi_1(t, x, \Omega) = -\frac{1}{\sigma_s} \Omega \cdot \nabla_x \psi_0(t, x, \Omega) + \int_{S^2} \psi_1(t, x, \Omega') d\Omega'.
$$

(3.11)

Third, we insert this into (3.10c) and integrate over $\Omega$. Since $\psi_0$ does not depend on $\Omega$ and $\int_{S^2} \Omega d\Omega = 0$, we obtain the diffusion equation after multiplying with $\frac{1}{|S^2|} = \frac{3}{4\pi}$

$$
\partial_t \psi_0(t, x) - \nabla_x \frac{1}{\sigma_s} \nabla_x \psi_0(t, x, \Omega) + \sigma_s \psi_0(t, x, \Omega) = \frac{3}{4\pi} \int_{S^2} q(t, x, \Omega) d\Omega.
$$

(3.12)

Finally, one can show that for this choice of $\psi_0$, $\psi_1$ and $\psi_2$, the remainder satisfies $\mathcal{R}_\varepsilon(t, x, \Omega) \rightarrow 0$ as $\varepsilon \rightarrow 0$ and $\psi_\varepsilon \rightarrow \psi_0$ follows immediately.

**Remark 3.1.** The derivation can easily be extended to non constant cross sections $\sigma_a$ and $\sigma_s$ or other velocity domains. Using, for instance, the two-dimensional flatland model above (3.1), one can easily see that we end up with (3.2). Moreover, the limit can be derived for an arbitrary scattering kernel $g(x, \Omega \cdot \Omega')$, but this requires to investigate the inverse of $\Omega$. 

70
3.2 The Numerical Method

In this section, we first reformulate the transport equation into a parity equation. Then, we describe the angular, the spatial, and the time discretization of the method in detail.

We begin with the transport equation in the diffusive scaling (3.1), restricted to two spatial dimensions \( x = (x, y) \). Following \[60\], the equation is split into four parts according to the quadrants of the velocity space. We obtain four equations with non-negative \( \xi, \eta \geq 0 \). This system can be rewritten if we define the even and odd parities

\[
\begin{align*}
    r^{(1)}(\xi, \eta) &= \frac{1}{2} [\psi(\xi, -\eta) + \psi(-\xi, \eta)], \\
    r^{(2)}(\xi, \eta) &= \frac{1}{2} [\psi(\xi, \eta) + \psi(-\xi, -\eta)], \\
    j^{(1)}(\xi, \eta) &= \frac{1}{2} \varepsilon [\psi(\xi, -\eta) - \psi(-\xi, \eta)], \\
    j^{(2)}(\xi, \eta) &= \frac{1}{2} \varepsilon [\psi(\xi, \eta) - \psi(-\xi, -\eta)]
\end{align*}
\]

leading to

\[
\begin{align*}
    \partial_t r^{(1)} + \xi \partial_x j^{(1)} - \eta \partial_y j^{(1)} &= -\frac{\sigma_s}{\varepsilon^2} (r^{(1)} - \rho) - \sigma_a r^{(1)} + q, \\
    \partial_t r^{(2)} + \xi \partial_x j^{(2)} + \eta \partial_y j^{(2)} &= -\frac{\sigma_s}{\varepsilon^2} (r^{(2)} - \rho) - \sigma_a r^{(2)} + q, \\
    \partial_t j^{(1)} + \frac{\xi}{\varepsilon^2} \partial_x r^{(1)} - \frac{\eta}{\varepsilon^2} \partial_y r^{(1)} &= -\frac{\sigma_s}{\varepsilon^2} j^{(1)} - \sigma_a j^{(1)}, \\
    \partial_t j^{(2)} + \frac{\xi}{\varepsilon^2} \partial_x r^{(2)} + \frac{\eta}{\varepsilon^2} \partial_y r^{(2)} &= -\frac{\sigma_s}{\varepsilon^2} j^{(2)} - \sigma_a j^{(2)},
\end{align*}
\]

where \( \rho = \frac{1}{2\pi} \int_\Omega \psi \, d\Omega \). It may seem that we have quadrupled the number of unknowns. Note, however, that due to symmetry we need to solve these equations for \( \xi, \eta \) in the positive quadrant only. Thus the number of unknowns is effectively the same.

The discretization of the angular variable uses Gaussian quadrature points. Let

\[
\xi(\lambda) = \cos(\lambda \pi/2) \quad \text{and} \quad \eta(\lambda) = \sin(\lambda \pi/2)
\]

for every \( 0 \leq \lambda \leq 1 \). Then, the density

\[
\rho = \frac{1}{2} \int_0^1 [r^{(1)}(\xi, \eta) + r^{(2)}(\xi, \eta)] \, d\lambda
\]

can be approximated by a Gaussian quadrature rule on \([0, 1]\), where the quadrature points \( \{\lambda_i\} \) are mapped to \( \{\xi_i\} \) and \( \{\eta_i\} \) by (3.15).
3. AP STAGGERED GRID METHOD

3.2.1 Spatial Discretization

First, we describe the spatial discretization, while keeping the time \(t\) continuous. We define a standard regular mesh with mesh size \(\Delta x \times \Delta y\) and place the unknowns in the following way (see Figure 3.1a):

- \(r^{(1)}\), \(r^{(2)}\), \(\rho\), and \(q\) are located at the vertices \((i, j)\) and the cell centers \((i + \frac{1}{2}, j + \frac{1}{2})\);
- \(j^{(1)}\) and \(j^{(2)}\) are located at the face centers \((i + \frac{1}{2}, j)\) and \((i, j + \frac{1}{2})\).

This choice enables us to approximate all spatial derivatives in the system (3.14) by half-grid centered finite differences and yields a closed system of equations.

\[
\begin{align*}
\partial_t r^{(1)}_{i,j} + & \frac{j^{(1)}_{i+1,j} - j^{(1)}_{i,j}}{\Delta x} - \frac{j^{(1)}_{i,j+1} - j^{(1)}_{i,j-1}}{\Delta y} = -\frac{\sigma_s}{\varepsilon^2}(r^{(1)}_{i,j} - \rho_{i,j}) - \sigma_a r^{(1)}_{i,j} + q_{i,j}, \\
\partial_t r^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} + & \frac{j^{(1)}_{i+1,j+\frac{1}{2}} - j^{(1)}_{i,j+\frac{1}{2}}}{\Delta x} - \frac{j^{(1)}_{i+\frac{1}{2},j+1} - j^{(1)}_{i+\frac{1}{2},j}}{\Delta y} = -\frac{\sigma_s}{\varepsilon^2}(r^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} - \rho_{i+\frac{1}{2},j+\frac{1}{2}})
- \sigma_a r^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} + q_{i+\frac{1}{2},j+\frac{1}{2}}, \\
& \text{(3.17)}
\end{align*}
\]

Figure 3.1: Staggered grids and control volumes: red circles (vertices and cell centers), \(r^{(1)}, r^{(2)}, \rho, q\); blue diamonds (face centers), \(j^{(1)}, j^{(2)}\).

In detail, the semidiscretized equations are defined as follows. Let \(i, j \in \mathbb{Z}\). Then, the parities \(r^{(1)}\) and \(r^{(2)}\) satisfy (3.14a) and (3.14b), which on the vertices \((i, j)\) and the cell centers \((i + \frac{1}{2}, j + \frac{1}{2})\) are given by
and integrate the system (3.14) over the control volumes. For instance, (3.14a) integrated over the volume $\hat{\Omega}_{i,j}$ is given by

$$
\partial_t r_{i,j}^{(1)} + \frac{1}{|\hat{\Omega}_{i,j}|} \int_{\hat{\Omega}_{i,j}} \left( \xi \partial_x j_{i,j}^{(1)} + \eta \partial_y j_{i,j}^{(1)} \right) \, d(x,y) = -\frac{\sigma_s}{\varepsilon^2} (r_{i,j}^{(1)} - \rho_{i,j}) + q_{i,j} - \frac{1}{\varepsilon^2} (r_{i,j}^{(1)} - \rho_{i,j}) + q_{i,j} + q_{i,j}.
$$

Similarly, the equations for the parities $j^{(1)}$ and $j^{(2)}$ on the face centers $(i + \frac{1}{2}, j)$ and $(i, j + \frac{1}{2})$ are derived from the equations (3.14c) and (3.14d):

$$
\partial_{i,j}^{(1)} + \frac{\xi}{\varepsilon^2} r_{i,j+\frac{1}{2},j+\frac{1}{2}} - \frac{\eta}{\varepsilon^2} r_{i+\frac{1}{2},j+\frac{1}{2}} - \frac{r_{i,j+\frac{1}{2},j+\frac{1}{2}} - r_{i+\frac{1}{2},j+\frac{1}{2}}}{\Delta x} = -\frac{\sigma_t}{\varepsilon^2} (r_{i,j+\frac{1}{2},j+\frac{1}{2}} - \rho_{i,j+\frac{1}{2},j+\frac{1}{2}}),
$$

and

$$
\partial_{i,j}^{(2)} + \frac{\xi}{\varepsilon^2} r_{i,j+\frac{1}{2},j+\frac{1}{2}} - \frac{\eta}{\varepsilon^2} r_{i+\frac{1}{2},j+\frac{1}{2}} - \frac{r_{i,j+\frac{1}{2},j+\frac{1}{2}} - r_{i+\frac{1}{2},j+\frac{1}{2}}}{\Delta x} = -\frac{\sigma_t}{\varepsilon^2} (r_{i,j+\frac{1}{2},j+\frac{1}{2}} - \rho_{i,j+\frac{1}{2},j+\frac{1}{2}}).
$$

Remark 3.2. The method can be interpreted as a finite volume method and is therefore conservative. To see this, note that all the unknowns are given on two regular grids: vertices and cell centers or vertical cell centers and horizontal cell centers. Merging the corresponding grids, we obtain two staggered nonregular grids as shown in Figure 3.1a. These can also be interpreted as control volumes of a finite volume method. The control volumes $\hat{\Omega}_{i,j}$ are defined as shown in Figures 3.1b, 3.1c, where $\hat{\Omega}_{i,j}$ is the diamond around the point $(x_i, y_j)$.

We define the volume averages corresponding to the control volumes by, e.g.,

$$
r_{i,j}^{(1)} = \frac{1}{|\hat{\Omega}_{i,j}|} \int_{\hat{\Omega}_{i,j}} r_{i,j}^{(1)} \, d(x,y) \quad \text{with} \quad |\hat{\Omega}_{i,j}| = \frac{1}{2} \Delta x \Delta y
$$

and integrate the system (3.14) over the control volumes. For instance, (3.14a) integrated over the volume $\hat{\Omega}_{i,j}$ is given by

$$
\partial_t r_{i,j}^{(1)} + \frac{1}{|\hat{\Omega}_{i,j}|} \int_{\hat{\Omega}_{i,j}} \left( \xi \partial_x j_{i,j}^{(1)} + \eta \partial_y j_{i,j}^{(1)} \right) \, d(x,y) = -\frac{\sigma_s}{\varepsilon^2} (r_{i,j}^{(1)} - \rho_{i,j}) + q_{i,j}.
$$
The integral can be simplified using Gauss’s theorem

\[
\int_{\partial_{i,j}} (\xi \partial_x j^{(1)} + \eta \partial_y j^{(1)}) \, d(x, y) = \int_{\partial_{i,j}} \left( \nabla \cdot \begin{bmatrix} \xi \\ \eta \end{bmatrix} j^{(1)} \right) \, d(x, y)
\]

\[
= \int_{\partial_{i,j}} \left( \begin{bmatrix} \xi \\ \eta \end{bmatrix} j^{(1)} \cdot n \right) \, d(x, y),
\]

where \( n \) is the outer normal vector. It remains to compute the integrals over the four edges of the diamond. These terms are approximated by the trapezoidal rule, e.g., the upper right part is then given by

\[
\int_{(x_i, y_j+\frac{1}{2})}^{(x_{i+1}, y_j+\frac{1}{2})} \left( \begin{bmatrix} \xi \\ \eta \end{bmatrix} j^{(1)} \cdot n \right) \, d(x, y) \\
\approx \frac{1}{2} \left( \xi \Delta y \left(j^{(1)}_{i,j+\frac{1}{2}} + j^{(1)}_{i+1,j+\frac{1}{2}} \right) + \eta \Delta x \left(j^{(1)}_{i,j+\frac{1}{2}} + j^{(1)}_{i+1,j+\frac{1}{2}} \right) \right).
\]

In total, we obtain the approximation for the integral

\[
\frac{2}{\Delta x \Delta y} \int_{\partial_{i,j}} \left( \begin{bmatrix} \xi \\ \eta \end{bmatrix} j^{(1)} \cdot n \right) \, d(x, y) \approx \frac{\xi}{\Delta x} \left(j^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} - j^{(1)}_{i-\frac{1}{2},j} \right) + \frac{\eta}{\Delta y} \left(j^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} - j^{(1)}_{i,j-\frac{1}{2}} \right)
\]

and therefore the same semidiscrete equation as before (3.17)

\[
\partial_t r^{(1)}_{i,j} + \frac{\xi}{\Delta x} \left(j^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} - j^{(1)}_{i-\frac{1}{2},j} \right) + \frac{\eta}{\Delta y} \left(j^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} - j^{(1)}_{i,j-\frac{1}{2}} \right) = -\frac{\sigma_s}{\varepsilon^2} (r^{(1)}_{i,j} - \rho_{i,j}) + q_{i,j}.
\]

In the same way, we obtain the semidiscretized equations (3.18)-(3.20).

In the case of a Cartesian mesh \( \Delta x = \Delta y \), the finite volumes are rotated by 45 degrees with respect to the axes, which seems to be odd. However, it seems that the choice of grid points and therefore the volumes are unique in the strategy we have adopted. In fact, it seems quite natural when we consider the fact that the quantity \( r^{(2)} \) describes the number of particles in \( \xi, \eta > 0 \) and \( \xi, \eta < 0 \). Thus a flow into the diagonal direction makes sense, and therefore interfaces at 45 degrees to the axes. This will be investigated further in future work.
Remark 3.3. Note that (3.13) can formally be inverted to obtain the density of the transport equation

\[ f(\xi, \eta) = \begin{cases} 
\frac{1}{\varepsilon}(|\xi|, |\eta|) + \varepsilon \text{sign}(\xi) j^{(1)}(|\xi|, |\eta|) & , \text{ for } \xi \eta < 0 , \\
\frac{1}{\varepsilon}(|\xi|, |\eta|) + \varepsilon \text{sign}(\xi) j^{(2)}(|\xi|, |\eta|) & , \text{ for } \xi \eta \geq 0 .
\end{cases} \] (3.27)

However, in the numerical scheme the parities \( r^{(1)} \), \( r^{(2)} \) and \( j^{(1)} \), \( j^{(2)} \) are not given on the same spatial grid and need to be interpolated. This is similar to other approaches which are based on parity decompositions.

### 3.2.2 Time Discretization

For simplicity, we consider again semidiscretized equations. This time, we keep the spatial variables \( x \) and \( y \) continuous and apply the time discretization technique from [60]. The idea is to introduce a relaxation parameter \( \phi = \phi(\varepsilon) \), such that we obtain a linear hyperbolic system with stiff relaxation. Then the linear hyperbolic part, which is nonstiff, can be separated from the stiff relaxation step.

First, we rewrite the system of (3.14) as the diffusive relaxation system

\[
\begin{align*}
\partial_t r^{(1)} + \xi \partial_x j^{(1)} - \eta \partial_y j^{(1)} &= -\frac{\sigma_s}{\varepsilon^2} (r^{(1)} - \rho) - \sigma_a r^{(1)} + q , \\
\partial_t r^{(2)} + \xi \partial_x j^{(2)} + \eta \partial_y j^{(2)} &= -\frac{\sigma_s}{\varepsilon^2} (r^{(2)} - \rho) - \sigma_a r^{(2)} + q , \\
\partial_t j^{(1)} + \phi \xi \partial_x r^{(1)} - \phi \eta \partial_y r^{(1)} &= -\frac{1}{\varepsilon^2} [\sigma_{ij}^{(1)} + (1 - \varepsilon^2 \phi) \xi \partial_x r^{(1)} - (1 - \varepsilon^2 \phi) \eta \partial_y r^{(1)}] - \sigma_a j^{(1)} , \\
\partial_t j^{(2)} + \phi \xi \partial_x r^{(2)} + \phi \eta \partial_y r^{(2)} &= -\frac{1}{\varepsilon^2} [\sigma_{ij}^{(2)} + (1 - \varepsilon^2 \phi) \xi \partial_x r^{(2)} + (1 - \varepsilon^2 \phi) \eta \partial_y r^{(2)}] - \sigma_a j^{(2)}
\end{align*}
\] (3.28)

with \( 0 \leq \phi \leq 1/\varepsilon^2 \). The condition \( \phi \geq 0 \) is necessary for the hyperbolicity of the left-hand side, whereas the condition \( \phi \leq 1/\varepsilon^2 \) ensures that the bracketed term on the right-hand side has a well-defined limit for \( \varepsilon \to 0 \). In Remark 3.4 below we comment on the role of \( \phi \) and how this choice differs from Klar’s scheme [62].
3. AP STAGGERED GRID METHOD

Thus, we split the equation into two parts, the transport step,

\[
\begin{align*}
\partial_t r^{(1)} + \xi \partial_x j^{(1)} - \eta \partial_y j^{(1)} &= -\sigma_a r^{(1)} + q, \\
\partial_t r^{(2)} + \xi \partial_x j^{(2)} + \eta \partial_y j^{(2)} &= -\sigma_a r^{(2)} + q, \\
\partial_t j^{(1)} + \phi \xi \partial_x r^{(1)} - \phi \eta \partial_y r^{(1)} &= -\sigma_a j^{(1)}, \\
\partial_t j^{(2)} + \phi \xi \partial_x r^{(2)} + \phi \eta \partial_y r^{(2)} &= -\sigma_a j^{(2)},
\end{align*}
\]  

and the relaxation step,

\[
\begin{align*}
\partial_t r^{(1)} &= -\frac{\sigma_s}{\epsilon^2} (r^{(1)} - \rho), \\
\partial_t r^{(2)} &= -\frac{\sigma_s}{\epsilon^2} (r^{(2)} - \rho), \\
\partial_t j^{(1)} &= -\frac{1}{\epsilon^2} [\sigma_s j^{(1)} + (1 - \epsilon^2 \phi) \xi \partial_x r^{(1)} - (1 - \epsilon^2 \phi) \eta \partial_y r^{(1)}], \\
\partial_t j^{(2)} &= -\frac{1}{\epsilon^2} [\sigma_s j^{(2)} + (1 - \epsilon^2 \phi) \xi \partial_x r^{(2)} + (1 - \epsilon^2 \phi) \eta \partial_y r^{(2)}].
\end{align*}
\]  

Finally, we apply the explicit Euler method to the first step and the implicit Euler method to the second step. Note that the implicit Euler method can be implemented explicitly, since \( \rho \) is preserved in the second step (which can be seen by adding the first two equations).

The fully discrete scheme is just splitting (3.17)-(3.20) into the two steps (3.29)-(3.30).

\textbf{Remark 3.4.} Klar [62] developed a similar decomposition, which corresponds to \( \phi = 0 \) in our framework. As mentioned in [60], there are two major differences. First, the resulting system is only weakly hyperbolic and therefore well-posedness is an issue. Second, computations are still performed on the hole velocity space. Using the symmetries of the parities, the computations can be performed on the first quadrant and the computational cost can be reduced.

\textbf{Remark 3.5.} To generalize the problem from a two-dimensional “flatland” to a full two-dimensional problem, the domain of the angular variable changes from the unit circle to the unit disc. In order to do this change, the Gauss integration rule needs to be substituted by a two-dimensional integration rule on the unit disc.
3.3 The AP Property

In this section, we analyze the AP property of the above scheme in two steps. First, we derive the discrete asymptotic limit. Second, we analyze stability.

3.3.1 The Discrete Diffusion Limit

In the same way as above, we consider the spatial and time discretization separately. The limit, as $\varepsilon \to 0$, of the time discretization is derived in [60]. Hence, it remains to investigate the discrete limit of the spatial discretization. To this end, we consider the diffusive limit $\varepsilon \to 0$ of the semidiscretized equations (3.17)-(3.20).

First, the limit of (3.17) and (3.19) for the parities $r^{(1)}$ and $j^{(1)}$ is given by

\begin{equation}
\begin{align*}
    r^{(1)}_{i,j} &= \rho_{i,j}, \\
    r^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} &= \rho_{i+\frac{1}{2},j+\frac{1}{2}}, \\
    j^{(1)}_{i+\frac{1}{2},j} &= -\frac{\xi}{\sigma_t} r^{(1)}_{i+1,j} - r^{(1)}_{i,j} + \frac{\eta}{\sigma_s} \frac{r^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} - r^{(1)}_{i+\frac{1}{2},j-\frac{1}{2}}}{\Delta y}, \\
    j^{(1)}_{i,j+\frac{1}{2}} &= -\frac{\xi}{\sigma_t} r^{(1)}_{i+\frac{1}{2},j+\frac{1}{2}} - r^{(1)}_{i-\frac{1}{2},j+\frac{1}{2}} + \frac{\eta}{\sigma_s} \frac{r^{(1)}_{i+1,j+1} - r^{(1)}_{i,j}}{\Delta y}.
\end{align*}
\end{equation}

(3.31)

Inserting these equations into (3.17) yields

\begin{equation}
\begin{align*}
    \partial_t \rho_{i,j} &= -\frac{\xi^2}{\sigma_t} \left( \rho_{i+1,j} - 2\rho_{i,j} + \rho_{i-1,j} \right) \frac{(\Delta x)^2}{\Delta y} + \frac{2\xi\eta}{\sigma_t} \left( \rho_{i+\frac{1}{2},j+\frac{1}{2}} - \rho_{i+\frac{1}{2},j-\frac{1}{2}} - \rho_{i-\frac{1}{2},j+\frac{1}{2}} + \rho_{i-\frac{1}{2},j-\frac{1}{2}} \right) \frac{\Delta x \Delta y}{\sigma_t} \\
    &\quad - \frac{\eta^2}{(\Delta y)^2} = -\sigma_a \rho_{i,j} + q_{i,j}, \\
    \partial_t \rho_{i+\frac{1}{2},j+\frac{1}{2}} &= -\frac{\xi^2}{\sigma_t} \left( \rho_{i+\frac{1}{2},j+\frac{1}{2}} - 2\rho_{i+\frac{1}{2},j+\frac{1}{2}} + \rho_{i-\frac{1}{2},j+\frac{1}{2}} \right) \frac{(\Delta x)^2}{\sigma_t} + \frac{2\xi\eta}{\sigma_t} \left( \rho_{i+1,j+1} - \rho_{i+1,j} - \rho_{i,j+1} + \rho_{i,j} \right) \frac{\Delta x \Delta y}{\sigma_t} \\
    &\quad - \frac{\eta^2}{(\Delta y)^2} = -\sigma_a \rho_{i+\frac{1}{2},j+\frac{1}{2}} + q_{i+\frac{1}{2},j+\frac{1}{2}}.
\end{align*}
\end{equation}

(3.32)
3. AP STAGGERED GRID METHOD

Treating (3.18) and (3.20) in the same way as above, we additionally obtain the following differential equations for \( \rho \):

\[
\partial_t \rho_{i,j} - \frac{\xi^2 \rho_{i+1,j} - 2 \rho_{i,j} + \rho_{i-1,j}}{(\Delta x)^2} - \frac{2 \eta \rho_{i+\frac{1}{2},j+\frac{1}{2}} - \rho_{i+\frac{3}{2},j+\frac{1}{2}} - \rho_{i-\frac{1}{2},j+\frac{1}{2}} + \rho_{i-\frac{3}{2},j+\frac{1}{2}}}{\Delta x \Delta y} - \frac{\eta^2 \rho_{i,j+1} - 2 \rho_{i,j} + \rho_{i,j-1}}{(\Delta y)^2} = -\sigma_a \rho_{i,j} + q_{i,j},
\]

(3.33)

Adding up the equations, the middle terms cancel and we obtain

\[
\partial_t \rho_{i,j} - \frac{\xi^2 \rho_{i+1,j} - 2 \rho_{i,j} + \rho_{i-1,j}}{(\Delta x)^2} - \frac{\eta^2 \rho_{i,j+1} - 2 \rho_{i,j} + \rho_{i,j-1}}{(\Delta y)^2} = -\sigma_a \rho_{i,j} + q_{i,j}.
\]

(3.34)

Integrating over \( \xi^2 + \eta^2 = 1 \) yields the semidiscretized diffusion equations on the vertices and the cell centers. Note that integrating (3.32) or (3.33) over \( \xi^2 + \eta^2 = 1 \), the middle terms cancel as well and we get the same result.

As expected, the spatial discretization with staggered grids leads to a compact five-point stencil for the diffusion equation (3.2). Together with the results [60] on the limit of the time discretization (3.29) – (3.30), this also shows that the formal limit of our scheme coincides with the diffusion equation.

3.3.2 Stability

We limit our discussion to the one-dimensional case (see Remark 3.11 for the two-dimensional case) and show uniform stability with \( \varepsilon \) using the von Neumann analysis [69, 103, 114].
3.3 The AP Property

In the following, we consider the transport equation in slab geometry and assume that the cross section \( \sigma_t = \sigma_s + \varepsilon^2 \sigma_a > 0 \) is independent of \( x \in \mathbb{R} \) (see Remark 3.10 for space-dependent scattering). Further, we consider a source-free two velocity model \( v \in \{ \pm 1 \} \). Then, the even and odd parities

\[
r(t, x) = \frac{1}{2} [f(t, x, 1) + f(t, x, -1)] \quad \text{and} \quad j(t, x) = \frac{1}{2\varepsilon} [f(t, x, 1) - f(t, x, -1)]
\]

fulfill

\[
\partial_t r + \partial_x j = -\sigma_a r \\
\partial_t j + \frac{1}{\varepsilon^2} \partial_x r = -\frac{1}{\varepsilon^2} \sigma_s j - \sigma_a j
\]

and the numerical scheme has the following update rule: For \( k = 0, 1, 2, \ldots \)

\[
r^{k+\frac{1}{2}} = r^k - \Delta t (D_x j^k + \sigma_a r^k) , \\
j^{k+\frac{1}{2}} = j^k - \Delta t (\phi D_x r^k + \sigma_a j^k) , \\
r^{k+1} = r^{k+\frac{1}{2}} , \\
j^{k+1} = \frac{\varepsilon^2}{\varepsilon^2 + \sigma_a \Delta t} j^{k+\frac{1}{2}} - \frac{\Delta t}{\varepsilon^2 + \sigma_a \Delta t} (1 - \varepsilon^2 \phi) D_x r^{k+1} ,
\]

where \( D_x \) denotes the half-grid centered finite difference of the spatial derivative. We place \( r \) on the half grid points \( (m + \frac{1}{2}) \Delta x \) and \( j \) on the full grid points \( m \Delta x \). For this scheme, we do not expect positivity, but we seek a uniform CFL condition.

For a von Neumann analysis of the scheme, we expand the parities in Fourier series:

\[
r(x, t) = \sum_{\ell = -\infty}^{\infty} a_\ell(t) e^{i\ell x} \quad \text{and} \quad j(x, t) = \sum_{\ell = -\infty}^{\infty} b_\ell(t) e^{i\ell x} .
\]

As no mixing between the Fourier modes occurs during the update of the solution, it is sufficient to consider the evolution of

\[
r(x, t) = a_\ell(t) e^{i\ell x} \quad \text{and} \quad j(x, t) = b_\ell(t) e^{i\ell x}
\]

for some \( \ell \) and to determine the growth factor matrix of the Fourier coefficients. First, we note that the staggered grid derivatives can be rewritten as

\[
(D_x r)(h(m + \frac{1}{2}), t) = a_\ell(t) \frac{e^{i\ell h(m+1)} - e^{i\ell hm}}{h} = 2 \frac{\ell}{h} \sin \left( \frac{\ell h}{2} \right) e^{i\ell hm} a_\ell(t) , \\
(D_x j)(hm, t) = b_\ell(t) \frac{e^{i\ell h(m+\frac{1}{2})} - e^{i\ell h(m-\frac{1}{2})}}{h} = 2 \frac{\ell}{h} \sin \left( \frac{\ell h}{2} \right) e^{i\ell hm} b_\ell(t) ,
\]
3. AP STAGGERED GRID METHOD

with \( h := \Delta x \). To shorten the notation, we define \( d_\ell := \frac{2i}{h} \sin \left( \frac{\ell h}{2} \right) \). Then, the first update step of the Fourier coefficients is given by

\[
\begin{bmatrix}
 a_k \\
 b_k
\end{bmatrix}(t + \Delta t) = \begin{bmatrix}
 1 - \sigma_a \Delta t & -\Delta t d_\ell \\
 -\Delta t \phi d_\ell & 1 - \sigma_a \Delta t
\end{bmatrix} \begin{bmatrix}
 a_k \\
 b_k
\end{bmatrix}(t) =: G_1
\]

and the second step is given by

\[
\begin{bmatrix}
 a_k \\
 b_k
\end{bmatrix}(t + \Delta t) = \begin{bmatrix}
 1 - \sigma_a \Delta t - \Delta t d_\ell & -\Delta t \phi d_\ell \\
 -\Delta t \phi d_\ell & 1 - \sigma_a \Delta t
\end{bmatrix} \begin{bmatrix}
 a_k \\
 b_k
\end{bmatrix}(t) =: G_2
\]

Thus, the growth factor matrix is \( G := G_2 \cdot G_1 \)

\[
G = \begin{bmatrix}
 1 - \sigma_a \Delta t & -\Delta t d_\ell \\
 -\Delta t \phi d_\ell & 1 - \sigma_a \Delta t
\end{bmatrix} = \begin{bmatrix}
 1 - \sigma_a \Delta t & -\Delta t d_\ell \\
 -\Delta t \phi d_\ell & 1 - \sigma_a \Delta t
\end{bmatrix} = \begin{bmatrix}
 1 - \sigma_a \Delta t & -\Delta t d_\ell \\
 -\Delta t \phi d_\ell & 1 - \sigma_a \Delta t
\end{bmatrix} = \begin{bmatrix}
 1 - \sigma_a \Delta t & -\Delta t d_\ell \\
 -\Delta t \phi d_\ell & 1 - \sigma_a \Delta t
\end{bmatrix}
\]

(3.43)

For stability, the eigenvalues of the matrix \( G \) are of main interest. They can be written as

\[
\lambda_{1,2} = g \pm \sqrt{g^2 - \det(G)}
\]

with \( g \) being the half trace and \( \det(G) \) being the determinant of \( G \):

\[
g = \frac{1}{2} \frac{\Delta t}{\varepsilon^2 + \sigma_a \Delta t} \left( \Delta t^2 \sigma_d^2 (1 - \varepsilon^2 \phi) + (1 - \sigma_a \Delta t)(1 - \sigma_a \Delta t) \right)
\]

and

\[
\det(G) = \frac{\varepsilon^2}{\varepsilon^2 + \sigma_a \Delta t} \left( (1 - \sigma_a \Delta t)^2 - \phi d_\ell^2 \Delta t^2 \right)
\]

(3.45)

**Proposition 3.6.** Let the time step \( \Delta t \) and the relaxation parameter \( \phi \) satisfy

\[
\Delta t \leq \min \left\{ \frac{1}{\sigma_a}, \max \left\{ \frac{1}{2} \varepsilon h, \frac{1}{4} h^2 \sigma_t \right\} \right\}
\]

(3.46)

and

\[
0 \leq \phi \leq \begin{cases}
 \frac{h \sigma_t}{2 \varepsilon}, & h \sigma_t \leq 2 \varepsilon \\
 \frac{1}{\varepsilon^2}, & \text{otherwise}
\end{cases}
\]

(3.47)

Then, the numerical scheme is \( L^2 \)-stable.

**Remark 3.7.** Note that the three terms in the time step restriction (3.46) can be interpreted separately. First, the \( \frac{1}{2} \varepsilon h \) term comes from the advection operator. Second, the \( \frac{1}{4} h^2 \sigma_t \) term corresponds to the Courant limit for explicit diffusion.
Third, the $\frac{1}{\sigma_a}$ term is a result of the explicit treatment of the corresponding relaxation term. In this case, we assume that this term is small, so that $\Delta t \leq \frac{1}{\sigma_a}$ is not the restrictive term. Otherwise the term could be treated implicitly, which would remove the restriction.

**Remark 3.8.** Note that the above restriction on $\phi$ is stricter than the one suggested in [60] $0 \leq \phi \leq \frac{1}{\varepsilon}$. Moreover, the condition $h\sigma_t \leq 2\varepsilon$ is satisfied, if and only if the hyperbolic condition $\Delta t \leq \max\{\frac{1}{2}\varepsilon h, \frac{1}{2}h^2\sigma_t\} = \frac{1}{2}\varepsilon h$ holds. This means, there are the following two cases:

\[
\begin{align*}
\text{Case } h\sigma_t &\leq 2\varepsilon : \quad \Delta t \leq \min \left\{ \frac{1}{\sigma_a}, \frac{1}{2}\varepsilon h \right\} \quad \text{and} \quad 0 \leq \phi \leq \frac{h\sigma_t}{2\varepsilon^2}, \\
\text{Case } h\sigma_t &> 2\varepsilon : \quad \Delta t \leq \min \left\{ \frac{1}{\sigma_a}, \frac{1}{4}h^2\sigma_t \right\} \quad \text{and} \quad 0 \leq \phi \leq \frac{1}{\varepsilon^2}.
\end{align*}
\]

In addition, as $\varepsilon \to 0$ the time step restriction becomes $\Delta t \lesssim \min\{\frac{1}{\sigma_a}, \frac{1}{4}h^2\sigma_t\}$, which does not vanish.

In the proof of the proposition, we use the von Neumann analysis. A complete overview of these stability conditions can be found in the lecture notes by Trefethen [114].

**Proof.** Stability follows from the von Neumann condition if we can show $|\lambda_{1,2}| \leq 1$ for $\lambda_1 \neq \lambda_2$ and $|\lambda_{1,2}| < 1$ for $\lambda_1 = \lambda_2$. To show these inequalities, we consider three different cases: two complex eigenvalues, two real eigenvalues, and one eigenvalue. Since $g$ and $\det(G)$ are real-valued (3.45), the cases are equivalent to: $g^2 < \det(G)$, $g^2 > \det(G)$, and $g^2 = \det(G)$.

**Case** $g^2 < \det(G)$ (two complex eigenvalues): If the eigenvalues $\lambda_{1,2}$ are complex, their real part is $g$ and their imaginary part is $\pm \sqrt{\det(G) - g^2}$. Thus, the stability condition $|\lambda_{1,2}|^2 \leq 1$ is satisfied if $\det(G) \leq 1$. For the determinant we have the following estimate

\[
\det(G) = \frac{\varepsilon^2}{\varepsilon^2 + \sigma_a \Delta t} \left( (1 - \sigma_a \Delta t)^2 - \phi d_t^2 \Delta t^2 \right) \leq \frac{\varepsilon^2}{\varepsilon^2 + \sigma_a \Delta t} \left( 1 - \sigma_a \Delta t + \phi \frac{4\Delta t^2}{h^2} \right),
\]

where we used that $-d_t^2 = \frac{4}{h^2} \sin^2\left( \frac{\Delta t h}{2} \right) \leq \frac{4}{h^2}$ and the CFL condition $\Delta t \leq \frac{1}{\sigma_a}$. It remains to show that the last term of (3.49) is bounded by 1. This is equivalent to

\[
\varepsilon^2 \phi \frac{4\Delta t}{h^2} \leq \sigma_s + \varepsilon^2 \sigma_a = \sigma_t,
\]

\[
(3.50)
\]
3. AP STAGGERED GRID METHOD

which in turn is satisfied under the condition $\Delta t \leq \max\{\frac{1}{2} \varepsilon h, \frac{1}{4} h^2 \sigma_t\}$ and the assumption (3.47). This is one of the reasons for the choice of the upper bound of $\phi$ in the assumption (3.47).

**Case** $g^2 > \det(G)$ (two real eigenvalues): The determinant of $G$ is always positive and therefore the eigenvalues are either both positive or both negative, and their sign changes with the sign of $g$. Thus, it is sufficient to show $\lambda_1 \leq 1$ if $g \geq 0$ and $\lambda_2 \geq -1$ if $g < 0$. In particular, one can show that this is equivalent to

$$\det(G) + 1 \mp 2g \geq 0.$$  

(3.51)

The first inequality is generic

$$\det(G) + 1 - 2g = \frac{\Delta t^2}{\varepsilon^2 + \sigma_s \Delta t} (\sigma_a^2 \varepsilon^2 + \sigma_s \sigma_t - d_e) \geq 0,$$  

(3.52)

since $d_e^2 \leq 0$. Whereas, the second inequality requires the CFL condition (3.46). More precisely, under the condition $0 \leq \phi \varepsilon^2 \leq 1$ and $\Delta t \leq \frac{1}{\sigma_a}$, we obtain

$$\det(G) + 1 + 2g > 1 + 2g = 1 + \frac{1}{\varepsilon^2 + \sigma_s \Delta t} (\Delta t^2 d_e^2 (1 - \varepsilon^2 \phi) + (1 - \sigma_a \Delta t) (2 \varepsilon^2 + \sigma_s \Delta t)) \geq \frac{1}{\varepsilon^2 + \sigma_s \Delta t} (\varepsilon^2 - \frac{4 \Delta t^2}{h^2} + \sigma_s \Delta t).$$  

(3.53)

On the one hand, this is obviously nonnegative under the condition $\Delta t \leq \frac{1}{2} \varepsilon h$. On the other hand, the second term can be rewritten as

$$\varepsilon^2 - \frac{4 \Delta t^2}{h^2} + \sigma_s \Delta t = \varepsilon^2 (1 - \sigma_a \Delta t) + \Delta t (\sigma_t - \frac{4 \Delta t^2}{h^2}),$$  

(3.54)

which is nonnegative under the condition $\Delta t \leq \frac{1}{\sigma_a}$ and $\Delta t \leq \frac{1}{4} h^2 \sigma_t$. Together, this yields the desired inequality $\det(G) + 1 + 2g > 0$.

**Case** $g^2 = \det(G)$ (one eigenvalue): The eigenvalue of $G$ is $\lambda_1 = \lambda_2 = g$. Thus, we need to show $|g| < 1$. But as $\det(G) + 1 + 2g > 0$ and $\det(G) \leq 1$ (see above cases) already imply $g > -1$, it remains to show $g < 1$. Since $\sigma_t = \sigma_s + \varepsilon^2 \sigma_a > 0$, at least one of the terms $\sigma_a \Delta t, \sigma_s \Delta t$ is positive and we obtain

$$g \leq \frac{1}{2} \frac{1}{\varepsilon^2 + \sigma_s \Delta t} ((1 - \sigma_a \Delta t) (2 \varepsilon^2 + \sigma_s \Delta t))$$

$$< \frac{1}{2} \frac{1}{\varepsilon^2 + \sigma_s \Delta t} ((1 - \sigma_a \Delta t + \sigma_a \Delta t) (2 \varepsilon^2 + \sigma_s \Delta t + \sigma_s \Delta t)) = 1.$$  

(3.55)

□
Remark 3.9. If there is neither scattering nor absorption and the conditions (3.46) and (3.47) hold, then the relaxation parameter satisfies \( \phi = 0 \) and the hyperbolic condition is always satisfied. Further, the determinant, \( \det(G) = 1 \), and the half trace, \( g = 1 + \frac{\Delta t^2 d^2}{2d^2} \), coincide only if \( d^2 = -\frac{h^2}{4} \sin^2 \left( \frac{4h}{T} \right) = 0 \), which is equivalent to \( \frac{4h}{T} \in \pi \mathbb{Z} \). In most cases, this does not occur and therefore the case \( g^2 = \det(G) \) does not arise. Then, we obtain \( g^2 < \det(G) \) and the eigenvalues are distinct and satisfy \( |\lambda_{1,2}| = 1 \), so that stability follows.

Remark 3.10. If the cross sections are space-dependent, the above analysis is not valid. In practice, the CFL condition is replaced by a worst-case condition. This means that we replace \( \sigma_a \) and \( \sigma_t \) in (3.46) and (3.47) by its maximum and minimum,

\[
\sigma_{a,\text{max}} = \max_x \sigma_a(x) \quad \text{and} \quad \sigma_{t,\text{min}} = \min_x \sigma_t(x),
\]

respectively.

Remark 3.11. In two dimensions, one can expect that the stability result from Proposition 3.6 carries over with the following changes. We replace \( h = \min(\Delta x, \Delta y) \) and add a factor of \( \frac{1}{2} \) in front of the time step to account for the presence of growth rates in each of the two spatial dimensions.

3.4 Numerical Results

In this section, we consider different numerical test cases to demonstrate the performance of our scheme. Since we did not examine boundary conditions, we only consider examples, where the solution is compactly supported away from the boundary. We implemented periodic boundary conditions, so that there is no influence of any discretization of boundary values.

The numerical calculations are performed using the two-dimensional scheme described in section 3.2 with the stability conditions from section 3.3. This means, we first choose the number of grid points \( (N \times N) \) for the staggered grids corresponding to the test case. Then, we determine the maximal time step (cf. Proposition 3.6, Remark 3.10, and Remark 3.11)

\[
\Delta t := 0.9 \cdot \frac{1}{2} \min \left\{ \frac{1}{\sigma_{a,\text{max}}}, \max \left\{ \frac{1}{2} \varepsilon h, \frac{1}{4} h^2 \sigma_{t,\text{min}} \right\} \right\}
\]
and define the relaxation parameter

\[ \phi := \begin{cases} 
    h \frac{\sigma_t, \min}{2 \varepsilon}, & h \sigma_t \leq 2 \varepsilon, \\
    \frac{1}{\varepsilon}, & \text{otherwise}
\end{cases} \tag{3.58} \]

with \( h := \frac{1}{N}, \sigma_{a, \max} := \max_x \sigma_a(x), \) and \( \sigma_{t, \min} := \min_x \sigma_t(x) \). The angular discretization uses a Gaussian quadrature with 16 points on the interval \([0, 1]\) for \( \lambda \). As the quadrature points are mapped to the directions \( \xi \) and \( \eta \) with (3.15), we obtain 16 points per quadrant. In all test cases, we compare the numerical solution on a grid where the parameter \( \varepsilon \) is resolved to a grid on which it is underresolved, thus demonstrating the AP property.

In the remainder of this section, we describe the test cases and the numerical results in detail. We consider four test cases to show different aspects of the AP property. First, we focus on the \( \varepsilon \)-dependence and investigate the convergence order in different regimes. In the second and third test cases, there are large spatial differences in the cross sections. The second test case is continuous and rotationally invariant, whereas in the third test case the material cross sections and the source term are discontinuous. These two test cases intend to demonstrate the performance in multiscale problems. The last test case investigates the stability of the scheme in dependence on the choice of the relaxation parameter \( \phi \).

### 3.4.1 Convergence Order

We examine the order of convergence with respect to the spatial variable. We expect first or second order convergence depending on the used CFL condition. If a hyperbolic condition is used, the time step is proportional to \( h \). As the explicit Euler method is used for the time discretization, we cannot expect more than first order convergence in \( h \). Whereas if the parabolic condition is used, the time step is proportional to \( h^2 \). Then, the explicit Euler method predicts \( \mathcal{O}(h^2) \) convergence. Moreover, centered differences, which are used for the spatial discretization, are as well a second order approximation in \( h \). Thus, we expect that the error is proportional to \( \mathcal{O}(h) \) when the hyperbolic condition is used, and \( \mathcal{O}(h^2) \), respectively, when the parabolic condition is used. To estimate the convergence order, we compute the \( L^2 \)-error \( E(N) \) between the solution computed...
3.4 Numerical Results

on a $N \times N$ grid and a reference solution. Using two different values $N_1$ and $N_2$, we then estimate the convergence order by

$$E_{N_2}^{N_1} = -\frac{\log(E(N_1)) - \log(E(N_2))}{\log(N_1) - \log(N_2)}. \quad (3.59)$$

3.4.1.1 Method of Manufactured Solutions

For the method of manufactured solutions (MMS), we first choose some function $\psi(t, x, y, \xi, \eta)$ and compute a corresponding source term and an initial condition, so that the chosen function is a solution of the transport equation. Let

$$\psi(t, x, y, \xi, \eta) = \exp(-t) \sin(2\pi x)^2 \sin(2\pi y)^2 (1 + \eta^2) \quad (3.60)$$

with $(x, y) \in [0, 1]^2$. Further, let the scattering cross sections be given by $\sigma_a = 0$ and $\sigma_s = 1$. Then, the corresponding source term is given by

$$q(t, x, y, \xi, \eta) = \partial_t \psi + \frac{\Omega}{\varepsilon} \cdot \nabla_x \psi - \frac{1}{\varepsilon^2} \left[ \frac{1}{2\pi} \int_{\Omega} \psi d\Omega' \right] \quad (3.61)$$

and the initial condition is given by

$$\psi(t = 0, x, y, \xi, \eta) = \sin(2\pi x)^2 \sin(2\pi y)^2 (1 + \eta)^2. \quad (3.62)$$

We use the source term and the initial condition to compute a solution with the above scheme. For different grid sizes and different values of $\varepsilon$, we compare the computed densities with the analytic density

$$\rho(t, x, y) = \frac{1}{2\pi} \int_{S^1} \psi d\Omega' \quad (3.63)$$

at time $t = 0.1$. The results are shown in Figure 3.2 and Table 3.1. They confirm second order convergence in the parabolic case. In the hyperbolic case, the convergence order is even slightly higher than expected.

3.4.1.2 Gauss Test

We consider an example case with a smooth initial condition and isotropic scattering

$$\psi(t = 0, x, y, \Omega) = \frac{1}{4\pi \cdot 10^{-2}} \exp\left(-\frac{x^2 + y^2}{4 \cdot 10^{-2}}\right) \quad \text{for} \quad (x, y) \in [-1, 1] \times [-1, 1],$$

$$q = 0, \quad \sigma_t = \sigma_s = 1, \quad \sigma_a = 0, \quad \text{and} \quad \varepsilon = 1, 10^{-1}, 10^{-2}. \quad (3.64)$$
3. AP STAGGERED GRID METHOD

Figure 3.2: Convergence order (MMS): $\ell^2$-error as a function of the spatial resolution. Hyperbolic (filled markers) or parabolic (empty markers) CFL condition.

<table>
<thead>
<tr>
<th>$\varepsilon_k$</th>
<th>$E_{16}^{32}$</th>
<th>$E_{32}^{64}$</th>
<th>$E_{64}^{128}$</th>
<th>$E_{128}^{256}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_0 = 1$</td>
<td>1.60</td>
<td>1.50</td>
<td>1.35</td>
<td>1.23</td>
</tr>
<tr>
<td>$\varepsilon_1 = 0.1$</td>
<td>1.98</td>
<td>1.93</td>
<td>1.86</td>
<td>1.76</td>
</tr>
<tr>
<td>$\varepsilon_2 = 0.01$</td>
<td>2.02</td>
<td>2.00</td>
<td>1.99</td>
<td>1.97</td>
</tr>
<tr>
<td>$\varepsilon_3 = 0.001$</td>
<td>2.02</td>
<td>2.01</td>
<td>2.00</td>
<td>2.00</td>
</tr>
</tbody>
</table>

Table 3.1: Convergence order (MMS): The term $E_{N_1}^{N_2}$ is the convergence rate when going from $N_1 \times N_1$ to $N_2 \times N_2$ grid points for a fixed mean free path $\varepsilon_k = 10^{-k}$, $k = 0, 1, 2, 3$. The dashed line indicates the switch from the hyperbolic to the parabolic CFL condition.

Then, we compute the density $\rho$ at time $t = 0.1$ for different grid sizes and different values of $\varepsilon$, so that the CFL condition (3.57) changes form hyperbolic to parabolic. As a reference solution, we use a highly resolved solution with $512 \times 512$ grid points. Table 3.2 and Figure 3.3 agree with the above assertion, showing first order convergence when the hyperbolic condition holds and second order, respectively, when the parabolic condition holds.
3.4 Numerical Results

Figure 3.3: Convergence order (Gauss test): $\ell^2$-error as a function of the spatial resolution. Hyperbolic (filled markers) or parabolic (empty markers) CFL condition.

Table 3.2: Convergence order (Gauss test): The term $E_{N_1}^{N_2}$ is the convergence rate when going from $N_1 \times N_1$ to $N_2 \times N_2$ grid points for a fixed mean free path $\varepsilon_k = 10^{-k}$, $k = 0, 1, 2$. The dashed line indicates the switch from the hyperbolic to the parabolic CFL condition.

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>$E_{16}^{32}$</th>
<th>$E_{32}^{64}$</th>
<th>$E_{64}^{128}$</th>
<th>$E_{128}^{256}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_0 = 1$</td>
<td>1.50</td>
<td>1.38</td>
<td>1.36</td>
<td>1.66</td>
</tr>
<tr>
<td>$\varepsilon_1 = 0.1$</td>
<td>1.96</td>
<td>1.37</td>
<td>1.15</td>
<td>1.50</td>
</tr>
<tr>
<td>$\varepsilon_2 = 0.01$</td>
<td>2.03</td>
<td>2.01</td>
<td>2.09</td>
<td>2.41</td>
</tr>
</tbody>
</table>

3.4.2 Variable Scattering

In this test case, we examine the performance of the scheme, when the scattering is space-dependent. Compared to the previous test case, we fix the scaling parameter $\varepsilon$ and modify the scattering cross section. Let

$$\psi(t = 0, x, y, \Omega) = \frac{1}{4\pi \cdot 10^{-2}} \exp\left(-\frac{x^2 + y^2}{4 \cdot 10^{-2}}\right) \quad \text{for} \quad (x, y) \in [-1, 1] \times [-1, 1],$$

$$\varepsilon = \frac{1}{100}, \quad q = 0, \quad \sigma_a = 0, \quad \text{and}$$

$$\begin{align*}
\sigma_t(x, y) &= \sigma_s(x, y) = \begin{cases} 
\frac{c^4(c + \sqrt{2})^2(c - \sqrt{2})^2}{c^2 + y^2 < 1}, \\
1, & \text{otherwise}.
\end{cases}
\end{align*}$$

(3.65)
Note that the total cross section \( \sigma_t(x, y) \) can be periodically extended to a \( C^2 \)-function and \( \frac{\sigma_t(x, y)}{\varepsilon} \) ranges from 0 to 100. This wide range compared to the size of the domain causes strong variations of the solution, which are a challenge for numerical schemes.

We compute the solution up to time \( t \) on two different grids. One of the grids underresolves the length scale \( \varepsilon = \frac{1}{100} \) (32 \( \times \) 32 grid points) and the other one resolves it (512 \( \times \) 512 grid points). Comparing the solution at different times (\( t = \frac{1}{10}\varepsilon, \frac{1}{2}\varepsilon, \varepsilon \); see Figure 3.4 and 3.5), we observe that the density, computed on the underresolved grid, matches the behavior of the density, computed on the resolved grid.

### 3.4.3 Two Material Test

The two material test case is a slight modification of the lattice test, which was proposed in [16]. It models a domain with different materials by discontinuous material cross sections and a discontinuous source term in space.

In this problem, the computational domain is a 5 \( \times \) 5 square. Most of the domain is purely scattering, except for some purely absorbing squares of size 0.5, which are distributed around an isotropic source in the middle of the domain

\[
q(x, y) = \begin{cases} 
1, & (x, y) \in [2, 3]^2, \\
0, & \text{otherwise}.
\end{cases} \tag{3.66}
\]

In the absorbing spots (cf. Figure 3.6a), the absorption coefficient jumps from 0 to 100, while the scattering coefficient jumps from 1 to 0. Thus, there are diffusive and kinetic regimes, although the scaling parameter satisfies \( \varepsilon = 1 \). We obtain a rapid change of the solution at the transition zones, which may cause difficulties in the numerics.

We compute the density up to time \( t = 1.7 \) on a coarse grid (64 \( \times \) 64) and on a fine grid (512 \( \times \) 512). The solutions are shown in Figure 3.6. Again, we observe that the solution on the underresolved grid resembles the solution on a grid that is resolved. In the case of the resolved solution, the oscillations near the beam edges are due to the angular discretization. They are the well-known ray
3.4 Numerical Results

Figure 3.4: Variable scattering: Density $\rho$ at different end times $\frac{t}{\varepsilon} = 0.1$, 0.5, 1.0, computed on a $32 \times 32$ grid (first column) or a $512 \times 512$ grid (second column).
Figure 3.5: Variable scattering: Density $\rho$ and error at different end times $\frac{t}{\varepsilon} = 0.1, 0.5, 1.0$ on a cut along $y = 0$, computed with $n \times n$ grid points.
effects for finite discrete velocity models (cf. [15] and references therein, as well as [68, 79]).

3.4.4 Relaxation Parameters and Stability

In our final test, we consider different relaxation parameters. Proposition 3.6 suggests an upper bound on the relaxation parameter $\phi$, which in the hyperbolic case is more restrictive than in the parabolic case. We expect that in certain example cases our scheme becomes unstable if $\phi$ is too large.

Similar to the Gauss test, let

$$\psi(t = 0, x, y, \Omega) = \frac{1}{4\pi \cdot 3 \times 10^{-10}} \exp(-\frac{x^2+y^2}{4 \cdot 5 \times 10^{-3}}) \quad \text{for} \quad (x, y) \in [-1, 1] \times [-1, 1],$$

$$\varepsilon = 1, \quad q = 0, \quad \sigma_t = \sigma_s = 1, \quad \sigma_a = 0, \quad N = 300, \quad \text{and} \quad t = 0.36.$$  \hfill (3.67)

Then, we compute the density on a $N \times N = 300 \times 300$ grid up to time $t = 0.36$ using different relaxation parameters

$$\phi_1 = \frac{h \sigma_t}{2 \varepsilon^2} = \frac{10}{3} \times 10^{-4} \quad \text{and} \quad \phi_2 = \frac{1}{\varepsilon^2} = 1.$$  \hfill (3.68)

In the first case, the relaxation parameter satisfies the assumption of Proposition 3.6 and the solution is stable (see Figure 3.7a), whereas in the second case, the assumption is violated and the solution starts to blow up (see Figure 3.7b). As a consequence, the upper bound on the relaxation parameter in Proposition 3.6 can in general not be substituted by the less restrictive upper bound $\phi \leq \frac{1}{\varepsilon^2}$.

3.5 Extensions

3.5.1 Boundary Conditions

In the numerical examples, we consider only periodic boundary conditions. However, one can use different approaches to implement, for instance, inflow boundary conditions. In the following, we discuss how to adapt two different methods [60, 62] to our scheme. Let the computational domain be given by $X = [0, L_x] \times [0, L_y]$
3. AP STAGGERED GRID METHOD

![Diagram](image)

(a) Geometry

(b) 64 × 64 grid

(c) 512 × 512 grid

(d) Error

**Figure 3.6:** Two material test: (a) geometry – source (orange), purely scattering $\sigma_t = \sigma_s = 1$ (white and orange), purely absorbing $\sigma_t = \sigma_a = 100$ (black); (b) and (c) density $\rho$ at $t = 1.7$, computed on a 64 × 64 grid or 512 × 512 grid; (d) absolute error in the coarse grid compared to the fine grid. Logarithmic scaling, values are limited to seven orders of magnitude.
3.5 Extensions

Figure 3.7: Stability: Density on a cut along \( y = 0 \), computed on a 300 \( \times \) 300 grid up to time \( t = 0.36 \) using different values of \( \phi \).

for some \( L_x, L_y > 0 \). Then, inflow boundary conditions on a continuous level are given by

\[
\psi(t, \mathbf{x}, \Omega) = \psi^{\text{in}}(t, \mathbf{x}, \Omega) \quad \text{for} \quad n \cdot \mathbf{x} > 0, \quad \mathbf{x} \in \partial \mathbf{X},
\]

(3.69)

where \( n \) is the outer normal vector and \( \psi^{\text{in}} \) describes the inflow.

Similar to [39, 61, 62], the solution of a kinetic half-space problem, describing the zeroth order diffusion boundary conditions, can be approximated to determine the outflow function \( \psi^{\text{out}}(t, \mathbf{x}, v) \):

\[
\psi(t, \mathbf{x}, \Omega) = \psi^{\text{out}}(t, \mathbf{x}, \Omega) \quad \text{for} \quad n \cdot \mathbf{x} < 0, \quad \mathbf{x} \in \partial \mathbf{X}.
\]

(3.70)

This determines completely the density \( f \) on the boundary and with this also the parities on the boundary. Considering, for instance, the left boundary \( x = 0 \), the boundary conditions of the parities \( r^{(1)}, j^{(1)} \) are given by

\[
\begin{align*}
  r^{(1)}(\xi, \eta) + \varepsilon j^{(1)}(\xi, \eta) &= \psi(\xi, -\eta) = \psi^{\text{in}}(\xi, -\eta), \\
  r^{(1)}(\xi, \eta) - \varepsilon j^{(1)}(\xi, \eta) &= \psi(-\xi, \eta) = \psi^{\text{out}}(-\xi, \eta).
\end{align*}
\]

(3.71)

In one dimension, the parity \( r^{(1)} \) can be placed on the grid \( i \Delta x, \ i = 0, \ldots, N_x \) with \( N_x = \frac{L_x}{\Delta x} \) and \( j^{(1)} \) on the staggered grid \( (i + \frac{1}{2}) \Delta x, \ i = -1, \ldots, N_x \). Then \( r^{(1)} \) lies on the boundary and \( j^{(1)} \) can be interpolated using the ghost points.
3. AP STAGGERED GRID METHOD

\(-\frac{1}{2}\) and \(N_x + \frac{1}{2}\) \([62]\). This can be extended to two dimensions if we use a one-dimensional interpolation orthogonal to the boundary and add suitable ghost points (see Figure 3.8). Then, the boundary condition (3.71) becomes for \(r^{(1)}\) and \(j = 1, \ldots, N_y - 1\) and \(N_y = \frac{L_y}{\Delta y}\)

\[ r^{(1)}_{0,j} + \varepsilon \left( j^{(1)}_{-\frac{1}{2},j} + j^{(1)}_{\frac{1}{2},j} \right) = \psi^{in}_{0,j} \quad \text{and} \quad r^{(1)}_{0,j} - \frac{\varepsilon}{2} \left( j^{(1)}_{-\frac{1}{2},j} + j^{(1)}_{\frac{1}{2},j} \right) = \psi^{out}_{0,j} \]  \( (3.72) \)

and for \(j^{(1)}\) and \(j = \frac{1}{2}, 1 + \frac{1}{2}, \ldots, N - \frac{1}{2}\)

\[ \frac{1}{2} \left( r^{(1)}_{-\frac{1}{2},j} + r^{(1)}_{\frac{1}{2},j} \right) + \varepsilon j^{(1)}_{0,j} = \psi^{in}_{0,j} \quad \text{and} \quad \frac{1}{2} \left( r^{(1)}_{-\frac{1}{2},j} + r^{(1)}_{\frac{1}{2},j} \right) - \varepsilon j^{(1)}_{0,j} = \psi^{out}_{0,j} \]  \( (3.73) \)

Note that the edges of the boundary can be dropped, because these points are needed neither for computing the inner points nor for the interpolation (due to the one-dimensional interpolation rule). This choice yields a system with the same number of equations and unknowns.

![Figure 3.8: Black solid line denotes the boundary of the domain \(\partial X\); filled markers denote inner points; empty markers denote boundary and ghost points.](image)

Alternatively, the boundary conditions presented in \([60]\) can be used. The key idea is to approximate the \(j\)-unknowns on the boundary by

\[ j^{(1)} = -\xi \partial_x r^{(1)} - \eta \partial_y r^{(1)} \quad \text{and} \quad j^{(2)} = -\xi \partial_x r^{(2)} - \eta \partial_y r^{(2)} \]  \( (3.74) \)

and insert this into the inflow boundary parts. For example, at the left boundary \(x = 0\), the boundary condition for \(r^{(1)}\) is given by

\[ \psi^{in}(\xi, -\eta) = r^{(1)}(\xi, \eta) + \varepsilon j^{(1)}(\xi, \eta) = r^{(1)}(\xi, \eta) - \varepsilon \left( \xi \partial_x r^{(1)} + \eta \partial_y r^{(1)} \right) \]  \( (3.75) \)
To implement the boundary conditions in our scheme the spatial derivatives can, for instance, be approximated by one-sided finite-differences. There are multiple choices to do this, which need to be tested.

3.5.2 Anisotropic Scattering

The simplicity of the scheme relies on the simplicity of the scattering operator, which is isotropic. In the following, we briefly consider the case of anisotropic scattering. Let the scattering operator be given by

\[
K\psi(\Omega) = \int_{S^1} g(\Omega, \Omega') \psi(\Omega') \, d\Omega'
\]  

(3.76)

for some symmetric scattering kernel \( g \), which depends on the cosine of the scattering angle \( \Omega \cdot \Omega' \).

To compute the parity equations, the transport equation is rewritten for \( \Omega = (\xi, \eta), (\xi, -\eta), (-\xi, \eta), \) and \((-\xi, -\eta)\) with nonnegative \( \xi, \eta \geq 0 \) and added in correspondence with the definition of the parities (3.13). This leads to a summation of the scattering operator of the form \( K\psi(\Omega) \pm K\psi(-\Omega) \), which has to be expressed in terms of the parities. Since the scattering kernel satisfies \( g(-\Omega, \Omega') = g(\Omega, -\Omega') \), we obtain

\[
K\psi(\Omega) \pm K\psi(-\Omega) = \int_{S^1} g(\Omega, \Omega') \psi(\Omega') \, d\Omega' \pm \int_{S^1} g(-\Omega, \Omega') \psi(\Omega') \, d\Omega'
\]

(3.77)

More precisely, we consider for \( \xi, \eta \geq 0 \)

\[
\frac{1}{2} (K\psi(\xi, -\eta) + K\psi(-\xi, \eta)) = \int_{S^1} g(\xi, \eta, \xi', \eta') \frac{1}{2} (\psi(\xi', -\eta') + \psi(-\xi', \eta')) \, d(\xi', \eta').
\]

(3.78)

In this term, we recover the parities \( r^{(1)} \) and \( r^{(2)} \)

\[
\frac{1}{2} (\psi(\xi', -\eta') + \psi(-\xi', \eta')) = \begin{cases} 
  r^{(1)} & \text{for } \xi', \eta' \geq 0 \text{ or } \xi', \eta' \leq 0, \\
  r^{(2)} & \text{otherwise}.
\end{cases}
\]

(3.79)
Splitting the integral into the four quadrants and changing to positive $\xi', \eta'$, we obtain
\[
\frac{1}{2} (K \psi(\xi, -\eta) + K \psi(-\xi, \eta)) = \int_{\xi,\eta>0} (g(\xi, \eta, \xi', \eta') + g(\xi, \eta, -\xi', -\eta')) r^{(1)}(\xi', \eta') d(\xi', \eta') \\
+ \int_{\xi,\eta>0} (g(\xi, \eta, \xi', -\eta') + g(\xi, \eta, -\xi', \eta')) r^{(2)}(\xi', \eta') d(\xi', \eta').
\] (3.80)

Similarly, the other summations can be expressed in the parities. To simplify the notation, we define the following operators:
\[
K^+ \psi(\xi, \eta) := \int_{\xi,\eta>0} (g(\xi, \eta, \xi', \eta') + g(\xi, \eta, -\xi', -\eta')) f(\xi', \eta') d(\xi', \eta'),
\]
\[
K^- \psi(\xi, \eta) := \int_{\xi,\eta>0} (g(\xi, \eta, \xi', -\eta') + g(\xi, \eta, -\xi', \eta')) f(\xi', \eta') d(\xi', \eta').
\] (3.81)

Then, the parity equations for anisotropic scattering are given by
\[
\partial_t r^{(1)} + \xi \partial_x j^{(1)} - \eta \partial_y j^{(1)} = -\frac{\sigma_s}{\varepsilon^2} (r^{(1)} - K^+ r^{(1)} - K^- r^{(2)}) - \sigma_a r^{(1)} + q,
\]
\[
\partial_t r^{(2)} + \xi \partial_x j^{(2)} + \eta \partial_y j^{(2)} = -\frac{\sigma_s}{\varepsilon^2} (r^{(2)} - K^+ r^{(2)} - K^- r^{(1)}) - \sigma_a r^{(2)} + q,
\]
\[
\partial_j^{(1)} + \frac{\xi}{\varepsilon^2} \partial_x r^{(1)} - \frac{\eta}{\varepsilon^2} \partial_y r^{(1)} = -\frac{\sigma_s}{\varepsilon^2} (j^{(1)} - K^+ j^{(1)} - K^- j^{(2)}) - \sigma_a j^{(1)},
\]
\[
\partial_j^{(2)} + \frac{\xi}{\varepsilon^2} \partial_x r^{(2)} + \frac{\eta}{\varepsilon^2} \partial_y r^{(2)} = -\frac{\sigma_s}{\varepsilon^2} (j^{(2)} - K^+ j^{(2)} - K^- j^{(1)}) - \sigma_a j^{(2)}.
\] (3.82)

One could now apply a similar splitting as above, but in contrast to the isotropic case the implementation of the relaxation step is not straightforward. It is necessary to invert the integral operators at a potentially expensive computational cost (cf. [58] for more details).

\textbf{Remark 3.12.} The case of anisotropic scattering is particularly important for applications, for instance, if we want to apply the method to the electron transport problems considered in Chapter 1.
Chapter 4

Two-Level Sampling Strategies for Uncertainty Quantification in Hyperbolic Relaxation Systems

4.1 Introduction

In this chapter, we introduce a new sampling strategy for hyperbolic relaxation systems to quantify the effect of uncertain input parameters on the solution. In general, uncertainties can be introduced into the system through model errors, numerical errors, measurement errors, and input uncertainties [108]. For instance, in radiation therapy (cf. Chapter 1), the model relies on measurement data from a CT scan of the patient and on material parameters, such as scattering cross sections. All parameters need to be modeled and precomputed, so that they introduce various uncertainties in the system.

We model the uncertain input parameters by a random variable and assume that their distribution function is known. Then the solution becomes a random variable as well, but with unknown distribution. The aim is to analyze how the uncertainties propagate through the model and how they effect a particular quantity of interest (QoI). This problem is known as forward uncertainty propagation. There are various approaches to solve the forward uncertainty propagation problem, including stochastic polynomial or sampling-based methods.
4. TWO-LEVEL SAMPLING STRATEGIES

Stochastic polynomial methods, such as stochastic Galerkin methods (also known as general polynomial chaos (gPC) [120]), use polynomial expansions to obtain an approximation of the moments of a QoI. As a spectral method, the stochastic Galerkin method achieves fast convergence rates for sufficiently smooth solutions. However, to apply the method, the Galerkin system has to be derived and analyzed and a code has to be implemented for the specific application (intrusive method).

Sampling-based methods, such as Monte Carlo (MC) or stochastic collocation (SC) methods, take samples from the distribution of the input parameter and solve the model for each one to obtain samples of the QoI. The samples can then be used to approximate the distribution of the QoI or its mean and variance. In particular, the methods do not require a modification of the numerical solver itself (non-intrusive). This is a big advantage over the stochastic polynomial methods.

MC methods use random samples and rely on the law of large numbers. For Monte Carlo integration, it is well known that the convergence rate is \( O \left( M^{-1/2} \right) \) with a constant that is given by the variance of the integrand [18] and \( M \) being the number of samples. Due to the slow convergence, usually many samples are needed. This is particularly expensive in the case of forward uncertainty propagation as for every sample the model, i.e. a partial differential equation, has to be solved. On the other hand, the convergence rate is independent of the dimension of the random variable, so that high-dimensional problems can be considered and there are various approaches to reduce the number of samples such as quasi-Monte Carlo methods [18, 86, 107] and Multilevel Monte Carlo (MLMC) methods [38, 48].

SC methods [9, 87] use deterministic samples. In many cases, the random variable is sampled at quadrature points, so that the expected value can be computed efficiently, but one may also sample at interpolation points. Although the methods can achieve a higher convergence rate, quadrature based methods suffer from the curse of dimensionality (see for instance [108]). In general, the convergence rate diminishes with increasing dimension and the computational cost rises dramatically. However, the methods can be improved by a multilevel approach leading to sparse grids. Sparse grid formulas were first introduced by Smolyak [109] and further developed and analyzed in [37, 89, 90]. They are
4.1 Introduction

extensions of one-dimensional quadrature (or interpolation) formulas to higher
dimensions via a linear combination of tensor products. We briefly introduce the
concept at the end of this section.

Our sampling strategy uses a multilevel sparse grid stochastic collocation ap-
proach. Typically, a multilevel approach uses a hierarchy of grids to reduce the
cost by balancing the different discretization and statistical error terms. This
idea is for instance used for the construction of MLMC methods and also for the
construction of multilevel SC methods [112]. The main difference to our method
is that we use the multilevel framework for a hierarchy of models instead of a
hierarchy of grids.

A hierarchy of models is induced by a hyperbolic relaxation system through
the reduced models of various order. When collisions dominate the dynamics, we
are in the asymptotic regime and the reduced models have a very high accuracy.
But we expect that also in transition regimes the reduced models give reasonable
information. The combination of the asymptotic limit and a correction has been
heavily used for deterministic settings, for instance, in a micro-macro decompo-
sition [69, 71]. Here, we apply the hybrid approach to the stochastic formulation.

The novelty of our method is to use this model hierarchy to construct a two-
level sparse grid SC method. Similar to MLMC or multilevel SC, the major idea
is to balance the different error terms. To balance the error terms in an optimal
way, a model for the error in the quadrature is needed, which we approximate by
asymptotic error bounds. We expect that only a few samples are needed for the
correction, as the term has a small absolute value, and more samples are needed
for the reduced model. However, since the reduced model is less expensive to
solve, a high accuracy can be expected at a reduced computational cost.

In what follows, we first briefly introduce sparse grids. Then, in section 4.2,
we describe and analyze our new sampling strategy and apply the method to
specific models in section 4.3 and 4.4. Although the method can be applied
to any hyperbolic relaxation system, we only consider linear kinetic equations
(discrete velocity model and $P_N$ system) for now. Various numerical test cases
are presented to confirm the expected cost savings of the new sampling method
compared to the standard SC method.
4. TWO-LEVEL SAMPLING STRATEGIES

4.1.1 Sparse Grids

In this subsection, we briefly introduce the concept of sparse grids (more details can be found in [17, 51, 108] and references therein). Sparse grid quadrature formulas are extensions of one-dimensional quadrature formulas to approximate the $d$-dimensional integral

$$\int_\Gamma \phi(x) \, dx$$

(4.1)

of a function $\phi : \Gamma \to \mathbb{R}$ with $\Gamma = \Gamma_1 \times \cdots \times \Gamma_d$ and $\Gamma_i \subset \mathbb{R}$, $i = 1, \ldots, d$.

In one dimension, a variety of quadrature formulas is well established. For instance, the Gauss quadrature formulas can be used to approximate the integral of a function $\phi : [0,1] \to \mathbb{R}$ by

$$\int_0^1 \phi(x) \, dx \approx \sum_{i=1}^{M_1} \omega_i \phi(\xi_i) =: Q^1 \phi,$$

(4.2)

where $\omega_i$, $\xi_i$ are the quadrature weights and nodes. The Gauss quadrature rules build a family of quadrature formulas of level $\ell = 1, 2, \ldots$ with the highest precision, meaning that a quadrature formula with $M_\ell$ points can integrate polynomials up to degree $2M_\ell - 1$ exactly. The main drawback of Gauss quadrature formulas is that they are not nested. This means that the nodes of the level $\ell$ quadrature formula are not contained in the nodes of the subsequent level $\ell + 1$, so that refinement strategies are costly. Nested one-dimensional quadrature formulas are, for example, given by the Clenshaw-Curtis formulas [22] and the Gauss-Patterson formulas [94]. Due to their nested structure, these formulas are better suited for the construction of higher dimensional quadrature formulas.

In higher dimensions, we can apply successively one-dimensional quadrature formulas for each dimension. This leads to a tensor product quadrature

$$\int_{[0,1]^d} \phi(x) \, dx \approx \sum_{i_1=1}^{M_{i_1}} \cdots \sum_{i_d=1}^{M_{i_d}} \omega_{i_1} \cdots \omega_{i_d} \phi(\xi_{i_1}, \ldots, \xi_{i_d}) =: Q^{\ell_1} \cdots Q^{\ell_d} \phi,$$

(4.3)

where $\phi : [0,1]^d \to \mathbb{R}$. Then the number of function evaluations is equal to the number of permutations of the input variables and therefore given by $M = \prod_{i=1}^d M_{\ell_i}$, so that the computational cost explodes for higher dimensions. This is a particular problem in the context of forward uncertainty propagation,
where each function evaluation requires the solution of a partial differential equation. Moreover, the convergence rate, expressed in the total number of degrees of freedoms, diminishes with increasing dimension, so that tensor product quadrature formulas are unfeasible for high dimensional problems. To alleviate this curse of dimensionality, sparse grids are constructed.

We define the surplus operators \( \Delta^0 = 0 \) and \( \Delta^\ell = (Q^\ell - Q^{\ell-1}) \) for \( \ell > 0 \). Then

\[
Q^\ell_d \phi := \sum_{|\mathbf{i}| \leq \ell + d - 1} \Delta^i_d \cdots \Delta^1_d \phi ,
\]

(4.4)

with \( \mathbf{i} = (i_1, \ldots, i_d) \) and \( |\mathbf{i}| = \sum_{k=1}^d i_k \), is a \( d \)-dimensional quadrature formula and defines a classical sparse grid \([109]\). Note that the above sum actually has a telescoping structure, so that it can be rewritten in terms of the underlying one-dimensional quadrature formulas as

\[
Q^\ell_d \phi = \sum_{\ell \leq |\mathbf{k}| \leq \ell + d - 1} (-1)^{\ell + d - |\mathbf{k}| - 1} \binom{d - 1}{|\mathbf{k}| - \ell} Q^{k_1} \ldots Q^{k_d} \phi .
\]

(4.5)

This already indicates the efficiency of nested quadrature formulas and the advantage over the tensor product formulas. More general, a sparse grid quadrature formula is given by

\[
G_\theta \phi := \sum_{\mathbf{i} \in \theta} \Delta^i_d \cdots \Delta^1_d \phi
\]

(4.6)

for a lower set \( \theta \). The index set \( \theta \) is a lower set if for any \( \mathbf{i} \in \theta \) we have

\[
\{ \mathbf{k} \in \mathbb{N}^d : k_1 \leq i_1, \ldots, k_d \leq i_d \} \subset \theta .
\]

(4.7)

In this case, the quadrature is still given as a telescoping sum expansion similar to (4.5), but the choice of \( \theta \) allows adaptations, for example if certain information about the importance of the dimensions is available. We denote a family of sparse grids with increasing accuracy by \( (G_\ell)_{\ell=1,2,\ldots} \), where \( \ell \) is the level of the quadrature formula.

While the accuracy of sparse grid quadrature formulas and tensor product quadrature formulas is comparable in the sense of the total polynomial degree, the number of points in a sparse grid is substantially reduced compared to the number of points in a tensor product grid. As an example we consider the two-dimensional
4. TWO-LEVEL SAMPLING STRATEGIES

Figure 4.1: Two-dimensional sparse grid and tensor product grid constructed with the one-dimensional Gauss-Patterson formulas with maximal seven points.

sparse grid and the tensor product grid constructed with the one-dimensional Gauss-Patterson formulas with maximal seven points (see Figure 4.1). While the tensor product grid has \( M = 7^2 = 49 \) points, the sparse grid has only \( M = 17 \) points.

Remark 4.1. For the approximation error, there are two structurally different asymptotic error bounds in the literature. The first one describes an algebraic decay of the form \[ M^{-\mu}, \] where \( M \) is the number of points and \( \mu > 0 \) depends on the regularity of the solution with respect to the random variable and the order of convergence of the one-dimensional quadrature formula. The second one is an exponential type estimate of the form \[ M \exp(-\mu M^{\frac{1}{d}}), \] where \( M \) is again the number of points and \( \mu > 0 \) is a constant independent of the dimension.

Remark 4.2. For the computations, we use the sparse grid library TASMANIAN [110] to generate the quadrature rules. The quadrature rules used in the computations are based on the one-dimensional Gauss-Patterson rules [94], since Gauss-Patterson rules are nested and have a high precision.
4.2 Sampling Strategy

In the following, we introduce a two-level sampling strategy to approximate the distribution of a QoI if some input parameters are uncertain. As mentioned before, we model the uncertainties by a $d$-dimensional random variable $\xi$. Further, let $\psi(\xi)$ be the solution of a hyperbolic relaxation system, where we suppress the deterministic arguments such as time, position, and velocity. Then $\psi(\xi)$ is a random variable and we want to determine the distribution of a QoI $f(\psi(\xi))$, where $f$ is point-wise defined in $\xi$. For instance, $f$ can be given by the $L^2$ norm (in space) of the first angular moment of $\psi$ at a certain time $t$.

Usually, the moments of the stochastic distribution are of main interest. Assuming that the random variable $\xi$ is independent and identically distributed (i.i.d.), we can, for example, approximate the expectation with a sparse grid quadrature formula

$$E[f(\psi)] \approx G_\ell f(\psi),$$

where $G_\ell$ is a suitable sparse grid of level $\ell$. Here, the underlying one-dimensional quadrature family has to be chosen in accordance with the distribution of $\xi$. To compute the sparse grid estimator $G_\ell(f(\psi))$, the hyperbolic relaxation system has to be solved for each point in the sparse grid. This is computationally expensive, so that for a large number of points the computational cost rises dramatically.

To reduce the computational cost, we use a multilevel framework, but instead of using different discretizations of the underlying system, we use a model hierarchy. A model hierarchy is given by a hyperbolic relaxation system and its reduced models of different order. For instance, we can use the linear transport equation in scaled variables

$$\varepsilon \partial_t \psi + \Omega \cdot \nabla_x \psi = \frac{1}{\varepsilon} \left[ \frac{\sigma_s}{2\pi} \int_{S^1} \psi \, d\Omega' - \sigma_t \psi \right] + \varepsilon q$$

and the limiting ($\varepsilon \to 0$) diffusion equation (discussed in Chapter 3)

$$\partial_t \rho = \frac{1}{2} \nabla_x \cdot \left( \frac{1}{\sigma_s} \nabla_x \rho \right) - \sigma_a \rho + q,$$

which gives an $O(\varepsilon^2)$ approximation of the transport equation. Further examples for the full and the reduced model include the Boltzmann equation and the Euler...
4. TWO-LEVEL SAMPLING STRATEGIES

or Navier-Stokes limit. For these equations, the reduced models are structurally different and lower dimensional compared to the full model. This leads to a great reduction of the computational cost for the reduced model compared to the full model.

The limiting equations only yield accurate solutions in the collisional regime \((\varepsilon \ll 1)\), i.e. when collisions dominate the dynamics. In this case, we can replace the full model by a reduced model in (4.10) and obtain the results at a much cheaper computational cost. On the other hand, in the rarefied case \((\varepsilon = \mathcal{O}(1))\), the full model is very different from the reduced models and we cannot expect to reduce the cost using a reduced model. However, many problems are in the transition, where the reduced models do not yield the desired accuracy, but still contain some reasonable information. In this case, we now want to combine a reduced model and a correction, so that we still obtain an accurate solution at a reduced computational cost.

From now on, the subscript \(F\) always refers to the hyperbolic relaxation system, which we also call the “full model”, and the subscript \(R\) refers to the \(p\)-th order reduced system, which leads to an \(\mathcal{O}(\varepsilon^p)\) approximation of the solution of the full model. Thus, let \(\psi_F(\xi)\) be the solution of the full model and \(\psi_R(\xi)\) be the solution of the reduced model with \(\psi_F(\xi) - \psi_R(\xi) = \mathcal{O}(\varepsilon^p)\). As mentioned above, it is usually much cheaper to solve the reduced model \(\psi_R(\xi_i)\) for a sample \(\xi_i\) than to solve the full model \(\psi_F(\xi_i)\).

To combine the two models, we use a multilevel approach. First, we rewrite the expectation as

\[
E[f(\psi_F)] = E[f(\psi_R)] + E[f(\psi_F) - f(\psi_R)].
\]

(4.13)

This is a trivial equality, but the main idea is to use different estimators for the two expected values on the right-hand side

\[
E[f(\psi_F)] \approx G_{\ell_R} f(\psi_R) + G_{\ell_D}(f(\psi_F) - f(\psi_R)),
\]

(4.14)

where the subscript \(D\) refers to the difference between the models, and \(G_{\ell_R}\) and \(G_{\ell_D}\) are sparse grids of different levels with \(M_R \geq M_D\) points. Essentially, we approximate the full model by the reduced model and a correction. While the
first term is less expensive, the second term has a smaller absolute value and with this a smaller variance. Similar to the Monte Carlo error estimate, where the convergence rate is $O(M^{-1/2})$ with a constant that is given by the variance of the integrand, this reduces the constant in the convergence rate and with this the number of required samples. Therefore, we expect that the full model has to be sampled less frequently and the total computational cost can be reduced. It remains to choose the number of points $M_R$ and $M_D$ of the sparse grids, or to balance the model error and the stochastic error.

### 4.2.1 Error and Cost Analysis

The aim is to choose the optimal number of grid points in (4.14) by minimizing the cost for a fixed error tolerance. To formulate this in a constrained optimization problem, we need to analyze the error and the cost.

Let the cost to solve the full model $\psi_F(\xi_i)$ and the reduced model $\psi_R(\xi_i)$ for a single sample $\xi_i$ be given by $\text{Cost}_F$ and $\text{Cost}_R$. If we assume that the main contribution to the cost is the evaluation of the models, the total computational cost can be approximated by

$$\text{Cost} = M_R \text{Cost}_R + M_D \text{Cost}_F . \quad (4.15)$$

Here, we additionally assume that the quadrature formula is nested and $M_R \geq M_D$, so that the samples to compute the first term $G_{t_R} f(\psi_R)$ can be reused to compute $G_{t_D} (f(\psi_R))$ in the second term in (4.14).

To model the error of the sparse grid quadrature formulas, we use asymptotic error bounds. We assume for now that the asymptotic error bounds introduced in Remark 4.1 are explicitly known (including all constants) for the sparse grid quadrature formulas in (4.14). Then we split the error into two parts

$$\| E[f(\psi_F)] - (G_{t_R} f(\psi_R) + G_{t_D} (f(\psi_F) - f(\psi_R))) \|$$

$$\leq \| E[f(\psi_R)] - G_{t_R} f(\psi_R) \| + \| E[f(\psi_F) - f(\psi_R)] - G_{t_D} (f(\psi_F) - f(\psi_R)) \|$$

(4.16)
and estimate each of the terms separately. Using the first error bound (4.8), we obtain
\[
\| \mathbb{E}[f(\psi_F)] - (G_{\ell_R}f(\psi_R) + G_{\ell_D}(f(\psi_F) - f(\psi_R))) \| 
\leq c_R M_R^{-\mu_R} + c_D M_D^{-\mu_D} \tag{4.17}
\]
for some constants \(c_R, c_D, \mu_R,\) and \(\mu_D.\) Analogously, using the second error bound (4.9), we obtain
\[
\| \mathbb{E}[f(\psi_F)] - (G_{\ell_R}f(\psi_R) + G_{\ell_D}(f(\psi_F) - f(\psi_R))) \|
\leq c_R M_R \exp(-\mu_R M_R^{\frac{1}{D}}) + c_D M_D \exp(-\mu_D M_D^{\frac{1}{D}}) \tag{4.18}
\]
for some different constants \(c_R, c_D, \mu_R,\) and \(\mu_D.\)

Finally, we determine the optimal number of grid points by minimizing the cost given some error tolerance \(\text{Error}_{\text{tol}}.\) Then the optimal numbers of grid points \(M_R\) and \(M_D\) are obtained by solving
\[
\min_{M_R, M_D} \text{Cost}_R M_R + \text{Cost}_F M_D \quad \text{s.t.} \quad c_R M_R^{-\mu_R} + c_D M_D^{-\mu_D} = \text{Error}_{\text{tol}} \tag{4.19}
\]
if the algebraic type error estimate (4.17) is used, or by solving
\[
\min_{M_R, M_D} \text{Cost}_R M_R + \text{Cost}_F M_D \quad \text{s.t.} \quad c_R M_R \exp(-\mu_R M_R^{\frac{1}{D}}) + c_D M_D \exp(-\mu_D M_D^{\frac{1}{D}}) = \text{Error}_{\text{tol}} \tag{4.20}
\]
if the exponential type error estimate (4.18) is used. The main problem is that the constants \(c_R, c_D, \mu_R,\) and \(\mu_D\) are usually not explicitly known and need to be approximated numerically.

**Example 4.3.** Let the convergence rate of the reduced model coincide with the convergence rate of the difference \(\mu := \mu_R = \mu_D.\) In this case, the optimization problem (4.19) can be solved analytically and the optimal numbers of grid points are given by
\[
M_R = \text{Error}_{\text{tol}}^{-\frac{1}{\mu}} \left( c_D^{\frac{1}{\mu+1}} \left( \frac{c_R}{\kappa} \right)^{\frac{\mu}{\mu+1}} + c_R \right)^{\frac{1}{\mu}} \quad \text{and} \quad \kappa := \frac{\text{Cost}_R}{\text{Cost}_F}, \\
M_D = \text{Error}_{\text{tol}}^{-\frac{1}{\mu}} \left( c_R^{\frac{1}{\mu+1}} \left( \kappa c_D \right)^{\frac{\mu}{\mu+1}} + c_D \right)^{\frac{1}{\mu}} \quad \text{with} \quad \kappa := \frac{\text{Cost}_R}{\text{Cost}_F}. \tag{4.21}
\]
For comparison with the standard method (4.10), we assume that the full model converges at the same rate $\mu$

$$\|E[f(\psi_F)] - G_{\ell_F}f(\psi_F)\| \leq c_F M_F^\mu,$$  \hfill (4.22)

so that the error tolerance is reached for

$$M_F = \left( \frac{c_F}{\text{Error}_{\text{tol}}} \right)^{\frac{1}{\mu}}.$$  \hfill (4.23)

Then the total cost ratio between the two methods is given by

$$\frac{\kappa M_R + M_D}{M_F} = \kappa c_F^{-\frac{1}{\mu}} \left( c_D^{\frac{1}{\mu}} \left( \frac{c_R}{\kappa} \right)^{\frac{\mu}{\mu+1}} + c_R \right)^{\frac{1}{\mu}} + c_F c_R \left( (\kappa c_D)^{\frac{\mu}{\mu+1}} + c_D \right)^{\frac{1}{\mu}}.$$  \hfill (4.24)

The cost ratio as a function of the convergence order $\mu$ and the single computational cost ratio $\kappa$ is shown in Figure 4.2 for $c_F = c_R = 1$, $c_D = \varepsilon^p$ with $p \in \{1, 2\}$, where $p$ is the order of the reduced model. We observe that potentially more cost savings can be achieved using the second-order reduced model instead of the first-order reduced model. Moreover, we observe that the cost savings decrease with increasing convergence rate.

**Remark 4.4.** For the error models (4.17) and (4.18), it is particularly important that the solution is sufficiently regular with respect to the random variable. The regularity usually depends on the regularity of the initial condition and might also depend on the decay of its Fourier coefficients, so we assume that the initial condition is analytic and the Fourier conditions fall off sufficiently fast (see [82, 83]). As the regularity can also depend on the regularity of the boundary conditions, we also choose them in a suitable way.

### 4.2.2 Numerical Strategy

Theoretically, optimal grids for the two-level sampling strategy can be chosen by solving one of the above optimization problems. However, the optimization problems are based on the asymptotic error models, where the constants are usually not explicitly known. Thus, it remains to develop a practical strategy to choose the numbers of points in the quadrature $M_R$ and $M_D$. We consider the following four approaches:
4. TWO-LEVEL SAMPLING STRATEGIES

Figure 4.2: Cost ratio as a function of the convergence order $\mu$ and the single computational cost ratio $\kappa$ (values are limited to $[0, 1]$). Results are computed for the reduced models of order $p = 1, 2$ with different relaxation parameters $\varepsilon = 0.1$ and $\varepsilon = 0.5$. 
4.2 Sampling Strategy

**Approach 1 (empirical)**: Independent of the asymptotic error models, we first determine any possible cost savings with a two-level approach over the standard SC approach. To this end, we compute estimates for the full and the reduced model with quadrature formulas of different levels. Then we consider all combinations of levels \( \ell_R \) and \( \ell_D \) and search among those combinations, which reached a given error tolerance, for the minimal cost. Obviously, this strategy is not practical, but yields a benchmark for the other approaches and investigates the general capabilities of the two-level sampling strategy. In addition, we approximate the accuracy of the reduced model in an appropriate norm

\[
\alpha_R = \| E[f(\psi_F) - f(\psi_R)] \| \tag{4.25}
\]

with a high level quadrature formula. We call this the *empirical* approach.

**Approach 2 (estimated)**: To validate the optimization problem, we estimate the constants in the error models with the help of fully converged solutions. Using the converged solutions as a reference, we can compute the separate errors in (4.16) for a number of levels \( \ell_R, \ell_D \in \{1, 2, \ldots\} \). Then we estimate the constants \( c_R, c_D, \mu_R, \) and \( \mu_D \) in the error models with a least square fit of the logarithmic rates in (4.17), which have the form

\[
\log(cM^\mu) = \log(c) - \mu \log(M) \tag{4.26}
\]

and are linear in \( \log(M) \), or the logarithmic rates in (4.18), which have the form

\[
\log(cM \exp(-\mu M^{1/2})) - \log(M) = \log(c) - \mu M^{1/2} \tag{4.27}
\]

and are linear in \( M^{1/2} \). We insert the estimates into the optimization problems (4.19) and (4.20) and solve for \( M_R \) and \( M_D \). Once, we have the number of points, we can construct sparse grids with at least \( M_R \) and \( M_D \) points and compute (4.14). This again is not a practical strategy, but is a plausibility check for the optimization problems. We call this the *estimated* approach.

**Approach 3 (iterative)**: Independent of the error models, but more practical, is an iterative approach. For the standard SC approach, we simply refine
4. TWO-LEVEL SAMPLING STRATEGIES

the quadrature formula, i.e. increase the level, until a given error tolerance is reached. As a measure for the error, we use the difference between two successive quadrature levels in an appropriate norm. Similarly, for the two-level approach, we use

\[ e_R = \| G_{\ell R} f(\psi_R) - G_{\ell R-1} f(\psi_R) \| \quad \text{and} \quad e_D = \| G_{\ell D} (f(\psi_F) - f(\psi_R)) - G_{\ell D-1} (f(\psi_F) - f(\psi_R)) \| \]  

for \( \ell_R, \ell_D = 1, 2, \ldots \) to successively estimate the separate error terms in (4.16). First, we increase the level \( \ell_D = 1, 2, \ldots \) until the error \( e_D \) drops below a given error tolerance. Second, we increase the level \( \ell_R = 1, 2, \ldots \) until the sum of the errors \( e_R + e_D \) achieves the error tolerance. Note that this approach assumes that it is always better to increase the level of the reduced model \( \ell_R \) instead of increasing the level of the difference \( \ell_D \). This means we assume that the reduced model is much cheaper to solve than the full model. We call this the iterative approach.

**Approach 4 (greedy)**: We consider another practical approach for the two-level sampling strategy. It is similar to the iterative approach, but we increase the levels \( \ell_R \) and \( \ell_D \) simultaneously using a greedy algorithm. We follow Algorithm 1 and compute in each step error estimates and weighted gain terms to decide whether the method is converged and which level \( \ell_R \) or \( \ell_D \) should be increased. As before, we estimate the separate error terms in (4.16) successively by \( e_R \) and \( e_D \) given in (4.28). For the weighted gain, we estimate the efficiency of each term by \( w_R = \frac{\tilde{e}_R}{C_R} \) and \( w_D = \frac{\tilde{e}_D}{C_D} \), where \( \tilde{e}_R, \tilde{e}_D, C_R, C_D \) are estimates for the cost and the error. We estimate the cost by

\[ C_R = (M_\ell - M_{\ell-1}) \text{Cost}_R \quad \text{and} \quad C_D = (M_\ell - M_{\ell-1}) \text{Cost}_F \]  

and the error either by \( \tilde{e}_R = e_R \) and \( \tilde{e}_D = e_D \) or by using the error models. If we use the error models, we change the error estimates \( \tilde{e}_R = e_R \) and \( \tilde{e}_D = e_D \) when the level is larger than two. In this case, we first compute the error in the lower levels compared to the last level. Second, we compute a model for the error using a least square fit as in (4.26) and (4.27). Third,
we use the error model to estimate the error in level $\ell_R$ or $\ell_D$. We call this the *greedy* approach (and indicate if the error models are used).

Initialize the levels: $\ell_R = \ell_D = 1$.

Initialize the error:

$$e_R = \|G_{\ell_R}F(\psi_R)\| \quad \text{and} \quad e_D = \|G_{\ell_D}(f(\psi_F) - f(\psi_R))\|.$$ 

Initialize the weighted gain: $w_R$ and $w_D$.

while $e_D + e_R > \text{Error}_{\text{tol}}$ do

if $w_D < w_R$ then

Increase level: $\ell_R = \ell_R + 1$.

Compute sparse grid estimate: $G_{\ell_R}F(\psi_R)$.

Update error: $e_R = \|G_{\ell_R}F(\psi_R) - G_{\ell_{R-1}}F(\psi_R)\|$.

**Update weighted gain**: $w_R = \frac{\tilde{e}_R}{C_R}$

Compute $\tilde{e}_D$ and $C_D$ as described in Approach 4 (*greedy*).

else

Increase level: $\ell_D = \ell_D + 1$.

Compute sparse grid estimate: $G_{\ell_D}(f(\psi_F) - f(\psi_R))$.

Update error:

$$e_D = \|G_{\ell_D}(f(\psi_F) - f(\psi_R)) - G_{\ell_{D-1}}(f(\psi_F) - f(\psi_R))\|.$$ 

**Update weighted gain** (using the error model): $w_D = \frac{\tilde{e}_D}{C_D}$

Compute $\tilde{e}_R$ and $C_R$ as described in Approach 4 (*greedy*).

end

end

Compute two-level estimate: $G_{\ell_R}F(\psi_R) + G_{\ell_D}(f(\psi_F) - f(\psi_R))$.

**Algorithm 1**: Greedy algorithm for the two-level sampling strategy.

### 4.3 The Discrete Velocity Model

In this section, we consider an important example for a hyperbolic relation system, namely, the discrete velocity model of the linear transport equation with
4. TWO-LEVEL SAMPLING STRATEGIES

relaxation (note that this is not the same scaling as in (4.11)). In this case, the
discrete velocity model (DVM) can be written as
\[
\partial_t \psi_k + \Omega_k \cdot \nabla_x \psi_k = \frac{1}{\varepsilon} L_k(\psi), \quad k = 1, \ldots, K
\]
(4.30)
where \(\psi_k(t, x)\) describes the density of particles at time \(t \in \mathbb{R}\) and position \(x \in \mathbb{R}^s\) moving with velocity \(\Omega_k \in \mathbb{R}^s, s = 1, 2, 3\). The transition rate of particles from velocity \(k\) to \(j\) is given by \(\lambda_{j,k} \geq 0\), which are gathered in the linear operators
\[
L_k(\psi) := \sum_{j=1}^{K} (\lambda_{k,j} \psi_j - \lambda_{j,k} \psi_k), \quad k = 1, \ldots, K
\]
(4.31)
with \(\psi = (\psi_1, \ldots, \psi_K)\). To simplify the notation, we define
\[
\langle \psi \rangle = \sum_{k=1}^{K} \psi_k.
\]
(4.32)
Further, we assume that the collision operator satisfies the detailed-balance condition [88]:
\[
\psi_{eq}^k = \frac{\lambda_{k,j}^eq}{\lambda_{j,k}} \psi_{eq}^j \quad \text{with} \quad k, j = 1, \ldots, K
\]
(4.33)
for any equilibrium \(\psi_{eq}^1, \ldots, \psi_{eq}^K\), i.e. for any density \(\psi_{eq}\) satisfying \(L_k(\psi_{eq}) = 0\) for \(k = 1, \ldots, K\). In particular, this implies that any equilibrium density satisfies
\[
\rho_{eq} := \sum_k \psi_{eq}^k = \sum_k \frac{\lambda_{k,j}}{\lambda_{j,k}} \psi_{eq}^j = m_j^{-1} \psi_{eq}^j \quad \text{with} \quad m_j^{-1} := \sum_k \lambda_{k,j} \lambda_{j,k}.
\]
(4.34)
In the following, we assume that the equilibrium density is not exactly known
and introduces some uncertainty in the model. We model the uncertainty by a
\(K\)-dimensional independent distributed random variable \(m = (m_1, \ldots, m_K)\).

Remark 4.5. If the detailed-balance equation (4.33) has a solution, one can show
that all equilibrium solutions of the system have to satisfy the detailed-balance
equation.

Remark 4.6. The detailed-balance equation (4.33) leads to a symmetrization of
the scattering operator
\[
s_{k,j} := \frac{\lambda_{k,j}}{m_k} = \frac{\lambda_{j,k}}{m_j} = s_{j,k},
\]
(4.35)
which can easily be seen by rewriting the equilibrium density
\[ m_j^{-1} \psi_j^{eq} = \rho^{eq} = m_k^{-1} \psi_k^{eq} = m_k^{-1} \frac{\lambda_{k,j}}{\lambda_{j,k}} \psi_j^{eq}. \] (4.36)

Additionally, one can show that \( \langle \mathbf{m} \rangle = 1 \) by rewriting the equilibrium solution
\[ \psi_j^{eq} = m_j m_j^{-1} \psi_j^{eq} = m_j \sum_k \frac{\lambda_{k,j}}{\lambda_{j,k}} \psi_j^{eq} = m_j \sum_k \psi_k^{eq} = m_j \rho \] (4.37)
and summing both sides over \( j \). This actually contradicts our assumption on \( \mathbf{m} \) being independent distributed. However, for simplicity we stick with the assumption and normalize \( \mathbf{m} \) later on.

### 4.3.1 The Reduced Models

To obtain the reduced models, we first expand the fluctuations around the equilibrium in a Chapman-Enskog expansion
\[ \psi = \psi^{[0]} + \varepsilon \psi^{[1]} + \varepsilon^2 \psi^{[2]} + \ldots, \] (4.38)
with \( \psi^{[0]} = \psi^{eq} \) and \( \langle \psi^{[\ell]} \rangle = 0 \) for all \( \ell > 0 \). Then we insert the expansion into (4.30) and obtain (in vector notation)
\[ \partial_t (\psi^{[0]} + \varepsilon \psi^{[1]} + \varepsilon^2 \psi^{[2]} + \ldots) + \mathbf{\Omega} \cdot \nabla_x (\psi^{[0]} + \varepsilon \psi^{[1]} + \varepsilon^2 \psi^{[2]} + \ldots) = \frac{1}{\varepsilon} L (\psi^{[0]} + \varepsilon \psi^{[1]} + \varepsilon^2 \psi^{[2]} + \ldots). \] (4.39)

Summing over all velocities, this yields
\[ \partial_t \rho^{eq} + \langle \mathbf{\Omega} \cdot \nabla_x (\psi^{[0]} + \varepsilon \psi^{[1]} + \varepsilon^2 \psi^{[2]} + \ldots) \rangle = 0. \] (4.40)

Finally, we truncate the expansion and solve for \( \rho^{eq} \) to obtain a reduced model of order \( \varepsilon \). We call the solution of the reduced model \( \rho \). Then the first reduced model is given by
\[ \partial_t \rho + \langle \mathbf{\Omega} \mathbf{m} \rangle \cdot \nabla_x \rho = 0 \] (4.41)
and the particle density is approximated by \( \psi \approx \mathbf{m} \rho \) with \( \mathbf{m} = (m_1, \ldots, m_K) \).

The second reduced model requires a linear solve. On the one hand, by matching the orders of \( \varepsilon \) in (4.39), we obtain
\[ L (\psi^{[1]}) = \partial_t \psi^{[0]} + \mathbf{\Omega} \cdot \nabla_x \psi^{[0]} = \mathbf{m} \partial_t \rho + \mathbf{m} \cdot \nabla_x \rho \]
\[ = -\mathbf{m} (\mathbf{\Omega} \cdot \mathbf{m} \nabla_x \rho) + \mathbf{\Omega} \cdot \mathbf{m} \nabla_x \rho. \] (4.42)
On the other hand $\psi^{[1]}$ has to satisfy $\langle \psi^{[1]} \rangle = 0$. Combining these two conditions, we obtain
\[
\psi_k^{[1]} = \tilde{L}^{-1}(-m_k\langle \Omega m \rangle + \Omega km_k) \cdot \nabla_x \rho ,
\] (4.43)
where $\tilde{L}^{-1}$ denotes the resulting pseudo-inverse defined on the range of $L$. Plugging this into (4.40), we obtain the second reduced model
\[
\partial_t \rho + \langle \Omega m \rangle \cdot \nabla_x \rho + \nabla_x \cdot \langle \Omega \tilde{L}^{-1}(-m\langle \Omega m \rangle + \Omega m) \rangle \cdot \nabla_x \rho = 0
\] (4.44)
with the reconstruction
\[
\psi \approx m \rho + \varepsilon \tilde{L}^{-1}(-m_k\langle \Omega m \rangle + \Omega km_k) \cdot \nabla_x \rho .
\] (4.45)
Due to the simplified structure in the reduced models and the lower dimensionality of the problems, the reduced models have a substantially lower computational cost than the full model.

Remark 4.7. The diffusion matrix in (4.44) is positive definite. This can be seen by rewriting the matrix as
\[
- \langle \Omega \tilde{L}^{-1}(-m\langle \Omega m \rangle + \Omega m) \rangle = -\langle g m^{-1} \tilde{L}^{-1}(g) \rangle = -\langle g \tilde{L}^{-1}(g) \rangle_{m^{-1}} > 0 ,
\] (4.46)
with $g := -m\langle \Omega m \rangle + \Omega m$, since $g$ satisfies $\langle \langle \Omega m \rangle \tilde{L}^{-1}(g) \rangle = 0$, $g$ is an element of the range of $L$ (as $\langle g \rangle = 0$), and $L$ is self-adjoint and negative definite on its range with respect to the weighted norm $\langle \cdot \rangle_{m^{-1}}$.

Remark 4.8. In one dimension, $x = x \in \mathbb{R}$ and $\Omega_k = \mu_k \in [-1, 1]$, $k = 1, \ldots, K$, the first reduced model is given by
\[
\partial_t \rho + \langle \mu m \rangle \partial_x \rho = 0
\] (4.47)
and the second reduced model is given by
\[
\partial_t \rho + \langle \mu m \rangle \partial_x \rho + \langle \mu \tilde{L}^{-1}(-m \langle \mu m \rangle + \mu m) \rangle \partial_{xx} \rho = 0 .
\] (4.48)
In two dimensions, $x = (x, y) \in \mathbb{R}^2$ and $\Omega_k = (\Omega_x, \Omega_y)_{k}^{T} \in \mathbb{S}^1$, the first reduced model is given by
\[
\partial_t \rho + \langle \Omega_x m \rangle \partial_x \rho + \langle \Omega_y m \rangle \partial_y \rho = 0
\] (4.49)
4.3 The Discrete Velocity Model

and the second reduced model is given by

\[
\partial_t \rho + \langle \Omega_x \mathbf{m} \rangle \partial_x \rho + \langle \Omega_y \mathbf{m} \rangle \partial_y \rho + \left( \langle \Omega_x \tilde{L}^{-1}(-M \langle \Omega_x \mathbf{m} \rangle + \Omega_x \mathbf{m} \rangle \rangle \right) \partial_{xx} \rho \\
+ \left( \langle \Omega_x \tilde{L}^{-1}(-\mathbf{m} \langle \Omega_y \mathbf{m} \rangle + \Omega_y \mathbf{m} \rangle \rangle \right) + \left( \langle \Omega_y \tilde{L}^{-1}(-\mathbf{m} \langle \Omega_y \mathbf{m} \rangle + \Omega_y \mathbf{m} \rangle \rangle \right) \partial_{xy} \rho \\
+ \left( \Omega_y \tilde{L}^{-1}(-\mathbf{m} \langle \Omega_y \mathbf{m} \rangle + \Omega_y \mathbf{m} \rangle \rangle \right) \partial_{yy} \rho = 0. \tag{4.50}
\]

4.3.2 Numerical Results

To compare the two-level sampling strategy with the original SC method, we consider several test cases. The test cases are sufficiently smooth and given on periodic domains, such that the sparse grid sampling converges nicely and the models can easily be solved in Fourier space. Solving the models in Fourier space leads to a linear system of ordinary differential equations, which can be solved with an eigenvalue decomposition. Thus, the fast Fourier transform (FFT) is used to transform the initial condition and an eigenvalue solver is used for the computation of the solution in Fourier space. Then the inverse FFT is used to obtain the solution in real space.

In the test cases, we assume that the velocities \( \Omega_k \) and the symmetrized collision kernel \( s_{k,j} \) are constant, while \( m_k \) is uncertain for \( k = 1, \ldots, K \). We model the uncertainty by a \( K \)-dimensional independent distributed random variable \( \mathbf{m} \), which is uniformly distributed around a prior guess \( \bar{\mathbf{m}} \). Note that the samples of \( \mathbf{m} \) might violate \( \langle \mathbf{m} \rangle = 1 \) (cf. Remark 4.6), so that an additional normalization is needed to guaranty consistency with the reduced models. Moreover, we use well-prepared data for the initial condition

\[
\psi|_{t=0} = \frac{\rho_0 \mathbf{m}}{\langle \mathbf{m} \rangle}, \tag{4.51}
\]

where \( \rho_0 \) is a function of the spatial variable.

In what follows, we consider a one-dimensional and a two-dimensional test case. For each test case, we compare the standard SC sampling with our new two-level sampling strategy by following the steps in section 4.2.2. As we expect that the cost ratio depends on the order of the model and the scale \( \varepsilon \) (cf. Example 4.3), we compare the first and the second reduced model for various relaxation parameters \( \varepsilon = 0.5, 0.1, 0.02 \).
4. TWO-LEVEL SAMPLING STRATEGIES

4.3.2.1 DVM in One Dimension with Ten Velocities

We investigate the performance of the two-level sampling strategy in one spatial dimension on a problem with a ten-dimensional random variable.

Let the computational domain be given by $[0, 2\pi)$, the initial condition be given by

$$\rho_0(x) = 1 + \cos(2x), \quad x \in [0, 2\pi),$$

and the velocity set be given by $K = 10$ discrete velocities, which are equally distributed on $[-1, 1]$.

$$\mu_k = -1 + \frac{2k}{K - 1}, \quad k = 0, \ldots, K - 1. \quad (4.53)$$

The transition rates are uncertain, which we model by a symmetric matrix $S$ and a ten-dimensional random variable $\mathbf{m}$, which is uniformly distributed on $[\bar{m} - \delta, \bar{m} + \delta]$ with $\bar{m}_k = \mu_k^2 + \frac{3}{4}$ and $\delta = \frac{1}{4}$. \quad (4.54)

For the spatial discretization, we use a highly resolved grid with 100 grid points, which corresponds to the number of waves in the Fourier transform. In the following, we want to determine the expected value of the $L^2$ norm of the flux

$$E[f(\psi)] = E\left[ \left( \int \langle \mu \psi \rangle^2 \, dx \right)^{1/2} \right], \quad (4.55)$$

where we use the trapezoidal rule to approximate the spatial integral.

First, we compute a reference solution using the standard SC method with a level 11 (about $2.3 \cdot 10^6$ points) quadrature formula. Then we use the reference solution to compare the two-level sampling with the standard SC approach as described in Approach 1 (empirical) and use the results as a benchmark. The reference solution can also be used to compute the accuracy of the reduced model (4.25), which gives a threshold on the error tolerance for simply using the reduced model instead of the full model. Note that solving a reduced model is about 100 (for $p = 1$) or 50 (for $p = 2$) times faster than solving the full model.
4.3 The Discrete Velocity Model

Second, we use reference solutions to estimate the constants in the asymptotic error bounds for the two different error models and then solve the resulting continuous optimization problems to obtain the optimal number of points \( M_R \) and \( M_D \) as described in Approach 2 (estimated). In addition, we use the error model to compute the optimal number of points \( M_F \) for the standard SC method. As an example, the error terms, which are used to estimate the convergence rates, are shown in Figure 4.3 for \( \varepsilon = 0.1 \).

![Figure 4.3: DVM in 1D with 10 velocities: Error in the full model, in the reduced model of order \( p = 1, 2 \), and in the difference between the two as a function of the number of points of the quadrature formula for \( \varepsilon = 0.1 \).](image)

From the solution of the optimization problem, we can directly compute the theoretical cost ratio given by

\[
\frac{\kappa M_R + M_D}{M_F} \quad \text{with} \quad \kappa = \frac{\text{Cost}_R}{\text{Cost}_F},
\]

where the single costs \( \text{Cost}_R \) and \( \text{Cost}_F \) are measured. However, this estimate is only of theoretical interest, since the number of the quadrature points cannot be chosen arbitrarily. To obtain a more realistic estimate, we map the number of points \( M_R, M_D, \) and \( M_F \) to a quadrature formula of level \( \ell_R, \ell_D, \) and \( \ell_F \), which has at least the required number of points. Then we compute the cost ratio for the actual number of quadrature points again. A comparison of the different cost ratios is given in Figure 4.4 for the first and second reduced model with \( \varepsilon = 0.1 \).
4. TWO-LEVEL SAMPLING STRATEGIES

Considering the first-order reduced model (see Figure 4.4a), we observe that the error models predict some cost savings beyond the accuracy of the reduced model $\alpha_R$ if we only consider the number of points predicted by the continuous optimization problem. However, for a quadrature with at least the required number of points we obtain an overhead and the cost savings cannot always be captured (the cost ratio jumps above one). This suggests that the theoretical cost savings are too small to be seen and indeed the empirical approach has a similar behavior. Note again that the accuracy of the reduced model $\alpha_R$ indicates the range of error tolerances, where the reduced model can be used instead of the full model, and our main interest is in the range $\text{Error}_{\text{tol}} > \alpha_R$.

Considering the second-order reduced model (see Figure 4.4c), the cost ratio is always significantly less than one, which means that the two-level sampling reduces the cost even if the error tolerance exceeds the accuracy of the reduced model $\alpha_R$. Although the error tolerance is not always reached with this approach (see Figure 4.4b and d), the overall behavior of the error is similar to the error of the empirical approach. Altogether, the results suggest that the error models and the solution of the resulting optimization problems give reasonable estimates.

Finally, we use the practical approaches described in Approach 3 (iterative) and Approach 4 (greedy) for the first and the second reduced model with various scales $\varepsilon = 0.5, 0.1, 0.02$. For each test case, we compare the results to the standard SC method computed with an iterative approach. The cost ratio (4.56) is shown in Figure 4.5 and 4.6. We observe that in all cases the cost ratio is in good agreement with the cost ratio computed with the empirical approach and there are only minor differences between the iterative and the different greedy approaches. However, it should be noted that there are a few points where the greedy approach combined with the error models is more restrictive, which leads to smaller cost savings. To confirm the validity of the estimates, we also compute the “true” error of the two-level sampling methods using the reference solution. Here, Figure 4.5 and 4.6 show that the given error tolerance is always reached.

Altogether, the practical approaches yield promising results compared to the empirical approach. However, we observe that the two-level sampling (no matter
how it is implemented) is not always advantageous. In particular, if the first-order reduced model is used, the cost reduction is only clearly evident for a small relaxation parameter $\varepsilon = 0.02$, whereas in the other cases the cost savings are too small to be resolved by the quadrature levels. On the other hand, for the second-order reduced model great cost savings can be achieved. The only restriction is that the relaxation parameter needs to be sufficiently small ($\varepsilon = 0.1, 0.02$), which was to be expected as the reduced models become meaningless for large $\varepsilon$.

![Figure 4.4: DVM in 1D with ten velocities: Reduced models of order $p = 1, 2$ with relaxation parameter $\varepsilon = 0.1$. Cost ratio and error as a function of the error tolerance $\text{Error}_{\text{tol}}$. The maximal accuracy of the reduced model $\alpha_R$, the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical and the estimated approach.](image-url)
Figure 4.5: DVM in 1D with ten velocities: First-order reduced model ($p = 1$) with various relaxation parameters $\varepsilon = 0.5, 0.1, 0.02$. Cost ratio and error as a function of the error tolerance $\text{Error}_{\text{tol}}$. The maximal accuracy of the reduced model $\alpha_R$, the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical, the iterative, and the greedy approach.
4.3 The Discrete Velocity Model

Figure 4.6: DVM in 1D with ten velocities: Second-order reduced model \((p = 2)\) with various relaxation parameters \(\varepsilon = 0.5, 0.1, 0.02\). Cost ratio and error as a function of the error tolerance \(\text{Error}_{\text{tol}}\). The maximal accuracy of the reduced model \(\alpha_R\), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical, the iterative, and the greedy approach.
4. TWO-LEVEL SAMPLING STRATEGIES

4.3.2.2 DVM in Two Dimensions with Eight Velocities

Similar to the last test case, we investigate the performance of the two-level sampling strategy on the discrete velocity model following the steps described in section 4.2.2, but we increase the spatial dimension. Here, we consider a two-dimensional problem in space with an eight-dimensional random variable.

Let the computational domain be given by a square $[0, 2\pi) \times [0, 2\pi)$, the initial condition be given by

$$\rho_0(\mathbf{x}) = 2 + \cos(2x) + \sin(3y), \quad \mathbf{x} = (x, y) \in [0, 2\pi) \times [0, 2\pi),$$

(4.57)

and the velocity set be given by eight equally distributed discrete velocities on the unit circle

$$\Omega_k = \left[ \begin{array}{c} \Omega_{x,k} \\ \Omega_{y,k} \end{array} \right] = \left[ \begin{array}{c} \text{real}(e^{2\pi i k/8}) \\ \text{imag}(e^{2\pi i k/8}) \end{array} \right], \quad k = 1, \ldots, K = 8.$$  

(4.58)

Further, let the transition rates be given by a symmetric matrix $S$ and an eight dimensional random variable $\mathbf{m}$, which is uniformly distributed on $[\bar{m} - \delta, \bar{m} + \delta]$ with

$$\bar{m}_k = \frac{\Omega_{x,k} + \Omega_{y,k}}{2} + \frac{3}{4} \quad \text{and} \quad \delta = \frac{1}{4}.$$  

(4.59)

For the spatial discretization, we use $50 \times 50$ grid points or Fourier waves, respectively. As before, we are interested in the expected value of the $L^2$ norm of the flux, but we average over the directions

$$E \left[ \frac{1}{2} \left( \left( \int (\Omega_x \psi)^2 \, d\mathbf{x} \right)^{1/2} + \left( \int (\Omega_y \psi)^2 \, d\mathbf{x} \right)^{1/2} \right) \right].$$

(4.60)

We briefly go through the same steps as in the one-dimensional test case.

First, we compute a reference solution with the standard SC method and a level 11 quadrature formula, which has about $4.1 \cdot 10^5$ points. Then we use Approach 1 (empirical) to obtain a benchmark and the accuracy of the reduced models $\alpha_R$. Note that solving a reduced models is about 270 (for $p = 1$) or 150 (for $p = 2$) times faster than solving the full model.
Second, we estimate the constants in the asymptotic error bounds and solve the continuous optimization problems with Approach 2 (estimated). As an example, the results for the first and second reduced model with $\varepsilon = 0.1$ are shown in Figure 4.7 and 4.8. We observe again that the approaches lead to a similar behavior in the error and the cost ratio, which underlines the validity of the optimization problems. Moreover, we observe that for the first-order reduced model, the cost savings are not noticeable due to the overhead in the quadrature, while for the second-order reduced model, the cost savings are significant. Compared to the one-dimensional test case, the effect is a little less pronounced and we can clearly observe that the cost ratio increases slightly with decreasing error tolerance.

![Figure 4.7: DVM in 2D with eight velocities: Error in the full model, in the reduced model of order $p = 1, 2$, and in the difference between the two as a function of the number of points of the quadrature formula for $\varepsilon = 0.1$.](image)

Finally, we compare the practical estimates from Approach 3 (iterative) and Approach 4 (greedy) with the standard SC method. As before, we observe that the cost ratio is similar to the one computed with the empirical approach (see Figure 4.9). In addition, Figure 4.10 shows that the error tolerance is almost always reached. However, there are a few exceptions, where the “true” error violates the given error tolerance. This indicates that the successive error estimate is not always a good approximation of the “true” error. This can also be seen...
in Figure 4.7 for $p = 2$ and $\varepsilon = 0.1$, where the error in the difference is not decreasing from the first to the second level. Note however that the greedy algorithm combined with the error models seems to be again more restrictive and the greedy with the exponential error model does not fail the error tolerance. Another difference compared to the one-dimensional case is that the cost ratio, computed with any of the practical approaches, has a kink at $\text{Error}_{\text{tol}} \approx 3 \cdot 10^{-7}$ for $\varepsilon = 0.1$ and $p = 2$, which means that we do not obtain cost savings on the hole possible range. As mentioned above, the cost savings are not as pronounced, so that slight variations in the number of levels might trigger a loss of cost savings. Nevertheless, in almost all other cases we obtain the possible cost savings using the practical approaches.
4.3 The Discrete Velocity Model

Figure 4.8: DVM in 2D with eight velocities: Reduced models of order $p = 1, 2$ with relaxation parameter $\varepsilon = 0.1$. Cost ratio and error as a function of the error tolerance $\text{Error}_{\text{tol}}$. The maximal accuracy of the reduced model $\alpha_R$, the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical and the estimated approach.
4. TWO-LEVEL SAMPLING STRATEGIES

Figure 4.9: DVM in 2D with eight velocities: First-order reduced model (p = 1) with various relaxation parameters \( \varepsilon = 0.5, 0.1, 0.02 \). Cost ratio and error as a function of the error tolerance \( \text{Error}_{\text{tol}} \). The maximal accuracy of the reduced model \( \alpha_R \), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical, the iterative, and the greedy approach.
4.3 The Discrete Velocity Model

Figure 4.10: DVM in 2D with eight velocities: Second-order reduced model \( p = 2 \) with various relaxation parameters \( \varepsilon = 0.5, 0.1, 0.02 \). Cost ratio and error as a function of the error tolerance \( \text{Error}_{\text{tol}} \). The maximal accuracy of the reduced model \( \alpha_R \), the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked. Results are computed with the empirical, the iterative, and the greedy approach.
4. TWO-LEVEL SAMPLING STRATEGIES

4.4 The Spherical Harmonic P\(_N\) Method

We consider the linear transport equation in a diffusive scaling with scaling parameter \(\varepsilon\) (cf. section 1.1.1 and section 3.1.1)

\[
\partial_t \psi + \frac{1}{\varepsilon} \Omega \cdot \nabla_x \psi + \sigma_a(x)\psi = \frac{\sigma_s(x)}{\varepsilon^2} \left( \int_{S^2} g(x, \Omega \cdot \Omega') \psi(\Omega')d\Omega' - \psi \right) + q
\]  

(4.61)

and its spherical harmonic P\(_N\) approximation (cf. section 1.3.1)

\[
\partial_t u_{P,N} + \frac{1}{\varepsilon} A \cdot \nabla_x u_{P,N} + \sigma_a u_{P,N} + \frac{1}{\varepsilon^2}\sigma_s Gu_{P,N} = q.
\]  

(4.62)

In this case, the P\(_N\) system can be interpreted as an asymptotic expansion of the kinetic equation, since a formal asymptotic analysis shows that the solution components are of order \(O(\varepsilon^\ell)\), where \(\ell\) is the degree of the corresponding spherical harmonic basis function [45]. This can be seen by first expanding the solution components in \(\varepsilon\) and then inserting the expansions into the P\(_N\) system and matching powers of \(\varepsilon\). In particular, any P\(_N_1\) approximation is a reduced model of any higher order approximation P\(_N_2\) with \(N_1 < N_2\).

In the following, we investigate the two-level sampling approach using two different P\(_N\) approximations as the full and the reduced model. This enables to determine the uncertainties in a highly resolved P\(_N\) approximation at a reduced computational cost. Uncertainties can be introduced into the model via the initial condition and the material parameters. Here, we only consider uncertainties in the scattering cross section.

Remark 4.9. We could also consider the discretization error of the P\(_N\) system to describe the error in the difference between the models (cf. Theorem 2.8 with \(G_f = 0\)).

4.4.1 Numerical Results

To determine the actual impact of the two-level sampling approach for the P\(_N\) method, we consider two different test cases. As mentioned before, we use the TASMANIAN library [110] to compute sparse Gauss-Patterson rules. To solve the P\(_N\) system, we use again the code StaRMAP [103] (see also section 1.3 and section 2.3). There is no need to modify the original code, since the scaling
4.4 The Spherical Harmonic P_\_N Method

parameter can be introduced by changing the matrices A and the scattering coefficient \( \sigma_s \). Note that the code preserves the asymptotic limit and can be used for any \( \varepsilon > 0 \). Moreover, the newer version \(^1\) is able to compute a maximal time step (instead of restricting to a hyperbolic CFL) and is asymptotic preserving. Thus, it might be advantageous to use the newer version if we are close to the asymptotic regime, but in the transition regime, which we are mainly interested in, we do not expect major differences.

The first test case investigates the cost savings of the two-level method in a transition regime \( \varepsilon = 0.1 \). On the other hand, the second test case investigates the benefits of the method if the model does not include an explicit scaling parameter \( \varepsilon \). We consider again the continuous slowing down approximation for electron transport with applications in radiation therapy from Chapter 1.

4.4.1.1 Hemisphere Test

We define a two-dimensional test case with a modified uncertain Henyey-Greenstein scattering kernel \(^8\) and an anisotropic source inside the domain. Let the computational domain be given by a square \([-3.5, 3.5] \times [-3.5, 3.5]\), the initial condition be zero, and the source be given by a Gaussian in space and a characteristic function in angle (similar to section 2.3.2.3)

\[
g(t, x, y, \Omega) = \frac{1}{4 \pi (3.2 \cdot 10^{-3})} \exp \left( -\frac{(x + 0.3)^2 + y^2}{4(3.2 \cdot 10^{-3})} \right) \chi_{\mathbb{R}+}(\Omega_x). \tag{4.63}
\]

Further, let the absorption coefficient \( \sigma_a(x) = 1 \), the scattering coefficient \( \sigma_s(x) = 1 \), and scattering kernel \( g(x, \mu) = g(\mu) \) be independent of the spatial variable and let the moments of the scattering kernel be given by the first moments of the Henyey-Greenstein kernel with some uncertainty

\[
g_\ell = \begin{cases} a \ell, & \ell \leq 3, \\ 0, & \text{otherwise}, \end{cases} \tag{4.64}
\]

where \( a = 0.9 \) is the anisotropy factor.

\(^1\)The newer version StaRMAP ver1p9 is able to compute a maximal time step. We use StaRMAP ver1p5 for the computations, which uses a hyperbolic CFL condition.
4. TWO-LEVEL SAMPLING STRATEGIES

We assume that the scattering kernel introduces some uncertainty in the system, which we model by a 4-dimensional random variable as the above scattering kernel is a polynomial of degree 3. The polynomial can uniquely be determined by the function values at four Gauss Legendre grid points. We assume that these function values are uncertain and are uniformly distributed around $g(\mu)$ with ±10% difference. As an example, 20 random samples of the scattering kernel $g(\mu)$ and the corresponding moments of the scattering kernel $g_\ell$ are shown in Figure 4.11. Note that it is not straightforward to see that the kernel $g(\mu)$ remains positive on the hole interval $[-1, 1]$, but we have not observed any negative values.

![Figure 4.11](image)

**Figure 4.11:** Hemisphere Test: (a) samples of the scattering kernel; (b) moments of the scattering kernel. The mean is shown as a black solid line and 20 random samples are shown as colored dotted lines. Black dashed lines indicate the position of the interpolation points and the order of the moments.

In the following, we want to determine the expected value of the zeroth order moment of the solution for each point in space. We compute the solution until $t = 0.5$ on a grid with $100 \times 100$ points with $\varepsilon = 0.1$ using $P_{11}$ as the full model and $P_5$ as the reduced model.

As before, we use Approach 1 (empirical) to compute a benchmark solution using level $\ell = 11$ as a reference. The expected value and the $L^2$ error in the quadrature of level $\ell = 1, \ldots, 10$ are shown in Figure 4.12. We observe that the error in the full model, the reduced model, and their difference has a similar
convergence rate. Moreover, the error in the difference is always substantially smaller than the other errors. As in addition the reduced model is about four times faster to solve than the full model, we expect to obtain cost savings using the two-level approach. Note that the quantity of interest is in the range of $[0, 0.017]$, so that we consider accordingly smaller error tolerances $\text{Error}_{\text{tol}} \in [10^{-11}, 10^{-3}]$.

![Figure 4.12](image_url)

(a) expected value  
(b) $L^2$ convergence

**Figure 4.12:** Hemisphere Test: (a) expected value of the zeroth order moment of the $P_{11}$ solution computed with a high level ($\ell = 11$) SC method; (b) $L^2$ error in the full model ($P_{11}$), the reduced model ($P_5$), and the difference between the two as a function of the number of points of the quadrature.

In the following, we compare the practical approaches described in Approach 3 (iterative) and Approach 4 (greedy) with the benchmark solution. For the practical approaches, we use the $L^2$ norm in space to measure the error in the quadrature formulas. We observe that the two-level approach yields cost savings over the hole range of considered error tolerances independent of the considered approach (see Figure 4.13a). Moreover, we observe that the error tolerance is always reached (see Figure 4.13b), so that the two-level sampling approach yields practically good results in this test case.

However, if we use the $P_3$ system as a reduced model, the cost savings are less pronounced (see Figure 4.14). Thus, further work is needed to choose the optimal moment order approximation for the reduced model.
4. TWO-LEVEL SAMPLING STRATEGIES

Figure 4.13: Hemisphere Test with $\varepsilon = 0.1$: (a) cost ratio between the two-level sampling approach (using $P_5$ and $P_{11}$) and the standard SC approach (using $P_{11}$); (b) $L^2$ error in the two-level sampling approach compared to the high level standard SC method. The maximal accuracy of the reduced model $\alpha_R$, the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked.

Figure 4.14: Hemisphere Test with $\varepsilon = 0.1$: (a) cost ratio between the two-level sampling approach (using $P_3$ and $P_{11}$) and the standard SC approach (using $P_{11}$); (b) $L^2$ error in the two-level sampling approach compared to a high level standard SC method. The maximal accuracy of the reduced model $\alpha_R$, the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked.
4.4.1.2 Water Phantom

In this test case we consider again the electron transport problem from Chapter 1, but we now investigate the propagation of uncertainties for this model. Compared to the other test cases this model does not include an explicit scaling parameter ε, but we expect that in dense regions a diffusion scaling is implied in the equation, similar to (4.61). To examine the benefits of the two-level sampling strategy, we consider once more the water phantom test case introduced in section 1.4.1.1. We use the continuous slowing down approximation (1.18), but with an additional uncertain input parameter σ(ξ, E) depending on a random variable ξ:

\[ - \partial_e (S(E) \rho(x) \psi(\xi, E, x, \Omega)) + \Omega \cdot \nabla_x \psi(\xi, E, x, \Omega) + \rho(x) \Sigma_t(E) \psi(\xi, E, x, \Omega) = \rho(x) \sigma(\xi, E) \int_{S^2} \Sigma_s(E, \Omega' \cdot \Omega') \psi(\xi, E, x, \Omega') d\Omega' + q(E, x, \Omega). \quad (4.65) \]

In our example, we model the uncertainty by a four dimensional random variable \( \xi = (\xi_1, \ldots, \xi_4) \) and

\[ \sigma(\xi, E) = 1 + \frac{1}{10} \sum_{k=1}^{4} \xi_k \cos \left( 2\pi (k-1) \frac{t}{T} \right) \]

with \( t = \phi(E) \) and \( T = \phi(0) \), where \( \phi \) is the energy transformation described in section 1.2.2 (time-energy marching). In the following we assume \( \xi_k \) is uniformly distributed on \([-1, 1]\) for \( k = 1, \ldots, 4 \). As an example, Figure 4.15a shows 10 random samples of \( \sigma(\xi, E) \) as a function of the energy \( E \).

To solve the equation we use the \( P_N \) method introduced in section 1.3. For the deterministic setting, we have seen that the method yields an accurate solution for the water phantom test case with \( E_{\text{max}} = 5 \text{MeV} \) for \( N = 21 \) and a spatial grid with \( \Delta x = 0.05 \text{ cm} \) or \( n = 150 \) grid points (see section 1.4.1.1). Thus, we use \( P_{21} \) as the “full” model, but we do a grid convergence study in \( n \) by doubling the number of points to \( n = 300 \) as the spatial resolution may affect the convergence in the random parameter (see Figure 4.16). For the reduced model, we use a lower order \( P_N \) approximation with \( N = 11 \). Then the reduced model is about three times faster to solve than the full model.

As a quantity of interest we chose the mean and the variance of the depth dose profile. Furthermore, to measure the error between the two models, we compute...
the $L^2$ norm in space of the mean and the variance on the interval $I = [1, 2.5]$ as indicated by the dashed lines in Figure 4.15b. In particular, we truncate the interval such that the build-up region (or initial layer) of the beam is excluded from the interval. Thus, we measure the error in terms of

$$\sqrt{\|E[\psi]\|_2^2 + \|\text{Var}[\psi]\|_2^2}.$$  

(4.67)

![Figure 4.15: Water phantom: (a) 10 random samples of the uncertain input parameter $\sigma(\xi, E)$; (b) mean and variance of the depth dose profile computed with SC ($P_{21}$ and level 11 quadrature formula). The $\gamma$-index is computed for $E \pm \text{Var}$ compared to $E$ on a 3%, 3 mm level.](image)

First, we compute a reference solution using the standard SC method with a high level quadrature formula $\ell = 11$ (see Figure 4.15b). To evaluate the results, we also compute the $\gamma$-index\(^1\) on a 3%, 3 mm level for the positive and negative deviations of the mean, i.e., we compute the $\gamma$-index for $E \pm \text{Var}$ compared to the mean. Although the considered uncertainties in the scattering cross section are quite substantial, we observe that the deviations of the mean seem to be tolerable ($\gamma(x) < 1$).

Secondly, we consider the convergence in the full model, the reduced model, and the difference between the two (see Figure 4.16). Here, we observe that

\(^1\)The $\gamma$-index is computed by first interpolating the data on a 100 times finer grid and then using (1.63), where we minimize over all grid points in a surrounding of $\bar{x}$ with radius $2R$. 

134
the full and the reduced model have a similar convergence behavior, while the difference has a smaller error. Note that this behavior is essential for our two-level approach. We also perform a grid convergence study (as mentioned above) and observe that the convergence in the sparse grid quadrature formula does not change substantially if we increase the number of spatial grid points.

![Graph](image.png)

**(a) n = 150**

**(b) n = 300**

**Figure 4.16:** Water phantom: $L^2$ error (4.67) in the full model ($P_{21}$), the reduced model ($P_{11}$), and the difference between the two as a function of the number of points of the quadrature for different spatial resolutions: (a) $n = 150$ grid points; (b) $n = 300$ grid points.

Similar to the previous test case, we then compute a benchmark solution for the two-level sampling method as described in Approach 1 (empirical) and compare the results to the practical estimates described in Approach 3 (iterative) and Approach 4 (greedy). The cost ratio (4.56) and the $L^2$ error, measured in terms of (4.67), are shown in Figure 4.17.

The results show that the given error tolerance is always reached with our method. In addition, the results suggest that the computational cost can be reduced. Although the cost savings are not very pronounced, which also comes from the fact that the reduced model is only about three-times faster to solve, we almost always capture the cost savings with the practical approaches. However, as we have also seen in the previous example, the cost savings heavily depend on the order of the reduced model and an optimal trade-off between the accuracy and the cost of the reduced model still has to be determined. For further reductions of
the computational cost, it would also be advantageous to introduce a multilevel sampling strategy. Especially in this application, where multiple reduced models are easily available, a multilevel sampling might be very efficient.

Figure 4.17: Water phantom: (a) cost ratio between the two-level sampling approach (using $P_{11}$ and $P_{21}$) and the standard SC approach (using $P_{21}$); (b) $L^2$ error (4.67) in the two-level sampling approach compared to a high level standard SC method. The maximal accuracy of the reduced model $\alpha_R$, the threshold “cost ratio = 1”, and the identity “true error = error tolerance” are marked.
In this work we have studied various numerical aspects of the linear transport equation with the focus on radiation therapy. The linear transport equation is difficult and computationally intensive to solve due to the high dimensionality of the problem, the integro-differential form, and the highly varying material parameters (in particular in the context of radiation therapy). Facing these challenges, we have achieved the following four points in the previous chapters:

- We have studied a model for electron transport and introduced a numerical method based on the spherical harmonic ($P_N$) approximation.

- We have investigated the filtered spherical harmonic ($FP_N$) equations, which reduce the Gibbs phenomena in a spectral method.

- We have studied an asymptotic preserving (AP) scheme as AP schemes are well suited for highly varying material parameters.

- We have introduced a sampling strategy to investigate the propagation of uncertainties as material parameters always introduce uncertainties in the model.

In what follows, we discuss the results and possible future tasks for each of the previous chapters in detail.
5. DISCUSSION AND CONCLUSIONS

5.1 Electron Transport

We have introduced a numerical method for electron transport in tissue based on the continuous slowing down (CSD) approximation. The CSD approximation assumes small energy deviations and a small mean free path and is a common simplification for electrons with energies above 1 keV. Our method relies on a radiation transport code, which uses the spherical harmonic (P\textsubscript{N}) approximation. In addition, we introduced physical parameters into the method, which have been extracted from the ICRU database [101].

We have shown how the CSD approximation can be transformed into a radiation transport equation with a space-dependent flux term similar to [11]. This enables the adaptation of an existing radiation transport code to our application. We have adapted the second order radiation transport code StaRMAP [103] to three spatial dimensions and changed to a space-dependent flux. Furthermore, we have implemented physical parameters, which have previously been extracted from Penelope and the ICRU database in [92]. Since the depth dose deposition is of main interest in applications, our method is also able to compute a weighted time/energy integral. Finally, we presented numerical results to quantify the modeling error compared to a well-established Monte Carlo code (Penelope [100]), which was evaluated against experiments.

Considering that we used a simplified physical model, the results are in good agreement with the MC results. For the 5 MeV phantom test cases, the results even pass the $\gamma$-criterion, but also for the higher energy (10 MeV) phantom test case the $\gamma$-criterion is only violated at the front of the beam, where the P\textsubscript{N} solution falls off faster than the MC solution. For the complex test cases on the two- and three-dimensional CT scan with 21 MeV, the results also pass the $\gamma$-criterion to a great extent.

Although the violations of the $\gamma$-criterion are still not tolerable in clinical practice, we have seen that the results are surprisingly good for the used physical model. In addition, compared to Monte Carlo methods, our method is relatively cheap and allows to use the same structure for the optimization of the dose. The aim of future work should be to improve the physical model by taking for instance
different material coefficient into account or Bremsstrahlung and secondary electrons. In addition, the implementation of low densities, which for now slows down the computations substantially, should be improved to reduce the computational cost. Then, with the principal aim of developing a treatment planning system, one could use the method for the optimization of the dose profile by using for instance the adjoint method.

5.2 Convergence of FP$_N$ Equations

We have proven global $L^2$ convergence properties for filtered spherical harmonic (FP$_N$) equations. These equations govern the evolution of the coefficients in a spectral approximation, with respect to the angular variable, of a radiative transport equation. The estimates derived here are based on the reformulation of the filter in [98] as an additional anisotropic scattering term in the transport equation which depends on the order of the spectral approximation.

We have shown, how the convergence rates depend on both the regularity of the underlying transport solution and the order of the filter. In particular, we observed that for problems with smooth solutions, the order of the filter determines the rate of convergence, while for non-smooth problems, it is the regularity of the transport equation. In addition, we have shown that sharper estimates are possible if the angular $L^2$ projection of the transport solution onto rotationally invariant subspaces satisfies additional mild conditions. Finally, we presented numerical results for several problems to demonstrate possible benefits of filtering and to demonstrate various aspects of the theoretical predictions.

The numerical experiments confirm that filtering dampens non-physical oscillations and can lead to a better overall accuracy of the solution. The results also confirm that the discrete filter steps may reduce the convergence order of the code, whereas the implementation of the continuous FP$_N$ equations is order preserving.

While most of the numerical convergence results agree with the theoretical predictions, we do observe one discrepancy for the lattice test case. There the convergence of the filtered solution is actually slightly better than what is predicted by the estimates on the closure error. Thus, either our estimates are not
5. DISCUSSION AND CONCLUSIONS

quite optimal or the numerical simulations used to approximate the decay in the moments of the true transport solution (upon which the estimates depend) are not resolved enough. Further investigations of this issue are ongoing.

At the end of the chapter, we have investigated the impact of filtering on radiation therapy. We have considered again two problems from the previous chapter to see if with filtering the accuracy of a lower order approximation can be increased. However, the filtering does not seem to be particular suitable for this applications.

The interplay between the filter order and the regularity of the transport solution brings to mind two important practical issues. First, the regularity should be known apriori, either from a mathematical analysis of the transport equation or from previous experience with similar problems. Indeed, for real applications, a numerical convergence study to extract Sobolev indices or moment decay rates is not practical. Second, while the convergence results suggest the use of high-order filters, it has been found in several examples [98] that, for moderate values of $N$, low-order filters yield better results around sharp transitions. This is clearly due to the fact that low-order filters suppress oscillations more strongly than higher-order ones. The trade-off, however, is that the low-order filter reduces the accuracy of the solution in smooth regions that are fully-resolved. Determining the “optimal” order and strength of a filter remains open, and it is likely that these parameters will have to be varied locally in space.

Finally, the analysis here has been done in the context of a global norm. To better understand the observed benefits of the filter, a local analysis seems useful and should be considered in future work.

5.3 AP Staggered Grid Method

We have introduced a two-dimensional AP scheme for the linear transport equation. The linear transport equation has the diffusion equation as an analytic asymptotic limit. For AP schemes the discretization has to be chosen such that the analytic limit is preserved at a discrete level and the scheme is uniformly stable with respect to the mean free path. Here, we used a parity-based time discretization combined with a staggered-grid spatial discretization.
We have shown that the spatial discretization has the desired AP property. In particular, due to the use of staggered grids, a compact five-point stencil can be achieved in the discrete diffusion limit. Furthermore, the parity-based time discretization is suitable for the use of staggered grids, as the coupling between the even and odd parities reduces the number of the required unknowns. In addition, we have presented a rigorous stability analysis for the same scheme in one dimension. This provides a condition on the relaxation parameter and a CFL condition. Finally, we have performed several numerical tests for the two-dimensional scheme, which demonstrate the AP property. Since staggered grids can easily be extended to three dimensions, there is a straightforward generalization of our method to three spatial dimensions. Although we did not test the method, we expect that it has similar properties.

In the future, it would be worthwhile to investigate the time discretization. Since our method uses a simple time-integration method (explicit Euler method), the convergence order is in general limited to one. To maintain a second order scheme, one could use some higher order IMEX time-integration method. Another possible topic of future work is to apply staggered grids in combination with a parity-based time discretization to other kinetic equations such as the continuous slowing down approximation.

5.4 Two-Level Sampling Strategies

We have derived a two-level sampling strategy for hyperbolic relaxation systems to reduce the computational cost of a sparse grid stochastic collocation method. Our method uses a model hierarchy given by the reduced models for the hyperbolic relaxation system. A reduced model for the hyperbolic relaxation system and a correction are combined in a two-level framework to balance the errors coming from the reduced model and the sparse grid integration. The method can be used to determine the effect of a random input parameter with known distribution on a particular quantity of interest.

We have analyzed the cost and the error of the sampling strategy and derived a constrained optimization problem to determine the optimal number of quadrature points in the two levels. The optimization problem relies on an error
model for the sparse grid integration formulas, which was taken from two different asymptotic error bounds [37, 118] and [10, 28]. Furthermore, we have introduced numerical strategies to confirm the theoretical predictions and to come up with practical approaches (as the constants in the error models are usually unknown). Finally, we presented numerical comparisons between the standard SC method and our two-level sampling strategy for various linear problems to demonstrate the advantages of the new method.

The results confirm the validity of the error models and the optimization problem, as well as the dependence of the cost savings on the scaling parameter and the order of the reduced models. In particular, we have seen that the second-order reduced model is of advantage in comparison to the first-order reduced model. Moreover, we have seen that the results computed with the practical approaches are in good agreement with the benchmark solutions, and the new sampling strategy can give good cost savings for problems in the transition regime.

Since the method relies heavily on the convergence behavior of the sparse grids for the two different models, the dependence of the convergence on the regularity of the initial condition and the impact of different quantities of interest should be investigated in future work. It would also be worthwhile to test the method on non-linear problems such as the Boltzmann equation and its Euler or Navier-Stokes limit. Another future topic is the extension of the method to a multilevel sampling strategy, which includes multiple reduced models of different orders, for instance different order $P_N$ approximations.
Appendix A

Real-Valued $P_N$ Equations

We consider the properties of the matrices $A_i$, which occur in the real-valued $P_N$ equations (1.43)

$$\partial_t u_{P_N} + A \cdot \nabla x u_{P_N} + \sigma_a u_{FP_N} - \sigma_s Gu_{FP_N} = s. \quad (A.1)$$

First, we turn our attention to the complex-valued spherical harmonic basis functions, which are analyzed in [16]. They fulfill the following recursion relation

$$\Omega Y^k = \frac{1}{2} \left[ -c_{\ell-1}^{k-1} Y_{\ell-1}^{k-1} + d_{\ell+1}^{k-1} Y_{\ell+1}^{k-1} + e_{\ell-1}^{k+1} Y_{\ell-1}^{k+1} - f_{\ell+1}^{k+1} Y_{\ell+1}^{k+1} \right] + i \left[ c_{\ell-1}^{k+1} Y_{\ell-1}^{k+1} - d_{\ell+1}^{k+1} Y_{\ell+1}^{k+1} + e_{\ell-1}^{k-1} Y_{\ell-1}^{k-1} - f_{\ell+1}^{k-1} Y_{\ell+1}^{k-1} \right], \quad (A.2)$$

where the coefficients are

$$a_{\ell}^k = \sqrt{\frac{\ell-k}{(2\ell+1)(2\ell+3)}} , \quad b_{\ell}^k = \sqrt{\frac{(\ell-k)(\ell+k)}{(2\ell+1)(2\ell+3)}} , \quad c_{\ell}^k = \sqrt{\frac{(\ell+1+k)(\ell+1+k)}{(2\ell+1)(2\ell+3)}} , \quad f_{\ell}^k = \sqrt{\frac{\ell+1+k}{(2\ell+1)(2\ell-1)}}. \quad (A.3)$$

The real-valued spherical harmonics are defined in terms of the complex-valued spherical harmonics (1.37) and the mapping between the two is bijective. Therefore, the above recursion relation can be translated into a real version. Since for any $\ell = 0, 1, 2 \ldots$ and $-\ell \leq k \leq \ell$ the coefficients are symmetric in $k$, i.e.,

$$a_{\ell}^k = a_{\ell}^{-k} , \quad b_{\ell}^k = b_{\ell}^{-k} , \quad c_{\ell}^k = c_{\ell}^{-k} , \quad d_{\ell}^k = f_{\ell}^{-k}, \quad (A.4)$$
it can be written in a compact form as
\[
\Omega m^k_\ell = \frac{1}{2} \left[ \Theta(k) \left(1 - \delta_{k-1}\right)\left(\hat{c}_{\ell-1}^{[k-1]}m^{k-1}_{\ell-1} - \hat{d}_{\ell+1}^{[k-1]}m^{k-1}_{\ell+1}\right) - \hat{c}_{\ell}^{[k+1]}m^{k+1}_{\ell-1} + \hat{f}_{\ell+1}^{[k+1]}m^{k+1}_{\ell+1}\right],
\]
\[
2(a^{k}_{\ell-1}m^{k}_{\ell-1} + b^{k}_{\ell+1}m^{k}_{\ell+1})
\]
(A.5)

where \(\delta_{i,j}\) denotes the Kronecker delta, and \(\Theta(k)\) is the sign function (with abuse of notation in zero)
\[
\delta_{i,j} = \begin{cases} 1, & k = j, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad \Theta(k) = \begin{cases} -1, & k < 0, \\ 1, & k \geq 0. \end{cases}
\]
(A.6)

The coefficients are given by
\[
k^+ = k + \text{sgn}(k), \quad k^- = k - \text{sgn}(k),
\]
\[
\hat{c}_\ell^k = \begin{cases} 0, & k < 0, \\ \sqrt{2}c_\ell^k, & k = 0, \\ c_\ell^k, & k > 0, \end{cases} \quad \hat{d}_\ell^k = \begin{cases} 0, & k < 0, \\ \sqrt{2}d_\ell^k, & k = 0, \\ d_\ell^k, & k > 0, \end{cases}
\]
(A.7)
\[
\hat{c}_\ell^k = \begin{cases} \sqrt{2}c_\ell^k, & k = 1, \\ c_\ell^k, & k > 1, \end{cases} \quad \hat{f}_\ell^k = \begin{cases} \sqrt{2}f_\ell^k, & k = 1, \\ f_\ell^k, & k > 1. \end{cases}
\]

The recursion relation is used in (1.40) to obtain the explicit formulation of the \(P_N\) equations
\[
\partial_t \mathbf{u}_{P_N} + A \cdot \nabla_x \mathbf{u}_{P_N} + \sigma_a \mathbf{u}_{FP_N} - \sigma_s \mathbf{G} \mathbf{u}_{FP_N} = s,
\]
(A.8)

with \(A_i = \langle \mathbf{m} \mathbf{m}^T \Omega_i \rangle\) and \(i \in \{x,y,z\}\). Moreover, it shows the existence of the matrices \(\mathbf{a}^{(i)}_\ell\) of size \((2\ell - 1) \times (2\ell + 1)\), which satisfy (1.45)
The first matrices $a^{(i)}_\ell$ are given by

\[
\begin{align*}
    a^{(x)}_1 &= \begin{bmatrix} 0 & 0 & \frac{1}{\sqrt{2}} f_1^1 \\ \frac{1}{\sqrt{2}} f_1^1 & 0 & 0 \end{bmatrix}, \\
    a^{(y)}_1 &= \begin{bmatrix} \frac{1}{\sqrt{2}} f_1^1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\
    a^{(z)}_1 &= \begin{bmatrix} 0 & b_1^0 & 0 \\ 0 & 0 & b_1^0 \end{bmatrix}, \\
    a^{(x)}_2 &= \frac{1}{2} \begin{bmatrix} f_2^1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{2} f_2^1 & 0 \\ 0 & 0 & -\sqrt{2} d_2^0 & 0 & f_2^2 \\ f_2^1 & 0 & 0 & 0 & 0 \\ 0 & \sqrt{2} f_2^1 & 0 & 0 & 0 \end{bmatrix}, \\
    a^{(y)}_2 &= \frac{1}{2} \begin{bmatrix} 0 & 0 & 0 & -\sqrt{2} d_2^0 & 0 & -f_2^2 \\ 0 & \sqrt{2} f_2^1 & 0 & 0 & 0 & 0 \\ f_2^1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & f_2^1 & 0 & 0 \\ 0 & 0 & 0 & f_2^1 & 0 & 0 \end{bmatrix}, \\
    a^{(z)}_2 &= \begin{bmatrix} 0 & b_2^1 & 0 & 0 & 0 \\ 0 & 0 & b_2^1 & 0 & 0 \\ 0 & 0 & 0 & b_2^1 & 0 \end{bmatrix}, \ldots
\end{align*}
\]

(A.10)

Since the coefficients $a_k^k, \ldots, f_k^k$ are bounded by 1, the entries of the matrices $a^{(i)}_\ell$ are in the interval $[-1, 1]$. Together with the recursion relation (A.5), this yields upper bounds for the $\infty$-norm and the 1-norm

\[
\|a^{(i)}_\ell\|_\infty \leq 1 \quad \text{and} \quad \|a^{(i)}_\ell\|_1 \leq 4.
\]

(A.11)

By implication, we also get an estimate for the 2-norm

\[
\|a^{(i)}_\ell\|_2 \leq \|a^{(i)}_\ell\|_\infty \|a^{(i)}_\ell\|_1 \leq 4.
\]

(A.12)

This is used to estimate the closure error in (2.42).
A. REAL-VALUED $P_N$ EQUATIONS
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


153


