Communicated by Prof. Dr. Manuel Torrilhon

Master Thesis

Inverse Modeling in Electron Probe Microanalysis based on Deterministic Transport Equations

Inverse Probleme in der Elektronenstrahlmikroanalyse mittels deterministischer Transportgleichungen

Submitted by: Niklas Mevenkamp
Dahmengraben 2-4, 52062 Aachen
niklas.mevenkamp@rwth-aachen.de

First supervisor: Prof. Dr. Manuel Torrilhon
Department of Mathematics
RWTH Aachen

Second supervisor: Prof. Dr. Martin Frank
Department of Mathematics
RWTH Aachen

External supervisor (GFE): Dr. Silvia Richter
Central Facility for Electron Microscopy
RWTH Aachen
To Aneta, Jan, Anette and Klemens
Abstract

The aim of this thesis is to develop advanced reconstruction algorithms for electron probe microanalysis, that allow for a significant increase of the spatial resolution and complexity of the analyzed specimens. We concentrate on two different aspects: the use of gradient based optimization techniques for the iterative reconstruction and the development of a deterministic X-ray prediction model for arbitrary specimens.

We introduce the Levenberg-Marquardt optimization algorithm and investigate its performance in combination with currently available Monte Carlo programs. Finite difference approximations of the necessary gradients and the influence of stochastic noise in Monte Carlo simulations on the error is investigated. Numerical results are presented that show the reconstruction of a particle and a layer on the nanometer scale, i.e. both features are significantly smaller than the excitation volume.

The method of moments is applied to the Boltzmann equation in continuous slowing-down approximation to retrieve a first order moment system for electron transport. A minimum entropy principle is chosen to close the set of partial differential equations, resulting in the $\mathcal{M}1$ model for electron transport. The model is reformulated in terms of an initial value problem in the energy variable with the configuration of the electron beam determining the initial condition. Based on the HLL flux in two spatial dimensions a first order finite volume scheme for the numerical solution of the initial value problem is derived. The characteristic velocities are approximated based on an analytical expression of the Jacobian of the mathematical fluxes and a rational function fit to the Eddington factor, that arises from the closure relation. Numerical results are presented that compare the $\mathcal{M}1$ solution to Monte Carlo simulations.
Acknowledgements

First of all I would like to thank Silvia Richter, who I have worked with as a student assistant for four years, and who has inspired my interest in the field of electron probe microanalysis. I am grateful for her effort to coordinate this project and especially thank her for giving me the opportunity to present my research at the 2013 European Microbeam Analysis Society workshop in Porto.

Special thanks belong to my supervisor Manuel Torrilhon, who has taught me a lot about mathematical modeling of physical problems. I thank him for his great interest in the topic and his immediate agreement to a collaboration with the Central Facility for Electron Microscopy. His constant encouragement and advice has greatly contributed to my ability to write this thesis.

I am greatly indebted to Philippe T. Pinard, who has given me deep insights into the concepts and physics behind Monte Carlo simulations and who was always available to discuss these topics. I am especially thankful for his collaboration and guidance in the implementation and usage of the interfaces between several Monte Carlo simulations and the mathematical program packages.

I am grateful to Martin Frank for introducing me to his work on electron and radiative transfer. I especially thank him for his advice on the study of the validity of the minimum entropy closure in electron probe microanalysis applications. I also thank him for sharing tabulated transport coefficients, that were gathered by his group, as well as a MATLAB module that helped verifying the calculation of the Jacobian matrix of the mathematical flux.
I thank Hendrix Demers, who I met at the workshop in Porto, for his advice on the choice and validity of analytical models for the scattering cross sections.

I also thank Torsten Trimborn, Pascal Richter and Graham Alldredge, who I have shared an office with at the MathCCES institute during my thesis. Their encouragement and practical advices have greatly helped me to sustain the motivation to bring this research to paper.
Selbstständigkeitserklärung


________________________________________________________________________
Ort, Datum                                                         Unterschrift
## Contents

### Introduction

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Physical background</td>
<td></td>
</tr>
<tr>
<td>1.1 A physical model of solid materials</td>
<td>7</td>
</tr>
<tr>
<td>1.2 Elastic and inelastic electron scattering processes</td>
<td>8</td>
</tr>
<tr>
<td>1.2.1 Elastic electron-nucleus scattering</td>
<td>9</td>
</tr>
<tr>
<td>1.2.2 Inelastic scattering of electrons</td>
<td>11</td>
</tr>
<tr>
<td>1.2.3 Cross-sections and stopping power for compounds</td>
<td>14</td>
</tr>
<tr>
<td>1.3 Photon absorption and fluorescence</td>
<td>18</td>
</tr>
<tr>
<td>1.3.1 Photoelectric absorption</td>
<td>18</td>
</tr>
<tr>
<td>1.3.2 The fluorescence effect</td>
<td>19</td>
</tr>
<tr>
<td>1.4 Definition of the inverse problem</td>
<td>20</td>
</tr>
<tr>
<td>1.4.1 Data acquired in electron probe microanalysis</td>
<td>20</td>
</tr>
<tr>
<td>1.4.2 Definition of the material parameter optimization</td>
<td>21</td>
</tr>
<tr>
<td>1.4.3 Discretization of the model parameters</td>
<td>23</td>
</tr>
<tr>
<td>1.5 A conventional reconstruction technique</td>
<td>24</td>
</tr>
<tr>
<td>1.5.1 The $ZAF$-correction factor</td>
<td>24</td>
</tr>
<tr>
<td>1.5.2 Reconstruction algorithm for homogeneous materials</td>
<td>25</td>
</tr>
<tr>
<td>1.6 The Boltzmann equation for electron transport</td>
<td>27</td>
</tr>
<tr>
<td>1.6.1 Balance law formulated in a velocity phase space</td>
<td>27</td>
</tr>
<tr>
<td>1.6.2 Transformation to the energy and direction phase space</td>
<td>32</td>
</tr>
<tr>
<td>1.6.3 Prediction of k-ratios based on the electron number density</td>
<td>34</td>
</tr>
</tbody>
</table>
## 2 Monte Carlo based reconstruction

2.1 Introduction to Monte Carlo methods .......................... 37  
  2.1.1 Basic concept behind the Monte Carlo approach .... 37  
2.2 A constrained Levenberg-Marquardt algorithm .............. 40  
  2.2.1 Defining the target function .......................... 40  
  2.2.2 The classical Gauss-Newton method ................. 42  
  2.2.3 The original Levenberg-Marquardt algorithm ....... 44  
  2.2.4 Enforcing the inequality constraint .................. 45  
2.3 Estimating partial derivatives ............................... 47  
  2.3.1 Numerical differentiation using finite differences .... 48  
  2.3.2 Influence of statistical fluctuations on finite differences 49  
2.4 Numerical results ........................................ 50  
  2.4.1 Analysis of a homogeneous compound ............... 51  
  2.4.2 Analysis of a half-spherical particle of unknown composition and diameter .......................... 53  
  2.4.3 Analysis of a thin layer of unknown composition .... 53  

## 3 A deterministic model for electron transport

3.1 The Boltzmann equation in continuous slowing-down approximation ........................................ 57  
3.2 The method of moments .................................... 59  
  3.2.1 Expansion in one of the independent variables ..... 59  
  3.2.2 The minimum entropy closure ....................... 60  
3.3 The $M_1$ model for electron transport .................... 62  
  3.3.1 Derivation of the partial differential equations ... 63  
  3.3.2 The minimum entropy closure ....................... 69  
  3.3.3 The final system of partial differential equations ... 72  

## 4 Numerical treatment of the $M_1$ model

4.1 The Godunov method for conservation laws ............... 75  
4.2 Reformulation as an initial value problem ................ 78  
  4.2.1 Transformation to a pseudo time variable ........... 78  
  4.2.2 The electron beam as an initial condition ........... 79
List of Figures

2.1 Residual of the target function for a homogeneous compound of chromium, iron and copper .......................... 52
2.2 Schematical illustration of a measurement of a chromium-carbide particle in an iron matrix ............................. 54
2.3 Schematical illustration of several measurements on thin nickel-chromium layers ........................................ 55

3.1 Normalized angular distribution of electrons in a homogeneous material ......................................................... 62
3.2 Angular distribution of electrons at a high depth .................. 63

4.1 Illustration of a simplified Riemann problem with three states 90
4.2 A rational function fit to the Eddington factor ....................... 101
4.3 Eigenvalues of the Jacobian of the mathematical flux ............ 102
4.4 Convergence behavior of the finite volume scheme in the case of linear advection ........................................ 108
4.5 Convergence behavior of the finite volume scheme in the case of the M1 equations with smooth initial conditions 109
4.6 Stopping power of electrons in copper .............................. 111
4.7 Elastic transport coefficient of electrons in water .................. 112
4.8 Spatial distribution of the M1 and Monte Carlo electron number densities in a homogeneous bulk (copper) at different energies .......................................................... 114
4.9 Normalized energy distribution of the $M_1$ and Monte Carlo electron number densities at different depths . . . . . . . . . . 115
4.10 Spatial distribution of the $M_1$ and Monte Carlo electron number densities in a binary structure (nickel and chromium) at different energies . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 117
Introduction

The main task in electron probe microanalysis (EPMA) is to analyze the chemical composition of solid materials. Often the materials of interest are heterogeneous with very fine structures (on a micro- or nanometer scale). In an electron microprobe specimens are exposed to a focused electron beam which penetrates the material surface, ionizes the atoms inside the material and causes the emission of characteristic X-rays. The intensities of these X-rays are measured by detectors inside the electron microprobe and can be used to quantify the mass concentrations of the elements.

Matrix correction methods

For composite materials no method for the direct and accurate calculation of the mass concentrations of the elements from the measured characteristic X-ray intensities is known. The most common class of methods to solve this problem are iterative matrix correction methods [33, pp. 407–426], [17, 402–420]. They combine analytical models for the prediction of X-ray intensities with iterative procedures to optimize the material parameters. In Chapter 1 we will briefly discuss the ZAF matrix correction method, which is widely used in practice. It can be used to setup a fix point iteration to optimize the material parameters.

In their classical form these methods are only accurate as long as the region of the material sample that interacts with the electron beam is homogeneous. Due to the broadening of the beam inside the specimen, caused by electron scattering, this region is usually magnitudes larger than the original beam...
INTRODUCTION

diameter. This imposes a severe restriction on either the material complexity or the resolution and accuracy of its analysis when using classical matrix correction methods.

Extended matrix correction methods have been developed for certain simple geometries. For the case of thin films on a substrate matrix correction methods based on an extended analytical model for the ionization depth distribution have been proposed [32].

Different iterative procedures and Monte Carlo methods

Over the last decades several Monte Carlo programs, that simulate the events in electron microprobe experiments, have been developed [3, 36, 15, 35]. These programs can be used to predict X-ray intensities for arbitrary materials. Therefore their use in iterative optimization algorithms could potentially increase the spatial resolution in electron probe microanalysis and extend the analysis to more complex geometries. In thin film analysis Monte Carlo simulations are used to calculate the ionization depth distribution, enabling the reconstruction of vertical layers on the submicrometer scale [3, 34]. In Chapter 2 an introduction to the basic concept behind Monte Carlo methods is given.

Apart from fix point iterations, the combination of prediction models with random search based algorithms has been suggested for the material parameter optimization. In contrast to fix point iterations this allows the use of differently many data points and parameters, especially the use of the information from multiple measurements. Such an algorithm is used in the thin film analysis reconstruction mentioned before [34]. Furthermore the exclusive use of Monte Carlo simulations in a simulated annealing algorithm for the non-destructive resolution of embedded particles of known composition at high lateral and depth resolution has been discussed [38].

In Chapter 2 we introduce a different iterative reconstruction method, namely
the Levenberg-Marquardt algorithm [9, pp. 222–224]. It uses the gradient of a target function, given by the differences between predicted and measured X-ray intensities, to determine an update of the material parameters. Compared to random search algorithms, gradient based methods usually require significantly fewer iterations to find an optimal choice of the parameters. This algorithm is designed for non-linear regression, which means that, like in random search algorithms, the information from several measurements can be combined to increase both the accuracy and the resolution of the reconstruction.

In Chapter 2 we discuss the combination of the Levenberg-Marquardt algorithm with Monte Carlo simulations. In the method proposed in [38] the composition was fixed to two known phases and only the spatial distribution of these two phases was reconstructed. The method proposed in this work does only require the knowledge of the elements present in the specimen, but not the compositions of the different phases. We present numerical experiments that show the reconstruction of particles and layers that are significantly smaller than the electron interaction volume.

A disadvantage of combining gradient based methods with Monte Carlo simulations is that the gradient has to be approximated by finite differences. In Chapter 2 we discuss how the statistical fluctuations inherent to Monte Carlo simulations increase the error in these finite differences. Furthermore each finite difference quotient requires an additional simulation, causing the number of necessary simulations per iteration to scale with the number of unknown parameters. In view of this we investigate an analytical model that allows the prediction of X-rays for arbitrary materials and could potentially eliminate these disadvantages.

**A deterministic model for electron transfer**

When neglecting the influence of fluorescence, the intensity of characteristic X-rays can be approximated based on the electron number density
distribution inside the specimen, using ionization cross sections for the generation of X-rays and an exponential absorption of the X-rays within the specimen.

The macroscopic electron number density distribution inside an arbitrary material can be described by the Boltzmann transport equation [33, pp. 110–112]. Unfortunately the numerical solution of this equation is difficult and associated with high computational costs. In the past there have been attempts to develop more practical physical models based on the Boltzmann equation [5, 6, 11], but all of them involve significant simplifications of the physical processes.

Over the last two decades a general framework for the reduction of complicated integro-differential equations to lower-dimensional partial differential equations, called the method of moments, has been developed [25, 23]. With the advance of according numerical schemes, a well developed mathematical framework for the solution of such equations is available today [18, 24, 37].

Very recently the application of the method of moments to the Boltzmann equation for electron and radiative transfer in medical radiotherapy has been discussed [20, 12, 14, 13, 29]. The conditions in radiotherapy applications differ significantly from those in electron probe microanalysis. In radiotherapy applications electrons usually have initial energies within the mega-electron-volt range, whereas the beam energy used in electron probe microanalysis is usually within the kilo-electron-volt range. Furthermore in radiotherapy applications the electrons travel through organic materials, which have very low densities, whereas the materials analyzed in electron probe microanalysis range from light elements like carbon to heavy metals.

In this thesis the first order moment system arising from the Boltzmann equation in continuous slowing-down approximation, called $M_1$ model for electron transport is discussed. The validity of this model under typical operating conditions in electron probe microanalysis is investigated. In Chapter
3 an extensive theoretical discussion of the derivation of the final system of partial differential equations for electron transport in the context of electron probe microanalysis is given. In Chapter 4 a finite volume scheme for the numerical solution of these partial differential equations is derived. Numerical results are presented that compare the solution of the $M1$ model to Monte Carlo simulations.
Chapter 1

Physical background

1.1 A physical model of solid materials

In this section we present a physical model with which arbitrarily heterogeneous materials can be represented. This model will be used to describe the microscopic interactions discussed in Section 1.2 and 1.3 and to define the unknown parameters of the inverse problem introduced in Section 1.4.

There are many quantities that have an influence on the properties of a chemical compound, like the relative positions and bonding structure of its atoms. In electron probe microanalysis solid materials can be described by their mass density distribution and the mass concentration distributions of the occurring elements. In electron probe microanalysis the different elements present inside a specimen are often known or can be measured. We denote them by $E_{l_1}, \ldots, E_{l_n}$, where $n \in \mathbb{N}$ is the total number of different elements that occur. In an experiment only a very small region of the specimen is investigated. Assuming the usual case of a flat material surface, we define

$$Q := [x_l, x_r] \times [y_l, y_r] \times [z_l, 0],$$

the region of interest within the specimen and

$$S := [x_l, x_r] \times [y_l, y_r] \times \{0\} \subset \partial Q,$$
its surface. We regard the material inside the region of interest as fully characterized, when the mass density distribution,

$$\rho \in L^2(Q; \mathbb{R}_{\geq 0}),$$

as well as the vector of the mass concentration distributions of the elements,

$$c \in L^2(Q; [0, 1]^n),$$

have been specified, where

$$L^2(Q; R) := \{ f : Q \to R \mid f \text{ is measurable}, \int_Q |f|^2 \, d\mu < \infty \}$$

for some space $R$. This function space has been chosen to allow for sharp heterogeneities of the material in the form of discontinuities of $\rho$ and $c$.

Note that the fractions $c_j$ have to add up to one at each point in the material sample. We account for this by defining

$$c_n(x) := 1 - \sum_{j=1}^{n-1} c_j(x) \text{ for } x \in Q,$$  \hspace{1cm} (1.1)

and demanding the non-negativity of $c_n$ almost everywhere in $Q$.

### 1.2 Elastic and inelastic electron scattering processes

This section gives an overview over the scattering processes of the incident electrons with the atoms of the solid material under investigation, that are most important in electron microprobe experiments. More detailed descriptions can be found in [17] and [36]. For detailed derivations of explicit formulae we refer to [33]. The probabilities of different effects of individual scattering events on the incident electron can be described by cross-sections.
1.2. ELASTIC AND INELASTIC ELECTRON SCATTERING PROCESSES

In this context, the differential cross-section

\[ \sigma_j(\epsilon, \epsilon', \Omega \cdot \Omega'), \]  

represent the probability that an electron, that is subject to a scattering event with an atom of the element \( El_j \), goes from one phase state \( (\epsilon, \Omega) \) before the collision to another phase state \( (\epsilon', \Omega') \) after the collision, where

- \( \epsilon = \text{electron energy} \) (Unit: J)
- \( \Omega = \text{direction of travel} \) (Unit: dimensionless),

with \( \Omega \in S^2 := \{ x \in \mathbb{R}^3 \mid |x| = 1 \} \). In an interaction between a single electron and a single atom we consider the forces created by both collision partners to be isotropic. Thus (1.29) only depends on \( \Omega \cdot \Omega' \), the cosine of the angle between \( \Omega \) and \( \Omega' \).

We differentiate between elastic and inelastic scattering events. In the former \( \epsilon' = \epsilon \) and in the latter \( \epsilon' < \epsilon \) holds. Therefore, elastic scattering cross-sections are just denoted:

\[ \sigma_{el,j}(\epsilon, \Omega \cdot \Omega'). \]  

At the end of this section we will discuss a general framework that allows the computation of cross-sections for compounds, given the cross-sections of the individual elements.

1.2.1 Elastic electron-nucleus scattering

A free electron that travels near the nucleus of an atom is attracted by the Coulomb force that arises from the positive charge of the nucleus. Due to the screening by the negative charges of the orbiting electrons the Coulomb force decreases from the nucleus towards the outer shells of the atom, beyond which it completely vanishes. The trajectory of a free electron, that travels through such a Coulomb field, is bend. This results in a change of
the direction of travel of the electron after leaving it. Because the mass of
the nucleus is several magnitudes larger than the mass of the electron, the
recoil of the nucleus can be neglected and thus the energy of the free electron
can be considered as being conserved.

In the following analytical expressions of the Rutherford differential cross-
sections for elastic scattering are detailed. They are used in this work for
their simplicity, that allows an introduction to the basic concepts behind the
modeling of scattering events.

**Unscreened Rutherford elastic differential cross-section**

As a first approximation, the differential cross-section for elastic scat-
tering can be derived using classical mechanics, neglecting the screening of
the atomic electrons. Its derivation introduces the basic concepts behind
the modeling of scattering events. A detailed derivation of the unscreened
Rutherford elastic differential cross-section is given in Appendix A.

Here we just give the analytical expression. We use the following notation:

\[ Z_j = \text{atomic number of element } El_j \quad \text{(dimensionless)}, \]
\[ \epsilon = \text{kinetic energy of the electron} \quad \text{(unit: J)}, \]

as well as the following physical constants:

\[ m_0 = 9.10938215 \times 10^{-31} \text{kg (electron rest mass)}, \]
\[ e = 1.602176565 \times 10^{-19} \text{C (elementary charge)}, \]
\[ \epsilon_0 = 8.854187817 \times 10^{-12} \text{F m}^{-1} \text{ (vacuum permittivity)}. \]

Then the unscreened Rutherford differential cross-section for scattering with
an element \( El_j \) is:

\[
\sigma_{R,j}(\epsilon, \theta) = \frac{e^4 Z_j^2}{16(4\pi\epsilon_0)\epsilon^2} \frac{1}{\sin^4(\theta/2)} \text{ m}^2.
\]  

(1.4)
Screened Rutherford elastic differential cross-section

In the literature one can find different equations where (A.18) has been altered in order to account for the mentioned screening effect. In [33, p. 68] the following analytical expression for the screened Rutherford elastic differential cross-section is given:

\[
\sigma_{R,j}(\epsilon, \theta) = \frac{e^4 Z_j^2}{16(4\pi \epsilon_0)^2 \epsilon^2 \left( \sin^2(\theta/2) + \sin^2(\theta_j(\epsilon)/2) \right)^2} \text{m}^2, \quad (1.5)
\]

where

\[
\theta_j(\epsilon) = \frac{\lambda(\epsilon)}{2\pi R_j} \quad \text{(screening angle)},
\]

\[
R_j = a_H Z_j^{-1/3} \text{m} \quad \text{(screening radius)},
\]

\[
\lambda(\epsilon) = \frac{h}{m_0 |v(\epsilon)|} = \frac{h}{\sqrt{2m_0 \epsilon}} \text{m} \quad \text{(de Broglie wavelength)},
\]

\[
a_H = \frac{\hbar^2 \epsilon_0}{\pi m_0 e^2} = 5.291 772 109 2 \times 10^{-11} \text{m} \quad \text{(Bohr radius)},
\]

\[
h = 6.626 069 57 \times 10^{-34} \text{J s} \quad \text{(Planck constant)}.
\]

Mott cross-sections

The Rutherford cross-sections are only recommended to model the scattering through small angles [33, p. 59]. More realistic models have been developed using quantum mechanics. However, the discussion of quantum physics is beyond the scope of this thesis. For a detailed description of the Mott cross-section, we refer to [33, pp. 69–73]. An analytical expression can be found in [10].

1.2.2 Inelastic scattering of electrons

In the following we will treat electron interactions in which incident electrons transfer energy to an atom during a collision. We will give no explicit inelastic cross-section for these events, as they will be modeled by a continuous slowing-down approximation.
Inner shell ionization & emission of characteristic X-rays

Atomic electrons, also called shell electrons, are arranged in orbitals having certain energy levels. These differ from element to element. Let $\epsilon_{O}^{E_{l}j}$ denote the energy of an electron $e_{O}^{E_{l}j}$ in an orbital $O$ of an atom of the element $E_{l}j$. For a detailed description of the orbitals, we refer to [17, pp. 279–283]. A shell electron $e_{O}^{E_{l}j}$ may become excited to a higher energy level by receiving energy from an incident electron or photon.

Immediately after such an excitation the atom relaxes to its ground state. Therefore it has to fill the vacancy in the orbital $O$ that was caused by the excitation of $e_{O}^{E_{l}j}$. This is achieved by the transition of an electron $e_{U}^{E_{l}j}$ from a more energetic orbital $U$, i.e. one with $\epsilon_{U}^{E_{l}j} > \epsilon_{O}^{E_{l}j}$, to the orbital $O$. Thereby $e_{U}^{E_{l}j}$ loses energy equivalent to the difference between the energy levels of $U$ and $O$. This energy loss is mainly preserved through two different processes. In the Auger process the energy is transmitted to an electron in a more energetic orbital, which is then completely ejected from the atom. This ejected electron is then called Auger electron. In the characteristic X-ray process the energy is transferred to a photon. Because the orbitals have a unique arrangement for each element, the energy of such a photon can be uniquely related to a transition $U \rightarrow O$ from the orbital $U$ to the orbital $O$ in an atom of the element $E_{l}j$. Therefore X-rays originating from this process are called characteristic (with respect to the corresponding element). The different energies corresponding to possible transitions within an element are called lines.

The probability that an electron of energy $\epsilon$ will ionize an atom of the element $E_{l}j$ upon collision, can be expressed as a differential ionization cross-section:

$$\sigma_{\text{ion},j}(\epsilon) \quad \text{(unit: m}^2\text{).}$$

(1.6)

For a detailed quantum-mechanical description of this probability density, see [33, pp. 78–81].
of a characteristic X-ray instead of an Auger electron, is called fluorescence yield. It depends on the line of the emitted X-ray. To simplify the notation we assume that only one characteristic X-ray line per element is regarded and denote the corresponding fluorescence yield with $\omega_j$. Then the probability, that an electron of energy $\epsilon$ will cause the emission of a characteristic X-ray of the regarded line upon collision with an atom of the element $El_j$, is given as:

$$\sigma_{\text{emiss},j}(\epsilon) := \omega_j \sigma_{\text{ion},j}(\epsilon) \quad \text{(unit: m}^2\text{).} \quad (1.7)$$

**Continuous deceleration & emission of bremsstrahlung**

In addition to the energy losses resulting in the emission of characteristic X-rays the incident electrons may lose energy due to a deceleration in the Coulomb field of the atoms. Because electrons may lose any amount of energy during this process, it is responsible for the production of a continuous spectrum of X-rays, which is often referred to as "bremsstrahlung" in the literature. This term origins from the German word "Bremsen" for deceleration and "Strahlung" for radiation.

The intensity of bremsstrahlung X-rays is usually very low and acts like a background noise in experimental measurements. Characteristic X-ray intensities are obtained from a measured spectrum by filtering out this background radiation. Because we will be interested in comparing only these characteristic X-rays, we neglect the production of bremsstrahlung X-rays in this work.

**Continuous slowing-down approximation (CSD)**

Despite the fact that electrons only loose discrete amounts of energy during the inelastic scattering processes, in practice the energy loss of the electrons is often approximated as a continuous process. The so-called stopping power describes a mean energy loss per unit length. This approximation is referred to as the continuous slowing-down approximation. Apart from the PENELOE [36] software, all Monte Carlo simulations that are common in
CHAPTER 1. PHYSICAL BACKGROUND

electron probe microanalysis, use this approximation.

In [33, p. 91], apart from a difference in units, the following expression for the stopping power is given:

\[ \tilde{S}_i(\epsilon) = \frac{2\pi e^4}{(4\pi\epsilon_0)^2\epsilon} \frac{Z_i}{A_i} \ln \left( \frac{b \epsilon}{J_i} \right) \text{J m}^2\text{kg}^{-1}, \quad (1.8) \]

where

\[ b = \sqrt{0.5e} = 1.165 \, 821 \, 990 \, 8 \]

is a relativistic constant and

\[ J_j = 1.602 \, 176 \, 565 \times 10^{-19} \begin{cases} (9.76 Z_i + 58.8 Z_i^{-0.19}) & , \quad Z_j > 6 \ J \\ 11.5 Z_i & , \quad Z_j \leq 6 \end{cases} \]

is the mean ionization potential. Note that compared to the expression given in [33, p. 91], we have eliminated Avogadro’s number, because we express

\[ A_j = \text{atomic mass of } El_j \quad (\text{unit: kg}), \]

in SI units.

In order to reformulate this as an energy loss per unit length, we have to multiply (1.8) by the mass density \( \rho_j \in L^2(Q;\mathbb{R}_{\geq0}) \) of \( El_j \) inside the material. This yields the following stopping power for the element \( El_j \) at \( x \in Q \):

\[ S_j(x,\epsilon) = \rho_j(x) \frac{2\pi e^4 Z_j}{(2\pi\epsilon_0)^2\epsilon} \ln \left( \frac{b \epsilon}{J_j} \right) \text{J m}^{-1}. \quad (1.9) \]

### 1.2.3 Cross-sections and stopping power for compounds

In the following we will first show how cross-sections for compounds can generally be computed from the cross-sections of the individual elements. Then we will apply this framework to the cross-sections and the stopping power that were presented in this section.
1.2. ELASTIC AND INELASTIC ELECTRON SCATTERING PROCESSES

A general framework to compute cross-sections for compounds

Let $A_{se}(x)$ denote the event that an observed particle traveling through a point $x \in \mathbb{R}^3$ of a background medium hits a scattering center (a point object) and let $N_V(x)$ denote the density of the scattering centers inside the medium at $x$. Then

$$P(A_{se}(x)) = N_V(x),$$

(1.10)

where $P(A)$ denotes the probability of the event $A$. Furthermore let $A(x, \eta, \eta')$ be the event that an observed particle changes its phase state from $\eta$ to $\eta'$ in a scattering event at $x$. In this work we use the term phase state to denote the vector of all quantities that are necessary to describe the physical configuration of a particle within the model (e.g. position and velocity). We then define

$$P(A(x, \eta, \eta') \mid A_{se}(x)) := f(x, \eta, \eta').$$

Then we have

$$P(A_{se}(x) \cap A(x, \eta, \eta')) = P(A(x, \eta, \eta') \mid A_{se}(x)) \cdot P(A_{se}(x))$$

$$= N_V(x) f(x, \eta, \eta').$$

(1.11)

Now we assume a differentiation between certain types of scattering centers $s_1, \ldots, s_n$ with densities $\rho_1(x), \ldots, \rho_n(x)$ in our background material, where $\sum_{j=1}^{n} N_{V,j}(x) = N_V(x)$. With $A_{se,i}(x)$ we denote the event that a particle traveling through $x$ hits a scattering center of type $s_i$. Then analogously to (1.10) we have:

$$P(A_{se,i}(x)) = N_{V,i}(x)$$

Note that $A_{se}(x) = \cup_{i=1}^{n} A_{se,i}(x)$. We also analogously model the probabilities for the different scattering events $A_i(x, \eta, \eta')$ using according densities $f_i(x, \eta, \eta')$. This yields:

$$P(A_i(x, \eta, \eta') \mid A_{se,j}(x)) = \begin{cases} f_i(x, \eta, \eta') & , j = i \\ 0 & , j \neq i \end{cases}$$

(1.12)
Finally we receive:

\[
P(A_{se}(x) \cap A(x, \eta, \eta')) = P(\cup_{i=1}^{n} A_{se,i}(x) \cap \cup_{j=1}^{n} A_j(x, \eta, \eta')) \\
= P(\cup_{i=1}^{n} \cup_{j=1}^{n} (A_{se,i}(x) \cap A_j(x, \eta, \eta'))) \\
= \sum_{i=1}^{n} \sum_{j=1}^{n} P(A_j(x, \eta, \eta') \mid A_{se,i}(x)) \cdot P(A_{se,i}(x)) \\
= \sum_{i=1}^{n} P(A_i(x, \eta, \eta') \mid A_{se,i}(x)) \cdot P(A_{se,i}(x)) \\
= \sum_{i=1}^{n} N_{V,i}(x) f_i(x, \eta, \eta').
\]

Dividing by \(N_V(x)\) and using (1.11) gives:

\[
f(x, \eta, \eta') = \sum_{j=1}^{n} \frac{N_{V,j}(x)}{N_V(x)} f_j(x, \eta, \eta'). \tag{1.13}
\]

Note that here

\[
\frac{N_{V,i}(x)}{N_V(x)} = \frac{P(A_{se,i}(x))}{P(A_{se}(x))} = P(A_{se,i}(x) \mid A_{se}(x))
\]

is the probability of hitting a scattering center of type \(s_i\) under the assumption that some scattering center has already been hit.

**Screened Rutherford cross-section for compounds**

In our application the scattering centers are the atoms inside the material. We use the following notation to denote

\[
N_V(x) = \frac{\rho(x)}{M} \text{ m}^{-3} \quad \text{(total number of atoms per m}^3\text{)},
\]

\[
M = \sum_{j=1}^{n} m_j A_j \text{ kg} \quad \text{(total molecular weight)},
\]

\[
N_{V,j}(x) = \frac{\rho_j(x)}{A_j} \text{ m}^{-3} \quad \text{(number of atoms of } E \text{ of } l_i \text{ per m}^3\text{)},
\]

16
where \( \rho_j \) denotes the partial mass density of \( El_j \) and

\[
m_j = \text{atomic fraction of } El_j \quad \text{(unit: dimensionless)}.
\]

In this work we assume the following relation between the partial and total mass density:

\[
\rho_j(x) = c_j(x) \rho(x) \text{ kg m}^{-3} \quad \text{(mass density of } El_i)\]

Note that due to the influence of the molecular structure on the partial mass densities, this model is not exact. By applying equation (1.13) to (1.5), we receive the following screened Rutherford elastic differential cross-section for composite materials:

\[
\sigma_R(\epsilon, \theta) = \sum_{i=j}^n N_{V,j}(x) \frac{\sigma_{R,j}(\epsilon, \theta)}{N_V(x)} = \frac{1}{N_V(x)} \sum_{j=1}^n c_j(x) \rho(x) A_j \sigma_{R,j}(\epsilon, \theta)
\]

\[
= \frac{1}{N_V(x)} \frac{\epsilon^4 \rho(x)}{16(4\pi\epsilon_0)^2 \epsilon^2} \sum_{j=1}^n c_j(x) \frac{Z_j^2}{A_j} \frac{\ln \left( \frac{b}{\epsilon J_j} \right)}{\left( \sin^2(\theta/2) + \sin^2(\theta_0,i(\epsilon)/2) \right)^2}.
\]

### Stopping power for compounds

Now we will apply (1.13) to the stopping power given in (1.9). Note that there \( N_{V,j}(x) = \frac{\rho_j(x)}{A_j^2} \) is already included in the expression. Thus application of (1.13) yields the following stopping power for a composite material (cf. [33, p. 92]):

\[
S(x, \epsilon) = \sum_{j=1}^n S_j(x, \epsilon) = \rho(x) \sum_{j=1}^n \frac{c_j(x) 2\pi e^4 Z_j}{A_j (2\pi\epsilon_0)^2 \epsilon} \ln \left( \frac{b \epsilon}{J_j} \right) \text{ J m}^{-1}.
\]

Alternatively, the stopping power can also be expressed as an integral over the inelastic scattering cross-section \( \sigma_{in}(x, \epsilon, \epsilon', \Omega \cdot \Omega') \). The following general
expression is given in [13]:

\[ S(x, \epsilon) = 2\pi \int_{0}^{\infty} \int_{-1}^{1} (\epsilon - \epsilon')\sigma_{\text{in}}(x, \epsilon, \epsilon', \mu) \, d\mu \, d\epsilon'. \] (1.16)

Using an analytical expression for the inelastic differential scattering cross-section one could compute a closed expression of this integral.

1.3 Photon absorption and fluorescence

In this section we explain the most important interactions of X-rays with the background material. For more details, see [36].

1.3.1 Photoelectric absorption

Similar to the inelastic scattering of an electron at an atom, as described in Section 1.2, photons may also transfer energy to an atom during a collision. The main difference is that photons will either lose all of their energy and become completely absorbed, or transfer no energy at all. This is crucial for quantitative X-ray analysis, because it means that the specific energies of characteristic X-rays are exactly retained.

The probability that an atom absorbs a given specific amount of energy in an interaction with an X-ray depends significantly on the structure of the atomic electrons and thereby on the according element. As discussed in Section 1.2, characteristic X-rays have specific energies that depend on the shell transition and element they origin from. Therefore the probabilities for photoelectric absorption events are dependent both on the line of the incident X-ray and the element of the incident atom.

Exponential absorption approximation

In practice a similar approach as the continuous slowing-down approximation for the energy loss of electrons is often applied to the overall loss of photons in the system. An exponential decrease of the amount of X-rays
surviving along a straight path with increasing path length has been found empirically [17, p. 412]:

\[ I = I_0 e^{-\mu x}. \] (1.17)

Here

\[ \mu_j = \text{MAC of } El_i \text{ X-rays absorbed by } El_j \quad \text{(unit: } \text{m}^2\text{kg}^{-1}). \]

is the so-called mass absorption coefficient (MAC), \( I_0 \) is the initial intensity of the X-rays and \( t \) the traveled path length. The dependence on the energy of the X-ray is expressed by the element whose transition caused its generation. Note that in favor of a shorter notation we omit the differentiation between different X-ray lines of the same element here.

These mass absorption coefficients have been empirically tabulated and are available for a wide range of common combinations of X-ray lines and absorbing elements.

The mass absorption coefficient for the absorption of \( El_i \) X-rays within a composite background material can be calculated as follows (cf. [17, pp. 291–292]):

\[ \mu_i(x) = \sum_{j=1}^{n} c_j(x)\mu_j, \quad x \in Q. \] (1.18)

1.3.2 The fluorescence effect

The energy, that is transferred to an atom during photoelectric absorption, is conserved by the same processes that occur when electrons excite atoms (see Section 1.2). Therefore X-rays with energies higher than the minimum excitation energies of the occurring elements can also be responsible for the generation of characteristic X-rays. This phenomenon is called fluorescence.

Depending on the occurring elements, the contribution of fluorescence to
the measured characteristic X-ray intensity can be more or less significant. In this work we focus on the cases where the direct approximation of the measured X-ray intensities based on the exponential absorption of the generated X-rays is sufficiently accurate.

1.4 Definition of the inverse problem

In this section we will first describe the data that is available about the material from electron microprobe experiments. Then we define the inverse problem that relates this data to the model parameters. Finally we will discretize the model parameters in order to reduce the dimensions of the unknowns and express the inverse problems in terms of a least squares fit.

1.4.1 Data acquired in electron probe microanalysis

A detailed description of electron microprobe operations can be found in [17]. Here we summarize the main steps from the initial electron beam to the measured characteristic X-ray intensities.

At first we regard the data acquired from a single measurement. The specimen is exposed to a focused electron beam of a specific energy $\epsilon_b$. It is usually directed orthogonally onto a spot on the sample surface. For the duration of the exposure the free electrons interact with the specimen, which causes a significant broadening of the beam as the electrons penetrate the material. Furthermore the production of characteristic X-rays are caused by inner shell ionizations, as discussed in Section 1.2.

X-rays, that leave the material in the direction of a detector inside the microprobe, are counted for a certain amount of time. After subtracting the background radiation only the counts of the characteristic X-rays remain, which can then be interpreted as net intensities. We assume that from $m$ measurements, taken at the points $x_1, \ldots, x_m \in S$, we receive $m \cdot n$ charac-
1.4. Definition of the Inverse Problem

Characteristic X-ray intensities

\[ I_{ij}^{\text{exp}} \in \mathbb{R}_{\geq 0}, \ i = 1, \ldots, m\ j = 1, \ldots, n, \]  \hspace{1cm} (1.19)

corresponding to the elements \( El_1, \ldots, El_n \). In practice the points \( x_1, \ldots, x_m \) are often chosen in equidistant-distant steps along a line or a two-dimensional grid.

Note that we have assumed here, that for each element only one characteristic X-ray line is regarded. The extension of (1.19) to more lines per element is straightforward.

1.4.2 Definition of the material parameter optimization

In Section 1.1 we have defined the model parameters \( \rho \in L^2(Q; \mathbb{R}_{\geq 0}) \) and \( c \in L^2(Q; [0,1])^n \), that characterize the region of interest inside the specimen. In the following we assume that \( \rho \) is either known or can be approximated by a simple physical model of the form \( \rho = \rho(c) \). An example of such an approximation for constant \( c \) is given in [15]. Using (1.1), the mass concentration distributions \( c_j, j = 1, \ldots, n-1 \) remain as the unknown model parameters.

Optimizing \( c_j \) for \( j = 1, \ldots, n-1 \) given the data \( I_{ij}^{\text{exp}}, i = 1, \ldots, m, j = 1, \ldots, n \) yields an inverse problem: we have to find a vector of mass concentration distributions such that the resulting model of the material in a theoretical experiment produces data that is identical or a best fit to the actually measured data. To assess this inverse problem we have to be able to solve the corresponding forward problem: given \( \rho \in L^2(Q; \mathbb{R}_{\geq 0}) \) and \( c \in L^2(Q; [0,1])^n \) predict the resulting characteristic X-ray intensities

\[ I_{ij} : L^2(Q; [0,1])^n \times L^2(Q; \mathbb{R}_{\geq 0}) \to \mathbb{R}_{\geq 0}, \]

\[ I_{ij} = I_{ij}(c_1, \ldots, c_{n-1}; c_n, \rho). \]  \hspace{1cm} (1.20)
CHAPTER 1. PHYSICAL BACKGROUND

In general the intensity $I_{ij}$ will depend on all $c_1, \ldots, c_{n-1}$. One reason is the dependence of the scattering behavior of the electrons on all mass concentrations, as seen in (1.14). Another is the dependence of the X-ray absorption behavior on all mass concentrations, as seen in (1.18).

Most models of the form (1.20) do not account for instrumental or operational parameters such as the efficiency of the detector or the duration of the beam exposure. Because the measured intensities $I_{ij}^{\text{exp}}$ change proportionally with these parameters, a normalization is possible. Therefore a standard intensity $I_{j,\text{std}}^{\text{exp}}$ is measured for each element $El_j$, using a homogeneous material with known density

$$\rho_{\text{std}} \in \mathbb{R}_{\geq 0},$$

and composition

$$c_{\text{std}}^n \in [0, 1]^n, \quad c_{\text{std}}^j > 0, \quad \sum_{i=1}^{n} c_{\text{std}}^i = 1.$$

Note that $I_{j,\text{std}}^{\text{exp}}$ does not depend on the position of the measurement, because $c_{\text{std}}$ is constant. To receive standard intensities for the theoretical model, we set:

$$I_{j,\text{std}} := I_{ij}(c_{\text{std}}^1, \ldots, c_{\text{std}}^{n-1}; c_{\text{std}}^n, \rho_{\text{std}}).$$

The resulting calibrated intensities are referred to as k-ratios. They are defined as:

$$k_{ij}(c_1, \ldots, c_{n-1}; c_n, \rho) := \frac{I_{ij}(c_1, \ldots, c_{n-1}; c_n, \rho)}{I_{j,\text{std}}} \quad \text{and} \quad k_{ij}^{\text{exp}} := \frac{I_{ij}^{\text{exp}}}{I_{j,\text{std}}}.$$

(1.21)

Assuming the presence of a model of the form (1.20), the inverse problem in equality form reads as follows:

**Problem 1 (Inverse problem in equality form).**

*Given a mass density $\rho \in L^2(Q; \mathbb{R}_{\geq 0})$ and $m \cdot n$ k-ratios $k_{ij}^{\text{exp}}, \ldots, k_{mn}^{\text{exp}}$ corresponding to the elements $El_1, \ldots, El_n$, measured at $m$ positions $x_1, \ldots, x_m$,.*
find mass concentrations $c_1, \ldots, c_{n-1} \in L^2(Q; [0,1])$ such that

$$k_{ij}(c_1, \ldots, c_{n-1}; c_n, \rho) = k_{ij}^{exp} \quad \text{for } j = 1, \ldots, n \text{ and } i = 1, \ldots, m,$$

$$1 - \sum_{j=1}^{n-1} c_j \geq 0 \quad \text{almost everywhere in } Q.$$

### 1.4.3 Discretization of the model parameters

In Problem 1 the unknowns are elements of $L^2$ and thus of infinite dimension whereas the set of available data is finite. Thus Problem 1 is heavily underdetermined. To reduce the unknowns to finite dimension we restrict $c$ to be a vector of piecewise constant functions:

$$c_{j|T_i} \equiv \text{const}, \text{ for } i = 1, \ldots, n_T, \text{ and } j = 1, \ldots, n,$$

where

$$T = \{T_1, \ldots, T_{n_T}\}, \quad \bigcup_{i=1}^{n_T} T_i = Q, \quad T_i \cap T_j \subset (\partial T_i \cup \partial T_j) \text{ for all } i \neq j, \quad (1.22)$$

is some finite partition of $Q$. We denote the space of such piecewise constant $L^2$ functions by:

$$X^0_T(R) := \{ c \in L^2(Q; R) \mid c_{|T} \equiv \text{const} \text{ for all } T \in \mathcal{T} \}.$$

In this work we will treat two different types of partitions, that justify this simplification of $c$. In an analytical partition the $T_i$ represent different regions of the material. Here the specimen consists of homogeneous regions, whose geometry is either known or is given as a parametrized analytical object (e.g. spherical particles or layers). If the geometry is completely unknown, a more generic partition can be used. Then $\mathcal{T}$ is a geometrical mesh over the sample constructed from simplices or quadrilaterals. The mesh size represents the resolution of the reconstruction and should be related to the information content of the given data.
In Problem 1 the model data predicted for the solution is required to exactly fit the measured data. This is not a realistic requirement, because all macroscopic models for the prediction of X-rays treated in this work introduce a model error arising from the simplification of physical processes in their derivation. Thus we will reformulate Problem 1 in terms of a least squares regression using the discretized mass concentration distributions:

**Problem 2** (Discretized inverse regression problem).

Given a mass density \( \rho \in X_T(\mathbb{R}_\geq 0) \) and \( m \cdot n \) \( k \)-ratios \( k_{11}^{\text{exp}}, \ldots, k_{mn}^{\text{exp}} \) corresponding to the elements \( E_1, \ldots, E_n \), measured at \( m \) positions \( x_1, \ldots, x_m \), and a partition \( T \) as defined in (1.22), solve

\[
\minimize_{c \in X_T^T([0,1])^{n-1}} \sum_{i=1}^M \sum_{j=1}^n \left( k_{ij}(c_1, \ldots, c_{n-1}; c_n, \rho) - k_{ij}^{\text{exp}} \right)^2,
\]

subject to

\[
1 - \sum_{j=1}^{n-1} c_{j|T_i} \geq 0, \text{ for } i = 1, \ldots, N.
\]

### 1.5 A conventional reconstruction technique

In this section we will briefly present the \( ZAF \)-correction method. This conventional reconstruction technique is commonly used for quantitative analysis in EPMA. A description of the historical development of this method can be found in [17, pp. 391–420]. For explicit formulae we refer to [33, pp. 407–424]. The method is only accurate for homogeneous specimens or, using suitable extensions, certain simple geometries like thin films on a substrate [32].

#### 1.5.1 The \( ZAF \)-correction factor

We treat the case of a homogeneous material here and adjust the notation introduced in Section 1.4 to a single measurement \( k_j^{\text{exp}} \) and assume that the \( c_j \) are constant.

Based on empirical observations, in first approximation a proportional re-
1.5.1 Relationship between mass fraction and characteristic X-ray intensity

The relationship between the mass fraction of an element and its corresponding characteristic X-ray intensity is assumed (Castaing’s first approximation) [17, p. 402]:

\[ k_j(c, \rho) = \frac{I_j(c, \rho)}{I_{j,\text{std}}} \approx \frac{c_j}{c_{j,\text{std}}} \cdot Z_j(c, \rho). \]

There are three major effects that can cause a significant deviation from this proportional relationship. First of all, as seen in Section 1.2, the mass concentrations of all elements have an influence on the electron scattering behavior and thus on the generation of X-rays. The so-called atomic number correction accounts for this effect:

\[ k_j(c, \rho) \approx \frac{c_j}{c_{j,\text{std}}} \cdot Z_j(c, \rho). \]

As discussed in Subsection 1.3.1, the intensity of the generated X-rays is weakened through photoelectric absorption, which also depends on all mass concentrations. Therefore an absorption correction is applied:

\[ k_j(c, \rho) \approx \frac{c_j}{c_{j,\text{std}}} \cdot Z_j(c, \rho) \cdot A_j(c, \rho). \]

Finally correction factors accounting for the fluorescence effect have been developed, resulting in the so-called \( ZAF \)-correction factor:

\[ ZAF_j(c, \rho) := Z_j(c, \rho) \cdot A_j(c, \rho) \cdot F_j(c, \rho). \]

Multiplication by the ratio between the unknown and standard mass concentration yields the following model for the prediction of calibrated intensities:

\[ k_j(c, \rho) = ZAF_j(c, \rho) \frac{c_j}{c_{j,\text{std}}}. \tag{1.23} \]

1.5.2 Reconstruction algorithm for homogeneous materials

Due to the form of (1.23) it is advantageous to formulate the optimization of \( c \) in terms of Problem 1. Formally equalizing the predicted and
experimental k-ratios yields:

\[ ZAF_j(c, \rho) \frac{c_j}{c_{j,\text{std}}} = k_j^{\text{exp}} \text{ for } j = 1, \ldots, n \]

\[ \Leftrightarrow c_j = \frac{k_j^{\text{exp}} c_{j,\text{std}}}{ZAF_j(c, \rho)} \text{ for } j = 1, \ldots, n. \]

This can be vectorized:

\[ c = \text{diag}(k_1^{\text{exp}} \frac{1}{ZAF_1(c, \rho)}, \ldots, k_n^{\text{exp}} \frac{1}{ZAF_n(c, \rho)}) c_{\text{std}}. \]

Now we can reformulate Problem 1 in terms of a fix point problem.

**Problem 3** (Conventional reconstruction problem).

Given \( n \) k-ratios \( k_1^{\text{exp}}, \ldots, k_n^{\text{exp}} \) corresponding to the elements \( E_1, \ldots, E_n \), retrieved from a single measurement on a completely homogeneous sample of mass density \( \rho \in \mathbb{R}_{\geq 0} \), find mass concentrations \( c \in [0, 1]^n \), such that:

\[ c = \text{diag}(k_1^{\text{exp}} \frac{1}{ZAF_1(c, \rho)}, \ldots, k_n^{\text{exp}} \frac{1}{ZAF_n(c, \rho)}) c_{\text{std}}, \]

\[ \sum_{j=1}^n c_j = 1. \]

When neglecting the equality constraint, this problem can be solved using a fix point iteration. In practice simple updates are used between iterations to correct the mass fractions if they violate the constraints. We will simply divide each mass fraction by the total sum of all mass fractions. The following fix point iteration represents a conventional reconstruction algorithm for homogeneous specimens that is commonly used in practice:

**Algorithm 1** (Conventional reconstruction algorithm).

Let \( k_1^{\text{exp}}, \ldots, k_n^{\text{exp}} \) and \( \rho \) as in Problem 3. Then given an initial guess \( c^0 \in [0, 1]^n \), \( \sum_{j=1}^n c_j^0 = 1 \) of the material composition and a threshold \( \epsilon > 0 \), compute:

\[ \text{function } \text{RECONSTRUCT}(\rho, c^0, \epsilon) \]

\[ i \leftarrow 0 \]
1.6. The Boltzmann equation for electron transport

In this section we will derive a deterministic transport equation that describes the dynamics of the electron number density in a six-dimensional phase space, i.e. in dependence of position (three dimensions), energy (one dimension) and direction of travel (two dimensions). The model is based on the assumption that the total amount of electrons is conserved and the only interactions between the free electrons and the solid material are binary collisions between the electrons and atoms. The cross-sections introduced in Section 1.2 are used to model these scattering events. A balance law is formulated that enables the description of the dynamics of the electrons on the macroscopic scale. At the end of this section we will also discuss how characteristic X-ray intensities (or k-ratios) can be predicted based on the electron number density.

1.6.1 Balance law formulated in a velocity phase space

The dynamics of the electrons can be described in terms of a balance law [33, pp. 110–112]. Instead of describing the history of individual electrons, we count the total number of electrons with a certain velocity inside a small control volume and balance the physical processes that influence this number. Based on the assumption that the total number of electrons

\[
\begin{array}{l}
\text{repeat} \\
\quad \text{for } j = 1, \ldots, n \text{ do} \\
\quad \quad c_{j}^{i+1} := c_{j}^{\text{std}} \cdot k_{j}^{\exp} / Z_{j} A \Phi_{j}(c_{j}^{i}, \rho) \\
\quad \text{end for} \\
\quad \hat{c} := \sum_{j=1}^{n} c_{j}^{i+1} \\
\quad c_{i}^{i+1} \leftarrow c_{i}^{i+1} / \hat{c} \\
\quad i \leftarrow i + 1 \\
\quad \text{until } \|c_{i}^{i} - c_{i}^{i-1}\|_{2} < \epsilon \\
\quad \text{return } c_{i}^{i} \\
\text{end function.}
\end{array}
\]
is conserved, this leads to a macroscopic description of the dynamics of the electrons.

Let

\[ n(x, v, t) = \text{electron number density, (unit: } J^{-1} m^{-3} \text{)} \]  

(1.24)

denote the density of the free electrons at a point \( x \in \mathbb{R}^3 \) in space and \( t \in \mathbb{R} \) in time, with velocity \( v \in \mathbb{R}^3 \) (as a directed quantity). This means that at a time \( t \), \( n(x, v, t)d^3x d^3v \) electrons with velocities \( v \) in \( d^3v \) occupy the volume \( d^3x \) around \( x \). Note that \( |v| \) can be expressed by the kinetic energy

\[ \epsilon = \frac{1}{2}m_0|v|^2 \]  

(1.25)

and \( \frac{v}{|v|} = \Omega \) is a dimensionless unit vector specifying the direction of travel. Thus \( d^3v \) can also be formulated in terms of \( d\epsilon d\Omega \) and hence the unit of \( n \) is \( J^{-1} m^{-3} \).

Let \( V \subset \mathbb{R}^3 \) be an arbitrary spatial volume. In the following we will derive a balance law that describes the change of the total number of electrons with velocity \( v \) in \( V \) over time:

\[ \frac{\partial}{\partial t} \int_V n(x, v, t) \, dx. \]  

(1.26)

Therefore we model all physical processes that have an influence on (1.26) based on the assumption that free electrons within the material only change their velocity \( v \) in binary collisions with atoms.

**Electron source**

Theoretically there could be some source that generates electrons inside \( V \). In our application however, the electrons enter the system from outside of the material. In this work the electron beam will be modeled as an initial condition and therefore we don’t model a source term here.
Flux of electrons through the boundary

Electrons might leave $V$ through its boundary $\partial V$ during their flight, without changing their velocity. The directed electron flux is:

$$J(x,v,t) = vn(x,v,t) \quad \text{(unit: } J^{-1}m^{-2}s^{-1}),$$

i.e. at a time $t$, $J(x,v,t)dSd^3vdt$ electrons move through a surface element $dS$ around $x$ with velocities $v$ in $d^3v$ per time interval $dt$. Let $\vec{n}$ denote the outward unit normal vector of $\partial V$. Then the flux of electrons over the boundary $\partial V$ is

$$-\int_{\partial V} J(x,v,t) \cdot \vec{n} \, ds.$$

Application of the divergence theorem (Gauss’s theorem) yields:

$$-\int_{\partial V} J(x,v,t) \cdot \vec{n} \, ds = -\int_V v \cdot \nabla_x n(x,v,t) \, dx. \quad (1.28)$$

Loss of electrons due to scattering out of the phase space

Let

$$\sigma(x,v,v'), \quad \text{(unit: } m^2 \text{ J}^{-1})$$

denote the differential cross-section for scattering of electrons of directed velocities $v$ into $v'$ at a point $x$. This notation is used for simplicity here. We will distinguish between the elastic and inelastic cross sections discussed in Section 1.2 later. Furthermore we define

$$\sigma^{\text{tot}}(x,v) := \int_{R^3} \sigma(x,v,v') \, dv', \quad \text{(unit: } m^2) \quad (1.30)$$

the combined total differential cross-section.

Now we regard a stationary scattering center, i.e. an atom, positioned at $x \in V$. The number of electrons with velocity $v$ that travel through the
point \( x \) within an infinitesimal time interval is given by
\[
|v| n(x, v, t)
\]
Thus the number of electrons that hit the scattering center at \( x \) per infinitesimal time interval is given by:
\[
N_V(x)|v|n(x, v, t).
\]
All of these electrons may scatter out of the phase space element \( dv^3 \), with a probability according to (1.29). Thus we receive the following contribution to the balance (with negative sign):
\[
\begin{align*}
- \int_V \int_{\mathbb{R}^3} N_V(x)|v|n(x, v, t)\sigma(x, v, v') dv'dx \\
= - \int_V N_V(x)\sigma_{\text{tot}}(x, v)|v|n(x, v, t) dx.
\end{align*}
\] (1.31)

**Gain of electrons due to scattering into the phase space**

Analogously an electron with an arbitrary velocity \( v' \in \mathbb{R}^3 \) that passes through \( x \in V \) might scatter and thereby change its velocity to \( v \). This gives the following contribution to the balance:
\[
\begin{align*}
\int_V \int_{\mathbb{R}^3} N_V(x)|v'|n(x, v', t)\sigma(x, v', v) dv'dx \\
= \int_V \int_{\mathbb{R}^3} N_V(x)\sigma(x, v', v)|v'|n(x, v', t) dv'dx.
\end{align*}
\] (1.32)

**Time dependent balance law**

Balancing the change of the total amount of electrons with velocity \( v \) inside \( V \) over time with the loss terms (1.28), (1.31) and the gain term
(1.32) gives the following time-dependent balance law:

\[
\frac{\partial}{\partial t} \int_V n(x, v, t) \, dx = - \int_V v \cdot \nabla_x n(x, v, t) \, dx \\
- \int_V N_V(x)\sigma_{\text{tot}}(x, v)|v|n(x, v, t) \, dx \\
+ \int_V \int_{\mathbb{R}^3} N_V(x)\sigma(x, v', v)|v'|n(x, v', t) \, dv' \, dx
\]

(1.33)

Stationary balance law

In an experiment there is a continuous flow of beam electrons into the specimen for the duration of the acquisition. Due to the small time scale of the physical processes we assume that shortly after the electron beam is switched on the distribution of the electrons inside the specimen reaches a constant state and remains in this state as long as the operating conditions are kept constant.

In view of this, we assume that the balance (1.33) is stationary in time:

\[
\frac{\partial}{\partial t} \int_V n(x, v, t) \, dx = 0.
\]

Thus we can now omit the time variable \( t \) from the notation and remove it from the phase space. This gives the following stationary balance law:

\[
\int_V v \cdot \nabla_x n(x, v) \, dx = \int_V \int_{\mathbb{R}^3} N_V(x)\sigma(x, v', v)|v'|n(x, v') \, dv' \, dx \\
- \int_V N_V(x)\sigma_{\text{tot}}(x, v)|v|n(x, v) \, dx.
\]

(1.34)

Because the initially chosen volume \( V \subset \mathbb{R}^3 \) was arbitrary, the fundamental lemma of the calculus of variations implies that (1.34) holds pointwise. This yields the following Boltzmann equation for electron transport in solid


This equation describes a time-stationary balance between the transport of electrons with velocity \( v \in \mathbb{R}^3 \) through a point \( x \in Q \), described by the term

\[
v \cdot \nabla_x n(x, v),
\]

and the scattering of electrons with velocity \( v \) at \( x \) to other velocities, described by the term

\[
N_V(x) \sigma_{\text{tot}}(x, v) |v| n(x, v),
\]

and the scattering of electrons with velocities \( v' \in \mathbb{R}^3 \) at \( x \) to the velocity \( v \), described by the term

\[
N_V(x) \int_{\mathbb{R}^3} \sigma(x, v', v) |v'| n(x, v') \, dv'.
\]

### 1.6.2 Transformation to the energy and direction phase space

Replacing \( v \) in the scalar product on the left-hand side of (1.35) by

\[
v = |v| \Omega, \quad \Omega \in S
\]

yields:

\[
\Omega \cdot \nabla_x |v| n(x, v) = N_V(x) \int_{\mathbb{R}^3} \sigma(x, v', v) |v'| n(x, v') \, dv' - N_V(x) \sigma_{\text{tot}}(x, v) |v| n(x, v).
\] (1.36)

As seen in (1.25), there is a direct relationship \( |v| = v(\epsilon) \) between the energy of an electron and its absolute velocity. Due to the form of (1.36) we define the following density to perform a transformation from the directed velocities
1.6. THE BOLTZMANN EQUATION FOR ELECTRON TRANSPORT

33

to the energy and direction variables:

\[ \psi(x, \epsilon, \Omega) := |v(\epsilon)| n(x, |v(\epsilon)| \Omega) \quad \text{(unit: } J^{-1} m^{-2} s^{-1}) \]  (1.37)

This can be interpreted as a fluence of the electrons, in the sense that per time interval \( dt \), \( \psi(x, \epsilon, \Omega) \cos \theta dS d\Omega d\epsilon dt \) electrons travel through the surface element \( dS \) around \( x \) into the element of solid angle \( d\Omega \) around \( \Omega \) with an energy between \( \epsilon \) and \( \epsilon + d\epsilon \). Here \( \theta \) is the angle between \( \Omega \) and \( \hat{n}_S \), the outer normal of \( dS \).

Applying the transformation (1.37) in (1.35) gives the following Boltzmann transport equation for the fluence of electrons in a six-dimensional phase space of position, energy and direction of travel:

\[
\Omega \cdot \nabla_x \psi(x, \epsilon, \Omega) = N_V(x) \int_0^\infty \int_{S^2} \sigma(x, \epsilon', \epsilon, \Omega' \cdot \Omega) \psi(x, \epsilon', \Omega') \, d\Omega' \, d\epsilon' \\
- N_V(x) \sigma_{\text{tot}}(x, \epsilon) \psi(x, \epsilon, \Omega). 
\]  (1.38)

Here we have used that the differential cross-sections

\[ \sigma(x, \epsilon, \epsilon', \Omega' \cdot \Omega'), \]

that are now expressed as they were introduced in Section 1.2, already contain the constants arising from the integral transformation

\[ \int_{\mathbb{R}^4} \cdots \, dv' = \int_0^\infty \int_{S^2} \cdots \, d\Omega' \, d\epsilon'. \]

Note that due to the transformation to the energy and direction variables,

\[ \sigma_{\text{tot}}(x, \epsilon) = 2\pi \int_0^\infty \int_{-1}^1 \sigma(x, \epsilon, \epsilon', \mu) \, d\mu \, d\epsilon', \quad \text{(unit: } m^2), \]  (1.39)

now depends only on the position and the energy of the electron before the collision.
Decomposing $\sigma(x, \epsilon, \epsilon', \Omega \cdot \Omega')$ into elastic and inelastic cross-sections

$$
\sigma(x, \epsilon, \epsilon', \Omega \cdot \Omega') = \begin{cases} 
\sigma_{\text{in}}(x, \epsilon, \epsilon', \Omega \cdot \Omega') & \text{for } \epsilon' < \epsilon, \\
\sigma_{\text{el}}(x, \epsilon, \Omega \cdot \Omega') & \text{for } \epsilon' = \epsilon, \\
0 & \text{for } \epsilon' > \epsilon.
\end{cases}
$$

(1.40)

finally gives the following Boltzmann transport equation:

$$
\Omega \cdot \nabla_x \psi(x, \epsilon, \Omega) = N_V(x) \int_{-\infty}^{\infty} \int_{S^2} \sigma_{\text{in}}(x, \epsilon, \epsilon', \Omega \cdot \Omega') \psi(x, \epsilon', \Omega') \, d\Omega' \, d\epsilon' \\
+ N_V(x) \int_{S^2} \sigma_{\text{el}}(x, \epsilon, \Omega \cdot \Omega') \psi(x, \epsilon, \Omega') \, d\Omega' \\
- N_V(x) \sigma_{\text{tot}}^{\text{in}}(x, \epsilon) \psi(x, \epsilon, \Omega) \\
- N_V(x) \sigma_{\text{tot}}^{\text{el}}(x, \epsilon) \psi(x, \epsilon, \Omega).
$$

(1.41)

### 1.6.3 Prediction of k-ratios based on the electron number density

In the following we derive an expression in terms of the electron fluence $\psi$ and the differential ionization cross-section $\sigma_{\text{ion}}$, that predicts the intensity of the generated characteristic X-rays and apply the exponential absorption approximation to retrieve a prediction for the theoretically measured characteristic X-rays. The formulae presented here are based on the physical procedures inside the Monaco [3] Monte Carlo program.

Let us regard an atom of the element $E_l_j$ as a stationary scattering center at a point $x \in Q$ within the background material. Then the number of electrons with energy $\epsilon$ and traveling in an arbitrary direction, that hit the atom per infinitesimal time interval, is given by

$$
N_{V,j}(x) \int_{S^2} \psi(x, \epsilon, \Omega) \, d\Omega.
$$
Multiplication with the differential ionization cross-section (1.6) corresponding to $El_j$ and integration over all energies yields the density of the inner shell ionizations of atoms of $El_j$ at $x$ per infinitesimal time interval (cf. [3]):

$$N_{V,j}(x) = \int_0^\infty \sigma_{\text{ion},j}(x,\epsilon) \int_{S^2} \psi(x,\epsilon,\Omega) \, d\Omega \, d\epsilon.$$  (1.42)

To calculate the net intensity of generated characteristic X-rays (no absorption) corresponding to $El_j$, we have to multiply (1.42) with the fluorescence yield, as discussed in Section 1.2. Using the definition (1.7), the net intensity of the characteristic X-rays generated at $x$ per infinitesimal time interval can be expressed as:

$$I_{0,ij}(x) = \omega_j N_{V,j}(x) \int_0^\infty \sigma_{\text{ion},j}(x,\epsilon) \int_{S^2} \psi_i(x,\epsilon,\Omega) \, d\Omega \, d\epsilon \, m^{-3} \, s^{-1}$$  (1.43)

$$= N_{V,j}(x) \int_0^\infty \sigma_{\text{emiss},j}(x,\epsilon) \int_{S^2} \psi_i(x,\epsilon,\Omega) \, d\Omega \, d\epsilon \, m^{-3} \, s^{-1}.$$

Here $\psi_i$ denotes the electron fluence determined by the Boltzmann equation (1.41) with initial conditions that correspond to the measurement $x_i$.

Now we apply the exponential absorption approximation, discussed in Section 1.3, to predict the intensity of the X-rays leaving the sample in the direction of the detector inside the electron microprobe, given their origin $x$ and initial intensity $I_{0,ij}(x)$. Let $x_d \in \mathbb{R}^3$ denote the position of the detector and

$$d(x) := \{\lambda x_d + (1 - \lambda)x \mid \lambda \in [0, 1], \lambda x_d + (1 - \lambda)x \in Q\},$$

the straight path from $x$ to the surface $S$ in the direction of the detector. Now the formula (1.17) for homogeneous materials can be translated for heterogeneous materials using an integral formulation:

$$I_{ij}(x) = I_{0,ij}(x)e^{-\int_{d(x)} \mu(y) \rho(y) \, dy} \, m^{-3} \, s^{-1}.$$  (1.44)

Finally integration over the whole domain $Q$ corresponding to the material sample results in the following model for the prediction of the total net inten-
sity (1.20) of the measured characteristic X-rays of $E l_j$, based on the mass concentration distributions $c$, the mass density $\rho$ and the electron fluence $\psi_i$:

$$I_{ij}(c, \rho) = \int_Q I_{0,ij}(x)e^{-\int_{0(x)}\mu(y)\rho(y)dy} dx \text{ s}^{-1}$$  \hspace{1cm} (1.45)$$

Using (1.21) the according k-ratios can be computed as well. Note that $I_{ij}(c, \rho)$ has the unit s$^{-1}$ here, because it is an intensity that is measured per infinitesimal time. One could multiply (1.45) by the acquisition time of each measurement, but as discussed in Section 1.4 we can omit this when using k-ratios.
Chapter 2

Monte Carlo based reconstruction

2.1 Introduction to Monte Carlo methods

In this section we discuss the basic concept behind Monte Carlo type methods and formulate a general algorithm that can be used to implement such a method. For a more practical introduction, see [22]. Finally we will discuss the possibilities and some limitations of Monte Carlo methods.

2.1.1 Basic concept behind the Monte Carlo approach

Monte Carlo methods can be used to approximate the solution of the Boltzmann equation (1.41) for theoretically arbitrary materials. Therefore the trajectories of a large amount of single electrons are simulated using random variables to decide the outcome of individual scattering events. The result is subject to statistical fluctuations with a standard deviation that is directly related to the amount of simulated electrons.

In the Monte Carlo simulations that allow for the most generic material structures, the sample can be divided into arbitrary regions $T_1, \ldots, T_{n_T}$ as in (1.22).
Assume that an electron $e$ with energy $\epsilon$ traveling in the direction $\Omega$ is at point $x$ and subject to a scattering event. The further direction of travel is then determined using a random variable:

$$\Omega' = X(x, \epsilon, \Omega), \quad P(X(x, \epsilon, \Omega) = \Omega') = \sigma(x, \epsilon, \Omega \cdot \Omega').$$ (2.1)

The mean distance $e$ is expected to travel in this direction, called the mean-free-path, is inversely proportional to the probability of hitting a scattering center and interacting with it:

$$\Lambda(x, \epsilon) = \frac{1}{N_V(x)\sigma^{\text{tot}}(x, \epsilon)}.$$ 

Note that because this mean free path is only valid inside the region $T$ with $x \in T$, it might be necessary to restrict the actual step length. Therefore we set

$$\Delta x = \min\{Y(x, \epsilon), \text{dist}_{\Omega'}(x, \partial T)\}, \quad E(Y(x, \epsilon)) = \Lambda(x, \epsilon),$$ (2.2)

where $E(Y)$ denotes the expectancy of $Y$ and $\text{dist}_{\Omega}(x, \partial T)$ is chosen such that

$$x + \text{dist}_{\Omega}(x, \partial T)\Omega \in \partial T.$$ 

Then we move $e$ accordingly:

$$x' = x + \Delta x\Omega'.$$ (2.3)

If due to this truncation $e$ ends at the boundary of a region, in the next iteration it will be regarded as being in the adjacent region in the direction $\Omega'$.

The energy loss of $e$ during the scattering events is approximated by the continuous slowing-down approximation discussed in Section 1.2, i.e. the energy of $e$ after traveling the distance $\Delta x$ is:

$$\epsilon' = \epsilon - \Delta xS(x, \epsilon).$$
In order to approximate the number density $n(x, \varepsilon)$ based on the trajectories of the electrons, a histogram $n_{kl}$ in position and energy can be used. Therefore $Q \times [\varepsilon_c, \varepsilon_b]$ is divided into $N_x \cdot N_\varepsilon$ bins

$$d x_k \times d \varepsilon_l \subset Q \times [\varepsilon_c, \varepsilon_b], \quad \bigcup_{k=1}^{N_x} \cup_{l=1}^{N_\varepsilon} (d x_k \times d \varepsilon_l) = Q \times [\varepsilon_c, \varepsilon_b] \quad (2.4)$$

Then with

$$t_{kl}(x, \Delta x, \varepsilon, \Omega), \quad k = 1, \ldots, N_x, \quad \text{and} \quad l \text{ with } \varepsilon \in d \varepsilon_l, \quad (2.5)$$

we denote the time that $e$ has spent inside the bin $d x_k \times d \varepsilon_l$ on its way from $x$ to $x + \Delta x \Omega$ taking into account the velocity related to its energy $\varepsilon$ through (1.25). After counting these time-weighted hits $n_{kl}$ for all trajectory segments of the trajectories of all electrons, we receive an electron number density

$$n(x, \varepsilon) \in L^2(Q \times [\varepsilon_c, \varepsilon_b]; \mathbb{R}_{\geq 0}),$$

$$n(x, \varepsilon) = n_{kl} \text{ for } x \in d x_k \text{ and } \varepsilon \in d \varepsilon_l. \quad (2.6)$$

A generic Monte Carlo simulation, that approximates this number density can now be formulated as follows:

**Algorithm 2** (Monte Carlo simulation).

Given a number $N_e \in \mathbb{N}$ of electrons to simulate, a measuring point $x_0 \in S$, a beam energy $\varepsilon_b > 0$, a cut-off-energy $0 < \varepsilon_c < \varepsilon_b$, a partition of the material into regions $\mathcal{T} = \{T_1, \ldots, T_{n_T}\}$ as in (1.22), with according densities $\rho_1, \ldots, \rho_{n_T}$ and mass concentrations $c_{11}, \ldots, c_{nT}$ of the occurring elements $El_1, \ldots, El_n$, as well as a histogram $n_{kl}$ with bins $d x_k \times d \varepsilon_l$ as in (2.4), compute:

**function** MONTECARLO($N_e, x_0, \varepsilon_b, \varepsilon_c, \mathcal{T}, \rho, c, El_j, n_{kl}, d x_k \times d \varepsilon_l$)

$$n_{kl} \leftarrow 0, \quad \text{for } k = 1, \ldots, N_x \text{ and } l = 1, \ldots, N_\varepsilon$$

**for** $i = 1, \ldots, N_e$ **do**

$$x \leftarrow x_0$$

$$\varepsilon \leftarrow \varepsilon_b$$
\[ \Omega \leftarrow (0, 0, -1)^T \]

**repeat**

1. \( T = T_i \) with \( x \in T_i \)
2. \( \Omega \leftarrow X(x, \epsilon, \Omega) \) with \( X \) as in (2.1)
3. \( \Delta x = \min \{ Y(x, \epsilon), \text{dist}_{\Omega}(x, \partial T) \} \) with \( Y \) as in (2.2)
4. \( x \leftarrow x + \Delta x \Omega \)
5. \( \epsilon \leftarrow \epsilon - \Delta x S(x, \epsilon) \)
6. \( n_{kl} \leftarrow n_{kl} + t_{kl}(x, \Delta x, \epsilon, \Omega) \), for \( k = 1, \ldots, N_x \) and \( l \) with \( \epsilon \in d\epsilon_l \)

**until** \( \epsilon < \epsilon_c \)

**end for**

**return** \( n \) as in (2.6)

**end function**

Combining this algorithm with the method to compute k-ratios from the electron number density, that was described at the end of Section 1.6, Monte Carlo methods yield a k-ratio prediction model of the form (1.21).

### 2.2 A constrained Levenberg-Marquardt algorithm

In this section we will present a mathematical algorithm that can be used to solve Problem 2 based on an according k-ratio prediction model. A detailed description of the derivation of the methods described here, as well as a theoretical discussion on their convergence, can be found in [9, pp. 213–224].

In its original form the Levenberg-Marquardt algorithm is not capable of enforcing the inequality constraint in Problem 2. Therefore we discuss a simple modification of the algorithm that enforces this constraint.

#### 2.2.1 Defining the target function

The Levenberg-Marquardt algorithm is designed to minimize the Euclidean norm \( \|F\|_2^2 = \sum_{i=1}^{M} F_i^2 \) of a vector-valued target function \( F : \mathbb{R}^N \rightarrow \mathbb{R}^M \), that takes a vector \( C \in \mathbb{R}^N \) as its arguments. In the following we introduce a notation that allows the reformulation of Problem 2 in terms of the
minimization of a vector-valued target function.

Let $\mathcal{T} = \{T_1, \ldots, T_{n_T}\}$ as in (1.22). Then for $\bar{c} \in \mathbb{R}^{0}_{\mathcal{T}}([0, 1])^{n-1}$ we define

\begin{align*}
c_{ij} &:= \bar{c}_{j|T_i} \text{ for } i = 1, \ldots, n_T \text{ and } j = 1, \ldots, n - 1, \\
c_{in} &:= 1 - \sum_{j=1}^{n-1} c_{ij} \text{ for } i = 1, \ldots, n_T.
\end{align*}

(2.7)

Now we vectorize the unknowns by defining

\begin{equation}
C_{(n-1)(i-1)+j} := c_{ij} \text{ for } i = 1, \ldots, n_T \text{ and } j = 1, \ldots, n - 1.
\end{equation}

(2.8)

Finally we define the following target function:

\begin{align*}
F : D &= \{ C \in [0, 1]^N \mid 1 - \sum_{j=1}^{n-1} c_{ij} \geq 0 \forall i \in \{1, \ldots, n_T\} \} \rightarrow \mathbb{R}^M, \\
F_{n(i-1)+j}(C) &= k_{ij}(C) - k_{ij}^{\exp}, \text{ for } i = 1, \ldots, m \text{ and } j = 1, \ldots, n,
\end{align*}

(2.9)

where $N := (n - 1)n_T$, $M := nm$ and $k_{ij}(C) = k_{ij}(\bar{c}_1, \ldots, \bar{c}_{n-1}; \bar{c}_n, \rho)$ in the sense of (2.7) and (2.8). For convenience we omit the density $\rho$ in this short notation.

By definition of $D$ we have included the inequality constraint of Problem 2 inside the target function’s valid domain. A short calculation yields:

\begin{equation}
\|F(C)\|_2^2 = \sum_{i=1}^{M} F_i(C)^2 = \sum_{i=1}^{n_T} \sum_{j=1}^{n} \left( k_{ij}(C) - k_{ij}^{\exp} \right)^2.
\end{equation}

Thus Problem 2 is equivalent to:

**Problem 4** (Non-linear least squares problem).

Let $F$ and $D$ as in (2.9). Then solve

\begin{equation}
\text{find } C^* \in D \text{ s.t. } \|F(C^*)\|_2 = \min_{C \in D} \|F(C)\|_2
\end{equation}

(2.10)
The rest of this section will be devoted to the solution of this Problem.

### 2.2.2 The classical Gauss-Newton method

The Levenberg-Marquardt algorithm has been developed on the basis of the classical Gauss-Newton method, which minimizes the Euclidean norm of non-linear target functions of the form $F: \mathbb{R}^N \rightarrow \mathbb{R}^M$. The restriction of $F$ to the domain $D$ as in Problem 4 will be discussed in Subsection 2.2.4. This algorithm heavily relies on the Jacobian matrix of the target function:

$$
F' : \mathbb{R}^N \rightarrow \mathbb{R}^{M \times N}, \quad F'(C) := \left( \frac{\partial}{\partial C_j} F_i(C) \right)_{i=1,\ldots,M, j=1,\ldots,N}.
$$

For $F'$ to be well defined all partial derivatives $\frac{\partial}{\partial C_j} F_i$ have to exist. Defining the target function on the basis of stochastic Monte Carlo simulations complicates the theoretical treatment of partial derivatives, because the target function is then a random variable instead of a well-defined function in the classical sense. However, we assume the existence of $F'$ for the rest of this section and treat this aspect in more detail in Section 2.3.

The linearization of $F$ is given by the first order Taylor polynomial, expanded around some point $\hat{C} \in \mathbb{R}^N$:

$$
F(C) = F(\hat{C}) + F'(\hat{C})(C - \hat{C}) + \mathcal{O}(\|C - \hat{C}\|^2). \quad (2.12)
$$

Using this we can linearize the extension of (2.10) to $\mathbb{R}^N$:

\[
\text{find } C^* \in \mathbb{R}^N \text{ that solves}
\]

\[
\min_{C \in \mathbb{R}^N} \|F'(\hat{C})(C - \hat{C}) + F(\hat{C})\|_2.
\]

(2.13)

Now the strategy is to start with an initial guess $C^0 \in \mathbb{R}^N$ and iteratively solve

\[
\text{find } S^k \in \mathbb{R}^N \text{ that solves}
\]

\[
\min_{S \in \mathbb{R}^N} \|F'(S^k)S + F(C^k)\|_2.
\]

(2.14)
Replacing \( \hat{C} \) by \( C^k \) in (2.13) gives \( S^k = C^* - C^k \). Thus the update is
\[
C^{k+1} = C^* = C^k + S^k
\]
Problem (2.14) is equivalent to the normal equations:
\[
solve \ A^T AS^k = A^T b, \quad \text{where} \quad A = F'(C^k) \quad \text{and} \quad b = -F(C^k),
\]
which can be solved using Gaussian elimination or more elaborate algorithms such as the Cholesky decomposition or QR decomposition.

This process is repeated until a stopping criterion is met. Common choices are
\[
\|C^k - C^{k-1}\|_2 \leq \text{eps}_C, \quad \|F'(C^k)^T F(C^k)\|_2 \leq \text{eps}_\nabla,
\]
with some thresholds \( \text{eps}_C, \text{eps}_\nabla > 0 \). The former criterion indicates that no significant change to the solution can be expected from further iterations. The latter implies that the necessary condition for a local optimum has approximately been met (see [9, p. 216]). This yields the following algorithm:

**Algorithm 3** (Gauss-Newton method).

Given an initial guess \( C^0 \in \mathbb{R}^N \), termination thresholds \( \text{eps}_C, \text{eps}_\nabla > 0 \) and a target function \( F \), compute:

```python
def GAUSSNEWTON(C0, F, epsC, epsN):
    k ← 0
    Ck ← C0
    repeat
        A = F'(Ck)
        b = -F(Ck)
        Sk ← solve \( A^T AS^k = A^T b \)
        Ck+1 = Ck + Sk
        k ← k + 1
    until \( \|C^k - C^{k-1}\|_2 \leq \text{eps}_C \) or \( \|F'(C^k)^T F(C^k)\|_2 \leq \text{eps}_\nabla \)
    return Ck
end function
```
2.2.3 The original Levenberg-Marquardt algorithm

In some cases where Algorithm 3 does not converge, the reason is that the step sizes $\|S^k\|_2$ are too large. Also note that (2.15) has a unique solution if and only if all columns of $A$ are linearly independent. Therefore it has been suggested [9, pp. 222–223] to replace (2.14) by:

$$\begin{align*}
\text{find } S_k^* \in \mathbb{R}^N \text{ that solves } \\
\min_{S \in \mathbb{R}^N} \| F'(C_k)S + F(C_k) \|^2_2 + \mu^2 \| S \|^2_2,
\end{align*}$$

(2.16)

where the parameter $\mu > 0$ penalizes large step sizes. We can rewrite (2.16) such that it is again of the form (2.14), which can be easily translated into the according normal equations:

$$\begin{align*}
\text{find } S_k^* \in \mathbb{R}^N \text{ that solves } \\
\min_{S \in \mathbb{R}^N} \left\| \begin{pmatrix} F'(C_k) & \mu I \end{pmatrix} s + \begin{pmatrix} F(C_k) \\ \emptyset \end{pmatrix} \right\|_2,
\end{align*}$$

(2.17)

where $I = \text{diag}(1, \ldots, 1) \in \mathbb{R}^{N \times N}$ denotes the identity matrix and $\emptyset$ indicates zero entries. Now for $\mu > 0$ the columns of $(F'(C_k)^T, \mu I)^T$ are always linearly independent.

A remaining question is the proper choice of $\mu$. In this work, we have used the strategy suggested in [9, p. 224]. There, the actual change in the residual is compared with the change in the residual of the linearized model:

$$\rho_{\mu}^k := \frac{\| F(C_k) \|^2_2 - \| F(C_k + S_k^*(\mu)) \|^2_2}{\| F(C_k) \|^2_2 - \| F'(C_k)S_k^*(\mu) + F(C_k) \|^2_2},$$

where $S_k = S_k(\mu)$ in the sense that the solution of (2.17) depends on $\mu$. If $\rho_{\mu}^k$ is too small, then the change in the residual is too low and (2.17) must be solved again, using a more restrictive choice of $\mu$. A large value of $\rho_{\mu}^k$ indicates a good convergence behavior that allows for a decrease of $\mu$. 

44
2.2.4 Enforcing the inequality constraint

Now we treat the restriction of the domain of \( F \) from \( \mathbb{R}^N \) to \( D \) as in (2.9). Therefore we enforce the inequality constraint of Problem 2. Note that there \( c \in \mathbb{R}_{+}^n(0,1)^{n-1} \) is part of the premise. Here this does not necessarily hold, as \( S^k \) might have negative entries. Thus we additionally enforce \( c^k_{ij} \geq 0 \) for \( i = 1, \ldots, n_T \) and \( j = 1, \ldots, n - 1 \).

To achieve this we calculate the step \( S^k \in \mathbb{R}^N \) as suggested in the original Levenberg-Marquardt algorithm, but before executing the update we contract \( S^k \) by \( \lambda \in [0,1] \) such that for \( i = 1, \ldots, n_T \):

\[
0 \leq C^k_{in} = 1 - \sum_{j=1}^{n-1} C^k_{ij} = 1 - \sum_{j=1}^{n-1} (C^k_{ij} + \lambda S^k_{ij}) = C^k_{in} - \lambda \sum_{j=1}^{n-1} S^k_{ij}, \tag{2.18}
\]

and

\[
C^k_{ij} = C^k_{ij} + \lambda S^k_{ij} \geq 0 \quad \text{for} \quad j = 1, \ldots, n - 1, \tag{2.19}
\]

where \( S^k_{ij} \) is related to \( S^k \) through analogous definitions as were introduced for \( c^k_{ij} \) and \( C \) in (2.7) and (2.8).

**Lemma 1** (Sufficient conditions for the damping parameter).

Let \( S^k_{ij} \in \mathbb{R}^n \) arbitrary and \( c^k_{ij} \geq 0 \) for \( i = 1, \ldots, n_T, \ j = 1, \ldots, n - 1 \) and \( C^k_{in} = 1 - \sum_{j=1}^{n-1} C^k_{ij} \geq 0 \) for \( i = 1, \ldots, n_T \). Set \( C^{k+1}_{ij} := C^k_{ij} + \lambda S^k_{ij} \) for \( i = 1, \ldots, n_T, \ j = 1, \ldots, n - 1 \) and \( C^{k+1}_{in} := 1 - \sum_{j=1}^{n-1} C^{k+1}_{ij} \) for \( i = 1, \ldots, n_T \) with

\[
\lambda = \min\{1, \lambda_1, \lambda_2\}, \tag{2.20}
\]

\[
\lambda_1 = \min\{-c^k_{ij}/S^k_{ij} | S^k_{ij} < 0, \ i = 1, \ldots, n_T, \ j = 1, \ldots, n - 1\}, \tag{2.21}
\]

\[
\lambda_2 = \min\{C^k_{in}/\sum_{j=1}^{n-1} S^k_{ij} | \sum_{j=1}^{n-1} S^k_{ij} > 0, \ i = 1, \ldots, n_T\}. \tag{2.22}
\]

Then \( \lambda \in [0,1] \) and the inequations (2.18) and (2.19) hold.

**Proof.** 1) We prove that \( \lambda_1, \lambda_2 \geq 0 \) and therefore (2.20) implies \( \lambda \in [0,1] \).

From the premise we have \( c^k_{ij} \geq 0 \) for all \( i \in \{1, \ldots, n_T\} \) and \( j \in \{1, \ldots, n - 1\} \). Set

\[
\lambda_1 = \min\{-c^k_{ij}/S^k_{ij} | S^k_{ij} < 0, \ i = 1, \ldots, n_T, \ j = 1, \ldots, n - 1\}, \tag{2.21}
\]

and

\[
\lambda_2 = \min\{C^k_{in}/\sum_{j=1}^{n-1} S^k_{ij} | \sum_{j=1}^{n-1} S^k_{ij} > 0, \ i = 1, \ldots, n_T\}. \tag{2.22}
\]

Then \( \lambda \in [0,1] \) and the inequations (2.18) and (2.19) hold.
1. Therefore $-c_{ij}^k/s_{ij}^k \geq 0$ for all $i \in \{1, \ldots, n_T\}$ and $j \in \{1, \ldots, n-1\}$ with $s_{ij}^k < 0$, which implies $\lambda_1 \geq 0$. The premise also states that $c_{in}^k \geq 0$ for all $i \in \{1, \ldots, n_T\}$. Hence $c_{in}^k/\sum_{j=1}^{n-1} s_{ij}^k \geq 0$ for all $i \in \{1, \ldots, n_T\}$ with $\sum_{j=1}^{n-1} s_{ij}^k > 0$, which implies $\lambda_2 \geq 0$.

2) We prove that (2.21) implies (2.19) for $\lambda \geq 0$.
Let $i \in \{1, \ldots, n_T\}$, $j \in \{1, \ldots, n-1\}$ arbitrary. From 1) we have that $\lambda \geq 0$. Assume $s_{ij}^k \geq 0$. This directly implies $c_{ij}^{k+1} = c_{ij}^k + \lambda s_{ij}^k \geq 0$ since $\lambda \geq 0$ and $c_{ij}^k \geq 0$ as stated in the premise. Now let $s_{ij}^k < 0$. Then (2.21) implies $\lambda \leq \lambda_1 \leq -c_{ij}^k/s_{ij}^k$. Thus $c_{ij}^k/s_{ij}^k + \lambda \leq 0$. Due to its negativity, multiplication with $s_{ij}^k$ yields $c_{ij}^k + \lambda s_{ij}^k = c_{ij}^{k+1} \geq 0$. Since $i \in \{1, \ldots, n_T\}$ and $j \in \{1, \ldots, n-1\}$ were arbitrary this gives (2.19).

3) We prove that (2.22) implies (2.18) for $\lambda \geq 0$.
Let $i \in \{1, \ldots, n_T\}$ arbitrary. Again from 1) we have that $\lambda \geq 0$. Assume $\sum_{j=1}^{n-1} s_{ij}^k \leq 0$. Then $c_{ij}^{k+1} = 1 - \sum_{j=1}^{n-1} c_{ij}^{k+1} = 1 - \sum_{j=1}^{n-1} c_{ij}^k - \lambda \sum_{j=1}^{n-1} s_{ij}^k = c_{in}^k - \lambda \sum_{j=1}^{n-1} s_{ij}^k \geq 0$ since $\lambda \geq 0$ and $c_{in}^k \geq 0$ as stated in the premise. Now let $\sum_{j=1}^{n-1} s_{ij}^k > 0$. Then (2.22) implies $\lambda \leq \lambda_2 \leq c_{in}^k/\sum_{j=1}^{n-1} s_{ij}^k$. Hence $c_{in}^k - \lambda \sum_{j=1}^{n-1} s_{ij}^k \geq 0$. Thus $c_{in}^{k+1} \geq 0$ because $c_{in}^{k+1} = c_{in}^k - \lambda \sum_{j=1}^{n-1}$ as calculated before. Since $i \in \{1, \ldots, n_T\}$ was arbitrary this gives (2.18) and concludes this proof.

Finally we state a constrained Levenberg-Marquardt algorithm, based on Algorithm 3. The normal equations have been modified according to (2.17), an adaptive strategy for the choice of the parameter $\mu$ is implemented and the inequality constraint of Problem 2 is enforced using the sufficient choice of a contraction parameter, as stated by Lemma 1.

Algorithm 4 (Levenberg-Marquardt with enforced inequality constraints).
Given an initial guess $C^0 \in \mathbb{R}^N$, an initial choice $\mu^0 > 0$, thresholds $\rho_0, \rho_1 > 0$ for its adaptive choice, termination thresholds $\epsilon_{SC}, \epsilon_{SV} > 0$ and a target function $F$ as in (2.9), compute:

function ConstrainedLevenbergMarquardt($C_0, F, \epsilon_{SC}, \epsilon_{SV}, \mu^0, \rho_0, \rho_1$)

$k \leftarrow 0$
2.3 Estimating partial derivatives

In this section we discuss how the Jacobian $F'$ in (2.11) can be approximated in the case that the target function $F$ is based on Monte Carlo simulations.

The increase $\text{Inc}(\mu)$ and decrease $\text{Dec}(\mu)$ of $\mu$ can, for example, be implemented as the multiplication of $\mu$ with 10 or 0.1 respectively.
2.3.1 Numerical differentiation using finite differences

In our approach we approximate all partial derivatives
\[ \frac{\partial}{\partial C_k} F_i(C), \ i = 1, \ldots, M, \ k = 1, \ldots, N. \]
individually. This reduces the problem of approximating the multi-dimensional Jacobian to the problem of approximating several derivatives of one-dimensional scalar functions. To clarify this, we introduce the following notation

\[ f_{ik}^C(x) : [0, 1 - \sum_{j=1}^{n} c_{ij}] \to \mathbb{R}_{\geq 0}, \]
\[ f_{ik}^C(x) := F_i(C_1, \ldots, C_{k-1}, x, C_{k+1}, \ldots, C_N), \ i = 1, \ldots, M, \ k = 1, \ldots, N, \]

where \( C \in D \) is arbitrary but fixed. This implies

\[ \frac{\partial}{\partial C_k} F_i(C) = \left. \frac{d}{dx} f_{ik}^C(x) \right|_{x = C_k}. \tag{2.23} \]

The following discussion will be independent of \( C \in D \) and \( i \in \{1, \ldots, M\} \) and \( k \in \{1, \ldots, N\} \) and we will therefore omit these to simplify the notation.

Let \( h > 0 \). Then analogous to (2.12), evaluating the first order Taylor polynomial of \( f_{ik}^C \) expanded around \( x \in D \) at \( x + h \) gives:

\[ f(x + h) = f(x) + h \frac{d}{dx} f(x) + \mathcal{O}(h^2). \tag{2.24} \]

Thus the partial derivative (2.23) can be approximated with first order accuracy by the finite difference quotient:

\[ \frac{d}{dx} f(x) \approx \frac{f(x + h) - f(x)}{h} \tag{2.25} \]

Equation (2.24) states that the expression (2.25) formally converges as \( h \to 0 \). In practice this will however not be the case, due to the fact that on a machine the nominator (as well as the denominator) can only be represented with finite precision. Therefore truncation errors in the nominator
get amplified with decreasing step size $h$ and corrupt the accuracy of the
finite difference approximation. In [9, pp. 290-292] an optimal bound for
the step size $h$ based on the expected error in the evaluation of $f$ and the
magnitude of $f''(x)$ is discussed.

### 2.3.2 Influence of statistical fluctuations on finite differences

Unfortunately in our case $f$ is not only subject to rounding errors, but
also influenced by statistical fluctuations which introduce errors that are
magnitudes larger than the machine precision. In [27, 28] a theoretical frame-
work for the analysis of such statistical noise has been developed. In the
following we briefly summarize the main ideas of this concept.

We can express $f$ as

$$f(x) = f_s(x) + \epsilon(x),$$

where $f_s$ is a smooth function that represents the model underlying $f$ and $\epsilon$
is a random variable representing the statistical fluctuations in $f$. Then

$$\epsilon_f := \sqrt{\text{Var}(\epsilon(x))}$$

is the noise level of $f$, where Var denotes the variance. The least squares
error in the finite difference, with respect to the underlying derivative $f'_s$,
based on the step size $h$, is given as:

$$\mathcal{E}(h) = \left( \frac{f(x+h) - f(x)}{h} - f'_s(x) \right)^2.$$  

Based on the minimization of the expected error $E\{\mathcal{E}(h)\}$ the following
nearly optimal choice of $h$ is suggested in [28]:

$$h^* = 8^{1/4} \left( \frac{\epsilon_f}{\mu} \right)^{1/2},$$  \hspace{1cm} (2.26)$$

where $\mu \approx |f''(x)|$ is an approximation of the second derivative of $f$. To ob-
tain this, another (second order finite difference) approximation is necessary,
which is subject to the same errors discussed before, but has less influence on the accuracy of (2.25). In [28] a heuristic is given that automatically chooses $\mu$ based on a few evaluations of $f$.

In Monte Carlo simulations the noise level $\epsilon_f$ depends mainly on the number of simulated electrons. In some codes (e.g. Penelope [36]) a threshold for $\epsilon_f$ can be given instead of the total number of simulated electrons. For other programs, where this feature is not available, the ECNoise [27] algorithm can be used. It automatically estimates $\epsilon_f$ based on a few evaluations of $f$.

To keep the computational cost caused by these extra evaluations of $f$ low, we suggest the estimation of $h^*$ at the initial guess $x_0$ in a precalculation step before initializing Algorithm 4. Then the step size $h$ is kept constant throughout the optimization. This is motivated by the fact that the statistical fluctuations in Monte Carlo simulations, that use analytical models like (1.45) to calculate the k-ratios based on the electron number density, only affect the calculation of the electron number density. Though this quantity depends on the material parameters, they should have a negligible influence on its noise level.

### 2.4 Numerical results

In this section we present several test cases where Algorithm 4 has been used to optimize the material parameters of both analytically parametrized and meshed geometries. Note that in all cases simulated k-ratios based on a reference geometry were used instead of experimentally acquired ones. Thereby we have neglected the model error between the used Monte Carlo simulations and real experiments. In return we can precisely analyze the convergence behavior of Algorithm 4, as the optimal solution is exactly known.

To enable the following tests, the program pyMonteCarlo\(^1\) [30] was used. It enables access to several freely available Monte Carlo codes from a Python

---

\(^1\)P. T. Pinard, pyMonteCarlo, http://pymontecarlo.bitbucket.org (Sept. 2013)
command line. During this work it was extended by program packages including implementations of Algorithm 4, the ECNoise [27] algorithm and the heuristic for the choice of $\mu$ in (2.26) that is suggested in [28]. To assess the correctness of these implementations, we first look at a simple test case of a homogeneous compound, where the convergence of the reconstruction algorithm is expected.

2.4.1 Analysis of a homogeneous compound

Let us regard a homogeneous material sample consisting of the elements chromium, iron and copper, i.e. $El_1 = \text{Cr}, El_2 = \text{Fe}, El_3 = \text{Cu}$. The unknown material parameters are given by the mass fractions $c_1, c_2 \in [0, 1]$ of chromium and iron. The mass concentration of copper is given by $1 - c_1 - c_2$. We have chosen $C^{\text{ref}} = (0.33, 0.33)^T$ as the reference solution, i.e. $k_{i1}^{\text{exp}} := k_{i1}(C^{\text{ref}})$ for $i = 1, 2, 3$. For each parameter configuration we have simulated only a single measurement using the DTSA – II [35] Monte Carlo program, $N_e = 5000$ electrons and an initial beam energy of 10 keV.

To assess the feasibility of the reconstruction, we have visualized the squared residual of the target function, $\|F(c_1, c_2)\|^2_2$, for $c_1, c_2 \in [0, 0.5] \times [0, 0.5]$. The plot in Figure 2.1 shows a smooth convex geometry. Hence Algorithm 4 should converge to the reference solution for any initial guess in $[0, 0.5] \times [0, 0.5]$.

We have performed the precalculation discussed in Section 2.3 to obtain a stable step size for the finite difference approximations (2.25). Using $x^0 = (0.1, 0.5)^T$, $n_f = 7$ and $h_0 = 1 \times 10^{-2}$, our implementation of the ECNoise [27] algorithm estimated a noise level of $\epsilon_f = 1.163 \times 10^{-5}$ and an optimal step size of $h^* = 0.147$. 

51
Finally we have initialized Algorithm 4 with

\[ C^0 = \begin{pmatrix} 0.1 \\ 0.5 \end{pmatrix}, \quad \text{eps}_{C} = 1 \times 10^{-8}, \quad \text{eps}_{\nabla} = 1 \times 10^{-8} \]

\[ \mu^0 = 1 \times 10^{-3}, \quad \rho_0 = 0.2, \quad \rho_1 = 0.8. \]  

Due to the knowledge of the optimal parameter choice, we have implemented the additional stopping criterion \( \|C^k - C_{\text{ref}}\|_2 / \|C_{\text{ref}}\|_2 < 0.01 \). Based on this criterion Algorithm 4 terminated after two iterations and returned \( C^2 = (0.33091809, 33085345)^T \).
2.4.2 Analysis of a half-spherical particle of unknown composition and diameter

After the first successful test we have tried to reconstruct a half-spherical particle embedded in a known substrate, again using only a single measurement. The center of the particle was fixed at the position $x_0 \in S$ of the measurement and it was parametrized by its diameter. We chose the particle to be a chromium-carbide and the substrate to be pure iron, i.e. $E_{l_1} = C, E_{l_2} = Cr, E_{l_3} = Fe$. The unknown material parameters are the diameter $d$ of the particle and the mass fraction $c$ of carbon inside the particle. We have chosen $(d^{\text{ref}}, c^{\text{ref}})^T = (200 \text{ nm}, 0.05)^T$ as the reference configuration. An illustration of the reference particle can be viewed in Figure 2.2. Simulated electron trajectories in the background of the image indicate the ratio between the size of the analyzed particle and the size of the electron interaction volume.

Note that the parameter $d$ does not fit within the framework presented in Section 2.2, because here we have a relation $k_{i1} = k_{i1}(c, d)$ for $i = 1, 2, 3$. Therefore in this case we have altered the constraints in Algorithm 4 such that $c \in [0, 1]$ and $d \geq 50 \text{ nm}$ holds in each iteration. Then we have initialized Algorithm 4 using $(d^0, c^0) = (400 \text{ nm}, 0.6)$. Again the Levenberg-Marquardt algorithm terminated after two iterations and returned $(d^2, c^2) = (199 \text{ nm}, 0.05)$. The optimization took approximately 300 seconds on an Intel Core 2 Duo processor with 2.53 GHz.

2.4.3 Analysis of a thin layer of unknown composition

Furthermore we have tested the reconstruction of a thin nickel-chromium layer using a meshed geometry. Therefore we have specified a small grid $\mathcal{T} = \{T_1, T_2, T_3, T_4\}$ of four cells, each 50 nm wide and infinitely extending in the $y$ and $z$ directions (see Figure 2.3). Each cell $T_j$ consists of a homogeneous compound of nickel and chromium with mass fractions $1 - c_j$ and $c_j$ respectively. They are surrounded by a pure nickel substrate. We have chosen a $100 \text{ nm}$ thick layer with $80\text{wt\%}$ chromium and $20\text{wt\%}$ nickel embedded
Figure 2.2: Schematical illustration of a measurement of a chromium-carbide particle in an iron matrix in a pure nickel substrate as the reference material. In the meshed geometry this corresponds to the reference parameters $C^\text{ref} = (0.0, 0.8, 0.8, 0.0)$. X-ray intensities were acquired from five simulated measurements at the positions $x = -120\,\text{nm}, -60\,\text{nm}, 0\,\text{nm}, 60\,\text{nm}, 120\,\text{nm}$, each with a beam energy of 10 keV and a beam diameter of 50 nm. In each simulation 5000 electrons were traced using the \textsc{dist} -- \textsc{ii} Monte Carlo program.

For the reconstruction, Algorithm 4 was performed with an initial guess of $C^0 = (0.5, 0.5, 0.5, 0.5)$. It terminated after five iterations due to $\|C^5 - C^4\|_2 \leq 10^{-8}$ and returned $C^5 = (0.0, 0.812, 0.769, 0.022)$. This yields a relative error $\|C^5 - C^\text{ref}\|_2/\|C^\text{ref}\|_2$ of 3.51%. 

Electron beam:
- energy: 10 keV;
- diameter: 50 nm;
- # of electrons: 5000

Sample surface

100 nm

Substrate: Fe

Cr$_{1-c}C_c$
2.4. NUMERICAL RESULTS

Electron beams: energy: 10 keV;
diameter: 50 nm;
# of electrons: 5000

Figure 2.3: Schematic illustration of several measurements on thin nickel-chromium layers

Ni $\rightarrow$ Ni$_{1-c_1}$Cr$_{c_1}$ $\rightarrow$ Ni$_{1-c_2}$Cr$_{c_2}$ $\rightarrow$ Ni$_{1-c_3}$Cr$_{c_3}$ $\rightarrow$ Ni$_{1-c_4}$Cr$_{c_4}$ $\rightarrow$ Ni

$T_1$ $\rightarrow$ $T_2$ $\rightarrow$ $T_3$ $\rightarrow$ $T_4$ $\rightarrow$

100 nm
Chapter 3

A deterministic model for electron transport

3.1 The Boltzmann equation in continuous slowing-down approximation

In Section 1.2 we have seen a continuous slowing-down approximation for the energy loss of the electrons and in Section 2.1 we have seen how this concept is used in Monte Carlo simulations. A similar approach can be made to simplify the Boltzmann equation (1.41).

Therefore an asymptotic analysis of the term

$$N_V(x) \int_\epsilon^\infty \int_{S^2} \sigma_{\text{in}}(x, \epsilon', \epsilon, \Omega' \cdot \Omega) \psi(x, \epsilon', \Omega') d\Omega' d\epsilon'$$

with respect to the energy is carried out. The asymptotic analysis is based on the assumption that the inelastic scattering cross-section is peaked about small energy losses. A detailed description of such an asymptotic analysis with respect to both the energy and angle, which leads to the Fokker-Planck equation, can be found in [31]. Because this analysis is very technical, we simply state the equation that results in the case of non negligible contributions by large-angle scattering events here, namely the Boltzmann equation.
in continuous slowing-down approximation (BCSD) [23, 10, 13]:

\[
\Omega \cdot \nabla_x \psi(x, \epsilon, \Omega) = N_V(x) \int_{S^2} \sigma_{in}^{CSD}(x, \epsilon, \Omega' \cdot \Omega) \psi(x, \epsilon, \Omega') \, d\Omega' \\
+ \frac{\partial}{\partial \epsilon} (S(x, \epsilon) \psi(x, \epsilon, \Omega)) \\
+ N_V(x) \int_{S^2} \sigma_{el}(x, \epsilon, \Omega' \cdot \Omega) \psi(x, \epsilon, \Omega') \, d\Omega' \\
- N_V(x) \sigma_{tot}^{el}(x, \epsilon) \psi(x, \epsilon, \Omega) \\
- N_V(x) \sigma_{tot}^{el}(x, \epsilon) \psi(x, \epsilon, \Omega),
\]

(3.1)

where

\[
\sigma_{in}^{CSD}(x, \epsilon, \mu) := \int_0^\infty \sigma_{in}(x, \epsilon, \epsilon', \mu) \, d\epsilon'.
\]

(3.2)

Like (1.41) this equation describes the spatial movement of the electrons by a transport term \( \Omega \cdot \nabla_x \psi \) and the scattering of the electrons to different energies and directions of travel by elastic and inelastic scattering integrals. Comparing with (1.41) we see that the electron fluence \( \psi \) does no longer depend on the integration variable \( \epsilon' \) in the inelastic in-scattering integral. This will be important for the method we discuss in the next section. Opposed to (1.41) there appears an additional term \( \frac{\partial}{\partial \epsilon} (S \psi) \) in (3.1), that has a derivative with respect to the energy. On the left-hand side of the equation it would have a negative sign and could be interpreted as a transfer from higher to lower energies within the energy space, with the stopping power governing the speed of this transfer.

Due to the high dimensionality and the remaining integral terms, the deterministic numerical solution of (3.1) is associated with high computational cost. However, for the computation of (1.43) only the average in angle

\[
\int_{S^2} \psi(x, \epsilon, \Omega') \, d\Omega'
\]

(3.3)

of the solution is necessary. This motivates further simplifications of (3.1) with respect to the angular variable \( \Omega \).
3.2 The method of moments

In this section we will discuss how a finite set of coupled partial differential equations can be derived through the method of moment approximation [26] of the Boltzmann equation in continuous slowing-down approximation. We will first introduce the general mathematical framework. Then we will discuss the properties and validity of a minimum entropy closure in our context. Finally we derive a finite set of partial differential equations, the \( M1 \) model [10], that approximate the original solution of (3.1) and can be solved directly using standard numerical methods.

3.2.1 Expansion in one of the independent variables

The method of moments is a mathematical concept for the reduction of the dimensions of the phase space in differential equations. Therefore the equation is expanded in one or more of its independent variables. In consideration of (3.3), we discuss the expansion with respect to the angular variable \( \Omega \) here.

Let

\[
\Omega^n := \Omega \otimes \cdots \otimes \Omega,
\]

(3.4)

where \( \otimes \) denotes the tensor product. Then \( \Omega^n \in \mathbb{R}^{3 \times \cdots \times 3} \) is a tensor of rank \( n \). Using this notation we define the \( n \)-th angular moment of \( \psi \):

\[
\psi^{(n)}(x, \epsilon) := \int_{S^2} \Omega^n \psi(x, \epsilon, \Omega) \, d\Omega.
\]

(3.5)

Note that \( \psi^{(n)}(x, \epsilon) \) is also a tensor of rank \( n \).

Let \( m_n(\Omega) \) denote the vector of all linearly independent entries of the tensors \( \Omega^0, \Omega^1, \ldots, \Omega^n \). Then it can be shown that for arbitrary \( n \in \mathbb{N}_0 \) multiplication of (3.1) with \( m_n(\Omega) \) and component-wise integration over the unit sphere yields a partial differential equation that contains the \((n + 1)\)st mo-
ment on the left-hand side and lower order moments on the right-hand side [26]. It is called the $n^{th}$ order moment system. In Subsection 3.3 we will prove this for $n \in \{0, 1\}$ and give explicit expressions for the resulting partial differential equations.

Note that these moment systems are underdetermined for arbitrary order $n \in \mathbb{N}_0$, because the linearly independent entries of $\psi^{(n+1)}$ appear only on the left-hand side of the last equation. To obtain a closed system of partial differential equations, that can be solved using standard numerical methods, a closure of the form

$$
\psi^{(n+1)} = \psi^{(n+1)}(\psi^{(0)}, \ldots, \psi^{(n)}),
$$

(3.6)

has to be defined for the highest order moment.

### 3.2.2 The minimum entropy closure

In this work we will follow the approach taken in [10] and design the closure (3.6) based on a minimum entropy principle.

By the second law of thermodynamics an isolated system in thermodynamic equilibrium is in the state of maximum entropy. In the context of the Boltzmann equation for electron transport the physical entropy can be interpreted as a measure for the number of possible different configurations of all individual electrons inside the entire system that represent the same macroscopic density $\psi(x, \epsilon, \Omega)$. In this sense, in thermodynamic equilibrium, the system (1.41) will take its most probable state.

**Mathematical formulation of the minimum entropy closure**

In general the Boltzmann entropy for a one particle system, described by the density $W$ in some phase-space of infinitesimal volumes $d\tau$ is given as (cf. [21]):

$$
H_B = \int W \log W \, d\tau.
$$
Based on this we define the following entropy of the angular distribution of the electrons [7, 10]:

\[ H(\psi) := - \int_{S^2} \psi \log \psi \, d\Omega. \] (3.7)

Note that the mathematical entropy is defined as the negative physical entropy here. We define the closure for \( \psi^{(n+1)} \) based on the distribution \( \psi_{ME} \) which minimizes (3.7) while representing the lower order moments:

\[
H(\psi_{ME}) = \min_{\psi} H(\psi),
\]

\[
s.t. \int_{S^2} \Omega^i \psi_{ME} \, d\Omega = \psi^{(i)}, \quad i = 0, \ldots, n. \] (3.8)

In Section 3.3 we will derive the relation (3.6) that results in the case \( n = 1 \).

**Validity of the minimum entropy closure**

In our case the assumption of a thermodynamic equilibrium for the dynamics of the electrons is not realistic, because we want to observe them from the moment they enter the specimen, where they have a well-structured configuration enforced by the electron beam and are therefore not in thermodynamic equilibrium. To assess the validity of the maximum entropy closure in our application, we have simulated electron trajectories using the DTSA – II [35] Monte Carlo program and computed histograms similar to those discussed in Section 2.1, only for the distribution in angle. The result is shown in Figure 3.1. Each cell of the spatial grid (black boxes) represents a histogram for the angular distribution of the electrons at that position. The weighted counts of the electron flight directions in each bin of a particular histogram are visualized as the length of a white line from the black center point in the direction corresponding to the angular bin. The lengths have been locally normalized. In the background of the image the mean energy distribution of the electrons is plotted for a better orientation. We observe a concentration of downward directions in the histograms near the entry point of the electron beam. Outside the highest energy region (red background) there is still a dominant direction in each histogram, but
the weighted counts decrease smoothly towards other angles. In Figure 3.2 the weighted counts of the histogram at \((x, z) = (0, -175 \text{ nm})\) are plotted against the corresponding angle between the direction of travel and the \(x\)-axis (red dots). They are compared with the simple and smooth function \(f(\phi) = 1.6 \times 10^{-13} e^{-(|\phi - \frac{\pi}{2}| - \pi)^2} + 10^{-13}\) (blue line). The comparison shows that in this example it is possible to represent the angular distribution based on few information. If the angular distributions were rich in details, that require much information to be represented, this would indicate a significant deviation from the minimum entropy state. The absence of such details in this example motivates the choice of a minimum entropy principle for the closure relation.

### 3.3 The \(M1\) model for electron transport

In this section we apply the framework discussed in Section 3.2 to the Boltzmann equation in continuous slowing-down approximation (3.1) for \(n =\)
3.3. THE M1 MODEL FOR ELECTRON TRANSPORT

Figure 3.2: Angular distribution of the electrons at \((x, z) = (0, -175 \text{ nm})\): weighted counts (red dots), \(f(\phi) = 1.6 \times 10^{-13} e^{-|\phi - \frac{\pi}{2}| - \pi} + 10^{-13}\) (blue line)

1, i.e. we derive partial differential equations for the zeroth and first order moments in angle of \(\psi\).

3.3.1 Derivation of the partial differential equations

As mentioned in Section 3.2, we will now calculate the first order moment system corresponding to (3.1). Note that \(m_1(\Omega) = (1, \Omega)^T\). Thus we will first integrate (3.1) over the unit sphere, receiving a partial differential equation of the form:

\[
\nabla_x \cdot \psi^{(1)}(x, \epsilon) = F_0(\psi^{(0)}(x, \epsilon), x, \epsilon).
\]

Then we integrate the product between \(\Omega\) and (3.1) over the unit sphere, which gives a partial differential equation of the form:

\[
\nabla_x \cdot \psi^{(2)}(x, \epsilon) = F_0(\psi^{(1)}(x, \epsilon), \psi^{(0)}(x, \epsilon), x, \epsilon).
\]
Partial differential equation for the first moment

In the following we will integrate the terms in (3.1) individually (with respect to $\Omega$). For the divergence term we receive:

$$\int_{S^2} \Omega \cdot \nabla_x \psi(x, \epsilon, \Omega) \, d\Omega = \nabla_x \cdot \int_{S^2} \Omega \psi(x, \epsilon, \Omega) \, d\Omega = \nabla_x \cdot \psi^{(1)}(x, \epsilon).$$

(3.9)

Integration of the stopping power term yields:

$$\int_{S^2} \frac{\partial}{\partial \epsilon} (S(x, \epsilon) \psi(x, \epsilon, \Omega)) \, d\Omega = \frac{\partial}{\partial \epsilon} \left( S(x, \epsilon) \int_{S^2} \psi(x, \epsilon, \Omega) \, d\Omega \right) = \frac{\partial}{\partial \epsilon} \left( S(x, \epsilon) \psi^{(0)}(x, \epsilon) \right).$$

(3.10)

For the total cross section terms we receive:

$$- \int_{S^2} N_V(x) \sigma_{\text{tot}}^{\text{in/el}}(x, \epsilon) \psi(x, \epsilon, \Omega) \, d\Omega$$

$$= - N_V(x) \sigma_{\text{tot}}^{\text{in/el}}(x, \epsilon) \int_{S^2} \psi(x, \epsilon, \Omega) \, d\Omega$$

$$= - N_V(x) \sigma_{\text{tot}}^{\text{in/el}}(x, \epsilon) \psi^{(0)}(x, \epsilon).$$

(3.11)

The integral over the elastic in-scattering term yields the same expression with positive sign:

$$\int_{S^2} N_V(x) \int_{S^2} \sigma_{\text{el}}(x, \epsilon, \Omega') \psi(x, \epsilon, \Omega, \Omega') \, d\Omega' \, d\Omega$$

$$= N_V(x) \int_{S^2} \int_{S^2} \sigma_{\text{el}}(x, \epsilon, \Omega') \, d\Omega' \, \psi(x, \epsilon, \Omega) \, d\Omega$$

$$= N_V(x) \sigma_{\text{el}}^{\text{tot}}(x, \epsilon) \int_{S^2} \psi(x, \epsilon, \Omega) \, d\Omega$$

$$= N_V(x) \sigma_{\text{el}}^{\text{tot}}(x, \epsilon) \psi^{(0)}(x, \epsilon).$$

(3.12)
We receive the same result for the inelastic in-scattering term:

\[
\int_{S^2} N_V(x) \int_{S^2} \sigma_{in}^{CSD}(x,\epsilon,\Omega'\cdot\Omega) \psi(x,\epsilon,\Omega') d\Omega' d\Omega
= \int_{S^2} N_V(x) \int_{S^2} \int_{0}^{\infty} \sigma_{in}(x,\epsilon,\epsilon',\Omega'\cdot\Omega) d\epsilon' d\Omega \psi(x,\epsilon,\Omega') d\Omega'
= N_V(x) \sigma_{in}^{tot}(x,\epsilon) \int_{S^2} \psi(x,\epsilon,\Omega) d\Omega
= N_V(x) \sigma_{in}^{tot}(x,\epsilon) \psi^{(0)}(x,\epsilon).
\] (3.13)

The terms (3.12) and (3.13) cancel against the terms in (3.11). Thus only the terms (3.9) and (3.10) remain and we receive the following partial differential equation for the first moment:

\[
\nabla_x \cdot \psi^{(1)}(x,\epsilon) = \frac{\partial}{\partial \epsilon} \left( S(x,\epsilon) \psi^{(0)}(x,\epsilon) \right).
\] (3.14)

**Partial differential equation for the second moment**

Now we multiply (3.1) by \( \Omega \) and then integrate over \( \Omega \). For the divergence term and the stopping power the calculations are analogous to the above:

\[
\int_{S^2} (\Omega \cdot \nabla_x \psi(x,\epsilon,\Omega) d\Omega = \nabla_x \cdot \int_{S^2} (\Omega \cdot \psi(x,\epsilon,\Omega) d\Omega
= \nabla_x \cdot \psi^{(2)}(x,\epsilon),
\] (3.15)

and

\[
\int_{S^2} \Omega \frac{\partial}{\partial \epsilon} \left( S(x,\epsilon) \psi(x,\epsilon,\Omega) \right) d\Omega = \frac{\partial}{\partial \epsilon} \left( S(x,\epsilon) \int_{S^2} \Omega \psi(x,\epsilon,\Omega) d\Omega \right)
= \frac{\partial}{\partial \epsilon} \left( S(x,\epsilon) \psi^{(1)}(x,\epsilon) \right).
\] (3.16)

To handle the in-scattering terms, we need the following result.

**Lemma 2.**
The following equality holds:

\[
\int_{S^2} \int_{S^2} \Omega \sigma(x, \epsilon, \Omega') \psi(x, \epsilon, \Omega') \, d\Omega \, d\Omega' = 2\pi \int_{-1}^{1} \mu \sigma(x, \epsilon, \mu) \, d\mu \int_{S^2} \Omega \psi(x, \epsilon, \Omega) \, d\Omega.
\] (3.17)

**Proof.** First we rewrite the initial expression as:

\[
\int_{S^2} \int_{S^2} \Omega \sigma(x, \epsilon, \Omega') \psi(x, \epsilon, \Omega') \, d\Omega \, d\Omega' = \int_{S^2} \int_{S^2} \Omega \sigma(x, \epsilon, \Omega') \, d\Omega \psi(x, \epsilon, \Omega') \, d\Omega'.
\]

Note that the innermost integrand

\[\Omega \sigma(x, \epsilon, \Omega') \cdot \Omega,\]

is a product of \(\Omega\) and a term depending only on the scalar product of \(\Omega \cdot \Omega'\). Thus averaging over all directions \(\Omega \in S^2\) yields a vector parallel to \(\Omega'\):

\[\int_{S^2} \Omega \sigma(x, \epsilon, \Omega') \, d\Omega = \gamma \Omega',\]

with some factor \(\gamma \in \mathbb{R}\). Since \(\Omega' \in S^2\) has unit length, we can calculate \(\gamma\) by multiplying both sides with \(\Omega'\):

\[\gamma = \gamma \Omega' \cdot \Omega' = \int_{S^2} \Omega \cdot \Omega' \sigma(x, \epsilon, \Omega \cdot \Omega') \, d\Omega\]

In order to simplify the right-hand side we perform a transformation to spherical coordinates:

\[\Omega = \begin{pmatrix} \sin \theta \cos \varphi \\ \sin \theta \sin \varphi \\ \cos \theta \end{pmatrix} \quad \text{and} \quad d\Omega = \sin \theta d\theta d\varphi.\]

Also note that in the inner integral over \(\Omega\) we can consider \(\Omega' \in S^2\) as an arbitrary but fixed direction. Using a coordinate transformation we can
achieve $\Omega' = (0, 0, -1)^T$ while integrating over $\Omega$ and afterwards transform back for the outer integration over $\Omega'$. This implies $\Omega \cdot \Omega' = -\cos \theta$ and together this gives:

$$
\gamma = \int_{S_2} \Omega \cdot \Omega' \sigma(x, \epsilon, \Omega \cdot \Omega') \, d\Omega \\
= \int_{-\pi}^\pi \int_{-1}^1 -\cos \theta \sigma(x, \epsilon, -\cos \theta) \sin \theta \, d\theta \, d\varphi.
$$

(3.18)

Now we substitute $\mu(\theta) = -\cos \theta$, which yields $\frac{d\mu(\theta)}{d\theta} = \sin \theta$ and thus $\sin \theta d\theta = d\mu$. Using this substitution we receive:

$$
\gamma = \int_{-\pi}^\pi \int_{-1}^1 \mu \sigma(x, \epsilon, \mu) \, d\mu \, d\varphi \\
= 2\pi \int_{-1}^1 \mu \sigma(x, \epsilon, \mu) \, d\mu.
$$

(3.19)

Overall this gives:

$$
\int_{S_2} \int_{S_2} \Omega \sigma(x, \epsilon, \Omega' \cdot \Omega) \psi(x, \epsilon, \Omega') \, d\Omega \, d\Omega' \\
= \int_{S_2} \int_{S_2} \Omega \sigma(x, \epsilon, \Omega' \cdot \Omega) \psi(x, \epsilon, \Omega') \, d\Omega \, d\Omega' \\
= \int_{S_2} \alpha \Omega' \psi(x, \epsilon, \Omega') \, d\Omega' \\
= \int_{S_2} \Omega' 2\pi \int_{-1}^1 \mu \sigma(x, \epsilon, \mu) \, d\mu \psi(x, \epsilon, \Omega') \, d\Omega' \\
= 2\pi \int_{-1}^1 \mu \sigma(x, \epsilon, \mu) \, d\mu \int_{S_2} \Omega \psi(x, \epsilon, \Omega) \, d\Omega,
$$

which is (3.17).

Using Lemma 2 and (3.2) we can reformulate the integral over the in-
elastic in-scattering term as:

\[
N_V(x) \int_{S^2} \int_{S^2} \Omega \sigma_{\text{in}}^{\text{CSD}}(x, \epsilon, \Omega') \psi(x, \epsilon, \Omega') \, d\Omega' \, d\Omega \\
= 2\pi \int_{-1}^{1} \mu \sigma_{\text{in}}^{\text{CSD}}(x, \epsilon, \mu) \, d\mu \int_{S^2} \Omega \psi(x, \epsilon, \Omega) \, d\Omega \\
= 2\pi \int_{-1}^{1} \mu \sigma_{\text{in}}^{\text{CSD}}(x, \epsilon, \mu) \, d\mu \psi^{(1)}(x, \epsilon) \\
= 2\pi \int_{0}^{\infty} \int_{-1}^{1} \mu \sigma_{\text{in}}(x, \epsilon, \mu') \, d\mu' \psi^{(1)}(x, \epsilon, \Omega).
\]  

Using the variable transformations performed in (3.18) and (3.19), integration over the total inelastic cross section term gives:

\[
- N_V(x) \int_{S^2} \Omega \sigma_{\text{in}}^{\text{tot}}(x, \epsilon) \psi(x, \epsilon, \Omega) \, d\Omega \\
= - N_V(x) 2\pi \int_{0}^{\infty} \int_{-1}^{1} \sigma_{\text{in}}(x, \epsilon, \epsilon', \mu) \, d\mu \int_{S^2} \psi(x, \epsilon, \Omega) \, d\Omega \\
= - N_V(x) 2\pi \int_{0}^{\infty} \int_{-1}^{1} \sigma_{\text{in}}(x, \epsilon, \epsilon', \mu) \, d\mu \, d\epsilon' \psi^{(1)}(x, \epsilon, \Omega) \\
= - N_V(x) 2\pi \int_{0}^{\infty} \int_{-1}^{1} \sigma_{\text{in}}(x, \epsilon, \epsilon', \mu) \, d\mu \, d\epsilon' \psi^{(1)}(x, \epsilon, \Omega).
\]

Combining (3.20) and (3.21) yields:

\[
N_V(x) \int_{S^2} \int_{S^2} \Omega \sigma_{\text{in}}^{\text{CSD}}(x, \epsilon, \Omega' \cdot \Omega) \psi(x, \epsilon, \Omega') \, d\Omega' \, d\Omega \\
- N_V(x) \int_{S^2} \Omega \sigma_{\text{in}}^{\text{tot}}(x, \epsilon) \psi(x, \epsilon, \Omega) \, d\Omega \\
= 2\pi \int_{0}^{\infty} \int_{-1}^{1} \mu \sigma_{\text{in}}(x, \epsilon, \epsilon', \mu) \, d\mu \, d\epsilon' \psi^{(1)}(x, \epsilon) \\
- N_V(x) 2\pi \int_{0}^{\infty} \int_{-1}^{1} \sigma_{\text{in}}(x, \epsilon, \epsilon', \mu) \, d\mu \, d\epsilon' \psi^{(1)}(x, \epsilon, \Omega) \\
= 2\pi N_V(x) \int_{0}^{\infty} \int_{-1}^{1} (1 - \mu) \sigma_{\text{in}}(x, \epsilon, \epsilon', \mu) \, d\mu \, d\epsilon' \psi^{(1)}(x, \epsilon, \Omega).
\]
integrals over the elastic scattering terms:

\[ N_V(x) \int_{S^2} \int_{S^2} \Omega \sigma_{el}(x, \epsilon, \Omega' \cdot \Omega') \psi(x, \epsilon, \Omega') d\Omega' d\Omega \]

\[ - N_V(x) \int_{S^2} \Omega \sigma_{el}^{tot}(x, \epsilon) \psi(x, \epsilon, \Omega) d\Omega \]

\[ = N_V(x) 2\pi \int_{-1}^{1} \mu \sigma_{el}(x, \epsilon, \mu) d\mu \psi^{(1)}(x, \epsilon, \Omega) \]

\[ - N_V(x) 2\pi \int_{-1}^{1} \sigma_{el}(x, \epsilon, \mu) d\mu \psi^{(1)}(x, \epsilon, \Omega) \]

\[ = 2\pi N_V(x) \int_{-1}^{1} (1 - \mu) \sigma_{el}(x, \epsilon, \mu) d\mu \psi^{(1)}(x, \epsilon, \Omega). \]

Finally we combine (3.15), (3.16), (3.22) and (3.23) to receive the following partial differential equation for the second moment:

\[ \nabla_x \cdot \psi^{(2)}(x, \epsilon) = -T_{el+in}(x, \epsilon) \psi^{(1)}(x, \epsilon) + \frac{\partial}{\partial \epsilon} \left( S(x, \epsilon) \psi^{(1)}(x, \epsilon) \right), \]  

(3.24)

where

\[ T_{el+in}(x, \epsilon) := T_{el}(x, \epsilon) + T_{in}(x, \epsilon), \]  

(3.25)

\[ T_{in}(x, \epsilon) := 2\pi N_V(x) \int_{0}^{\infty} \int_{-1}^{1} (1 - \mu) \sigma_{in}(x, \epsilon, \epsilon', \mu) d\mu d\epsilon', \]  

(3.26)

\[ T_{el}(x, \epsilon) := 2\pi N_V(x) \int_{-1}^{1} (1 - \mu) \sigma_{el}(x, \epsilon, \mu) d\mu. \]  

(3.27)

### 3.3.2 The minimum entropy closure

In the following we derive a closure for \( \psi^{(2)} \) based on the minimum entropy solution, discussed in Section 3.2. For \( n = 1 \) the minimization problem (3.28) reads:

\[ H(\psi_{ME}) = \min_{\psi} H(\psi), \]

\[ \text{s.t. } \int_{S^2} \psi_{ME} d\Omega = \psi^{(0)} \text{ and } \int_{S^2} \Omega \psi_{ME} d\Omega = \psi^{(1)}. \]  

(3.28)
We can eliminate the equality constraints by introducing a constrained entropy function:

$$
\tilde{H}(\psi) := H(\psi) + \lambda_0 \left( \int_{S^2} \psi \, d\Omega - \psi^{(0)} \right) + \lambda_1 \cdot \left( \int_{S^2} \Omega \psi \, d\Omega - \psi^{(1)} \right),
$$

(3.29)

where $\lambda_0 : \mathbb{R}^3 \times [0, \infty) \to \mathbb{R}$ and $\lambda_1 : \mathbb{R}^3 \times [0, \infty) \to \mathbb{R}^3$ are the Lagrange multipliers. Note that like $\psi$ the Lagrange multipliers are functions in space and energy. Now (3.28) is equivalent to:

$$
\tilde{H}(\psi_{ME}) = \min_{\psi} \tilde{H}(\psi),
$$

(3.30)

To solve this problem analytically, we derive the corresponding Euler-Lagrange equation. The first variation of (3.30) is:

$$
\delta \tilde{H}(\psi, \epsilon) := \frac{\partial}{\partial \epsilon}_{|\epsilon=0} \tilde{H}(\psi_{ME} + \epsilon \psi)
= \int_{S^2} \frac{\partial}{\partial \psi}_{|\psi=\psi_{ME}} \left[ -\psi \log \psi + \lambda_0 (\psi - \psi^{(0)}) + \lambda_1 \cdot (\Omega \psi - \psi^{(1)}) \right] \psi \, d\Omega
= \int_{S^2} [-\log \psi_{ME} - 1 + \lambda_0 + \lambda_1 \cdot \Omega] \psi \, d\Omega.
$$

For $\psi_{ME}$ to be a solution of (3.30), it is necessary that the first variation vanishes:

$$
\delta \tilde{H}(\psi, \epsilon) = 0 \quad \forall \psi
\Leftrightarrow \int_{S^2} [-\log \psi_{ME} - 1 + \lambda_0 + \lambda_1 \cdot \Omega] \psi \, d\Omega = 0 \quad \forall \psi.
$$

Because $\psi$ is arbitrary the fundamental lemma of the calculus of variations yields the following Euler-Lagrange equation:

$$
- \log \psi_{ME} - 1 + \lambda_0 + \lambda_1 \cdot \Omega = 0.
$$

(3.31)
Thus we have

\[
\psi_{\text{ME}}(x, \epsilon, \Omega) = e^{-1 + \lambda_0(x, \epsilon)} e^{\lambda_1(x, \epsilon) \cdot \Omega} = a_0(x, \epsilon) e^{-a_1(x, \epsilon) \cdot \Omega},
\]  

(3.32)

where

\[
a_0(x, \epsilon) = e^{-1 + \lambda_0(x, \epsilon)} \quad \text{and} \quad a_1(x, \epsilon) = -\lambda_1(x, \epsilon),
\]  

(3.33)

are functions in space and energy, arising from the Lagrange multipliers. This means that by applying the minimum entropy closure we assume that the angular distribution of the electrons can be approximated by the ansatz function \( \psi_{\text{ME}} \). The coefficients \( a_0 \) and \( a_1 \), that determine the spatial and energy dependence of the electron fluence, are directly related to the Lagrange multipliers, which are related to the zeroth and first order moment through the constraints (3.28).

In our case of the M1 model, i.e. for \( n = 1 \), this relation can be used to directly obtain a closure relation of the form (3.6). Using (3.32) to evaluate the right-hand side in the ansatz

\[
\psi^{(n)} = \int_{S^2} \Omega^n \psi_{\text{ME}} \, d\Omega,
\]  

(3.34)

for \( n = 0, 1 \) gives expressions for \( \psi^{(0)} \) and \( \psi^{(1)} \) in terms of \( a_0 \) and \( a_1 \). The following results are given in [10]:

\[
\psi^{(0)} = 4\pi a_0 \frac{\sinh |a_1|}{|a_1|},
\]

\[
\psi^{(1)} = 4\pi a_0 a_1 \frac{\sinh |a_1|(1 - |a_1| \coth |a_1|)}{|a_1|^3}.
\]  

(3.35)

Now the right-hand side of (3.34) can be evaluated for \( n = 2 \). In the resulting expression the anisotropy parameter

\[
\alpha := \frac{\psi^{(1)}}{\psi^{(0)}} = \frac{1 - |a_1| \coth |a_1|}{|a_1|^2} a_1,
\]  

(3.36)
plays an important role. In [10] the following closure relation is given:

$$\psi^{(2)} = \psi^{(0)} \left( \frac{1 - \chi(|\alpha|)}{2} I + \frac{3 \chi(|\alpha|) - 1}{|\alpha|} \frac{\alpha \otimes |\alpha|}{|\alpha|} \right), \quad (3.37)$$

where

$$\chi = \frac{|a_1|^2 - 2 |a_1| \coth |a_1| + 2}{|a_1|^2}, \quad (3.38)$$

is called Eddington factor. It is a function of $|\alpha|$ in the sense that (3.35) implies

$$|\alpha| = \frac{|\psi^{(1)}{\psi^{(0)}}|}{|\psi^{(0)}|} = \frac{|a_1| \coth |a_1| - 1}{|a_1|}, \quad (3.39)$$

which has a unique solution $|a_1|$ for given $|\alpha|$. For detailed calculations concerning the evaluation of the right-hand sides in (3.34), we refer to [7]. Note that the same entropy as in (3.7) is used there, but without the negative sign. Furthermore note that in [10] the tensor $\alpha \otimes \alpha$ appears without normalization in the expression (3.37). However, from the calculations in [7] it is clear that $\alpha \otimes \alpha$ must appear in normalized form. Also, based on the expression (3.37) as it is given in this work, a calculation of the spectrum of the Jacobian of $F$ was performed. The results presented in Section 4.4 agree with the results in [7, 10].

### 3.3.3 The final system of partial differential equations

To summarize the equations that we have derived above, we formulate them as one system of coupled partial differential equations.

First we denote the three components of $\psi^{(1)}$ with $\psi_x^{(1)}$, $\psi_y^{(1)}$, $\psi_z^{(1)}$ and define

$$U(x, \epsilon) := \begin{pmatrix} \psi^{(0)}(x, \epsilon) \\ \psi_x^{(1)}(x, \epsilon) \\ \psi_y^{(1)}(x, \epsilon) \\ \psi_z^{(1)}(x, \epsilon) \end{pmatrix}.$$  

(3.40)

Now we define a mathematical flux $F$ such that the vector of the divergence terms in (3.14) and (3.24) can be written as the divergence of the tensor
3.3. THE M1 MODEL FOR ELECTRON TRANSPORT

Let 

\[ F(U) \rightarrow \mathbb{R}^{4 \times 3}, \]

\[ F(U) := \begin{pmatrix} U_1 & U_2 & U_3 \\ U_0 \left( \frac{1-\chi(|\alpha|)}{2} I + \frac{3\chi(|\alpha|)-1}{2} \frac{\alpha}{|\alpha|} \otimes \frac{\alpha}{|\alpha|} \right) \end{pmatrix}, \quad (3.41) \]

with the anisotropy parameter

\[ \alpha = \frac{\psi(1)}{\psi(0)} = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_0 \end{pmatrix}. \quad (3.42) \]

Then we have

\[ \nabla_x \cdot F(U) = \nabla_x \left( \begin{pmatrix} U_1 & U_2 & U_3 \\ U_0 \left( \frac{1-\chi(|\alpha|)}{2} I + \frac{3\chi(|\alpha|)-1}{2} \frac{\alpha}{|\alpha|} \otimes \frac{\alpha}{|\alpha|} \right) \end{pmatrix} \right) \]

\[ = \begin{pmatrix} \partial_x U_1 + \partial_y U_2 + \partial_z U_3 \\ \nabla_x \cdot U_0 \left( \frac{1-\chi(|\alpha|)}{2} I + \frac{3\chi(|\alpha|)-1}{2} \frac{\alpha}{|\alpha|} \otimes \frac{\alpha}{|\alpha|} \right) \end{pmatrix} \]

\[ = \begin{pmatrix} \nabla_x \cdot \psi(1) \\ \nabla_x \cdot \psi(2) \end{pmatrix}. \]

Furthermore with \( I = \text{diag}(1, \ldots, 1) \in \mathbb{R}^{3 \times 3} \) we define

\[ T(x, \epsilon) := \begin{pmatrix} 0 & 0 \\ 0 & -T_{\text{el-in}}(x, \epsilon) I \end{pmatrix} \in \mathbb{R}^{4 \times 4}, \quad (3.43) \]

which yields

\[ T(x, \epsilon) U(x, \epsilon) = \begin{pmatrix} 0 \\ -T_{\text{el-in}}(x, \epsilon) \psi(x, \epsilon) \\ -T_{\text{el-in}}(x, \epsilon) \psi(x, \epsilon) \psi(x, \epsilon) \\ -T_{\text{el-in}}(x, \epsilon) \psi(x, \epsilon) \psi(x, \epsilon) \end{pmatrix}. \]
Thus together the partial differential equations (3.14) and (3.24) can be rewritten as the following equivalent system:

\[-\frac{\partial}{\partial \epsilon} (U(x, \epsilon)S(x, \epsilon)) + \nabla_x \cdot F(U(x, \epsilon)) = -T(x, \epsilon)U(x, \epsilon). \tag{3.44}\]

The form of this system is similar to a conservation law. For the solution of such partial differential equations well established numerical methods do exist, one of which will be presented in the following chapter.
Chapter 4

Numerical treatment of the M1 model

4.1 The Godunov method for conservation laws

The numerical method for solving (3.44) that we present in this thesis will be based on the conservative form of classical finite volume schemes. In this section we motivate the further proceeding throughout this chapter by introducing the basic concepts behind these schemes using an ordinary, spatially one-dimensional, scalar conservation law and the classical Godunov method as an example.

The prototype of a scalar conservation law, with conserved variable

\[ u : [0,1] \times [0,T] \rightarrow \mathbb{R}, \]  

is of the form

\[ \frac{\partial}{\partial t} u(x,t) + \frac{\partial}{\partial x} f(u(x,t)) = 0 \quad \text{for } x \in [0,1] \times [0,T], \]  

\[ u(x,0) = u^0(x) \quad \text{for } x \in [0,1], \]  

\[ (4.1) \]
with some mathematical flux \( f : \mathbb{R} \to \mathbb{R} \), an initial condition \( u^0 : \mathbb{R} \to \mathbb{R} \) and an end time \( T > 0 \).

Let \( N_x \in \mathbb{N} \) and \( \Delta t^n > 0 \) denote some time steps for \( n \in \mathbb{N} \). Then define:

\[
\Delta x := \frac{1}{N_x} \\
x_j := (j + \frac{1}{2}) \Delta x \quad \text{for } j = 1, \ldots, N_x \\
I_j := [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \quad \text{for } j = 1, \ldots, N_x \\
t^{n+1} := t^n + \Delta t^n \quad \text{for } n = 1, 2, \ldots \text{ and } t^0 := 0.
\]

Integrating (4.2) over \( I_j \) and from \( t^n \) to \( t^{n+1} \) yields:

\[
0 = \int_{I_j} \int_{t^n}^{t^{n+1}} \frac{\partial}{\partial t} u(x,t) \, dt \, dx + \int_{I_j} \int_{t^n}^{t^{n+1}} \frac{\partial}{\partial x} f(u(x,t)) \, dx \, dt \\
= \int_{I_j} u(x,t^{n+1}) - u(x,t^n) \, dx \\
+ \int_{t^n}^{t^{n+1}} f(u(x_{i+\frac{1}{2}}, t)) - f(u(x_{i-\frac{1}{2}}, t)) \, dt.
\]

Defining

\[
u^n_j := \frac{1}{\Delta x} \int_{I_j} u(x,t^n) \, dx,
\]

\[
f^{n+\frac{1}{2}}_{j+\frac{1}{2}} := \frac{1}{\Delta t^n} \int_{t^n}^{t^{n+1}} f(u(x_{j+\frac{1}{2}}, t)) \, dt,
\]

we can rewrite (4.4) equivalently as:

\[
u^{n+1}_j = \nu^n_j - \frac{\Delta t^n}{\Delta x} \left( f^{n+\frac{1}{2}}_{j+\frac{1}{2}} - f^{n+\frac{1}{2}}_{j-\frac{1}{2}} \right).
\]

This is the conservative form of a finite volume scheme. If \( u(\cdot, t^n) \) can be exactly represented by its means on all cells, then (4.6) yields an exact time update scheme, otherwise a first order approximation.

One remaining problem is the calculation of \( f^{n+\frac{1}{2}}_{j+\frac{1}{2}} \). In the classical Godunov
method this is done by exactly solving the following Riemann problem:

\[
\frac{\partial}{\partial t} w + \frac{\partial}{\partial x} f(w) = 0 \quad \text{on } \mathbb{R} \times [t^n, t^{n+1}],
\]

\[
w(x, t^n) = w_{j+\frac{1}{2}}^0(x) \quad \text{for } x \in \mathbb{R},
\]

where

\[
w_{j+\frac{1}{2}}^0(x) := \begin{cases} u_n^j, & x < x_{j+\frac{1}{2}} \\ u_{n+1}^j, & x > x_{j+\frac{1}{2}} \end{cases}.
\]

Then \( f_{j+\frac{1}{2}}^{n+\frac{1}{2}} \) can be calculated using the following equation (cf. [24]):

\[
f_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{\Delta t^n} \int_{t^n}^{t^{n+1}} f(u(x_{j+\frac{1}{2}}, t)) \, dt = f(w(x_{j+\frac{1}{2}}, t^{n+1})). \tag{4.8}
\]

Other numerical schemes try to approximate the solution of (4.7) using so-called approximate Riemann solvers. One of these, which is based on the Harten-Lax-van Leer flux, will be introduced in Section 4.3.

The solution of (4.7) only depends on how the initial information \( u^0 \) is propagated in time by the characteristics, i.e. curves of the form \( x = x(t) \) along which the solution is constant:

\[
\frac{d}{dt} u(x(t), t) = 0. \tag{4.9}
\]

Applying the chain rule on (4.9) yields:

\[
\frac{\partial}{\partial t} + \dot{x} \frac{\partial}{\partial x} u = 0.
\]

Comparing with (4.2), we see that in the scalar case these curves are always straight lines with constant characteristic speed \( \dot{x} = f'(u) \).

When two characteristics originating from different cells (with different initial values) meet, the solution will be discontinuous at their intersection. For the numerical stability of the scheme (4.6) one therefore needs to ensure that
characteristics from adjacent cells cannot intersect each other. This results in the so-called Courant-Friedrichs-Lewy (CFL) condition [24]:

\[
\nu := \frac{\Delta t^n}{\Delta x} \max \{|f'(u)| \mid \min u(\cdot, t^n) \leq u(\cdot, t^n) \leq \max u(\cdot, t^n)\} \leq 1, \quad (4.10)
\]

which imposes a constraint on the time step in dependence of the spatial mesh size and the derivative of the flux.

Note that if, in the Riemann problem (4.7), the characteristic at \( x = x_{j+\frac{1}{2}} \) has speed \( \dot{x} = 0 \), then, in the Riemann solution \( w \), the discontinuity in the initial condition \( u_{j+\frac{1}{2}} \) at \( (x_{j+\frac{1}{2}}, t^n) \) is propagated to \( (x_{j+\frac{1}{2}}, t^{n+1}) \). However, one can show that for a (weak) solution \( u \) of (4.2), the following so-called Rankine-Hugoniot condition holds (cf. [24]):

\[
f(u_L) - f(u_R) = s(u_L - u_R), \quad (4.11)
\]

where \( u_L, u_R \) are the states to the left and right of a characteristic with speed \( \dot{x} = s \). For \( \dot{x} = 0 \) this gives

\[
f(u_L) - f(u_R) = 0,
\]

and therefore the expression \( f(w(x_{j+\frac{1}{2}}, t^{n+1})) \) in (4.8) is well defined.

4.2 Reformulation as an initial value problem

Our aim is to derive a numerical finite volume scheme that directly solves (3.44). Comparing with (4.2), we see that the energy variable \( \epsilon \) can be interpreted as a time variable. In this section we reformulate (3.44) such that it becomes an initial value problem of a similar form as (4.2).

4.2.1 Transformation to a pseudo time variable

The initial information that is known about the distribution \( \psi(x, \epsilon, \Omega) \) stems from the configuration of the electron beam, which has a specific beam
energy $\epsilon_b$. As the electrons will immediately lose energy when entering the specimen, the phase states of the electrons with energy $\epsilon = \epsilon_b$ can be directly related to the properties of the electron beam. In order to formulate this as an initial condition at time $t = 0$, we perform the following variable transformation:

$$\epsilon(t) := \epsilon_b - t.$$  \hfill (4.12)

Note that $t$ is a pseudo time variable that has the same unit as $\epsilon$. This yields

$$\frac{\partial \epsilon(t)}{\partial t} = -1,$$  \hfill (4.13)

and thus we receive

$$\frac{\partial}{\partial t} (S(x, \epsilon(t))U(x, \epsilon(t)))
= S(x, \epsilon(t)) \frac{\partial}{\partial t} U(x, \epsilon(t)) + U(x, \epsilon(t)) \frac{\partial}{\partial \epsilon} S(x, \epsilon(t))
= -S(x, \epsilon(t)) \frac{\partial}{\partial \epsilon} U(x, \epsilon) - U(x, \epsilon(t)) \frac{\partial}{\partial \epsilon} S(x, \epsilon)
= - \frac{\partial}{\partial \epsilon} |_{\epsilon = \epsilon(t)} (S(x, \epsilon)U(x, \epsilon)).$$

Using (4.12) we now express the quantities in (3.44) in the new variable $t$:

$$\hat{U}(x,t) := U(x, \epsilon(t)),
\hat{S}(x,t) := S(x, \epsilon(t)),
\hat{T}(x,t) := T(x, \epsilon(t)).$$

Then (3.44) is equivalent to:

$$\frac{\partial}{\partial t} \left( \hat{S}(x,t)\hat{U}(x,t) \right) + \nabla_x \cdot F(\hat{U}(x,t)) = -\hat{T}(x,t)\hat{U}(x,t).$$  \hfill (4.14)

### 4.2.2 The electron beam as an initial condition

In the following we present an initial condition that models the phase states of the electrons in the initial electron beam and is suitable for the method that will be introduced in Section 4.3.
In a real experiment the electron beam has a non-negligible width that can be approximated by a Gaussian distribution of the electrons in the plane parallel to the sample surface. While traveling through vacuum the energy and direction of travel of the electrons does not change. The initial direction of travel of the electrons is nearly exactly orthogonal to the material surface. Therefore we model their initial angular distribution as a Dirac. Based on the assumption that the electrons will immediately lose energy after entering the material sample, we could formulate the initial condition for the distribution $\psi$ as follows:

$$\psi(x, \epsilon_b, \Omega) = \tilde{\psi}(x, \Omega) = \begin{cases} e^{-c_1|x-x_b|^2} \delta(\Omega \cdot \Omega_b - 1) & , x \in S, \\ 0 & , x \in Q \setminus S, \end{cases}$$

where $\Omega_b = (0, 0, -1)^T$, $x_b \in S$ is the measuring point and $\delta(\cdot)$ denotes the Dirac delta function.

A mathematical inconvenience of this model is that the support of $\tilde{\psi}$ has zero measure with respect to three-dimensional space. For this reason we have decided to assume a narrow Gaussian distribution of the electrons with energy $\epsilon = \epsilon_b$ along the $z$-axis and with $x^b$ as its center. Thus for $x \in Q$ and $\Omega \in S^2$ we define the following initial distribution of the electron fluence:

$$\psi_b(x, \Omega) := e^{-c_1[(x_1, x_2)^T - (x_b^1, x_b^2)^T]^2} e^{-c_2(x_3 - x_b^3)^2} \delta(\Omega \cdot \Omega_b - 1). \quad (4.15)$$

Physically this means that some electrons start within the material. From this we can calculate the boundary condition in terms of the first two moments. Using the integration properties of the $\delta$ function, we receive:

$$\hat{U}^0(x) := \left( \frac{\int_{S^2} \psi_b(x, \Omega) d\Omega}{\int_{S^2} \Omega \psi_b(x, \Omega) d\Omega} \right) = \begin{pmatrix} e^{-c_1[(x_1, x_2)^T - (x_b^1, x_b^2)^T]^2} e^{-c_2(x_3 - x_b^3)^2} \\ \Omega_b e^{-c_1[(x_1, x_2)^T - (x_b^1, x_b^2)^T]^2} e^{-c_2(x_3 - x_b^3)^2} \end{pmatrix}.$$

80
4.3. Formulation of the initial value problem

We now combine (4.14) and (4.16) to formulate the \( \mathcal{M}1 \) model for electron transport as an initial value problem: find \( \hat{U} \) such that \( \forall x \in \mathbb{R}^3 \) and \( \forall t \in [0, T] \)

\[
\frac{\partial}{\partial t} \left( \hat{S}(x,t)\hat{U}(x,t) \right) + \nabla_x \cdot F(\hat{U}(x,t)) = -\hat{T}(x,t)\hat{U}(x,t),
\]

\[
\hat{U}(x,0) = \hat{U}^0(x),
\]

(4.17)

where \( T := \epsilon_b - \epsilon_c \) and \( \epsilon_c \) is the cut-off energy, i.e. a threshold below which the energy of the electrons is not sufficient for the generation of characteristic x-rays. Note that \( \epsilon(T) = \epsilon_b - (\epsilon_b - \epsilon_c) = \epsilon_c \).

4.3 A finite volume scheme based on the HLL flux

4.3.1 Discretization of the phase space

To simplify the derivation and the implementation of the method presented here, we assume that the material is homogeneous in the \( y \) direction. This simplification allows us to deal with a spatially two-dimensional problem only. The extension to three spatial dimensions is theoretically straightforward, but complicates the implementation. In this regard we refer to the lateral axis as the \( x \)-axis and the vertical axis (former \( z \)-axis) as the \( y \)-axis in this chapter and denote the material sample as:

\[
\hat{Q} = [x_l, x_r] \times [y_l, 0],
\]

(4.18)
We discretize this domain by dividing it into $N_x \cdot N_y$ cells:

$$Q_{ij} := [x_{i-\frac{1}{2},x_{i+\frac{1}{2}}}] \times [y_{j-\frac{1}{2},y_{j+\frac{1}{2}}}],$$

$$x_i := x_l + (i + \frac{1}{2}) \Delta x,$$

$$y_j := y_l + (j + \frac{1}{2}) \Delta y,$$

$$\Delta x := \frac{x_r - x_l}{N_x},$$

$$\Delta y := \frac{0 - y_l}{N_y} = \frac{-y_l}{N_y}. \quad (4.19)$$

We assume that the mesh $\mathcal{T}$ for the material properties, as introduced in (1.22), is chosen according to the mesh $Q_{ij}$, in the sense that

$$c|_{Q_{ij}} \equiv \text{const} \quad \text{for} \quad i = 1, \ldots, N_x \quad \text{and} \quad j = 1, \ldots, N_y. \quad (4.20)$$

This allows us to define

$$S_{ij}(t) := S(x,t) \quad \text{for} \quad x \in Q_{ij}$$

$$T_{ij}(t) := T(x,t) \quad \text{for} \quad x \in Q_{ij}. \quad (4.21)$$

Furthermore let $\Delta t^n$ for $n = 1, 2, \ldots$ denote the time steps. For the moment we assume that they are properly chosen. In Section 4.4 we will discuss in detail how to choose these time steps such that they fulfill a similar condition as the one introduced in (4.10).

### 4.3.2 Derivation of the conservative form of the IVP

In the following we will derive a finite volume scheme for the spatially two-dimensional variant of the initial value problem (4.17). Our derivation will be based on the concepts that were briefly discussed in Section 4.1.

In the following we integrate (4.17) over $Q_{ij}$ and from $t^n$ to $t^{n+1}$. We treat
the integration of the three terms

$$\frac{\partial}{\partial t} \left( \hat{S}(x, t) \hat{U}(x, t) \right),$$

(4.22)

$$\nabla_x \cdot F(\hat{U}(x, t)),$$

(4.23)

$$- \hat{T}(x, t) \hat{U}(x, t),$$

(4.24)

separately. Using (4.21) the integration of (4.22) yields:

$$\hat{t}_n^{+1} \hat{Q}_{ij} \frac{\partial}{\partial t} \left( \hat{S}(x, t) \hat{U}(x, t) \right) dx dt$$

$$= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{t}_n^{+1} \hat{Q}_{ij} \frac{\partial}{\partial t} \left( \hat{S}_{ij}(t) \hat{U}(x, y, t) \right) dt dy dx$$

$$= \hat{S}_{ij}(t^{n+1}) \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{U}(x, y, t^{n+1}) dy dx$$

$$- \hat{S}_{ij}(t^n) \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{U}(x, y, t^n) dy dx.$$

Analogous to (4.5), we define:

$$U_{ij}^n := \frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{U}(x, y, t^n) dy dx.$$

(4.25)

Furthermore let $$S_{ij}^n := \hat{S}_{ij}(t^n)$$. Then we receive:

$$\int_{t^n}^{t^{n+1}} \int_{Q_{ij}} \frac{\partial}{\partial t} \left( \hat{S}(x, t) \hat{U}(x, t) \right) dx dt$$

$$= \Delta x \Delta y \left( S_{ij}^{n+1} U_{ij}^{n+1} - S_{ij}^n U_{ij}^n \right).$$

(4.26)
Let $F_1, F_2$ denote the first and second column of $F$ respectively. Then integration of (4.23) yields:

\[
\int_{t^n}^{t^{n+1}} \int_{Q_{ij}} \nabla_x \cdot F(\hat{U}(x,t)) \, dx \, dt \\
= \int_{t^n}^{t^{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial}{\partial x} F_1(\hat{U}(x,y,t)) \, dy \, dx \, dt \\
+ \int_{t^n}^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \frac{\partial}{\partial y} F_2(\hat{U}(x,y,t)) \, dy \, dx \, dt \\
= \int_{t^n}^{t^{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} F_1(\hat{U}(x_{i+\frac{1}{2}},y,t)) \, dy \, dt \\
- \int_{t^n}^{t^{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} F_1(\hat{U}(x_{i-\frac{1}{2}},y,t)) \, dy \, dt \\
+ \int_{t^n}^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} F_2(\hat{U}(x,y_{j+\frac{1}{2}},t)) \, dy \, dt \\
- \int_{t^n}^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} F_2(\hat{U}(x,y_{j-\frac{1}{2}},t)) \, dy \, dt.
\]

Defining

\[
\hat{F}^{n+\frac{1}{2}}_{i+\frac{1}{2},j} := \frac{1}{\Delta t^n \Delta y} \int_{t^n}^{t^{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} F_1(\hat{U}(x_{i+\frac{1}{2}},y,t)) \, dy \, dt, \\
\hat{G}^{n+\frac{1}{2}}_{i,j+\frac{1}{2}} := \frac{1}{\Delta t^n \Delta x} \int_{t^n}^{t^{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} F_2(\hat{U}(x,y_{j+\frac{1}{2}},t)) \, dy \, dt. 
\tag{4.27}
\]

this can be equivalently expressed as:

\[
\int_{t^n}^{t^{n+1}} \int_{Q_{ij}} \nabla_x \cdot F(\hat{U}(x,t)) \, dx \, dt \\
= \Delta t^n \Delta y \left( \hat{F}^{n+\frac{1}{2}}_{i+\frac{1}{2},j} - \hat{F}^{n+\frac{1}{2}}_{i-\frac{1}{2},j} \right) + \Delta t^n \Delta x \left( \hat{G}^{n+\frac{1}{2}}_{i,j+\frac{1}{2}} - \hat{G}^{n+\frac{1}{2}}_{i,j-\frac{1}{2}} \right). 
\tag{4.28}
\]
Finally we integrate the right-hand side (4.24) of (4.17). Using (4.21) we receive:

\[
- \int_{t^n}^{t^{n+1}} \int_{Q_{ij}} \hat{T}(x, t) \hat{U}(x, t) \, dx \, dt
= - \int_{t^n}^{t^{n+1}} \hat{T}_{ij}(t) \int_{Q_{ij}} \hat{U}(x, t) \, dx \, dt.
\]

Defining

\[
\hat{T}_{ij}^{n+\frac{1}{2}} := \frac{1}{\Delta x \Delta y} \int_{t^n}^{t^{n+1}} \hat{T}_{ij}(t) \int_{Q_{ij}} \hat{U}(x, t) \, dx \, dt,
\]

this can be written as:

\[
- \int_{t^n}^{t^{n+1}} \int_{Q_{ij}} \hat{T}(x, t) \hat{U}(x, t) \, dx \, dt
= - \Delta t^n \Delta x \Delta y \hat{T}_{ij}^{n+\frac{1}{2}}.
\]

Combining (4.26), (4.28) and (4.30), integration of (4.17) over \( Q_{ij} \times [t^n, t^{n+1}] \) results in the following equation:

\[
\Delta x \Delta y \left( S_{ij}^{n+1} U_{ij}^{n+1} - S_{ij} U_{ij}^n \right) + \Delta t^n \Delta y \left( \hat{F}_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - \hat{F}_{i-\frac{1}{2},j}^{n+\frac{1}{2}} \right)
+ \Delta t^n \Delta x \left( \hat{G}_{i+,j+\frac{1}{2}}^{n+\frac{1}{2}} - \hat{G}_{i+,j-\frac{1}{2}}^{n+\frac{1}{2}} \right)
= -\Delta t^n \Delta x \Delta y \hat{T}_{ij}^{n+\frac{1}{2}}.
\]

Dividing by \( \Delta x \Delta y \) yields the following two-dimensional analogon of (4.6) for (4.17):

\[
S_{ij}^{n+1} U_{ij}^{n+1} = S_{ij} U_{ij}^n - \frac{\Delta t^n}{\Delta x} \left( \hat{F}_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - \hat{F}_{i-\frac{1}{2},j}^{n+\frac{1}{2}} \right)
- \frac{\Delta t^n}{\Delta y} \left( \hat{G}_{i+,j+\frac{1}{2}}^{n+\frac{1}{2}} - \hat{G}_{i+,j-\frac{1}{2}}^{n+\frac{1}{2}} \right) - \Delta t^n \hat{T}_{ij}^{n+\frac{1}{2}}.
\]

Recalling (3.43), note that \( \hat{T}_{1,} = 0 \). Thus the first component of \( \hat{T} \hat{U} \) is zero. The definition (4.29) implies that the first component of \( \hat{T}_{ij}^{n+\frac{1}{2}} \), which
represents the right-hand side in (4.17), is also zero. Therefore in (4.31) the first component of \( \hat{S} \hat{U} \) is conserved.

However, to receive a scheme that yields an update formula for the unknown variable \( \hat{U} \), we divide by \( S_{ij}^{n+1} \):

\[
U_{ij}^{n+1} = U_{ij}^n - \frac{\Delta t}{\Delta x} \left( \frac{\hat{F}_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - \hat{F}_{i-\frac{1}{2},j}^{n+\frac{1}{2}}}{S_{ij}^{n+1}} \right) - \frac{\Delta t}{\Delta y} \left( \frac{\hat{G}_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} - \hat{G}_{i,j-\frac{1}{2}}^{n+\frac{1}{2}}}{S_{ij}^{n+1}} \right) \\
+ \left( \frac{S_{ij}^n}{S_{ij}^{n+1}} - 1 \right) U_{ij}^n - \Delta t \frac{\hat{T}_{ij}^{n+\frac{1}{2}}}{S_{ij}^{n+1}}. 
\]

(4.32)

In order to give an explicit time update scheme, we have to specify numerical fluxes that approximate the mathematical fluxes in (4.27) and depend only on \( U_{ij}^n \). For (4.29) we simply use:

\[
\hat{T}_{ij}^{n+\frac{1}{2}} \approx T_{ij}^n U_{ij}^n, \quad (4.33)
\]

where \( T_{ij}^n := T_{ij}(t^n) \).

### 4.3.3 The HLL flux on a regular two-dimensional grid

In the following we derive the numerical HLL flux that approximates the mathematical fluxes in (4.27). It is named after Harten, Lax and van Leer [19], who have introduced this flux and have contributed much to the general development and understanding of finite volume schemes. For extensive background information about the theory behind hyperbolic conservation laws and numerical finite volume schemes, we refer to [24].

The HLL flux is based on an approximate solver for the Riemann problems that dominate the propagation of information. Before we start the derivation we therefore briefly discuss the behavior of the characteristics in linear and non-linear hyperbolic systems.
4.3. A FINITE VOLUME SCHEME BASED ON THE HLL FLUX

Characteristics in linear hyperbolic systems

In the case that the mathematical flux is scalar, the characteristic velocities are given by $\dot{x} = f'(u)$. Let us now regard a linear system of conservation laws for a vector quantity $U : \mathbb{R} \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^3$:

$$\frac{\partial}{\partial t} U + A \frac{\partial}{\partial x} U = 0.$$  \hfill (4.34)

Here the mathematical flux is linear:

$$F : \mathbb{R}^3 \rightarrow \mathbb{R}^3, \; F(U) = AU,$$  \hfill (4.35)

where $A \in \mathbb{R}^{3 \times 3}$. If $A$ is diagonalizable and all eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of $A$ are real, the system (4.34) is called hyperbolic. This means, that there exists an orthogonal eigenvector basis $R \in \mathbb{R}^3$ such that

$$R^{-1}AR = \text{diag}(\lambda_1, \lambda_2, \lambda_3).$$  \hfill (4.36)

Multiplying (4.34) and using (4.36) yields a decoupled system of scalar conservation laws in terms of the characteristic variables $R^{-1}U$:

$$0 = R^{-1} \frac{\partial}{\partial t} U + R^{-1}A \frac{\partial}{\partial x} U$$
$$= \frac{\partial}{\partial t} R^{-1} U + R^{-1}R \text{diag}(\lambda_1, \lambda_2, \lambda_3) R^{-1} \frac{\partial}{\partial x} U$$
$$= \frac{\partial}{\partial t} R^{-1} U + \text{diag}(\lambda_1, \lambda_2, \lambda_3) \frac{\partial}{\partial x} R^{-1} U.$$

This means that in such a linear system the initial information, when represented in the basis $R^{-1}$, is propagated along characteristics with constant velocities $\dot{x}_1 = \lambda_1, \dot{x}_2 = \lambda_2, \dot{x}_3 = \lambda_3$.

Characteristics in non-linear hyperbolic systems

In the case of a non-linear system of hyperbolic conservation laws, a similar characterization of the propagation speeds is possible, though the theory behind it is more involved. For a detailed discussion, we refer to [24]. The
CHAPTER 4. NUMERICAL TREATMENT OF THE M1 MODEL

The basic idea is that, due to the implicit function theorem, a Riemann problem in a non-linear hyperbolic system can be approximated by its linearization:

$$\frac{\partial}{\partial t}U + DF(\bar{U}) \frac{\partial}{\partial x} U = 0,$$

(4.37)

where $F: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ may now be non-linear and $\bar{U}$ is an expansion point for the Taylor expansion of $F$. The characteristic velocities in the linearized system are then given by the eigenvalues of the Jacobian matrix $DF(\bar{U})$ of $F$.

**A separation approach for two-dimensional regions**

The classical HLL flux can only be applied directly to spatially one-dimensional Riemann problems. Therefore we split our spatially two-dimensional problem into one-dimensional Riemann problems in $x$ direction based on the mathematical $F_1$ flux and one-dimensional Riemann problems in $y$ direction based on the mathematical $F_2$ flux.

Apart from the right-hand side, which is not regarded in the Riemann problem, our original partial differential equation (4.17) reads:

$$\frac{\partial}{\partial t} \left( \hat{S} \hat{U} \right) + \nabla_x \cdot F(\hat{U}) = 0.$$

(4.38)

The corresponding one-dimensional Riemann problem in $x$ direction at $(x_{i+\frac{1}{2}}, y_j)$ is:

$$\frac{\partial}{\partial t} \left( \hat{S}(x, y_j, t) \hat{U}(x, y_j, t) \right) + \frac{\partial}{\partial x} F_1(\hat{U}(x, y_j, t)) = 0 \quad \text{on } \mathbb{R} \times [t^n, t^{n+1}],

\hat{U}(x, y_j, t^n) = U_{i+\frac{1}{2}, j}^{n, 1}(x) \quad \text{for } x \in \mathbb{R},$$

(4.39)

where

$$U_{i+\frac{1}{2}, j}^{n, 1}(x) := \begin{cases} U_{ij}^n, & x < x_{i+\frac{1}{2}} \\ U_{i+1,j}^n, & x > x_{i+\frac{1}{2}} \end{cases}.$$
The corresponding one-dimensional Riemann problem in $y$ direction at $(x_i, y_{j+\frac{1}{2}})$ is:

$$\frac{\partial}{\partial t} \left( \hat{S}(x_i, y, t) \hat{U}(x_i, y, t) \right) + \frac{\partial}{\partial y} F_2(\hat{U}(x_i, y, t)) = 0 \quad \text{on } \mathbb{R} \times [t^n, t^{n+1}],$$

$$\hat{U}(x_i, y, t^n) = U_{i,j+\frac{1}{2}}^{n,2}(y) \quad \text{for } y \in \mathbb{R},$$

where

$$U_{i,j+\frac{1}{2}}^{n,2}(y) := \begin{cases} U_{ij}^n & , y < y_{j+\frac{1}{2}} \\ U_{i,j+1}^n & , y > y_{j+\frac{1}{2}}. \end{cases}$$

**Derivation of the HLL flux**

In the following we will derive the HLL flux for the above Riemann problems. Due to their similarity we do this exemplarily for the Riemann problem (4.39). The translation of the HLL flux to (4.40) is then straightforward. We base the following discussion on [37].

To simplify the notation let us shift the coordinate system in (4.39) such that $x_{i+\frac{1}{2}} = 0$ and $t^n = 0$ and drop the $y$ variable from the notation, i.e. we define

$$\hat{U}(x, t) := \hat{U}(x + x_{i+\frac{1}{2}}, y_j, t + t^n), \quad \hat{S}(x, t) := \hat{S}(x + x_{i+\frac{1}{2}}, y_j, t + t^n).$$

Furthermore we use the following notation for the initial states of the variables $\hat{S}, \hat{U}$:

$$U_L := U_{ij}^n, \quad U_R := U_{i+1,j}^n,$$

$$S_L := S_{ij}^n, \quad S_R := S_{i+1,j}^n.$$  

(4.41)

Now let us denote with $s_L, s_R$ the fastest characteristic speeds occurring in (4.39), in the sense that the characteristic with speed $s_L$ is left-most and the characteristic with speed $s_R$ is right-most with respect to all characteristics that have their origin at $x = 0$. Then let $T > 0$ and $x_L < x_R$ such that $x_L \leq Ts_L$ and $x_R \geq Ts_R$. 

89
The basic principle behind the HLL flux is the simplification of the Riemann problem (4.39) to three different states of the solution $\hat{S} \hat{U}$: the state $S_L U_L$ to the left side of the characteristic $\dot{x} = s_L$, the state $S_R U_R$ to the right side of the characteristic $\dot{x} = s_R$ and an intermediate state $U^*_C$ between the two characteristics (see Figure 4.1). To approximate the flux at this intermediate state, we will first derive an expression for $U^*_C$ by determining the mean value of $\hat{S}(\cdot,T)\hat{U}(\cdot,T)$ on the intermediate region $[T_{SL}, T_{SR}]$:

$$U^*_C := \frac{1}{T_{SR} - T_{SL}} \int_{T_{SR}}^{T_{SL}} \hat{S}(x,T)\hat{U}(x,T) \, dx. \quad (4.42)$$

After that we discuss how the flux at this state has to be chosen such that it fulfills the Rankine-Hugoniot condition (4.11) for $s = s_L$ and $s = s_R$.

Analogous to (4.4), we receive the following consistency condition by in-
4.3. A FINITE VOLUME SCHEME BASED ON THE HLL FLUX

Integrating (4.38) over the control volume \([x_L, x_R] \times [0, T]\):

\[
\int_{x_L}^{x_R} \tilde{S}(x,T) \tilde{U}(x,T) \, dx = \int_{x_L}^{x_R} \tilde{S}(x,0) \tilde{U}(x,0) \, dx + \int_0^T F_1(\tilde{U}(x_L,t)) \, dt - \int_0^T F_1(\tilde{U}(x_R,t)) \, dt.
\]

(4.43)

Using the initial conditions of the Riemann problem (4.39) as denoted in (4.41), we obtain

\[
\int_{x_L}^{x_R} \tilde{S}(x,0) \tilde{U}(x,0) \, dx = \int_{x_L}^{x_R} \tilde{S}(x,0) \tilde{U}(x,0) \, dx + \int_0^{x_L} S_L U_L \, dx + \int_0^{x_R} S_R U_R \, dx
\]

\[
= x_R S_R U_R - x_L S_L U_L.
\]

(4.44)

Here we have used (4.20), which implies that \(\tilde{S}(x,0) \equiv S^n_{ij} = S_L \) on \([x_L, 0]\) and \(\tilde{S}(x,0) \equiv S^n_{i+1,j} = S_R \) on \([0, x_R]\).

Let us define:

\[
F_{L/R} := \frac{1}{T} \int_0^T F_1(\tilde{U}(x_L/R,t)) \, dt
\]

(4.45)

Then

\[
\int_0^T F_1(\tilde{U}(x_L/R,t)) \, dt = T F_{L/R}.
\]

(4.46)

Combining (4.43), (4.44) and (4.46) yields:

\[
\int_{x_L}^{x_R} \tilde{S}(x,T) \tilde{U}(x,T) \, dx = x_R S_R U_R - x_L S_L U_L + T(F_L - F_R).
\]

(4.47)

Due to the construction of \(s_{L/R}\), to the left/right side of \(\dot{x} = s_{L/R}\) only \(S_{L/R} U_{L/R}\) is propagated. Due to the choice \(x_L \leq T s_L\) and \(x_R \geq T s_R\) this
imply
\[ \bar{S}(x_{L/R}, t) \bar{U}(x_{L/R}, t) \equiv S_{L/R} U_{L/R} \text{ for } t \in [0, T]. \] (4.48)

Hence splitting the integral on the left-hand side of (4.43) gives the following expression:
\[ \int_{x_L}^{x_R} \bar{S}(x, T) \bar{U}(x, T) \, dx = \int_{x_L}^{T_{s_L}} \bar{S}(x, T) \bar{U}(x, T) \, dx + \int_{T_{s_L}}^{T_{s_R}} \bar{S}(x, T) \bar{U}(x, T) \, dx + \int_{T_{s_R}}^{x_R} \bar{S}(x, T) \bar{U}(x, T) \, dx \] (4.49)

Equalizing (4.47) and (4.49) gives:
\[ \int_{T_{s_L}}^{T_{s_R}} \bar{S}(x, T) \bar{U}(x, T) \, dx = (T_{s_L} - x_L) S_L U_L + (x_R - T_{s_R}) S_R U_R + \int_{T_{s_L}}^{T_{s_R}} \bar{S}(x, T) \bar{U}(x, T) \, dx. \]

Dividing by \( T(s_R - s_L) \) finally yields an expression for the mean of \( \bar{S}(\cdot, T) \bar{U}(\cdot, T) \) on \([T_{s_L}, T_{s_R}]\) as defined in (4.42):
\[ U_C^* = s_R S_R U_R - s_L S_L U_L + F_R - F_L. \] (4.50)

To find an expression \( F_C \) for the flux \( F_1 \) at the intermediate state \( U_C^* \), we evaluate the Rankine-Hugoniot condition (4.11) across either of the characteristics with the fastest speed, e.g. \( \dot{x} = s_L \):
\[ F_C - F_L = s_L (U_C^* - S_L U_L). \]
Replacing $U_C^*$ by the right-hand side of (4.50) gives:

$$F_C = s_L \left( \frac{s_RS_RU_R - s_LS_LU_L + F_L - F_R - s_LU_L}{s_R - s_L} \right) + F_L$$

$$= s_LS_RS_RU_R - s_L^2S卢L + s_LF_L - s_LF_R - s_L(s_R - s_L)S卢L + (s_R - s_L)F_L$$

$$= \frac{s_RF_L - s_LF_R + s_Ls_R(S卢R - S卢L)}{s_R - s_L}. \quad (4.51)$$

To approximate the mathematical flux in (4.27), we decide which one of the three states $S卢L, U_C^*, S卢R$ is propagated along the characteristic with $\dot{x} = 0$. in the simplified Riemann problem that has only three different states (see Figure 4.1). If $s_L \geq 0$ then $\dot{x} = 0$ is to the left of the characteristic $\dot{x} = s_L$ and thus we choose $F_{HLL} = F_L$. If $s_R \leq 0$ then $\dot{x} = 0$ is to the right of the characteristic $\dot{x} = s_R$ and thus we choose $F_{HLL} = F_R$. Provided that $s_L < 0 < s_R$, then $\dot{x} = 0$ lies within the intermediate region between $\dot{x} = s_L$ and $\dot{x} = s_R$ and we choose $F_{HLL} = F_C$.

Based on these considerations, the HLL flux is defined as:

$$F_{HLL} = \begin{cases} 
F_L, & s_L \geq 0 \\
\frac{s_RF_L - s_LF_R + s_Ls_R(S卢R - S卢L)}{s_R - s_L}, & s_L < 0 < s_R \\
F_R, & s_R \leq 0
\end{cases} \quad (4.51)$$

To obtain an explicit expression for $F_{HLL}$ we still need to evaluate the mean fluxes in (4.45). The problem is that (4.48) does not directly imply that $\tilde{U}(x_{L/R}, \cdot)$ is constant, which would directly imply

$$F_{L/R} = F_1(U_{L/R}).$$

We still suggest the use of this value, which can be interpreted as a first order approximation in the sense that we assume $S(x_{L/R}, t) \equiv S_{L/R}$ for $t \in [0, T]$ at this point.
CHAPTER 4. NUMERICAL TREATMENT OF THE M1 MODEL

Application of the HLL flux to the two-dimensional problem

Separately replacing the mathematical fluxes $\hat{F}_{i+\frac{1}{2},j}^{n+\frac{1}{2}}$ and $\hat{G}_{i,j+\frac{1}{2}}^{n+\frac{1}{2}}$ in (4.32) with the numerical HLL flux (4.51) yields:

\[
F_{i+\frac{1}{2},j}^{n+\frac{1}{2}} := \begin{cases} 
F_1(U_{i,j}^n) & , s_L > 0 \\
\frac{s_R F_1(U_{i+1,j}^n) - s_L F_1(U_{i+1,j}^n) + s_L s_R (S_{i+1,j}^n U_{i+1,j}^n - S_{i,j}^n U_{i,j}^n)}{s_R - s_L} & , s_L \leq 0 \leq s_R \\
F_1(U_{i+1,j}^n) & , s_R < 0 
\end{cases}
\]

\[
G_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} := \begin{cases} 
F_2(U_{i,j}^n) & , s_L > 0 \\
\frac{s_R F_2(U_{i,j+1}^n) - s_L F_2(U_{i,j+1}^n) + s_L s_R (S_{i,j+1}^n U_{i,j+1}^n - S_{i,j}^n U_{i,j}^n)}{s_R - s_L} & , s_L \leq 0 \leq s_R \\
F_2(U_{i,j+1}^n) & , s_R < 0 
\end{cases}
\]

Note that this is not equivalent to the direct approximation of a two-dimensional Riemann problem. However, on numerical grids where the faces of the grid cells are aligned with the flux directions (which is true in our case due to the choice of a rectangular grid and $x$ and $y$ fluxes), such a flux separation yields a first order accurate method [24, p. 201].

4.3.4 The final numerical scheme

Finally we replace $\hat{T}_{i,j}^{n+\frac{1}{2}}, \hat{F}_{i+\frac{1}{2},j}^{n+\frac{1}{2}}$ and $\hat{G}_{i,j+\frac{1}{2}}^{n+\frac{1}{2}}$ in (4.31) according to (4.33) and (4.52) and again divide by $S_{i,j}^{n+1}$ to receive the following explicit time update scheme for the solution $\hat{U}$ of (4.17):

\[
U_{i,j}^{n+1} = U_{i,j}^n + \frac{\Delta t^n}{\Delta x} \left( \frac{F_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - F_{i-\frac{1}{2},j}^{n+\frac{1}{2}}}{S_{i,j}^{n+1}} \right) - \frac{\Delta t^n}{\Delta y} \left( \frac{G_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} - G_{i,j-\frac{1}{2}}^{n+\frac{1}{2}}}{S_{i,j}^{n+1}} \right) + \left( \frac{S_{i,j}^n}{S_{i,j}^{n+1}} - 1 \right) U_{i,j}^n - \Delta t^n \frac{T_{ij}^n}{S_{ij}^{n+1}} U_{i,j}^n,
\]
or

\[
U_{ij}^{n+1} = U_{ij}^n - \frac{\Delta t^n}{\Delta x} \left( \frac{F_{i+1/2,j}^{n+1} - F_{i-1/2,j}^{n+1}}{S_{ij}^{n+1}} \right) - \frac{\Delta t^n}{\Delta y} \left( \frac{G_{i+1/2,j}^{n+1} - G_{i-1/2,j}^{n+1}}{S_{ij}^{n+1}} \right) - \Delta t^n \left( \frac{T_{ij}^n}{S_{ij}^{n+1}} I - \left( I + \Delta t^n T_{ij}^n \right) \right) U_{ij}^n
\]

(4.53)

where \( I = \text{diag}(1, \ldots, 1) \in \mathbb{R}^3 \) denotes the unity matrix. Note that a stable choice of the time step \( \Delta t^n \) and an approximation of the signal speeds \( s_L, s_R \) in (4.52) remain to be discussed.

### 4.4 Characteristic velocities and the CFL condition

In this section we discuss the estimation of the characteristic velocities in the Riemann problems (4.39) and (4.40).

#### 4.4.1 The spectrum of the Jacobian

As discussed in Section 4.3, the characteristic velocities in a Riemann problem can be approximated by the eigenvalues of the Jacobian of the corresponding mathematical flux. In the Riemann problem (4.39) this is given by:

\[
\frac{\partial}{\partial U^*} F_1(U) = \frac{\partial U}{\partial U^*} F_1'(U) = \frac{1}{S} F_1'(U),
\]

(4.54)

where \( U^* = SU \). We will treat the scaling with the stopping power later and first discuss the spectrum of the Jacobian itself. Let

\[
J_k(U) := \left( \frac{\partial}{\partial U_j} F_{i,k}(U) \right)_{i=1,\ldots,3}^{j=1,\ldots,3}, \text{ } k = 1, 2,
\]

(4.55)
where $F$ denotes the spatially two-dimensional version of the mathematical flux in (3.41), which is given by:

$$F(U) = \begin{pmatrix} U_2 & U_3 \\ U_1 \left( \frac{1 - \chi(|\alpha|)}{2} I + \frac{3\chi(|\alpha|) - 1}{|\alpha|^2} \alpha \otimes \alpha \right) \end{pmatrix},$$

$$= \begin{pmatrix} U_2 \\ \frac{1 - \chi(|\alpha|)}{2} U_1 + \frac{\alpha_1^2 \chi(|\alpha|) - 1}{|\alpha|^2} U_1 \\ \frac{\alpha_1 \alpha_2 \chi(|\alpha|) - 1}{|\alpha|^2} U_1 \end{pmatrix},$$

where again

$$\alpha = \begin{pmatrix} U_2/U_1 \\ U_3/U_1 \end{pmatrix},$$

denotes the anisotropy parameter, that was introduced in (3.42). Before we address the explicit calculation of the Jacobian, we prove that the symmetry in $F$ yields a direct relation between $J_1$ and $J_2$.

Lemma 3 (Symmetry of the mathematical flux).

Let $U \in \mathbb{R}^3$ arbitrary. Then

$$J_2(U_1, U_2, U_3) = J_1(U_1, U_3, U_2)$$

holds.

Proof. Due to the symmetry between $F_1(U)$ and $F_2(U)$, one can construct $F_2(U)$ by permuting the second and third row in $F_1(U)$ and then replacing $U_2$ by $U_3$ and vice versa, which, by definition, also replaces $\alpha_1$ by $\alpha_2$ and vice versa. This means:

$$F_2(U) = RF_1(RU),$$

where

$$R := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$
4.4. CHARACTERISTIC VELOCITIES AND THE CFL CONDITION

Note that \( R^2 = \text{diag}(1, 1, 1) \). This implies:

\[
J_2(U) = DF_2(U) = DRF_1(RU) = R^2DF_1(U)_{|U=RU} = J_1(RU).
\]

Lemma 3 implies that the spectrum of \( J_2 \) can be obtained by rotating the spectrum of \( J_1 \) by 90 degrees in the \( U_2 - U_3 \) plane. We can therefore restrict the following discussion to \( J_1 \) without loss of generality.

Using the chain rule on \( \chi \) one can calculate an analytical expression for \( J_1(U) \). However, this calculation is very technical and we will just give the result here:

\[
J_1(U) = \left( \frac{\partial}{\partial U_j} F_{i,1}(U) \right)_{j=1,\ldots,3} = \begin{pmatrix}
0 & \frac{1-t_2(|\alpha|)}{2} + \frac{\alpha_1^2}{|\alpha|^2} \frac{3\tau_2(|\alpha|)-1}{2} & \frac{\alpha_1 \alpha_2}{|\alpha|^2} \frac{3\tau_2(|\alpha|)-1}{2} \\
\frac{\alpha_1}{|\alpha|} \left( t_4(|\alpha|) + \alpha_1^3 \frac{3\tau_2(|\alpha|)-1}{2} \right) & \frac{\alpha_2}{|\alpha|^2} \frac{3\tau_2(|\alpha|)-1}{2} & \frac{\alpha_2}{|\alpha|^2} \left( t_1(|\alpha|) - \frac{\alpha_2^2}{|\alpha|^2} t_3(|\alpha|) \right) \\
0 & \frac{\alpha_2}{|\alpha|^2} \left( -\chi'(|\alpha|) \right) - \frac{\alpha_2^2}{|\alpha|^3} t_3(|\alpha|) & \frac{\alpha_1}{|\alpha|^2} \left( t_1(|\alpha|) - \frac{\alpha_2^2}{|\alpha|^2} t_3(|\alpha|) \right)
\end{pmatrix}^T,
\]

where

\[
t_1(|\alpha|) := \frac{3\chi(|\alpha|) - 1}{2},
\]

\[
t_2(|\alpha|) := \chi(|\alpha|) - |\alpha|\chi'(|\alpha|),
\]

\[
t_3(|\alpha|) := \frac{6\chi(|\alpha|) - 3|\alpha|\chi'(|\alpha|) - 2}{2},
\]

\[
t_4(|\alpha|) := \frac{\alpha_1^2}{|\alpha|^2} \frac{3\chi'(|\alpha|)}{2} - \frac{\chi'(|\alpha|)}{2}.
\]

For a detailed derivation of this expression, see Appendix B. Note that \( J_1(U) \) does not depend on \( U_1 \) or \( U_2, U_3 \) explicitly, but only on the anisotropy parameter \( \alpha \). Due to Lemma 3 the same holds for \( J_2(U) \). In view of this the notations \( J_k(\alpha) \) and \( J_k(U) \) can be used equivalently and Lemma 3 directly
implies
\[ J_2(\alpha_1, \alpha_2) = J_1(\alpha_2, \alpha_1). \] (4.57)

**Bounds for the anisotropy parameter**

We will now prove bounds for \( \alpha \) that enable us to visualize the complete spectrum of \( J_1 \).

**Lemma 4** (Realizability condition).

*Let \( \psi(x, \epsilon, \Omega) \geq 0 \) for all \((x, \epsilon, \Omega) \in Q \times (0, \infty) \times S^2\). Then*

\[ \psi(0)(x, \epsilon) \geq 0 \quad \text{for all } (x, \epsilon) \in Q \times (0, \infty), \] (4.58)
\[ |\alpha| = \frac{|\psi(1)(x, \epsilon)|}{|\psi(0)(x, \epsilon)|} \leq 1 \quad \text{for all } (x, \epsilon) \in Q \times (0, \infty). \] (4.59)

**Proof.** Let \( x \in Q \) and \( \epsilon \in (0, \infty) \) arbitrary. Then by assumption \( \psi(x, \epsilon, \cdot) \geq 0 \) on \( S^2 \). This directly implies
\[ \psi(0)(x, \epsilon) = \int_{S^2} \psi(x, \epsilon, \Omega) d\Omega \geq 0. \]

Now let \( \Omega' \in S^2 \) arbitrary. Then \( 1 - \Omega \cdot \Omega' \geq 0 \) for \( \Omega \in S^2 \). This implies
\[ 0 \leq \int_{S^2} (1 - \Omega \cdot \Omega') \psi(x, \epsilon, \Omega) d\Omega = \psi(0)(x, \epsilon) - \psi(1)(x, \epsilon) \cdot \Omega'. \]

Now we choose \( \Omega' := \psi(1)(x, \epsilon) / |\psi(1)(x, \epsilon)| \), which is in \( S^2 \) due to the normalization. Using the positivity of \( \psi(0)(x, \epsilon) \), we finally receive
\[
1 \geq \frac{\psi(1)(x, \epsilon) \cdot \Omega'}{\psi(0)(x, \epsilon)} = \frac{\psi(1)(x, \epsilon) \cdot \psi(1)(x, \epsilon)}{|\psi(1)(x, \epsilon)| |\psi(0)(x, \epsilon)|} \frac{1}{|\psi(0)(x, \epsilon)|} = \frac{|\psi(1)(x, \epsilon)|}{|\psi(0)(x, \epsilon)|} = |\alpha|.
\]

Lemma 4 implies that the total range of the eigenvalues of \( J_1 \) (and \( J_2 \)) can be obtained by calculating the eigenvalues of \( J_k(\alpha) \) for \( \alpha \in \mathbb{R}^2 \) with
4.4. CHARACTERISTIC VELOCITIES AND THE CFL CONDITION

\[ |\alpha| \leq 1. \]

Explicit approximation of the Eddington factor

In order to retrieve explicit numerical values of the eigenvalues of \( J_k(\alpha) \), we have to be able to evaluate the Eddington factor \( \chi \) and its derivative. The Eddington factor could be approximated by first solving (3.39) for \(|a_1|\) for a given \( \alpha \) using a Newton iteration and then evaluating (3.38). However this method cannot be used to evaluate \( \chi' \). Furthermore it results in relatively high computational costs in view of the necessity to approximate the eigenvalues and thus \( \chi \) and \( \chi' \) on each cell in each time step of the scheme (4.53).

We therefore follow the approach in [10] and approximate the Eddington factor by a rational function. The advantage is that a rational function can both be efficiently evaluated and differentiated. Let

\[
\hat{\chi}(|\alpha|) := \frac{a_6|\alpha|^6 + a_4|\alpha|^4 + a_2|\alpha|^2 + a_0}{|\alpha|^4 + b_2|\alpha|^2 + b_0}.
\] (4.60)

Then

\[
\hat{\chi}'(|\alpha|) = \frac{6a_6|\alpha|^5 + 4a_4|\alpha|^3 + 2a_2|\alpha|}{|\alpha|^4 + b_2|\alpha|^2 + b_0} - \frac{(a_6|\alpha|^6 + a_4|\alpha|^4 + a_2|\alpha|^2 + a_0) \cdot (4|\alpha|^3 + 2b_2|\alpha|)}{(|\alpha|^4 + b_2|\alpha|^2 + b_0)^2}.
\]

Now we choose equally spaced points \( \alpha_1, \ldots, \alpha_m \in [-1, 1] \setminus \{0\} \) for some \( m \in \mathbb{N} \) and calculate the data set \( \{(\alpha_i, \chi(\alpha_i)) \mid i = 1, \ldots, m\} \) using a Newton iteration to evaluate \( \chi \) as described above. The coefficients \( a_0, \ldots, a_6, b_0, b_2 \) of \( \hat{\chi} \) are then optimized using an algorithm similar to Algorithm 3.

We have performed this optimization in Mathematica using \( N = 1000 \), re-
resulting in the following coefficients:

\[
\begin{align*}
    a_0 &= 0.621529, & a_2 &= 0.348509, \\
    a_4 &= -0.139318, & a_6 &= 0.720371, \\
    b_0 &= 1.87095, & b_2 &= -1.32002.
\end{align*}
\]

In Figure 4.2 the approximate Eddington factor \( \hat{\chi} \) is compared with the corresponding data set from the newton iteration. The results show a precision of about \( 10^{-4} \). Let \( \lambda_{1k}(\alpha), \lambda_{2k}(\alpha), \lambda_{3k}(\alpha) \) denote the eigenvalues of \( J_k(\alpha) \) for \( k = 1, 2 \). Assuming that all eigenvalues are real, we choose the indexing such that \( \lambda_{1k}(\alpha) \leq \lambda_{2k}(\alpha) \leq \lambda_{3k}(\alpha) \). We have calculated the eigenvalues of \( J_1 \) using the analytical expression of the Jacobian calculated in Appendix B and the explicit approximation (4.60) of the Eddington factor. The results are plotted in Figure 4.3. They indicate that

\[
\lambda_{i1}(\alpha) \in \mathbb{R} \text{ and } |\lambda_{i1}(\alpha)| \leq 1 \text{ for all } \alpha \in \mathbb{R}^2, \ |\alpha| \leq 1 \text{ and } i = 1, 2, 3. \quad (4.61)
\]

Figure 4.3(d) also shows the spectrum of \( J_1 \) for \( \alpha \in [-1, 1] \times 0 \). The according graph agrees well with the results given in [7, 10]. The property (4.61) can be generally proven for the \( M1 \) model. For a detailed theoretical discussion, we refer to [25], where extensive theory on the structure of moment closures for kinetic equations is presented, along with a proof of the hyperbolicity of entropy moment closures. In particular, this means that the system (4.17) is hyperbolic.

### 4.4.2 Approximate bounds for the signal speeds

Now we derive approximate bounds for the characteristic velocities in the Riemann problem (4.39). As discussed in Section 4.3, the eigenvalues of \( J_1(\alpha) \) determine the velocities of the characteristics in the corresponding linearized Riemann problem. However, due to (4.54) we have to take into account the scaling with the stopping power. In the Riemann problem there are two reasonable points around which \( F \) can be expanded into its Taylor expansion at \( t = t^n \) on each cell, namely the left and right initial state
Figure 4.2: A rational function fit to the Eddington factor
Figure 4.3: Eigenvalues of the Jacobian of the mathematical flux $U^n_{ij}, U^n_{i+1,j}$. Therefore we suggest the following bounds for the signal speeds in the HLL flux (4.52):

$$s_L = s^{n}_{i+\frac{1}{2},j,L} := \frac{1}{S^{n+1}_{ij}} \min \{ \lambda_{11}(\alpha^n_{ij}), \lambda_{11}(\alpha^n_{i+1,j}) \},$$

$$s_R = s^{n}_{i+\frac{1}{2},j,R} := \frac{1}{S^{n+1}_{i+1,j}} \max \{ \lambda_{31}(\alpha^n_{ij}), \lambda_{31}(\alpha^n_{i+1,j}) \},$$

for the flux in $x$ direction, and

$$s_L = s^{n}_{i,j+\frac{1}{2},L} := \frac{1}{S^{n+1}_{ij}} \min \{ \lambda_{12}(\alpha^n_{ij}), \lambda_{12}(\alpha^n_{i+1,j}) \},$$

$$s_R = s^{n}_{i,j+\frac{1}{2},R} := \frac{1}{S^{n+1}_{i,j+1}} \max \{ \lambda_{32}(\alpha^n_{ij}), \lambda_{32}(\alpha^n_{i+1,j}) \},$$
for the flux in $y$ direction, where
\[
\alpha_{ij}^n := \begin{pmatrix} U_{ij,2}^n / U_{ij,1}^n \\ U_{ij,3}^n / U_{ij,1}^n \end{pmatrix}.
\]

Note that without loss of generality we assume the sorting $\lambda_{1j}(\alpha) \leq \lambda_{2j}(\alpha) \leq \lambda_{3j}(\alpha)$ of the eigenvalues for all $\alpha \in \mathbb{R}^2$ with $|\alpha| \leq 1$ here.

### 4.4.3 Stable choice of the time step

The CFL condition corresponding to (4.53) is a multi-dimensional version of (4.10). Again the characteristic velocities are given by the eigenvalues of $J_1(\alpha)$ scaled by the inverse of $S_{ij}^{n+1}$. Let $\sigma(U)$ denote the union of the spectra of $J_1(U)$ and $J_2(U)$. Then the following CFL condition guarantees that the characteristics of the linearized Riemann problems at $t = t^n$ do not intersect:

\[
\nu = \frac{\Delta t^n}{\Delta x} \max \{ \frac{|\lambda|}{S_{ij}^{n+1}} \mid \lambda \in \sigma(\bar{U}(x, y, t^n)) \} \leq 1.
\]

Let
\[
S_{n,\max} = \max_{i=1, \ldots, N_x, j=1, \ldots, N_y} \left\{ \frac{|\lambda_{11}(\alpha_{ij}^n)|}{S_{ij}^{n+1}}, \frac{|\lambda_{31}(\alpha_{ij}^n)|}{S_{ij}^{n+1}}, \frac{|\lambda_{12}(\alpha_{ij}^n)|}{S_{ij}^{n+1}}, \frac{|\lambda_{32}(\alpha_{ij}^n)|}{S_{ij}^{n+1}} \right\}
\]

This can be reformulated in terms of an explicit choice of the step size $t^n$:

\[
\Delta t^n := \frac{\nu \max \{ \Delta x, \Delta y \}}{S_{n,\max}^n},
\]

where $\nu$ is the so-called CFL number. Choosing $\nu < 1$ creates a safety margin that increases the numerical stability in the sense that even the fastest characteristics retain a positive distance to the cell boundaries. This is especially important in our case, because the bounds for the one-dimensional Riemann problems do not necessarily need to be bounds for the characteristics in the full two-dimensional Riemann problem. It can be shown that a choice of $\nu \leq \frac{1}{2}$ is required to obtain a numerically stable time update in our case [24, p. 201].
4.5 Summary and practical issues

In this section we briefly introduce the concept of ghost cells and present a pseudocode that summarizes the numerical scheme.

If we set \( i = N_x \) and assume \( s_L \leq 0 \) then the value \( U_{N_x+1,j}^n \) is required to calculate \( F_{i+\frac{1}{2},j}^{n+\frac{1}{2}} \) in the update (4.53) for \( U_{ij}^{n+1} \). This value is not defined since the index \( N_x + 1 \) is outside of the numerical grid. The same problem arises for all cells that have at least one edge that is on the boundary of \( Q \). This problem can be resolved by either imposing boundary conditions that define the fluxes over these edges directly, or explicit values for the solution on imaginary cells, that are directly adjacent to and outside of the numerical grid, have to be defined. Due to the fact that these additional cells are introduced artificially, they are called ghost cells. In this work we have used a simple constant extension to define the values on the ghost cells:

\[
\begin{align*}
U_0^n, j &= U_1^n, j, & U_{N_x+1}^n, j &= U_{N_x}^n, j, & j &= 1, \ldots, N_y, \\
U_i^0 &= U_{i,1}^n, & U_{i,N_y+1}^n &= U_{i,N_y}^n, & i &= 1, \ldots, N_x.
\end{align*}
\]

Now the finite volume scheme can be summarized as follows:

\[\text{Algorithm 5 (Finite volume scheme).}\]

\[\text{Given a beam energy } \epsilon_b, \text{ a cut-off-energy } \epsilon_c, \text{ constants } c_1, c_2 \text{ for the dimensions of the electron beam in the initial condition (4.15), grid dimensions } N_x, N_y, \text{ a corresponding numerical grid } Q_{ij} \text{ as in (4.19) and a partition } T \text{ of } Q \text{ for the different material regions as in (1.22), which is aligned with the grid } Q_{ij}, \text{ compute:}\]

\[\text{function } \text{FiniteVolumeScheme}(\epsilon_b, \epsilon_c, c_1, c_2, x_l, x_r, y_l, y_r, Q, N_x, N_y, T)\]

\[\text{for } i = 1, \ldots, N_x \text{ do}\]

\[\text{for } j = 1, \ldots, N_y \text{ do}\]

\[\text{Initialize}\]

\[U_i^0 \leftarrow \hat{U}_i^0(x_i, y_j)\]

\[S_i^0 \leftarrow S(x_i, y_j, \epsilon_b)\]

\[T_i^0 \leftarrow T(x_i, y_j, \epsilon_b)\]

\[\text{end for}\]
end for 
\[ t \leftarrow 0 \]
\[ n \leftarrow 0 \]
while \( t < \epsilon_b - \epsilon_c \) do 
\[ S_{\text{max}}^n \leftarrow \max_{i=1,\ldots,N_x} \{ \frac{|\lambda_{11}(\alpha_{ij}^n)|}{S_{ij}^{n+1}}, \frac{|\lambda_{31}(\alpha_{ij}^n)|}{S_{ij}^{n+1}}, \frac{|\lambda_{12}(\alpha_{ij}^n)|}{S_{ij}^{n+1}}, \frac{|\lambda_{32}(\alpha_{ij}^n)|}{S_{ij}^{n+1}} \} \]
\[ \Delta t^n \leftarrow \frac{\nu \max(\Delta x, \Delta y)}{S_{\text{max}}^n} \]  \[ t^{n+1} \leftarrow t^n + \Delta t^n \]  \( \triangleright \) Update time step 
for \( j = 1,\ldots,N_y \) do  
\[ U_0^n, j \leftarrow U_1^n, j \]  
\[ U_{N_x+1}^n, j \leftarrow U_{N_x}^n, j \]  \( \triangleright \) Update left and right ghost cells 
end for 
for \( i = 1,\ldots,N_x \) do  
\[ U_{i, 0}^n \leftarrow U_{i, 1}^n \]  
\[ U_{i, N_y+1}^n \leftarrow U_{i, N_y}^n \]  \( \triangleright \) Update upper and lower ghost cells 
end for 
for \( i = 1,\ldots,N_x \) do 
for \( j = 1,\ldots,N_y \) do  
\( \triangleright \) Update material coefficients 
\[ S_{ij}^{n+1} \leftarrow \hat{S}(x_i, y_j, t^{n+1}) \]  
\[ T_{ij}^{n+1} \leftarrow \hat{T}(x_i, y_j, t^{n+1}) \] 
end for 
end for 
for \( i = 1,\ldots,N_x \) do 
for \( j = 1,\ldots,N_y \) do  
\( \triangleright \) Perform time update 
\[ U_{ij}^{n+1} \leftarrow U_{ij}^n - \frac{\Delta t^n}{\Delta x} \left( \frac{F_{ij}^{n+\frac{1}{2}}}{S_{ij}^{n+1}} - \frac{F_{ij}^{n-\frac{1}{2}}}{S_{ij}^{n+1}} \right) - \frac{\Delta t^n}{\Delta y} \left( \frac{G_{ij}^{n+\frac{1}{2}}}{S_{ij}^{n+1}} - \frac{G_{ij}^{n-\frac{1}{2}}}{S_{ij}^{n+1}} \right) \]
\[ + \left( \frac{S_{ij}^n}{S_{ij}^{n+1}} - 1 \right) U_{ij}^n - \Delta t^n \frac{T_{ij}^n}{S_{ij}^n} U_{ij}^n \] 
end for 
end for 
\( n \leftarrow n + 1 \) 
end while 
end function
Note that in each iteration only $U_{ij}^{n+1}$ and $U_{ij}^n$ is used for $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$. Thus for efficient memory handling only two two-dimensional arrays $U_{ij}^{\text{new}}$ and $U_{ij}^{\text{old}}$ should be used in a real implementation. The same goes for the material coefficients $S_{ij}^n$ and $T_{ij}^n$. Furthermore note that $F_{(i+1)-\frac{1}{2},j}^{n+\frac{1}{2}} = F_{i+\frac{1}{2},j}^{n+\frac{1}{2}}$. The same goes for the flux in $y$-direction. Thus the fluxes over all edges in the interior of the domain are computed twice in Algorithm 5. These unnecessary calculations should be avoided in a real implementation by storing the according flux in a temporary variable and reusing it when the same flux is required again.

4.6 Numerical results

In this section we first show the numerical convergence behavior of the scheme (4.53) for simple test cases. Then we present a comparison between the numerical solution of the $\mathcal{M}1$ model and Monte Carlo simulations for a homogeneous material and a simple heterogeneous geometry.

4.6.1 Order of convergence of the finite volume scheme

Conservative finite volume schemes like (4.53) are first order accurate (cf. [24, p. 201]), i.e. if the mesh size is decreased by a certain factor, the error between the numerical and the exact solution of the partial differential equation decreases approximately by the same factor. To test our implementation against this expectation, we have performed two numerical experiments.

Linear advection equation

First we have replaced the mathematical flux (4.56) in our implementation by the linear flux:

$$F(U) := \begin{pmatrix}
    a_{11} U_1 & a_{12} U_1 \\
    a_{21} U_2 & a_{22} U_2 \\
    a_{31} U_3 & a_{32} U_3
\end{pmatrix}$$

(4.63)
with
\[ a_{11} = 1 \quad a_{12} = 1 \\
\quad a_{21} = 1 \quad a_{22} = 0 \\
\quad a_{31} = 0 \quad a_{32} = -1. \]

This allows us to derive an explicit expression for the exact solution that the numerical solution can be compared to. Let
\[ \hat{U}_i(x, y, t) := \hat{U}^0_i(x - a_{i1}t, y - a_{i2}t). \]  
(4.64)

Then
\[
\begin{align*}
\frac{\partial}{\partial t} \hat{U}_i(x, y, t) + \frac{\partial}{\partial x} F(\hat{U}(x, y, t)) + \frac{\partial}{\partial y} F(\hat{U}_i(x, y, t)) &= \frac{\partial}{\partial t} \hat{U}^0_i(x - a_{i1}t, y - a_{i2}t) \\
& \quad + A_{i1} \frac{\partial}{\partial x} \hat{U}^0_i(x - a_{i1}t, y - a_{i2}t) + a_{i2} \frac{\partial}{\partial y} \hat{U}^0_i(x - a_{i1}t, y - a_{i2}t) \\
& = - A_{i1} \frac{\partial}{\partial x} \hat{U}^0_i(x - a_{i1}t, y - a_{i2}t) - a_{i2} \frac{\partial}{\partial x} \hat{U}^0_i(x - a_{i1}t, y - a_{i2}t) \\
& \quad + A_{i1} \frac{\partial}{\partial x} \hat{U}^0_i(x - a_{i1}t, y - a_{i2}t) + a_{i2} \frac{\partial}{\partial y} \hat{U}^0_i(x - a_{i1}t, y - a_{i2}t) \\
& = 0.
\end{align*}
\]

Hence \( \hat{U} \) in (4.64) solves (4.17) with \( F \) as in (4.63), \( \hat{S} \equiv 1 \) and \( \hat{T} \equiv 0 \). We have chosen the domain \( \hat{Q} = [0, 1] \times [0, 1] \) and the initial condition
\[ \hat{U}^0_i(x, y) := \sin(2\pi x) + \cos(2\pi x), \quad (x, y) \in Q, \quad i = 1, \ldots, 3. \]

We have implemented periodic boundary conditions for this test:
\[
\begin{align*}
U^0_{0,j} &:= U^0_{N_x,j}, \quad U^0_{N_x+1,j} := U^0_{1,j}, \quad j = 1, \ldots, N_y, \\
U^0_{i,0} &:= U^0_{i,N_y}, \quad U^0_{i,N_y+1} := U^0_{i,1}, \quad i = 1, \ldots, N_x.
\end{align*}
\]
Figure 4.4: Convergence behavior of the finite volume scheme in the case of linear advection: $L^1$ error of the numerical solution on different meshes (blue solid line), first order convergence (red dashed line).

Then due to $a_{ij} \in \{-1, 1\}$ for $j = 1, 2$ and $i = 1, 2, 3$, we receive

$$
\hat{U}_i(x, y, 1) = \hat{U}_i^0(x - a_{i1}, y - a_{i2}) = U_i^0(x - a_{i1}, y - a_{i2})
= \sin(2\pi x - 2\pi a_{i1}) + \cos(2\pi x - 2\pi a_{i2})
= U_i^0(x, y).
$$

We have solved (4.17) from $t = 0$ to $t = 1$ using different mesh sizes $N_y = N_x \in \{16, 32, 64, 128\}$. Let $\hat{n}$ denote the value of $n$ at the end of the last iteration (where $t = 1$). In Figure 4.4 the $L^1$ error

$$
\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \int_{Q_{ij}} |\hat{U}_{ij}^n - \hat{U}(x, 1)| \, dx
= \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \int_{Q_{ij}} |\hat{U}_{ij}^n - U_i^0(x)| \, dx
$$

is plotted against the chosen mesh size. Comparing with the reference line of slope 1 with respect to the logarithmic axes, we see that the convergence rate of the scheme in this case is indeed of first order.
4.6. NUMERICAL RESULTS

Figure 4.5: Convergence behavior of the finite volume scheme in the case of the M1 equations with smooth initial conditions: $L^1$ error of the numerical solution on different meshes (blue solid line), first order convergence (red dashed line)

The M1 flux with smooth initial conditions

As a second test we have performed several simulations using the M1 mathematical flux (4.56), the domain $\hat{Q} = [-300 \text{ nm}, 300 \text{ nm}] \times [-300 \text{ nm}, 300 \text{ nm}]$, the periodic boundary conditions (4.65) and a smooth initial condition based on a Gaussian distribution in $x$ and $y$ direction:

$$U^0(x, y) = \begin{pmatrix} e^{-cx^2} e^{-cy^2} \\ 0 \\ -e^{-cx^2} e^{-cy^2} \end{pmatrix},$$

where $c = 50 \times 10^{-9}$. Because the exact solution is unknown, we have compared the numerical solutions on the meshes with $N_y = N_x \in \{16, 32, 64\}$ to a reference solution on the finest mesh $N_y = N_x = 128$ to approximate the $L^1$ error. In Figure 4.5 this $L^1$ error estimate is plotted against the chosen mesh size. Again a first order convergence is observed.
4.6.2 The $\mathcal{M}1$ solution for a homogeneous bulk

Finally we present numerical results that show the $\mathcal{M}1$ solution for a realistic test case. We have solved the Boltzmann equation (3.1) using both the numerical scheme presented in Section 4.3 and the DTSA − II [35] Monte Carlo simulation.

Material parameters

We have chosen a homogeneous bulk of copper as the material sample, i.e. $n = 1$, $E_{l_1} = \text{Cu}$. The material parameters were chosen as follows:

$$Z_1 = 29, \quad A_1 = 63.546 \times 10^{-3} / N_A \text{ kg}$$

$$\rho \equiv 8.96 \times 10^3 \text{kg m}^{-3}, \quad c_1 \equiv 1.$$  \hfill (4.66)

The bounded domain $\hat{Q} = [-300 \text{ nm}, 300 \text{ nm}] \times [-300 \text{ nm}, 0]$ was used in both computations.

Setup for the numerical scheme

For the initial value problem (4.17) we have chosen the initial condition (4.15) with

$$c_1 = 50 \times 10^{-9} / 6, \quad c_2 = 25 \times 10^{-9} / 6,$$

which results in a Gaussian with a width of roughly 50 nm in $x$-direction and 25 nm in $y$-direction. For the numerical scheme we have discretized the region $\hat{Q}$ using $N_x = N_y = 50$. We have chosen an initial energy of $\epsilon_b = 10 \text{ keV}$ and a cut-off-energy of $\epsilon_c = 6 \text{ keV}$ and set the end time to $T_{\text{end}} := \epsilon_b - \epsilon_c$. Then (4.12) implies $\epsilon(T_{\text{end}}) = \epsilon_b - (\epsilon_b - \epsilon_c) = \epsilon_c$. The initial values $U_{0,j}^n$ for the update formula (4.53) were then set to $\tilde{U}_j^0((x_i, y_j))$ for $i = 1, \ldots, N_x$, $j = 1, \ldots, N_y$. Note that in (4.53) ghost cells are necessary for the fluxes over the boundary of the domain. Here we have chosen a simple constant extension of the values from the previous time step:

$$U_{0,j}^n := U_{1,j}^n, \quad U_{N_x+1,j}^n := U_{N_x,j}^n, \quad j = 1, \ldots, N_y,$$

$$U_{i,0}^n := U_{i,1}^n, \quad U_{i,N_y+1}^n := U_{i,N_y}^n, \quad i = 1, \ldots, N_x.$$
For the stopping power we have used the expression (1.16) with the material coefficients (4.66). In Figure 4.6 the resulting function is compared with values from the ESTAR database [4], which is based on the ICRU report 37 [1].

We have neglected the inelastic transport coefficient, i.e. we have set $T_{\text{in}} \equiv 0$ in (3.25). For the elastic transport coefficient, we have replaced $\sigma_{\text{el}}(x, \epsilon, \mu)$ with the screened Rutherford cross-section (1.14) in (3.27) and
evaluated the resulting integral expression using Mathematica:

\[
2\pi N_V(x) \int_{-1}^{1} (1 - \mu)\sigma_{el}(x, \epsilon, \mu) \, d\mu = 2\pi e^4 \rho(x) \frac{1}{16(4\pi \epsilon_0)^2 \epsilon^2} \sum_{j=1}^{n} \frac{c_j(x)Z_j^2}{A_j} \int_{-1}^{1} \frac{1 - \mu}{\sin^2(\arccos(\mu)/2) + \sin^2(\theta_{0,j}(\epsilon)/2)} \, d\mu
\]

\[
= 2\pi e^4 \rho(x) \frac{1}{16(4\pi \epsilon_0)^2 \epsilon^2} \sum_{j=1}^{n} \frac{c_j(x)Z_j^2}{A_j} \cdot \left( \frac{8}{\cos(\theta_{0,j}(\epsilon)) - 3} + 4(\ln[3 - \cos(\theta_{0,j}(\epsilon))] - \ln[1 - \cos(\theta_{0,j}(\epsilon))] \right) .
\]

In Figure 4.7 this expression is compared to tabulated data that was published in [29] and is based on the ICRU report 77 [2]. Due to the author’s interest in radiotherapy applications, water was used as the background material. The calculations are based on the representation of water as a pure
material with an effective atomic number:

\[ Z_{\text{eff}} = 7.51, \quad \rho = 1000 \text{ kg m}^{-3}, \quad A = 0.01801528/N_A \text{ kg}. \]

**Setup for the Monte Carlo simulation**

In \textsc{DTSA} – II we have chosen a Gaussian beam with a diameter of 50 nm and an energy of 10 keV, as above, and simulated 100 000 electrons. Inside the code the predefined screened Rutherford model was used for the elastic scattering cross-section and the Bethe-loss formula for the stopping power. The electron number density was approximated by calculating a histogram as described in Algorithm 2, using 50 bins to discretize the energy interval \([0, 10 \text{ keV}]\) and a 50 \times 50 equidistant grid to discretize the spatial domain \(\hat{Q}\). The \textsc{DTSA} – II program is designed for three-dimensional geometries. We have decided to simply project all calculated trajectories onto the \(x\)-\(y\) plane to receive a two-dimensional geometry.

**Spatial distribution of the electron number densities**

Let \(n(x, y, \epsilon)\) and \(n_{\text{ref}}(x, y, \epsilon)\) denote the electron number densities, integrated over all angles and computed using the deterministic model and the \textsc{DTSA} – II Monte Carlo simulation respectively. In Figure 4.8 both electron number densities are plotted in the \(x\)-\(y\) plane for certain energies. Qualitatively the electron densities \(n\) and \(n_{\text{ref}}\) show a similar dynamic with changing energy.

**Energy distribution at different depths**

For a further comparison of the distribution of the electrons at different depths, we have calculated the mean distribution of the electron number density \(\int_{x_i}^{x_f} \hat{n}(x, y, \epsilon) \, dx\) at different depths and plotted them against the energy. The results are shown in Figure 4.9. Here we denote \(n, n_{\text{ref}}\) as \(\hat{n}, \hat{n}_{\text{ref}}\) to indicate that we have normalized the distribution in each plot to be able to
Figure 4.8: Spatial distribution of the $M_1$ and Monte Carlo electron number densities in a homogeneous bulk (copper) at different energies
4.6. NUMERICAL RESULTS

Figure 4.9: Normalized energy distribution of the $M_1$ and Monte Carlo electron number densities at different depths: $\int_{x_1}^{x_2} \hat{n}(x,y,\epsilon) \, dx$ (red dashed line), $\int_{x_1}^{x_2} \hat{n}_{\text{ref}}(x,y,\epsilon) \, dx$ (blue solid line)

directly compare both densities. While this does not allow for a quantitative comparison, we can still see a very similar trend of all energy distribution.

4.6.3 The $M_1$ solution for a binary structure

In addition to the first test with a homogeneous material, we have calculated the solution of the $M_1$ model for a simple binary structure of nickel and chromium, i.e. we have set $n = 2$, $E_{l_1} = \text{Ni}$, $E_{l_2} = \text{Cr}$ and chosen the
following material parameters:

\[ Z_1 = 28, \quad A_1 = 58.6934 \times 10^{-3}/N_A \text{ kg} \]

\[ \rho_1 \equiv 8.9 \times 10^{3} \text{kg m}^{-3}, \quad c_1(x, y) = \begin{cases} 
1, & x \leq 0 \\
0, & x > 0 
\end{cases}, \]

for nickel and

\[ Z_2 = 24, \quad A_2 = 51.9961 \times 10^{-3}/N_A \text{ kg} \]

\[ \rho_2 \equiv 7.2 \times 10^{3} \text{kg m}^{-3}, \quad c_2(x, y) = \begin{cases} 
0, & x \leq 0 \\
1, & x > 0 
\end{cases}. \]

for chromium. All other parameters are the same as in the previous test case and again a reference electron number density was calculated using the \textsc{dtsa}–\textsc{ii} Monte Carlo software. A comparison of the spatial distribution of the electron number densities at different energies is presented in Figure 4.10. For both the \( M1 \) and the Monte Carlo solution a difference in the dynamic of the electrons inside the two different materials can be observed. Both plots show a slightly deeper penetration of the electrons inside the chromium region \((x > 0)\) at the same energy. Furthermore the overall shape of the spatial distribution of the \( M1 \) solution looks similar compared to the spatial distribution of the electron number density computed by Monte Carlo simulations.
4.6. NUMERICAL RESULTS

Figure 4.10: Spatial distribution of the $M_1$ and Monte Carlo electron number densities in a binary structure (nickel and chromium) at different energies.

(a) $n(x, y, \epsilon), \epsilon = 9.2$ keV

(b) $n_{\text{ref}}(x, y, \epsilon), \epsilon = 9.2$ keV

(c) $n(x, y, \epsilon), \epsilon = 8.6$ keV

(d) $n_{\text{ref}}(x, y, \epsilon), \epsilon = 8.6$ keV

(e) $n(x, y, \epsilon), \epsilon = 7.6$ keV

(f) $n_{\text{ref}}(x, y, \epsilon), \epsilon = 7.6$ keV
Conclusions

The goal of this thesis was the development of a reconstruction method for electron probe microanalysis that is capable of overcoming the limitations to the spatial resolution of conventional matrix correction methods.

Monte Carlo based reconstruction

The gradient based Levenberg-Marquardt optimization algorithm was combined with X-ray predictions obtained from Monte Carlo simulations for heterogeneous material structures. In the presented numerical test cases this method has turned out to enable the accurate reconstruction of features on the nanometer scale. Because in both examples the structures were significantly smaller than the interaction volume of the electrons, it would be impossible to reconstruct them using conventional matrix correction methods. In this sense the goal of this thesis has been achieved, though the proposed method has not been applied to real experiments yet.

A disadvantage of the combination of gradient based methods with Monte Carlo simulations is that the gradient has to be approximated by finite differences. Though the statistical fluctuations inherent to Monte Carlo simulations introduce errors that are much larger than the machine precision, the presented numerical results indicate that a sufficient accuracy can be achieved by a proper choice of the finite difference step size. However, approximating gradients of a Monte Carlo based target function is very time consuming, because computing a full gradient based on this approach requires a number of simulations that scales with the amount of unknown
parameters. Though this process can be scheduled in parallel, an application of the proposed method to large-scale problems like full two-dimensional mappings might be impracticable at present.

The $M_1$ model for electron transport

In view of these challenges a deterministic partial differential equation model for the prediction of the electron number density distribution inside arbitrary solid materials was investigated. Numerical results for the cases of a homogeneous material and a binary structure show a good qualitative agreement in comparison to Monte Carlo simulations, though there are non-negligible differences in the energy distributions of the electrons at different depths. A possible source of errors is the choice of the initial condition. The initial lateral distribution of the electrons in the beam can be setup comparable to the Monte Carlo simulation, i.e. as a lateral Gaussian distribution in both cases. The problem is that physically the electrons with beam energy are still outside of the specimen but must be initialized on the numerical grid discretizing the specimen. Therefore a narrow Gaussian along the vertical axis was chosen to obtain an initial condition that has non-zero measure within the specimen. The choice of this initial condition might not be ideal and should be further investigated.

One advantage of the deterministic model over Monte Carlo simulations is the lack of stochastic noise in the deterministic solution.
Outlook

In the following we will point out some ideas for further research and development, that arose during the work on this thesis, but were not discussed or investigated in detail, because they were beyond its scope.

Improving the accuracy of the deterministic model

For the elastic transport coefficient we have used the screened Rutherford cross-section in this work. This model is known to be inaccurate for large scattering angles [33]. A more accurate model is given by the Mott cross section [33, pp. 69–73]. An expression for the elastic transport coefficient based on this cross section is given in [20]. Furthermore we have completely omitted the inelastic transport coefficient in our implementation. To include this, one could integrate the Möller cross section [33, p. 77] for inelastic scattering. An analytical expression of the resulting inelastic transport coefficient is given in [20]. Another possibility is to interpolate tabulated transport coefficients. For example the Monte Carlo program Penelope [36] has a very extensive and accurate built-in data base of various cross sections, material and transport coefficients.

As mentioned before the initial condition proposed in this work should be further investigated. Aside from adjusting the initial spatial distribution of the electrons, the initial condition might be completely replaced by a boundary condition that models a flux of electrons through the surface of the specimen. However, the same problem arises that directly after the first time step no electrons with beam energy are observed any more. Thus the
boundary condition would also have to be extended to range of energies with positive measure.

**Reconstruction based on the deterministic model**

Before the $M_1$ model can be used to predict X-ray intensities, the formula (1.45) has to be implemented. Therefore a data base of mass absorption coefficients and either an analytical expression or tabulated values for the ionization cross section have to be incorporated in the code.

In computational fluid dynamics adjoint methods are used for aerodynamic shape optimization [16]. With this method only a single simulation is required to estimate the full gradient with respect to the model parameters. This enables a much more efficient gradient based optimization and allows for a significantly higher complexity of the regarded model, in our case the material structure, at nearly the same computational costs. A long-term goal of the approach to develop a deterministic model for electron transport is the derivation of an adjoint-based optimization method for the material parameters in electron probe microanalysis.

**Improving and extending the numerical scheme**

The main objectives for the implementation of the scheme were source code readability and the minimization of error sources. Eliminating unnecessary calculations and optimizing inefficient implementation constructs might considerably increase the speed of the simulations.

The most time consuming subroutine is the calculation of the numerical fluxes over all cell boundaries. Instead of calculating the entries of the Jacobian of the mathematical flux and then its eigenvalues, one could approximate the spectrum of the Jacobian in dependence of the anisotropy parameter by an ansatz function as it was done for the Eddington factor in this work. This is expected to speed up the estimation of the characteristic velocities considerably and the HLL flux is designed to work with rough
estimates anyway. An even faster flux evaluation can be achieved by the constant approximation of the eigenvalues by the numerically verified lower and upper bounds $-1$ and $1$. This will in most cases decrease the size of the time steps and therefore require more iterations though.

Furthermore the scheme could be extended to three-dimensional geometries. For that purpose one would simply add a third numerical flux in the direction orthogonal to the $x$-$y$ plane and extend all coordinates and quantities to three spatial dimensions.

Higher accuracy and larger time steps, reducing the number of necessary iterations for each simulation, could be achieved by the use of higher order schemes. One possibility would be the reconstruction through slope or flux limiters in the finite volume scheme [24, pp. 176–191], [37, 459–487]. Another possibility is the replacement of the finite volume scheme by a discontinuous Galerkin method [8].
Appendix A

Derivation of the unscreened Rutherford cross-section

The following derivation is based on [33, pp. 57–61]. To describe the movement of the free electron, we denote its position vector relative to the nucleus with \( r(t) \in \mathbb{R}^3 \) for \( t \in [t_0, t_1] \) with \( t_0, t_1 \) such that \( r(t_0) = A \) (and \( r(t_1) = C \)) are points before (and after) interaction between the free electron and the Coulomb field takes place and \( B \) is the point where the normal to the trajectory of the free electron goes through the nucleus.

Because the nucleus has a much greater mass than the electron, we assume that the nucleus stays at rest during the interaction and thus \( r \) can be considered the position of the electron in a coordinate system with the nucleus as its origin. We now model the free electron as a particle of mass \( m_0 \) with electric charge

\[
Q_e = -e, \quad (A.1)
\]

The nucleus has charge

\[
Q_n = eZ, \quad (A.2)
\]

and thus by Coulomb’s law produces an electric field with intensity

\[
E(t) = \frac{Q_n}{4\pi\varepsilon_0} \frac{r(t)}{|r(t)|^3} = \frac{eZ}{4\pi\varepsilon_0} \frac{r(t)}{|r(t)|^3}. \quad (A.3)
\]
As mentioned before, we neglect the screening of the electric field by the charges of the atomic electrons here. Thus we can consider the free electron to be attracted to the nucleus by the Coulomb force:

\[
F(t) = QeE(t) = -e \frac{Ze}{4\pi\epsilon_0 |r(t)|^3} = -\frac{e^2Z}{4\pi\epsilon_0 |r(t)|^3}.
\]  (A.4)

By Newton’s law we have:

\[
m_0 \frac{\partial^2 r(t)}{\partial t^2} = F(t) \iff m_0 \frac{\partial^2 r(t)}{\partial t^2} = -\frac{e^2Z}{4\pi\epsilon_0 |r(t)|^3}.
\]  (A.5)

Let \(v(t) \in \mathbb{R}^3\) denote the velocity of the electron as a directed quantity. Then its linear momentum is defined as:

\[
p(t) := m_0v(t) = m_0 \frac{\partial r(t)}{\partial t},
\]  (A.6)

and its angular momentum is:

\[
L(t) = r(t) \times p(t).
\]  (A.7)

From (A.5) and (A.6) we receive

\[
\frac{\partial p(t)}{\partial t} = -\frac{e^2Z}{4\pi\epsilon_0 |r(t)|^3} r(t)
\]

and thus we have

\[
p(t) \parallel \frac{\partial r(t)}{\partial t} \quad \text{and} \quad \frac{\partial p(t)}{\partial t} \parallel r(t), \quad \text{for all} \quad t \in [t_0, t_1].
\]  (A.8)

Using these relationships we can show that the angular momentum is constant:

\[
\frac{\partial L(t)}{\partial t} = \frac{\partial r(t)}{\partial t} \times p(t) + r(t) \times \frac{\partial p(t)}{\partial t} = 0, \quad \text{for all} \quad t \in [t_0, t_1],
\]  (A.9)

which physically means, that the electron moves only in the \(y\)-\(z\) plane. Using polar coordinates \((|r|, \varphi)\) in this plane, the position vector \(r\) can be expressed
as:
\[ r(t) = |r(t)| \begin{pmatrix} \sin \varphi(t) \\ \cos \varphi(t) \end{pmatrix}. \] (A.10)

Here \( \varphi(t) \) is the angle between \( r(t) \) and the \( z \)-axis. Let \( b \) denote the impact parameter, i.e. the smallest distance between the line parallel to the electron’s initial flight direction and the position of the nucleus. Then

\[ b = r_1(t_0) = |r(t_0)| \sin \varphi(t_0). \]

Using this equation we can represent the magnitude of the angular momentum at \( t = t_0 \) by:

\[ |L(t_0)| = |r(t_0) \times p(t_0)| = ||r(t_0)||p(t_0)|| \sin \phi(t_0)||p(t_0)||| \sin \phi(t_0)|
= m_0|v(t_0)||r(t_0)|| \sin \phi(t_0)||p(t_0)||| \sin \phi(t_0)|
= m_0|v(t_0)||b|
= m_0|v(t_0)|b. \] (A.11)

Here \( \phi(t) \) denotes the angle between \( r(t) \) and \( p(t) \). From (A.6) we have that \( p(t) \parallel v(t) \). Because the initial flight direction \( v(t_0) \) is in negative \( z \)-direction, we have \( \phi(t_0) + \varphi(t_0) = \pi \). Thus \( \sin \phi(t_0) = \sin(\varphi(t_0) - \pi) = -\sin \varphi(t_0) \).

The magnitude of the angular momentum can also be described by:

\[ |L(t)| = mr(t)^2 \frac{\partial \varphi(t)}{\partial t}, \] (A.12)

where \( mr(t)^2 \) is the moment of inertia and \( \frac{\partial \varphi(t)}{\partial t} \) is the angular velocity. Because of (A.9) the equality in (A.11) holds for all \( t \in [t_0, t_1] \). Therefore combining (A.11) and (A.12) yields:

\[ \frac{1}{|r(t)|^2} = \frac{1}{|v(t_0)|b} \frac{\partial \varphi(t)}{\partial t}. \]
Using this equation and (A.10) we can find the following representation of the first component of (A.5):

\[
\frac{m_0}{\partial t} \frac{\partial^2 r_1(t)}{\partial t} = -\frac{e^2 Z}{4\pi \epsilon_0} \frac{r_1(t)}{|r(t)|^3}
\]

\[
\Leftrightarrow \frac{\partial v_1(t)}{\partial t} = -\frac{e^2 Z}{4\pi \epsilon_0} \frac{\sin \varphi(t)}{m_0 |r(t)|^2} \frac{\partial \varphi(t)}{\partial t}
\]

\[
= -\frac{e^2 Z}{4\pi \epsilon_0} \frac{\sin \varphi(t)}{m_0 b |v(t_0)|} \frac{\partial \varphi(t)}{\partial t}.
\]

(A.13)

Let \( \theta \) denote the scattering angle, i.e. the angle between the initial flight direction of the electron and the one after the collision. In regard of the dimensions of an atom, we assume that \( A \) and \( C \) are so far away from \( B \), that the influence of the collision parameter \( b \) on the angle \( \varphi(t) \) can be neglected at \( t = t_0 \) and \( t = t_1 \). Thus we can use \( \varphi(t_0) = 0 \) and \( \varphi(t_1) = \pi + \theta \) in the following. We also have \( v_1(t_0) = 0 \) and \( v_1(t_1) = |v(t_1)| \sin \theta \). Integrating (A.13) from \( t = t_0 \) to \( t = t_1 \) yields:

\[
\int_{t_0}^{t_1} \frac{\partial v_1(t)}{\partial t} \, dt = \int_{t_0}^{t_1} -\frac{e^2 Z}{4\pi \epsilon_0} \frac{1}{m_0 b |v(t)|} \frac{\partial \varphi(t)}{\partial t} \, dt
\]

\[
\Leftrightarrow v_1(t_1) - v_1(t_0) = -\frac{e^2 Z}{4\pi \epsilon_0} \frac{1}{m_0 b |v(t_0)|} \int_{t_0}^{t_1} \sin \varphi(t) \frac{\partial \varphi(t)}{\partial t} \, dt
\]

\[
\Leftrightarrow |v(t_1)| \sin \theta = -\frac{e^2 Z}{4\pi \epsilon_0} \frac{1}{m_0 b |v(t_0)|} \int_{\varphi(t_0)}^{\varphi(t_1)} \sin \varphi \, d\varphi
\]

\[
\Leftrightarrow |v(t_1)| \sin \theta = -\frac{e^2 Z}{4\pi \epsilon_0} \frac{1}{m_0 b |v(t_0)|} [-\cos(\pi + \theta) + 1]
\]

\[
\Leftrightarrow |v(t_1)| \sin \theta = -\frac{e^2 Z}{4\pi \epsilon_0} \frac{1}{m_0 b |v(t_0)|} (1 + \cos \theta).
\]

The electron is accelerated from \( A \) to \( B \) and decelerated from \( B \) to \( C \). Due to the conservation of energy in elastic scattering, \( |v(t_0)| = |v(t_1)| \) holds. Using this and \( \cot(\theta/2) = (1 + \cos \theta)/\sin \theta \) we get the following expression for the impact parameter in dependence of \( \theta \) and the initial velocity \( |v(t_0)| \):

\[
b = -\frac{e^2 Z}{4\pi \epsilon_0} \frac{1}{m_0 |v(t_0)|^2} \cot \frac{\theta}{2}.
\]

(A.14)
Thus we also have:

\[
\frac{d b}{d \theta} = -\frac{e^2 Z}{4\pi \epsilon_0 m_0 |v(t_0)|^2} \frac{1}{2} d\theta |\theta = \frac{\theta}{2}\cot \theta
\]

\[
= \frac{e^2 Z}{2(4\pi \epsilon_0) m_0 |v(t_0)|^2} \frac{1}{\sin^2(\theta/2)},
\]

which can be written as:

\[
db = \frac{e^2 Z}{2(4\pi \epsilon_0) m_0 |v(t_0)|^2} \frac{d\theta}{\sin^2(\theta/2)}. \tag{A.15}
\]

The differential cross-section for scattering from an area \(d\sigma\) through an angle \(\theta\) into a cone of solid angle \(d\Omega\) is defined by the ratio \(d\sigma/d\Omega\). The fraction of the in-scattering area that lies in the \(z-y\) plane is an area \(db\) with distance \(b\) to the center. Due to the radial symmetry, the whole in-scattering area is thus given by an infinitesimal annulus with inner radius \(b\) and width \(db\):

\[
d\sigma = 2\pi b db.
\]

The same argument can be applied to relate \(d\Omega\) to \(d\theta\). In this case (after normalization) the radius is given as \(\sin \theta\) and the width is \(d\theta\):

\[
d\Omega = 2\pi \sin \theta d\theta.
\]

Thus the ratio \(d\sigma/d\Omega\) becomes:

\[
\frac{d\sigma}{d\Omega} = \frac{2\pi b db}{2\pi \sin \theta d\theta} = \frac{e^4 Z^2}{2(4\pi \epsilon_0) m_0^2 |v(t_0)|^4} \frac{1}{\sin^2(\theta/2)} \cot(\theta/2) \sin \theta
\]

\[
= \frac{e^4 Z^2}{4(4\pi \epsilon_0) m_0^2 |v(t_0)|^4} \sin^4(\theta/2). \tag{A.16}
\]

Here we have used that \(\cot(\theta/2) = \sin \theta/(2 \sin^2(\theta/2))\). Now we reformulate (A.16) in terms of the energy of the free electron instead of its initial velocity. As in electron probe microanalysis the beam electrons usually have energies in a range where relativistic effects can be neglected, we relate the energy \(\epsilon\) of the free electron before (and after) the collision to its corresponding velocity.
according to the classical model for the kinetic energy:

\[ \epsilon = \frac{1}{2} m_0 |v(t_0)|^2. \]  

Finally we receive the following unscreened Rutherford elastic differential cross-section:

\[ \sigma_R(\epsilon, \theta) = \frac{e^4 Z^2}{16(4\pi\epsilon_0)e^2 \sin^4(\theta/2)} m^2. \]
Appendix B

Calculation of the Jacobian of the mathematical flux

In the following we will calculate an expression of the Jacobian of the mathematical flux (4.56) in terms of $\chi$, $\chi'$ and $\alpha$. To simplify the calculation of the derivative with respect to $U_0$, note that

$$\frac{\alpha}{|\alpha|} = \frac{\psi^{(1)}}{\psi^{(0)}} = \frac{\psi^{(1)}}{|\psi^{(1)}|} \frac{|\psi^{(0)}|}{\psi^{(1)}} = \frac{\psi^{(1)}}{|\psi^{(0)}|} = \frac{\psi^{(1)}}{|\psi^{(0)}|}. \quad (B.1)$$

Here we have used that $\psi^{(0)} \geq 0$ as discussed in Section 4.4. Expressing (B.1) in terms of $U$ as in the two-dimensional analogon to (3.40) yields

$$\frac{\alpha_i}{|\alpha|} = \frac{U_{i+1}}{\sqrt{U_2^2 + U_3^2}} \text{ for } i = 1, 2.$$

Thus we can express the mathematical flux in (4.56) as:

$$F(U) = \left( \begin{array}{ccc}
\left( \frac{1-\chi(|\alpha|)}{2} + \frac{U_2}{U_2^2 + U_3^2} 3\chi(|\alpha|)-1 \right) U_1 & \frac{U_2 U_3}{U_2^2 + U_3^2} 3\chi(|\alpha|)-1 U_1 \\
\frac{U_2 U_3}{U_2^2 + U_3^2} 3\chi(|\alpha|)-1 U_1 & \left( \frac{1-\chi(|\alpha|)}{2} + \frac{U_3}{U_2^2 + U_3^2} 3\chi(|\alpha|)-1 \right) U_1
\end{array} \right).$$

131
We will only give a detailed calculation of the Jacobian for $F_{-1}$. Due to the similarity of $F_{-1}$ and $F_{-2}$ the calculation of the Jacobian of $F_{-2}$ is analogous.

The first row of $F_{-1}$ is easy to differentiate:

\[
\frac{\partial}{\partial U_1} F_{1,1}(U) = \frac{\partial}{\partial U_1} U_2 = 0,
\]
\[
\frac{\partial}{\partial U_2} F_{1,1}(U) = \frac{\partial}{\partial U_2} U_2 = 1,
\]
\[
\frac{\partial}{\partial U_3} F_{1,1}(U) = \frac{\partial}{\partial U_3} U_2 = 0.
\]

Now we will differentiate the second and third row of $F_{-1}$ with respect to $U_1$. To simplify the calculation, we first evaluate the following derivatives:

\[
U_1 \frac{\partial}{\partial U_1} |\alpha| = U_1 \frac{\partial}{\partial U_1} \left( \frac{U_2^2}{U_1^2} + \frac{U_3^2}{U_1^2} \right)^{1/2} = \frac{U_1}{2|\alpha|} \left( U_2^2 + U_3^2 \right) \frac{\partial}{\partial U_1} U_1^{-2}
\]
\[
= -\frac{1}{|\alpha|} \left( \frac{U_2^2}{U_1^2} + \frac{U_3^2}{U_1^2} \right) = -\frac{|\alpha|^2}{|\alpha|} = -|\alpha|,
\]

\[
U_1 \frac{\partial}{\partial U_1} \frac{1 - \chi(|\alpha|)}{2} = -\frac{\chi'(|\alpha|)}{2} U_1 \frac{\partial}{\partial U_1} |\alpha| = \frac{|\alpha|\chi'(|\alpha|)}{2},
\]

\[
U_1 \frac{\partial}{\partial U_1} \frac{3\chi(|\alpha|) - 1}{2} = \frac{3\chi'(|\alpha|)}{2} U_1 \frac{\partial}{\partial U_1} |\alpha| = -\frac{3|\alpha|\chi'(|\alpha|)}{2}.
\]
This gives
\[
\frac{\partial}{\partial U_1} F_{2,1}(U) = \frac{\partial}{\partial U_1} \left[ U_1 \left( \frac{1 - \chi(|\alpha|)}{2} + \frac{3\chi(|\alpha|) - 1}{U_2^2 + U_3^2} \right) \right]
= \frac{1 - \chi(|\alpha|)}{2} + \frac{\alpha_1^2}{|\alpha|^2} \frac{3\chi(|\alpha|) - 1}{2}
\]
\[
+ U_1 \frac{\partial}{\partial U_1} \frac{1 - \chi(|\alpha|)}{2} + \frac{U_2^2}{U_2^2 + U_3^2} U_1 \frac{\partial}{\partial U_1} \frac{3\chi(|\alpha|) - 1}{2}
= \frac{1 - \chi(|\alpha|) - |\alpha|\chi'(|\alpha|)}{2} + \frac{\alpha_1^2}{|\alpha|^2} \frac{3(\chi(|\alpha|) - |\alpha|\chi'(|\alpha|)) - 1}{2}
\]
and
\[
\frac{\partial}{\partial U_1} F_{3,1}(U) = \frac{\partial}{\partial U_1} \left[ U_1 \frac{3\chi(|\alpha|) - 1}{2} - \frac{U_2 U_3}{U_2^2 + U_3^2} \right]
= \frac{\alpha_1 \alpha_2}{|\alpha|^2} \frac{3\chi(|\alpha|) - 1}{2} + \frac{U_2 U_3}{U_2^2 + U_3^2} U_1 \frac{\partial}{\partial U_1} \frac{3\chi(|\alpha|) - 1}{2}
\]
\[
= \frac{\alpha_1 \alpha_2}{|\alpha|^2} \frac{3\chi(|\alpha|) - 1}{2} - \frac{\alpha_1 \alpha_2}{|\alpha|^2} \frac{3|\alpha|\chi'(|\alpha|)}{2}
\]
\[
= \frac{\alpha_1 \alpha_2}{|\alpha|^2} \frac{(\chi(|\alpha|) - |\alpha|\chi'(|\alpha|)) - 1}{2}.
\]
Before we calculate the corresponding derivatives with respect to $U_2$, we again differentiate the most important terms with respect to $U_2$ individually:

$$U_1 \frac{\partial}{\partial U_2} |\alpha| = U_1 \frac{\partial}{\partial U_2} \left( \frac{U_2^2}{U_1^2} + \frac{U_3^2}{U_1^2} \right)^{1/2} = \frac{U_1}{2|\alpha|} \frac{2U_2}{U_1^2} = \frac{U_2}{U_1} \frac{1}{|\alpha|} = \frac{\alpha_1}{|\alpha|},$$

$$U_1 \frac{\partial}{\partial U_2} \frac{1 - \chi(|\alpha|)}{2} = -\frac{\chi'(|\alpha|)}{2} U_1 \frac{\partial}{\partial U_2} |\alpha| = -\frac{\alpha_1}{|\alpha|} \frac{\chi'(|\alpha|)}{2},$$

$$U_1 \frac{\partial}{\partial U_2} \frac{3\chi(|\alpha|) - 1}{2} = \frac{3\chi'(|\alpha|)}{2} U_1 \frac{\partial}{\partial U_2} |\alpha| = \frac{\alpha_1}{|\alpha|} \frac{3\chi'(|\alpha|)}{2},$$

$$U_1 \frac{\partial}{\partial U_2} U_2 \frac{U_2^2}{U_2^2 + U_3^2} = 2 U_1 U_2 \frac{U_2^2}{U_2^2 + U_3^2} + U_1 U_2^2 \frac{\partial}{\partial U_2} (U_2^2 + U_3^2)^{-1}$$

$$= \frac{2 U_1 U_2}{U_2^2 + U_3^2} - 2 \frac{U_1 U_2^3}{(U_2^2 + U_3^2)^2}$$

$$= 2 \frac{1}{\sqrt{U_2^2 + U_3^2}} \frac{U_2}{U_2^2 + U_3^2} \frac{1}{\sqrt{U_2^2 + U_3^2}} - 2 \frac{1}{\sqrt{U_2^2 + U_3^2}} \frac{U_2}{U_2^2 + U_3^2} \frac{U_2}{U_2^2 + U_3^2}$$

$$= \frac{2 \alpha_1^3}{|\alpha|^2} - 2 \frac{\alpha_1^3}{|\alpha|^4} = 2 \frac{\alpha_1^3}{|\alpha|^2} \left( 1 - \frac{\alpha_1^2}{|\alpha|^2} \right),$$

$$U_1 \frac{\partial}{\partial U_2} U_2 U_3 \frac{U_2 U_3}{U_2^2 + U_3^2} = \frac{U_1 U_3}{U_2^2 + U_3^2} + U_1 U_2 U_3 \frac{\partial}{\partial U_2} (U_2^2 + U_3^2)^{-1}$$

$$= \frac{\alpha_2}{|\alpha|^2} - 2 \frac{U_1 U_2 U_3}{(U_2^2 + U_3^2)^2} = \frac{\alpha_2}{|\alpha|^2} - 2 \frac{\alpha_2 \alpha_2}{|\alpha|^4} = \frac{\alpha_2}{|\alpha|^2} \left( 1 - \frac{\alpha_2^2}{|\alpha|^2} \right).$$
Using these replacement rules we calculate:

\[
\frac{\partial}{\partial U_2} F_{2,1}(U) = \frac{\partial}{\partial U_2} \left[ U_1 \left( \frac{1 - \chi(|\alpha|)}{2} + \frac{3\chi(|\alpha|) - 1}{2} \frac{U_2^2}{U_2^2 + U_3^2} \right) \right] = U_1 \frac{\partial}{\partial U_2} \left( \frac{1 - \chi(|\alpha|)}{2} \right) + \frac{U_2^2}{U_2^2 + U_3^2} \frac{\partial}{\partial U_2} \left( \frac{3\chi(|\alpha|) - 1}{2} \right) + 3\chi(|\alpha|) \frac{\partial}{\partial U_2} \left( \frac{U_2^2}{U_2^2 + U_3^2} \right) \\
= -\frac{\alpha_1}{|\alpha|} \chi'(|\alpha|) + \frac{\alpha_1^2}{|\alpha|^2} \frac{\alpha_1}{|\alpha|} \chi'(|\alpha|) + 2\alpha_1 \left( 1 - \frac{\alpha_1^2}{|\alpha|^2} \right) \frac{3\chi(|\alpha|) - 1}{2} \\
= \frac{\alpha_1}{|\alpha|} \left( \frac{\alpha_1^2}{|\alpha|^2} \chi'(|\alpha|) - \chi'(|\alpha|) \right) + 2\alpha_1 \left( 1 - \frac{\alpha_1^2}{|\alpha|^2} \right) \frac{3\chi(|\alpha|) - 1}{2},
\]

and

\[
\frac{\partial}{\partial U_2} F_{3,1}(U) = \frac{\partial}{\partial U_2} \left[ U_1 \frac{3\chi(|\alpha|) - 1}{2} \frac{U_2 U_3}{U_2^2 + U_3^2} \right] = \frac{\alpha_1}{|\alpha|^2} \frac{\alpha_1^2}{|\alpha|} \frac{3\chi(|\alpha|) - 1}{2} \frac{U_2 U_3}{U_2^2 + U_3^2} + \frac{\alpha_2}{|\alpha|^2} \left( 1 - 2\frac{\alpha_1^2}{|\alpha|^2} \right) \frac{3\chi(|\alpha|) - 1}{2} \\
= \frac{\alpha_2}{|\alpha|^2} \left( \frac{3\chi(|\alpha|) - 1}{2} - \frac{\alpha_1^2}{|\alpha|^2} \frac{6\chi(|\alpha|) - 3|\alpha|\chi'(|\alpha|) - 2}{2} \right). \]
Due to the symmetry with respect to $U_2$ and $U_3$ in most terms we directly get the following results by swapping the indices 2 and 3:

\[
U_1 \frac{\partial}{\partial U_3} |\alpha| = \frac{\alpha_2}{|\alpha|},
\]

\[
U_1 \frac{\partial}{\partial U_3} \frac{1 - \chi(|\alpha|)}{2} = -\frac{\alpha_2}{|\alpha|} \chi'(|\alpha|),
\]

\[
U_1 \frac{\partial}{\partial U_3} \frac{3\chi(|\alpha|) - 1}{2} = \frac{\alpha_2}{|\alpha|} 3\chi'(|\alpha|),
\]

\[
U_1 \frac{\partial}{\partial U_3} \frac{U_2U_3}{U_2^2 + U_3^2} = \frac{\alpha_1}{|\alpha|^2} \left( 1 - 2 \frac{\alpha_2^2}{|\alpha|^2} \right),
\]

\[
U_1 \frac{\partial}{\partial U_3} \frac{U_2^2}{U_2^2 + U_3^2} = U_1U_2 \frac{\partial}{\partial U_3} \left( U_2^2 + U_3^2 \right)^{-1} = -2 \frac{U_1U_2U_3}{(U_2^2 + U_3^2)^2} = -2 \frac{\alpha_1^2 \alpha_2}{|\alpha|^4}.
\]

Finally we calculate the derivatives of $F$ with respect to $U_3$:

\[
\frac{\partial}{\partial U_3} F_{2,1}(U) = \frac{\partial}{\partial U_3} \left[ U_1 \left( \frac{1 - \chi(|\alpha|)}{2} + \frac{3\chi(|\alpha|) - 1}{2} \frac{U_2^2}{U_2^2 + U_3^2} \right) \right]
\]

\[
= U_1 \frac{\partial}{\partial U_3} \frac{1 - \chi(|\alpha|)}{2} + \frac{U_2^2}{U_2^2 + U_3^2} U_1 \frac{\partial}{\partial U_3} \frac{3\chi(|\alpha|) - 1}{2}
\]

\[
+ \frac{3\chi(|\alpha|) - 1}{2} U_1 \frac{\partial}{\partial U_3} \frac{U_2^2}{U_2^2 + U_3^2}
\]

\[
= -\frac{\alpha_2}{|\alpha|} \frac{\chi'(|\alpha|)}{2} + \frac{\alpha_1^2}{|\alpha|^2} \frac{\alpha_2}{|\alpha|} 3\chi'(|\alpha|) - 2 \frac{\alpha_1^2 \alpha_2}{|\alpha|^4} \frac{3\chi(|\alpha|) - 1}{2}
\]

\[
= \frac{\alpha_2}{|\alpha|} \left( -\frac{\chi'(|\alpha|)}{2} + \frac{\alpha_1^2}{|\alpha|^3} 6\chi(|\alpha|) - 3|\alpha|\chi'(|\alpha|) - 2 \right),
\]
and
\[
\frac{\partial}{\partial U_3} F_{3,1}(U) = \frac{\partial}{\partial U_3} \left[ U_1 \frac{3\chi(|\alpha|) - 1}{2} U_2 U_3 \right]
\]
\[
= \frac{\alpha_1 \alpha_2}{|\alpha|^2} \left( \frac{3\chi(|\alpha|) - 1}{2} U_2 U_3 \right) + \frac{3\chi(|\alpha|) - 1}{2} U_1 \frac{\partial}{\partial U_3} \frac{U_2 U_3}{U_2^2 + U_3^2}
\]
\[
= \frac{\alpha_1 \alpha_2}{|\alpha|^2} \frac{3\chi(|\alpha|) - 1}{2} + \frac{\alpha_1}{|\alpha|^2} \left( 1 - \frac{\alpha_2^2}{|\alpha|^2} \right) \frac{3\chi(|\alpha|) - 1}{2}
\]
\[
= \frac{\alpha_1}{|\alpha|^2} \left( \frac{3\chi(|\alpha|) - 1}{2} - \frac{\alpha_2}{|\alpha|^2} \left( 6\chi(|\alpha|) - 3|\alpha|\chi'(|\alpha|) - 2 \right) \right).
\]

Collecting all results and using a shorter notation
\[
t_1(|\alpha|) := \frac{3\chi(|\alpha|) - 1}{2},
\]
\[
t_2(|\alpha|) := \chi(|\alpha|) - |\alpha|\chi'(|\alpha|),
\]
\[
t_3(|\alpha|) := 6\chi(|\alpha|) - 3|\alpha|\chi'(|\alpha|) - 2,
\]
\[
t_4(|\alpha|) := \frac{\alpha_1^2}{|\alpha|^2} \frac{3\chi(|\alpha|) - 1}{2} - \frac{\alpha_2}{|\alpha|^2} \chi'(|\alpha|),
\]
we can express the Jacobian of \( F_{-,1} \) as:
\[
J_1(U) = \left( \frac{\partial}{\partial U_j} F_{i,1}(U) \right)_{i=1,\ldots,3}^{j=1,\ldots,3}
\]
\[
= \begin{pmatrix}
0 & \frac{1-t_2(|\alpha|)}{2} + \frac{\alpha_1}{|\alpha|^2} \frac{3t_2(|\alpha|) - 1}{2} & \frac{\alpha_1 \alpha_2}{|\alpha|^2} \frac{3t_2(|\alpha|) - 1}{2} \\
\frac{\alpha_1}{|\alpha|} & t_4(|\alpha|) + \left( \frac{2\alpha_1}{|\alpha|^2} \right) t_1(|\alpha|) & \frac{\alpha_1}{|\alpha|^2} \left( t_1(|\alpha|) - \frac{\alpha_2}{|\alpha|^2} t_3(|\alpha|) \right) \\
0 & \frac{\alpha_2}{|\alpha|} \left( -\frac{\chi'(|\alpha|)}{2} - \frac{\alpha_1}{|\alpha|^2} t_3(|\alpha|) \right) & \frac{\alpha_1}{|\alpha|^2} \left( t_1(|\alpha|) - \frac{\alpha_2}{|\alpha|^2} t_3(|\alpha|) \right)
\end{pmatrix}^T.
\]
Bibliography


