Matrix Compression Methods for the Numerical Solution of Radiative Transfer in Scattering Media

Peter Schlosser

Vom Fachbereich Mathematik
der Universität Kaiserslautern
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Doctor rerum naturalium, Dr. rer. nat.)
genehmigte Dissertation

1. Gutachter: Prof. Dr. H. Neunzert
2. Gutachter: Prof. Dr. S. Sauter

Datum der Disputation: 11. Februar 2003

D 386

First of all, I wish to thank all those who have contributed to the success of this work. Especially, I thank Prof. Dr. Dr. h.c. Helmut Neunzert for posing the problem and supervising my studies, and Prof. Dr. Stefan Sauter for his willingness to be co-referee of this work.
For many conversations on various topics concerning this work and for proof reading this thesis, I wish to thank Dr. Norbert Siedow. Furthermore, I thank Erik Meinköhn for providing numerical results which are used for a comparison in this thesis.
I wish to thank Christina and my brother Markus for their friendship and patience, and Markus also for proof reading parts of this thesis.
Finally, I thank the Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM) and the 'Stiftung Rheinland-Pfalz für Innovation' (Rheinland-Pfalz, Germany) for the financial support of this work.

Table of Contents

Introduction

Radiation plays an important role in many physical processes and technical applications, e.g. in the investigation of stellar atmospheres, in illumination problems in computer graphics and in the inverse radiation therapy planing for patients with malignant tumors. Depending on the situation, radiation has many facets. These facets are reflected in the different models used for its description.

In case of thermal radiation, one distinguishes mainly two different situations: radiation in participating and non-participating media. Non-participating media, e.g. vacuum, do not influence the radiation. Therefore, the radiative transfer can be described by only considering the surfaces surrounding the medium. On the other hand, a participating medium does have an influence on the radiation. The radiation may be absorbed and scattered during its passage through the medium. The physical effects of absorption, emission, scattering and reflection at the boundary can be described precisely by using the quantum mechanical model for photons or the wave model respectively. Unfortunately, both models are too complicated to be handled numerically. Therefore, this dissertation is based on the ray model, which is much easier to handle. Furthermore, the above effects can nevertheless be included in a simplified way.

The ray model describes the energy transport in terms of a single scalar quantity, the intensity I. This quantity fulfills the radiative transfer equation, an integro-differential equation which can be interpreted as a combination of transport equations along rays coupled by an integral over all directions, which models the scattering.

Thermal radiation typically has to be combined with other heat transfer mechanisms, such as conduction and convection, which are very slow compared with the speed of light. Hence, to a very good approximation, the radiative heat transfer can be regarded as being instantaneous and can, therefore, be modeled using stationary equations. To simplify the situation further, we also neglect the dependency on the wavelength leading to the monochromatic radiative transfer equation. Apart from its own physical relevance, this model also serves as a solution step for more complex models where the dependency on the wavelength is included.

When the dependency on the time and on the wavelength are omitted, I only depends on the location inside the medium and on the direction of the ray, i.e. there are five independent variables: three space coordinates and two coordinates describing the direction of the ray. Therefore, a discretization still leads to a very large number of unknowns. This implies large requirements on storage and computing time, which is the major difficulty for the numerical treatment of radiation.

The radiative transfer equation describes the energy transport in terms of the angular dependent variable I. However, in many applications one is not interested in the exact angular dependence of the intensity but rather in some angle integrated quantities. Hence, it would be advantageous if it was possible to compute these quantities without computing the angular

dependent intensity I in a first step. Indeed, this is possible under some further simplifying assumptions which are readily applicable in many real world problems, e.g. isotropic scattering. In this case, the radiative transfer equation can be reformulated into a system of weakly singular integral equations of the second kind, where the position is the only independent variable. When a discretization method is applied to this equation, the number of unknowns is reduced tremendously compared with that of a discretization of the original integro-differential equation. Therefore, in this thesis the integral equation is used as a basis for the development of efficient and reliable algorithms to solve radiative transfer problems in participating media.

Besides the reduction of the number of unknowns, the integral equation has some further advantageous features. First, the operator is coercive, therefore, a standard Galerkin discretization is guaranteed to be stable. This is opposed to the direct discretization of the radiative transfer equation. Since the latter involves a transport operator, some kind of up-winding or artificial diffusion has to be used to obtain a stable discretization.

Second, the operator is self-adjoint. This leads to a symmetric matrix when the Galerkin method is used for the discretization. Therefore, the amount of storage needed is nearly reduced by one half.

Third, the resulting system is well conditioned in most cases, i.e. all eigenvalues are positive and well separated from zero independently of the grid size used. This property implies that iterative methods can be used for an efficient solution of the linear system; e.g. Krylov subspace methods yield very fast convergence. These methods do not need an explicit knowledge of the entries of the matrix. Only the possibility to perform a matrix-vector multiplication is required. As shown below, this fact is essential for an efficient implementation of the discretization of the integral equation.

Removing the angular dependency in the integral equation, leads to a tremendous reduction of the number of unknowns in the discrete system. On the other hand, the differential operator, which is local in space, has been replaced by an integral operator, which couples every point in the domain with every other point. Therefore, a discretization of the equation leads to a full matrix. For practical applications the number of discretization points must typically be of an order of magnitude of 10000 or more. Therefore, it is not recommendable to assemble and store the entire matrix.

The so-called matrix compression methods circumvent the assembly of the whole matrix. Instead, the matrix-vector multiplication needed by the iterative solver is performed only approximately. By this, the effort compared with that when the full matrix is used is reduced tremendously. Since the discrete solution is only an approximation to the solution of the continuous problem, the additional error introduced by approximating the discrete system need not diminish the overall accuracy if it is smaller than the discretization error.

As opposed to the fast Fourier transform, the matrix compression methods do not exploit an algebraic structure of the matrix. Instead, an approximation of the integral operator is constructed based on geometrical and analytical considerations. The integral kernels have a singularity for $x = y$ and are very smooth in regions at a sufficient distance from the diagonal. This corresponds to the physical fact that two points which are near together generally have a stronger influence on each other than two points further apart. On those parts of the domain where the kernel is smooth an approximation can be used, which is easier to handle numerically.

The matrix compression methods have been developed in context of the boundary element method, where a boundary value problem in a domain is transformed into an integral equation

on its surface. The kernels of these integral equations are very similar to the kernels of the integral equation describing the radiative transfer. Therefore, with some modifications, the matrix compression methods are readily applicable to the latter equation.

This thesis concentrates on the discussion of numerical solution methods for the stationary radiative transfer in scattering media based on the integral formulation of the problem. To this end the whole process from the derivation of the equation, and the discretization to the efficient implementation of the solution of the resulting linear system is considered. The outline of this thesis is as follows:

Chapter 1 introduces the mathematical model of the problem. Furthermore, some numerical methods used for the discretization of this model are presented.

Chapter 2 outlines the derivation of the integral formulation in case of isotropic scattering and diffuse reflecting boundaries. The mapping properties of the involved integral operators such as compactness and contractivity are examined in detail. Based on these properties, it is shown that the integral equation has a unique solution under some weak assumptions on the physical parameters which are readily applicable in real world problems.

The integral formulation in case of isotropic scattering is well known. By employing the so-called momentum equation, we are able to derive an integral equation which is also valid in case of linear anisotropic scattering. This equation is very similar to the equation for the isotropic case. No additional unknowns are introduced and the involved integral operators have very similar mapping properties. Therefore, the modified integral equation can be handled numerically with approximately the same effort as the equation for the isotropic case.

Chapter 3 addresses the numerical treatment of the integral equation. First of all, the Galerkin method, which is used to discretize the equation, is introduced. The convergence of the method is easily proven by using the mapping properties of the integral operators discussed in Chapter 2. The practical realization of the method involves the evaluation of integrals which cannot be performed analytically, because of the weak singularity of the integral kernels. Thus, some regularizing coordinate transformations have to be used before standard cubature formulae can be applied. These coordinate transformations are introduced in Section 3.2.

Furthermore, the solution of the linear system by Krylov methods is discussed. Since the integral equation is of the second kind, it is reasonable to expect that the spectrum is well suited for a Krylov method. This proves to be true for a wide range of parameters describing the problem, but if the medium is strongly scattering, the system is ill-conditioned. In this case, applying a suitable preconditioner still leads to fast convergence.

The chapter is concluded with some numerical experiments which compare the solution obtained by a discretization of the integral equation with that of a method which discretizes the radiative transfer equation directly. For the latter the so-called discrete ordinate method is used.

In Chapter 4 the concept of hierarchical matrices is introduced. It represents a common basis for many of the matrix compression methods used for matrices resulting from the discretization of integral equations. Thereafter, this concept is applied to the equation describing radiative transfer. As mentioned above, the matrix compression methods have been originally applied in context of the boundary element method, therefore, special emphasize is placed on the differences between the integral equations describing radiative transfer and the ones occurring in the boundary element method.

Radiative transfer in participating media changes its character in dependence on the rate of extinction due to absorption and scattering. For large extinction coefficients, radiative transfer is an almost local effect due to the short range of the radiation whereas for small extinction coefficients, radiative transfer is a global effect since the radiation spreads through the whole medium almost uninfluenced by the medium itself. One can make use of the different character by choosing the approximation rank of the matrix compression method in dependence on the rate of extinction, e.g. in case of large extinction coefficients a small approximation rank can be used on parts of the domain which are sufficiently apart.

The chapter is concluded with the presentation of some results concerning the difference in the computational complexity between the discretization of the integral equation and the discrete ordinate method.

As a conclusion the final chapter recapitulates the most important considerations presented throughout this thesis.

Finally, the appendix contains some definitions and derivations which have been omitted in the text to improve the readability.

Chapter 1

Radiative Transfer

The starting point for all considerations in this thesis is the stationary radiative transfer equation. This equation including possible boundary conditions is introduced in Section 1.1. Special emphasis is placed on the term that models the scattering. In Section 1.2 two well known methods are presented, which are widely used to handle the problem numerically. Further details about radiative transfer can be found in the monographs [40], [42], [54] and [16].

1.1 The Mathematical Model

There are different physical models describing thermal radiation: the quantum mechanical model for photons, the wave model and the ray model. This thesis is based on the ray model.

According to Planck's hypothesis radiative energy travels in the form of discrete photons, and according to Maxwell's classical electro-magnetic theory it travels in the form of electro-magnetic waves. Both of these concepts have been successfully applied in the investigation of radiative heat transfer. For example, the results of the quantum theory have been used to determine the amount of energy emitted by any kind of matter at any given wavelength because of its temperature; on the other hand, results obtained from electro-magnetic wave theory have been used to predict the radiative properties of materials, such as reflexivity and emissivity.

When radiation is treated as an electro-magnetic wave, its propagation can be described by the solution of Maxwell's equations. These equations describe the radiation in terms of the magnetic and electric field vectors H and E. The simplest solutions of Maxwell's equations are so-called plane waves. These plane waves are used to derive boundary conditions for the ray model. The magnitude and direction of the transfer of electro-magnetic energy is given by the Poynting vector $S = E \times H$. In case of thermal radiation, i.e. a wavelength $\lambda \sim 0.1 - 100\mu m$, a numerical treatment of the wave model is not feasible since the grid size would need to be chosen smaller than the wavelength to avoid aliasing effects.

In the special case of non-polarized radiation, the ray model can be used. This model describes the energy transport by a single scalar quantity: the intensity I, where I is defined as amount of radiative energy transfered per unit time, solid angle, wavelength and area normal to the pencil of rays. When polarization effects are to be considered, Chandrasekhar treats the radiation intensity as a four-component vector in terms of the four Stokes parameters (see [42] for further details).

The radiative transfer equation describes the radiative intensity field I within an enclosure as a function of time t, location x, direction $\Omega \in S^2$ and electro-magnetic wavelength λ. Since photons travel with light speed (in vacuum: $c_0 \approx 3 \cdot 10^8 \frac{m}{s}$), a steady state is reached after a very short period of time. Hence, it is almost always possible to neglect the time dependence.

In case of thermal radiation the intensities for different wavelengths λ are in general not coupled through the scattering phase function (see below). This leads to the monochromatic radiative transfer equation, in which the wavelength λ only occurs as a parameter. In the following, the dependency on λ is omitted for notational convenience.

Let $D \subset \mathbb{R}^3$ be an open, bounded domain with a piecewise smooth boundary, i.e. $\partial D \in C^{0,1}$ (see Appendix A.1 for a characterization of surfaces). Then the intensity $I(x, \Omega)$ is the solution of an integro-differential equation, the so-called radiative transfer equation (RTE)

$$\Omega \cdot \nabla_x I(x,\Omega) + (\kappa(x) + \sigma(x)) I(x,\Omega) = \frac{\sigma(x)}{4\pi} \int_{S^2} \Phi(\Omega, \Omega') I(x,\Omega') d\Omega' + \kappa(x) B(T(x)), \quad x \in D, \Omega \in S^2. \quad (1.1)$$

Remark 1.1. The dependency on λ is omitted for notational convenience, but we should keep in mind that the coefficients κ and σ as well as the scattering phase-function Φ and the source term B in the equation depend on λ. In order to obtain a finite number of transport problems in practical applications, the λ-domain $[0, \infty)$ is decomposed into finitely many spectral bands, and the λ-dependent quantities are somehow averaged over these bands. The resulting equations are then solved for every spectral-band.

The RTE describes the propagation of the intensity along a ray and can be interpreted as a balance of radiative energy. During the transport through the medium in direction Ω (described by $\Omega \cdot \nabla_x$) the intensity is attenuated by absorption and scattering into other directions. These are taken into account for by the absorption coefficient $\kappa(x) \geq 0$ and the scattering coefficient $\sigma(x) \geq 0$. The total extinction is denoted by $\gamma = \kappa + \sigma$. We assume that the functions κ and σ are in L^∞.

The two remaining terms describe the gain of the radiative energy in direction Ω. On the one hand, intensity from other directions Ω' is scattered into the considered direction Ω. This "in-scattering" has contributions from all directions and has to be calculated by integrating over all possible directions $\Omega' \in S^2$. The scattering phase function $\Phi(\Omega, \Omega')$ describes the according probability that a ray from direction Ω' is scattered into direction Ω. This function is described in more detail below.

On the other hand, due to Planck's hypothesis every medium continuously emits electromagnetic radiation uniformly into all directions at a rate depending on the local temperature and on the properties of the material. The radiative energy due to this effect can be split into a material dependent and a material independent part: $\kappa(x, \lambda) B(T, \lambda)$. It is a thermodynamical constraint that the proportionality constant $\kappa(x, \lambda)$ for emission is the same as for absorption. The factor which is independent of the material is given by the Planck function

$$B(T, \lambda) = \frac{2c^2 h_p}{\lambda^5 (e^{\frac{c h_p}{\lambda k_B T}} - 1)}, \quad (1.2)$$

where h_p denotes the Planck constant, $c = \frac{c_0}{n_m}$ the speed of light in the considered material, c_0 the speed of light in vacuum, n_m the refraction index of the material, and k_B the Boltzmann constant. B depends on the considered wavelength λ, but is independent of the direction, i.e.

it is isotropic. In the following, the temperature distribution is assumed to be known, hence, $\kappa(x)B(T(x))$ is a given source term.

The total emitted intensity can be computed by integrating over λ. This yields Stefan's law

$$B_{\text{total}}(T) = \frac{1}{\pi} n_m^2 \sigma_B T^4 \qquad \text{with } \sigma_B = \frac{2\pi^5 k_B^4}{15 h_p^3 c_0^2} = 5.670 \cdot 10^{-8} \frac{W}{m^2 K^4}. \tag{1.3}$$

This relation shows that radiation becomes an important mode of heat transfer compared with conduction and convection when the temperature is high since the magnitude of conduction and convection is linear in the temperature T.

The RTE describes radiation in terms of the intensity I, but in practical applications the radiative energy G and the radiative flux q are of higher interest. These quantities are related to the intensity in the following way

$$G(x) := \int_{S^2} I(x,\Omega) d\Omega, \tag{1.4}$$

$$q(x) := \int_{S^2} \Omega I(x,\Omega) d\Omega. \tag{1.5}$$

For a point $x \in \partial D$ the total incoming heat flux q_{in} and heat loss q_{out} are defined as

$$q_{\text{in}}(x) := \int_{n(x)\cdot\Omega<0} |n(x)\cdot\Omega| I(x,\Omega) d\Omega, \tag{1.6}$$

$$q_{\text{out}}(x) := \int_{n(x)\cdot\Omega>0} |n(x)\cdot\Omega| I(x,\Omega) d\Omega, \tag{1.7}$$

where $n(x)$ denotes the outer normal in the point $x \in \partial D$, i.e. in the first equation the integral is taken over all incoming directions and in the second over all outgoing directions.

Equation (1.1) is a differential equation of first-order. As such, it requires boundary conditions which prescribe the intensity for all incoming directions

$$I(x,\Omega) = I_b(x,\Omega), \quad (x,\Omega) \in (\partial D, S^2)_-, \tag{1.8}$$

where the set $(\partial D, S^2)_-$ is defined as

$$(\partial D, S^2)_- = \{(x,\Omega) \,|\, x \in \partial D,\, \Omega \in S^2|_{n(x)\cdot\Omega<0}\}. \tag{1.9}$$

At the boundary of the domain D the intensity is partially reflected and partially transmitted. The reflection can be modeled by the bidirectional reflection function $R(x,\Omega,\Omega')$ which describes the according probability that a ray from the outgoing direction Ω' is reflected into the incoming direction Ω:

$$I_{\text{refl}}(x,\Omega) = \int_{n(x)\cdot\Omega'>0} |n(x)\cdot\Omega'| R(x,\Omega,\Omega') I(x,\Omega') d\Omega'. \tag{1.10}$$

As a physical constraint, no radiative energy is produced by reflection, i.e.

$$\int_{n(x)\cdot\Omega>0} |n(x)\cdot\Omega| I(x,\Omega) d\Omega \geq$$

$$\int_{n(x)\cdot\Omega<0} |n(x)\cdot\Omega| \int_{n(x)\cdot\Omega'>0} |n(x)\cdot\Omega'| R(x,\Omega,\Omega') I(x,\Omega') d\Omega' d\Omega \quad \forall I(x,\Omega),$$

which results in the following condition for R

$$\int\limits_{n(x)\cdot\Omega<0} |n(x)\cdot\Omega| R(x,\Omega,\Omega')d\Omega \leq 1 \quad \forall x \in \partial D,\ \Omega' \in S^2|_{n(x)\cdot\Omega'>0}. \tag{1.11}$$

Furthermore, the reflection kernel has to be non-negative and, due to Kirchoff's law (refer to [40]), it should fulfill the following law of reciprocity

$$R(x,\Omega,\Omega') = R(x,-\Omega',-\Omega) \quad \forall x \in \partial D,\ \Omega \in S^2|_{n(x)\cdot\Omega<0},\ \Omega' \in S^2|_{n(x)\cdot\Omega'>0}. \tag{1.12}$$

Since the incident energy is distributed among the reflected and transmitted rays without loss, the transmissivity i.e. the fraction of transmitted radiation is

$$\tau(x,\Omega) = 1 - \int\limits_{n(x)\cdot\Omega'>0} |n(x)\cdot\Omega'| R(x,\Omega,\Omega')d\Omega' \geq 0. \tag{1.13}$$

Two cases have to be distinguished in the following: the surrounding material is absorbing or non-absorbing. If it is strongly absorbing, the thickness of the surface layer over which absorption of radiation from the inside occurs is very small. The same holds for emission of within the surrounding material which escapes into the considered domain. Therefore, it is customary to speak of absorption by and emission from a "surface", although a thin surface layer is implied. At local thermodynamical equilibrium the emissivity ϵ of the surface is again equal to the absorptivity α. Since all the radiation from the inside is absorbed it holds: $\epsilon = \alpha = \tau$. Summing up the emitted and reflected radiation, leads to the following boundary condition for the RTE:

$$I_b(x,\Omega) = \tau(x,\Omega)B(T_{\text{out}}(x)) + \int\limits_{n(x)\cdot\Omega'>0} |n(x)\cdot\Omega'| R(x,\Omega,\Omega')I(x,\Omega')d\Omega', \tag{1.14}$$

where T_{out} denotes the exterior temperature. Note that the speed of light used for the Planckian B in (1.14) is that inside the surrounding material.

Remark 1.2. The boundary conditions depend on the outer normal $n(x)$. If x lies on an edge or vertex of ∂D, $n(x)$ and thus also the boundary conditions are undefined. A weak formulation of the problem is introduced in Section 2.2, which circumvents this problem.

If the surrounding material is non-absorbing, radiation from the outside is transmitted into the medium. Due to the reciprocity condition (1.12) the transmissivity in the direction from the outside to the inside is the same as in the opposite direction. Therefore, assuming that the radiation from the exterior is a Planckian corresponding to the exterior temperature T_{out}, again leads to the boundary condition (1.14).

If the reflection function is independent of Ω and Ω':

$$R(x,\Omega,\Omega') = \frac{1}{\pi}\rho(x), \tag{1.15}$$

the surface reflects equal amounts of energy into all directions, regardless of the incoming direction. Such a surface is called a diffuse reflector. Using Equation (1.13) and (1.14) yields the boundary conditions

$$I_b(x) = (1-\rho(x))B(T_{\text{out}}(x)) + \frac{\rho(x)}{\pi}q_{\text{out}}(x). \tag{1.16}$$

Note that in case of diffuse reflecting boundaries the prescribed incoming intensity $I_b(x)$ is independent of Ω. The condition (1.11) leads to

$$0 \leq \rho(x) \leq 1 \quad \forall x \in \partial D. \tag{1.17}$$

In case of $\rho \equiv 0$, the surface is called a perfect absorber or a black boundary.

A diffuse reflector is a good approximation for rough surfaces. On the other hand, if the surface is smooth the reflection function R can be derived from the Maxwell equations. In the following, we assume for simplicity that both media are non-absorbing. The slightly more complicated situation of an absorbing medium is handled in [40].

Consider the situation where a plane wave hits a plane surface, i.e. the local radius of curvature is much greater than the wavelength. This wavefront is partially reflected and partially transmitted. Let θ_1 denote the angle between the propagation direction of the wavefront and the normal of the interface between the medium and the surroundings (see Figure 1.1).

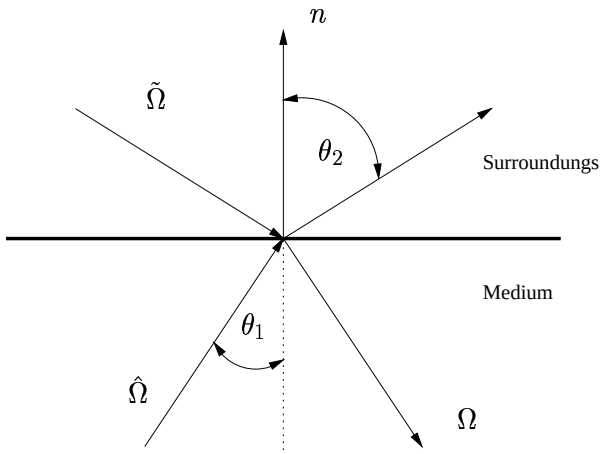

Figure 1.1: Specular reflecting boundary conditions

Furthermore, let n_m and n_s denote the refractive indices of the medium and the surroundings respectively. The speed of light inside the two media is given by $c_s = \frac{c_0}{n_s}$ and $c_m = \frac{c_0}{n_m}$. This gives a relationship between the directions of the incident and transmitted wave

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_m}{n_s}, \tag{1.18}$$

which is known as Snell's law. Note that, if $n_m > n_s$, θ_2 reaches a value of 90° for an angle of incidence θ_c, called the critical angle

$$\sin\theta_c = \frac{n_s}{n_m}. \tag{1.19}$$

For all $\theta_1 > \theta_c$ the radiation is totally reflected.

Besides the direction of reflection and transmission also the amount of reflected and transmitted radiative energy is of interest. The boundary conditions for the Maxwell equations can be used to derive the following expression for the amount of reflected radiation

$$\rho(\cos\theta_1) = \begin{cases} \frac{1}{2}\left[\left(\frac{n_m\cos\theta_2 - n_s\cos\theta_1}{n_m\cos\theta_2 + n_s\cos\theta_1}\right)^2 + \left(\frac{n_m\cos\theta_1 - n_s\cos\theta_2}{n_m\cos\theta_1 + n_s\cos\theta_2}\right)^2\right] & \text{if } \theta_1 \leq \theta_c, \\ 1 & \text{else.} \end{cases} \tag{1.20}$$

This relationship is known as Fresnel law for unpolarized light.

Using the reciprocity condition (1.12) leads to the following boundary condition

$$I_b(x,\Omega) = (1-\rho(n(x)\cdot\Omega))B(T_{\text{out}}(x)) + \rho(n(x)\cdot\Omega)I(x,\hat{\Omega}), \tag{1.21}$$

where ρ denotes the reflection index calculated by the Fresnel law and $\hat{\Omega}$ denotes the reflection direction belonging to the incoming direction Ω (see Figure 1.1). In later chapters of this thesis, we only consider the case of diffuse reflecting boundaries.

For later use, an operator notation is introduced for the case of homogeneous Dirichlet boundary conditions, i.e.

$$I_b(x,\Omega) = 0. \tag{1.22}$$

The operator H_Ω below is defined on the function space

$$\{I(x,\Omega)\,|\,x\in\bar{D},\ \Omega\in S^2,\ I(x,\Omega)=0 \text{ for } x\in\partial D \text{ with } n(x)\cdot\Omega<0\}, \tag{1.23}$$

i.e. the boundary conditions are included into the definition of the domain of the operator H_Ω,

$$\begin{aligned} H_\Omega &= \Omega\cdot\nabla_x + \gamma Id, \qquad H = \bigotimes_\Omega H_\Omega, \\ \Sigma_\Omega &= \frac{1}{4\pi}\int_{S^2}\Phi(\Omega,\Omega')\cdot' d\Omega', \qquad \Sigma = \bigotimes_\Omega \Sigma_\Omega, \\ T &= H - \sigma\Sigma. \end{aligned} \tag{1.24}$$

With these abbreviations, the RTE (1.1) reads

$$TI = \kappa B. \tag{1.25}$$

In the following the scattering phase function Φ is examined in more detail. It has to satisfy the conditions: $\forall\Omega,\ \Omega'\in S^2$

$$\Phi(\Omega,\Omega') \geq 0, \tag{1.26}$$

$$\Phi(\Omega,\Omega') = \Phi(-\Omega',-\Omega), \tag{1.27}$$

$$\frac{1}{4\pi}\int_{S^2}\Phi(\Omega,\Omega')d\Omega = 1. \tag{1.28}$$

The non-negativity requirement (1.26) is necessary since the intensity I is a non-negative quantity. Condition (1.27) ensures that the scattering phase function preserves symmetry of radiation between forward and backward propagation and Condition (1.28) guarantees that the total radiative energy is conserved during a scattering event. The intensity which is lost in direction Ω' due to out-scattering is gained in some other direction Ω by in-scattering.

The above properties of the phase function are necessary to derive the important momentum equation. Recalling the definition of G and q, the integration of the RTE (1.1) over all directions $\Omega\in S^2$ results in

$$\nabla\cdot q(x) + (\kappa(x)+\sigma(x))G(x) = \sigma(x)\int_{S^2}\frac{1}{4\pi}\int_{S^2}\Phi(\Omega,\Omega')I(x,\Omega')d\Omega' d\Omega + 4\pi\kappa(x)B(T(x)).$$

After interchanging the order of integration, $I(x, \Omega')$ is independent of the inner integration variable Ω and can be taken out of the inner integral. Due to condition (1.28), the inner integral is one, independent of Ω'. Hence, the following momentum equation is obtained

$$\nabla \cdot q(x) + \kappa(x)G(x) = 4\pi\kappa(x)B(T(x)). \tag{1.29}$$

In most applications the scattering phase function only depends on the angle θ between the direction before and after the scattering event

$$\Phi(\Omega, \Omega') = \Phi(\Omega \cdot \Omega'). \tag{1.30}$$

In this case, the reciprocity condition (1.27) is automatically satisfied. Furthermore, the kernel is symmetric and rotationally invariant

$$\Phi(\Omega, \Omega') = \Phi(\Omega', \Omega), \tag{1.31}$$

$$\Phi(\Omega, \Omega') = \Phi(\mathbf{R}\Omega, \mathbf{R}\Omega') \qquad \forall \text{ orthogonal matrices } \mathbf{R}. \tag{1.32}$$

The simplest phase function is $\Phi \equiv 1$, i.e. equal amounts of energy are scattered into all directions. This case is called isotropic scattering.

For a given material the phase function can be determined either by theoretical considerations or by experiments. In both cases, it is useful to expand it as a series of Legendre polynomials (see Appendix A.2) and truncate this series for k large enough

$$\Phi(\theta) = 1 + \sum_{p=1}^{k-1} A_p P_p(\cos\theta).$$

(Note, the coefficient A_0 has to be equal to one to satisfy condition (1.28) since the Legendre Polynomials are orthogonal.)

It remains to determine the coefficients A_p. For the theoretical investigations it is assumed that the material is adequately modeled as a cloud of spherical particles. The scattering phase function of a single particle can be computed solely in dependence on the complex index of refraction of the material and the size of the sphere in comparison with the considered wavelength. The rather complicated theory behind these computations is based on viewing radiation as electro-magnetic waves and is called Mie theory (refer to [32]). The calculations for a given set of particle parameters are rather involved, except for the particles being very small. In this case, the situation is simplified by taking the appropriate limits in the general solution to Mie's equations, resulting in Rayleigh scattering

$$\Phi(\theta) = \frac{3}{4}(1 + \cos^2\theta). \tag{1.33}$$

If the particle size is very large, geometrical optics can be used instead of the Mie theory, which also simplifies the problem. To get the scattering phase function of the whole particle cloud, one has to assume that the scattering occurs independently for the individual particles. This is the case if the distance between two particles is large compared with their diameter. Figure 1.2 shows a typical phase function computed by Mie theory. This figure is taken from [40] (page 396). The considered particles are non-absorbing with a complex index of refraction $m = 2$. The particles are either in clouds of constant radius $a = 5.0\mu m$, or in clouds with particles of different size ($a = 0.0 - 10.0\mu m$). The Mie calculations have been carried out for a typical wavelength of $\lambda = 3.1416\mu m$, resulting in a size parameter of $x = \frac{2\pi a}{\lambda} = 10$ and $0 \leq x \leq 20$ respectively. Besides the phase functions for these two clouds,

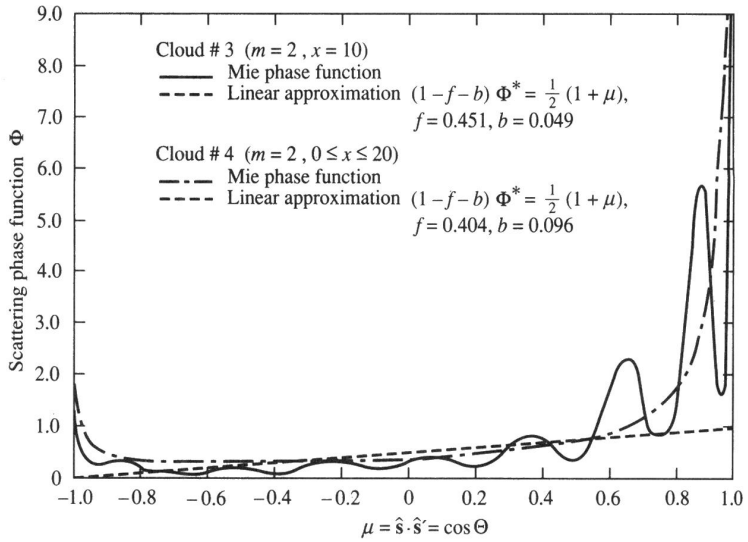

Figure 1.2: Mie scattering phase function for clouds of dielectric particles

a linear approximation of each is displayed. These approximations are considered later in this section.

The typical properties of a phase function are a very strong peak in the forward direction and, for dielectrical particles, a relative strong peak in the backward direction. In the region between the forward and backward peak the function is highly oscillatory, but if the cloud consists of particles of different sizes, which is a reasonable assumption in practical applications, these oscillations are smoothed (see Figure 1.2).

Numerical solution methods for the RTE either cannot handle complicated phase functions at all, or have to be carried out to unacceptably high orders. It is, therefore, common practice to approximate oscillatory phase functions by simpler expressions. In [41] the following double Dirac-delta phase function approximation is proposed

$$\frac{1}{4\pi}\int_{S^2}\Phi(\Omega\cdot\Omega')I(x,\Omega')d\Omega' \simeq fI(x,\Omega)+bI(x,-\Omega)+$$

$$(1-f-b)\frac{1}{4\pi}\int_{S^2}\Phi^*(\Omega\cdot\Omega')I(x,\Omega')d\Omega', \qquad \Phi^*(\mu)=1+A_1^*\mu.$$

That is, the forward scattering is treated as transmission and the backward peak is treated as being scattered into the opposite direction. In between the phase function is approximated by a linear function.

The coefficients f, b and A^* are determined heuristically. One chooses cutoff angles θ_f and θ_b for forward and backward scattering respectively. Then the slope A^* is chosen such that the linear function is a good approximation of the original function over the midrange of the phase function. The peak fractions f and b are given by the integrals of the original phase function over the cutoff intervals. Examples for the choice of phase function approximations following [41] are included in Figure 1.2. The numerical experiments for a one dimensional slab geometry in [41] show that the above approximations yield results of adequate accuracy for nearly all practically important cases in radiative heat transfer.

1.2 Approximations and Numerical Methods

Roughly speaking there are two kinds of computational models for radiative transfer. The first ones are based on certain assumptions, under which the physics involved can be simplified and therefore modeled by an equation which is easier to handle numerically. Some examples are the Rosseland approximation [48], the modified diffusion approximations [37], [58] and the diffusion approximation described below. On the other hand, these methods are inaccurate if the assumptions do not hold.

The second kind of methods handles the radiative transfer Equation (1.1) directly. They are therefore expensive. In this section, one representative of each kind is presented. The two selected methods are of interest in later chapters of this thesis.

The Discrete Ordinate Method

The discrete ordinate method (DOM) is a tool to transform the RTE into a set of simultaneous partial differential equations. It can be carried out to any arbitrary order and accuracy. The DOM is based on a discrete approximation of the directional variation of the radiative intensity, i.e. one chooses a finite set of directions instead of considering the whole unit sphere S^2. Like this, spherical integrals, e.g. for the evaluation of the scattering integral or for computing the energy or the heat flux, are approximated by a quadrature rule

$$\int_{S^2} f(\Omega) d\Omega \approx \sum_{i=1}^{m} w_i f(\Omega_i), \tag{1.34}$$

where w_i is the quadrature weight associated with the direction Ω_i. Thus, Equation (1.1) is approximated by a set of m coupled, linear convection equations for the unknown quantities $I_i(x) = I(x, \Omega_i)$, $i = 1, \ldots, m$

$$\underbrace{\Omega_i \cdot \nabla I_i + \gamma I_i}_{=H_{\Omega_i} I_i} = \frac{\sigma}{4\pi} \sum_{j=1}^{m} w_j \Phi(\Omega_i, \Omega_j) I_j + \kappa B(T), \tag{1.35}$$

subjected to the boundary conditions

$$I_b(x) = (1-\rho) B(T_{\text{out}}(x)) + \frac{\rho(x)}{\pi} \sum_{n(x)\cdot\Omega_j < 0} w_j I_j |n(x) \cdot \Omega_j|. \tag{1.36}$$

Once the intensities have been determined, the desired direction-integrated quantities, like the energy G, are readily calculated using the quadrature formula (1.34).

A straight forward approach for the selection of the ordinate directions is based on the parameterization of S^2 over $[0, 2\pi] \times [0, \pi]$ and uses a tensor product subdivision of this parameter space. This so-called longitude-latitude mesh has two distinct directions at the poles. This causes a non-physical symmetry axis in the solution and deteriorates convergence, since the cells near the poles have degenerate angles. To avoid such effects, several conditions which a set of quadrature points and weights should fulfill have been proposed

- rotational invariance under any rotations of 90°,
- exact computation of zeroth, first, and second moments,

- exact computation of half-space moments in coordinate directions (This might be important for the approximation of diffuse reflecting boundary conditions.),
- exact integration of spherical harmonics up to a certain degree,
- low discrepancy of the discrete measure defined by the directions and weights and the surface measure on the sphere,
- nearly equi-distribution of the ordinate directions on S^2.

Some of these requirements exclude each other so that one has to restrict oneself to a few conditions, which are most important for the desired application. A more detailed discussion of the selection of the directions and weights under the viewpoint of radiative transport can be found in [18] and [22].

For the numerical calculations presented in later chapters of this thesis, the following approach, which tries to equi-distribute the quadrature points on S^2, is used (see also Kanschat [31]). A triangulation of S^2 is constructed by using a successive subdivision of an icosahedron. The cell centers projected on S^2 are the quadrature points and spherical cell volumes serve as weights. The method can be interpreted as a Galerkin discretization with piecewise constant elements. Due to a super convergence result the error of the L^2-projection onto the space of piecewise constant elements is of second order in the cell centers (see [31]). A refined icosahedron with $m = 20, 80, 320, 1280$ cells can be used.

Several finite difference, finite volume and finite element methods for the discretization of the transport operators in Equation (1.35) have been proposed. Two of them should be mentioned here: the diamond difference method and the streamline diffusion finite element method.

The diamond difference method, developed in [12], uses piecewise trilinear ansatz functions on a grid consisting of hexahedral cells. The discretized equation is obtained by integrating Equation (1.35) over a cell. The major drawback of this method is that it requires very small grid sizes to guaranty that the computed intensities are non-negative, especially if the total extinction coefficient γ is large.

The second method is the streamline diffusion finite element method. It is applied to the RTE in [31]. Standard Galerkin discretizations produce spurious oscillations when applied to a transport operator like H_{Ω_i} in Equation (1.35). This behavior is due to the fact that these methods are stable only in L^2. The streamline diffusion finite element method is a Petrov-Galerkin scheme for the transport operator H_{Ω_i}, where $\psi + \delta\Omega_i \cdot \nabla_x \psi$ (with a suitable chosen small parameter δ) is used as a test function. This corresponds to adding a small amount of diffusion in transport direction only. Since this is done in a consistent way, there is no loss of accuracy. The method is stable and it has been proven that the discretization error is of order $O(h^{\frac{3}{2}})$ if piecewise linear elements on a general grid are used, but there is evidence for second order convergence on nearly all computationally interesting grids. Unfortunately, a mathematical proof for second order is available only on Cartesian grids (see [31]).

Since both, the discretization of the integral operator and the discretization of the differential operator, employ Galerkin methods, the two methods can be combined to yield a Galerkin discretization of the whole equation. This has two advantages. First, one is not restricted to meshes based on the standard tensor product splitting into $L^2(D) \otimes L^2(S^2)$, but can use different spatial grids for different directions. This might be useful in the transport dominated case, i.e. for optically thin media. Second, the whole method possesses a property

called Galerkin orthogonality: let I and I_h denote the solution the radiative transfer equation on the whole space and on the finite dimensional ansatz space V_h respectively. Then

$$\langle T(I - I_h), w_h \rangle = 0 \quad \forall w_h \in W_h, \tag{1.37}$$

where W_h denotes the test space. This property is very useful to prove higher order accuracy and to obtain an a posteriori error estimator. A posteriori error control techniques can be applied to the whole method. This is done in [31] to adaptively refine the spatial grid such that a given error tolerance is met.

When modeling $3D$ radiative transfer problems, stability and accuracy are not the only criteria for the usage of a particular code. Reasonable memory and CPU requirements are equally important. These requirements shall be estimated in the following.

Let n denote the degrees of freedom of the spatial grid and m the number of ordinates in S^2. Hence, the total number of unknowns is $n \cdot m$, which is very high in practical applications, e.g. $n = 50000$ and $m = 320$. The discrete system is of the form

$$\mathbf{A}\mathbf{u} = (\mathbf{H} - \mathbf{\Sigma})\mathbf{u} = \mathbf{f}, \tag{1.38}$$

with the vector $\mathbf{u}$ containing the discrete intensities and the vector $\mathbf{f}$ the values of the source term. Both vectors are of length $n \cdot m$ and the ordering of the unknowns is as follows:

$$(x_1, \Omega_1), \ldots, (x_n, \Omega_1), \ldots, (x_1, \Omega_m), \ldots, (x_n, \Omega_m).$$

The matrices $\mathbf{H}$ and $\mathbf{\Sigma}$ have the following block structure

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{H}_m \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \omega_{11}\mathbf{S}_1 & \cdots & \omega_{1m}\mathbf{S}_1 \\ \vdots & \ddots & \vdots \\ \omega_{m1}\mathbf{S}_m & \cdots & \omega_{mm}\mathbf{S}_m \end{pmatrix},$$

where $\omega_{ij} = w_j \Phi(\Omega_i, \Omega_j)$ (the w_j's are the weights of the quadrature rule (1.34)). The matrix $\mathbf{H}_i$ corresponds to the discretization of the operator H_{Ω_i} and the matrix $\mathbf{S}_i$ to the discretization of the identity operator. In case of the streamline diffusion method the used test functions depend on the considered direction Ω_i and, hence, the matrix $\mathbf{S}_i$ depends on i. In case of the diamond difference method $\mathbf{S}_i = \mathbf{S}$ is independent of i. The sparseness of the matrices $\mathbf{H}_i$ and $\mathbf{S}_i$ depends on the used spatial discretization, in case of streamline diffusion there are up to 27 entries per row (epr), in case of diamond difference, it holds: $epr = 8$.

Since, in case of the Petrov-Galerkin discretization, the memory requirements for the storage of the matrices $\mathbf{H}_i$ and $\mathbf{S}_i$ are quite high, they are not stored explicitly. Instead, the matrix entries are only computed for some large reference elements. The matrix entries for the refined elements are then determined by a scaling argument. To obtain the result of a matrix-vector multiplication with the entire matrix, each element matrix is multiplied with the corresponding part of the vector and the results are summed up (for more details refer to [31]). Therefore, the number of operations needed to perform one matrix-vector multiplication is approximately twice as high compared with a standard implementation of the matrix, but the storage requirements are reduced tremendously. For the finite difference method one simply stores the stencils.

Concerning the CPU time, there are two different tasks which should be considered:

- the costs for assembling the system matrix. They are negligible since the involved matrices are not stored explicitly.

- the costs for an iterative solution of the linear system (1.38). They are mainly determined by the costs for a matrix-vector multiplication. Iterative solution methods are discussed in Section (3.3).

The costs for a matrix-vector multiplication with the sparse matrices $\mathbf{H}_i$, $i = 1, \ldots, m$, are easily determined to be $m \cdot n \cdot epr$. To explain how the multiplication of the matrix $\mathbf{\Sigma}$ with the vector $\mathbf{u}$ is performed, $\mathbf{u}$ is decomposed into m vectors $\mathbf{u}_j$: $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)^T$ where $\mathbf{u}_j \in \mathbb{R}^n$ corresponds to $(x_1, \Omega_j), \ldots, (x_n, \Omega_j)$. This decomposition implies

$$\mathbf{\Sigma u} = \Big(\mathbf{S}_i \underbrace{(\sum_j \omega_{ij} \mathbf{u}_j)}_{=:\mathbf{v}_i} \Big)_{i=1}^{m}. \tag{1.39}$$

Hence, the costs are $m^2 \cdot n$ for the computation of $\mathbf{v}_i$, $i = 1, \ldots, m$ and $m \cdot n \cdot epr$ for the computation of $\mathbf{S}_i \mathbf{v}_i$, $i = 1, \ldots, m$. That is, the number of operations grows quadratically with the number of ordinates.

If the scattering phase function Φ has a special structure, these costs can be reduced tremendously. Assume that the scattering is isotropic. Then it holds $\omega_{ij} = w_j$ and, hence, $\mathbf{v}_i = \mathbf{v}$ is independent of i and the computation needs only $m \cdot n$ operations. This is the most simple case, but the costs can be reduced also for more complicated phase functions: assume that Φ can be expanded into a sum of Legendre polynomials

$$\Phi(\Omega, \Omega') = \sum_{p=0}^{k-1} A_p P_p(\Omega \cdot \Omega').$$

Using the addition theorem for spherical harmonics (see Appendix A.2) yields

$$\Phi(\Omega, \Omega') = \sum_{p=0}^{k-1} \frac{4\pi A_p}{2p+1} \sum_{q=-p}^{p} Y_p^q(\Omega) Y_p^{-q}(\Omega'),$$

and, hence,

$$\omega_{ij} = w_j \Phi(\Omega_i, \Omega_j) = w_j \sum_{p=0}^{k-1} \frac{4\pi A_p}{2p+1} \sum_{q=-p}^{p} Y_p^q(\Omega_i) Y_p^{-q}(\Omega_j) =: \sum_{l=1}^{L} \alpha_i^{(l)} \beta_j^{(l)}, \quad L = k^2.$$

Thus, the computation of $\mathbf{v}_i$, $i = 1, \ldots, m$, simplifies to

$$\mathbf{v}_i = \sum_j \omega_{ij} \mathbf{u}_j = \sum_{l=1}^{L} \alpha_i^{(l)} \underbrace{\sum_{j=1}^{m} \beta_j^{(l)} \mathbf{u}_j}_{=:\mathbf{w}^{(l)}},$$

and the overall costs for the computation of the matrix-vector-product with $\mathbf{\Sigma}$ reduce to $m \cdot n \cdot (2L + epr)$.

If the scattering matrices $\mathbf{S}_i = \mathbf{S}$ are independent of i (e.g. in case of the diamond difference method), the computation can be simplified further:

$$(\mathbf{\Sigma u})_i = \mathbf{S} \mathbf{v}_i = \sum_{l=1}^{L} \alpha_i^{(l)} \mathbf{S} \mathbf{w}^{(l)}.$$

In this case the overall costs are $m \cdot n \cdot 2L + L \cdot n \cdot epr$.

These costs are of interest in later chapters of this thesis when the discrete ordinate method is compared with a method for the discretization of the integral formulation of the RTE introduced in Chapter 2.

The Diffusion Approximation

The characteristics of radiation strongly depend on the optical parameters. If the optical thickness (product of the total extinction coefficient γ and a typical length scale l) is large, radiation is an almost local effect since two points which have a sufficiently large optical distance have nearly no influence on each other due to the large damping. For small optical thickness, radiation is a global effect since the radiation spreads out over the whole medium almost uninfluenced by the medium itself.

If the scattering coefficient σ is zero, the transport of the radiative energy in the interior takes place along straight lines, and the intensities in different directions are only coupled through the boundary conditions. On the other hand, if σ is large, the mean free path between two scattering events is very small, i.e. the direction of the ray changes very rapidly. Therefore, there is no distinguished direction of transport anymore, and the energy transport can be modeled as a diffusion process. Indeed, under the above assumptions it is possible to derive a diffusion equation from the RTE by asymptotic analysis (see [5]).

For simplicity we assume that the coefficients κ and σ do not depend on x. Furthermore, we restrict ourselves to homogeneous Dirichlet boundary conditions, i.e. $I_b(x,\Omega) = 0$, because this situation is examined in Section 3.3 when the preconditioning of the discrete system stemming from the discretization of the integral equation (see Chapter 2) is discussed. For the treatment of anisotropic and reflecting boundary conditions refer to [4] and [8] respectively.

To be able to apply an asymptotic analysis to the RTE (1.1), it is necessary to bring the equation in a dimensionless form first. To this end, we introduce the notation

$$\begin{aligned} x' &= \frac{x}{l_{\text{ref}}}, \qquad \kappa' = \frac{\kappa}{\kappa_{\text{ref}}}, \qquad \sigma' = \frac{\sigma}{\sigma_{\text{ref}}}, \\ I' &= \frac{I}{B_{\text{ref}}}, \qquad (B(T))' = \frac{(B(T))}{B_{\text{ref}}}, \qquad \text{with } B_{\text{ref}} = \frac{1}{\pi} n_m^2 \sigma_B T_{\text{ref}}^4, \end{aligned}$$

where l_{ref}, κ_{ref}, σ_{ref} and I_{ref} denote the reference scales and the primed quantities denote the non-dimensional quantities.

The assumptions about the size of the optical parameters are specified more accurately by the following definition.

Definition 1.1. The parameters of the problem are in the asymptotic diffusion limit if there is a small parameter $\varepsilon \ll 1$ such that

$$\varepsilon^2 = \frac{\kappa_{\text{ref}}}{\sigma_{\text{ref}}} \qquad \text{and} \qquad l_{\text{ref}}\sigma_{\text{ref}} = \frac{1}{\varepsilon}. \tag{1.40}$$

The first equation says that the scattering is dominant when compared with absorption, and the second says that the medium is optically thick.

Assuming that the problem parameters are in the asymptotic diffusion limit, yields the following dimensionless form of the RTE (To simplify the notation the primes for the dimensionless quantities are neglected.)

$$\frac{1}{\varepsilon}\Omega \cdot \nabla_x I(x,\Omega) + \kappa I(x,\Omega) + \frac{\sigma}{\varepsilon^2}[I(x,\Omega) - (\Sigma I)(x,\Omega)] = \kappa B(T(x)). \tag{1.41}$$

Using an ansatz of the form

$$I(x,\Omega) = I_0(x,\Omega) + \varepsilon I_1(x,\Omega) + \varepsilon^2 I_2(x,\Omega) + \ldots$$

and collecting terms of the same order in ε, gives

$$I_0 - \Sigma I_0 = 0, \tag{1.42}$$

$$\Omega \cdot \nabla_x I_0 + \sigma(I_1 - \Sigma I_1) = 0, \tag{1.43}$$

$$\Omega \cdot \nabla_x I_1 + \sigma(I_2 - \Sigma I_2) + \kappa I_0 = \kappa B, \tag{1.44}$$

$$\Omega \cdot \nabla_x I_{i+1} + \sigma(I_{i+2} - \Sigma I_{i+2}) + \kappa I_i = 0, \quad \text{for } i \geq 1. \tag{1.45}$$

Assuming that the phase function is of the form (1.30), yields that Σ is a self-adjoint operator w.r.t. the scalar product $\langle \cdot, \cdot \rangle_{L^2(S^2)}$. Furthermore, Σ is strictly positive, hence, its principal eigenvalue is simple (Harris-Krein-Rutman theorem, see [30]) . The relation

$$\frac{1}{4\pi} \int_{S^2} \Phi(\Omega, \Omega') d\Omega' = 1, \quad \Phi(\Omega, \Omega') \geq 0,$$

implies that the principal eigenvalue is one, and that the corresponding eigenvector is the function $1 : \Omega \to 1$. Therefore, the kernel of the operator $Id - \Sigma$ is the one dimensional space spanned by a constant. Hence, Equation (1.42) implies that $I_0(x, \Omega) = I_0(x)$ is independent of Ω.

The Fredholm alternative states that the equation

$$(Id - \Sigma) I = g$$

is solvable if, and only if, $g \in Ker(Id - \Sigma)^{\perp}$, i.e. if g has mean value zero, $\int_{S^2} g(\Omega) d\Omega = 0$. Since $\int_{S^2} \Omega_i d\Omega = 0$, there exists a function f_i such that

$$f_i(\Omega) - \Sigma f_i(\Omega) = \Omega_i, \quad \int_{S^2} f_i(\Omega) d\Omega = 0. \tag{1.46}$$

The second equation is a normalizing condition to make the solution unique. Let f denote the vector valued function with components f_i $(1 \leq i \leq 3)$. Then the solution of Equation (1.43) is given by

$$I_1(x, \Omega) = -\frac{f(\Omega)}{\sigma} \cdot \nabla_x I_0(x) + w(x), \tag{1.47}$$

where the function $w(x)$, which is independent of Ω, is to be determined later.

The function I_2 is given by relation (1.44). This equation admits a solution if, and only if,

$$\int_{S^2} \Big(\Omega \cdot \nabla_x I_1 + \kappa I_0 - \kappa B \Big) d\Omega = 0.$$

Inserting the expression for I_1 and exploiting the fact that I_0, B and w are independent of Ω yields

$$-\frac{1}{\sigma} \sum_{i,j} \underbrace{\int_{S^2} \Omega_i f_j(\Omega) d\Omega}_{=:a_{ij}} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} I_0 + 4\pi\kappa I_0 = 4\pi\kappa B. \tag{1.48}$$

The symmetry and rotation-invariance of the phase function Φ is transfered to the matrix $\mathbf{A} = (a_{ij})_{i,j=1}^3$ (see [5]). Hence, $\mathbf{A} = \alpha \mathbf{I}, \alpha > 0$ (where $\mathbf{I}$ denotes the identity matrix) and (1.48) is an isotropic diffusion equation

$$-\frac{\alpha}{\sigma} \Delta I_0 + 4\pi\kappa I_0 = 4\pi\kappa B. \tag{1.49}$$

The homogeneous Dirichlet boundary conditions of the RTE result in homogeneous Dirichlet conditions for I_0.

The Equation (1.44) can be written as

$$\sigma(I_2 - \Sigma I_2) = \underbrace{\kappa B - \kappa I_0 - \Omega \cdot \nabla(\frac{f(\Omega)}{\sigma}\nabla I_0)}_{=:\mathrm{rhs}^{(1)}} + \underbrace{\Omega \cdot \nabla w}_{=:\mathrm{rhs}^{(2)}}.$$

Since the equation is linear, the solution can be split into two parts $I_2 = I_2^{(1)} + I_2^{(2)}$. f is an odd function of Ω. Therefore, $\mathrm{rhs}^{(1)}$ is even. Due to the properties of the scattering phase function Φ, Σ maps even functions of Ω to even functions, and odd functions to odd functions. Furthermore, constant functions are the only eigenfunctions corresponding to the eigenvalue one. Therefore, $I_2^{(1)}$ is an even function of Ω. From the definition of f it follows that $I_2^{(2)} = -\frac{f(\Omega)}{\sigma} \cdot \nabla w$.

Putting this expression for I_2 into Equation (1.45) yields

$$\sigma(I_3 - \Sigma I_3) = -\kappa\Big(- \frac{f(\Omega)}{\sigma}\nabla I_0 + w\Big) - \Omega \cdot \nabla\Big(I_2^{(1)} - \frac{f(\Omega)}{\sigma}\nabla w\Big).$$

Since f is an odd and $I_2^{(1)}$ is an even function of Ω this equation is solvable if, and only if,

$$-\frac{\alpha}{\sigma}\Delta w + 4\pi\kappa w = 0. \tag{1.50}$$

Note that up to now, no boundary condition is prescribed for w.

With w satisfying (1.50) a solution of I_3 exists, and an approximate solution of Equation (1.41) can be defined as

$$\tilde{I} = I_0 - \varepsilon\Big(- \frac{f(\Omega)}{\sigma}\nabla I_0 + w\Big) + \varepsilon^2 I_2 + \varepsilon^3 I_3.$$

By construction, the error $r_\varepsilon = I - \tilde{I}$ satisfies the equation

$$\begin{aligned}\frac{1}{\varepsilon}\Omega \cdot \nabla_x r_\varepsilon + \kappa r_\varepsilon + \frac{\sigma}{\varepsilon^2}(Id - \Sigma)r_\varepsilon &= -\varepsilon^2(\kappa I_2 + \Omega \cdot I_3 + \varepsilon\kappa I_3) =: \varepsilon^2\delta_i(x,\Omega),\\ r_\varepsilon &= \varepsilon\Big(\frac{(f(\Omega)\cdot n)}{\sigma}\frac{\partial I_0}{\partial n} - w\Big) + \varepsilon^2\delta_b(x,\Omega), \qquad (x,\Omega) \in (\partial D, S^2)_-,\end{aligned}$$

where the fact that I_0 is constant on the boundary and, hence, its tangential derivatives are zero, has been exploited. n denotes the outer normal of D.

Since the RTE is stable, the error r_ε is bounded from above by the maximum of the consistency error in the interior (i.e. $\varepsilon^2\delta_i(x,\Omega)$) and the error on the boundary. It is easy to see that the dominant error comes from the boundary. The error of order ε depends on Ω. Hence, an approximation of order ε^2 cannot be obtained by prescribing a suitable boundary condition for w since w is independent of Ω. Therefore, a boundary layer term has to be introduced. This leads to a so-called Milne problem (see [5]). It can be shown that the solution of such a problem converges exponentially fast to a constant C. To match the boundary layer term with the solution in the interior, this constant, which is independent of Ω, is used to define a boundary condition for w

$$w(x) = \frac{C}{\sigma}\frac{\partial I_0}{\partial n}, \quad x \in \partial D. \tag{1.51}$$

Putting the equations for I_0 and w together, shows that the function $\hat{I}_\varepsilon := I_0 + \varepsilon w$ satisfies the following diffusion equation with Robin boundary conditions:

$$-\frac{\alpha}{\sigma}\Delta\hat{I}_\varepsilon + 4\pi\kappa\hat{I}_\varepsilon = 4\pi\kappa B, \quad x \in D, \tag{1.52}$$

$$\hat{I}_\varepsilon + \varepsilon\frac{C}{\sigma}\frac{\partial\hat{I}_\varepsilon}{\partial n} = 0, \quad x \in \partial D. \tag{1.53}$$

In [4] the following convergence result is shown

$$\|\int_{S^2} I(\cdot,\Omega)d\Omega - \hat{I}_\varepsilon\|_{L^2(D)} \leq \varepsilon^2 C. \tag{1.54}$$

Note that due to the boundary layer a similar estimation as the one above does not hold for the L^∞-norm.

It remains to determine the diffusion constant α and the constant C from the Milne problem. α is given by $\alpha = \int_{S^2} \Omega_1 f_1(\Omega)d\Omega$, i.e. f_1, which depends on the phase function Φ, has to be computed first. The case of isotropic and linear anisotropic scattering are considered in the following.

$\int_{S^2} \Omega_1 d\Omega = 0$, yields $f_1(\Omega) = \Omega_1$ in case of isotropic scattering and therefore

$$\alpha = \int_{S^2} \Omega_1^2 d\Omega = \frac{4\pi}{3}.$$

For linear anisotropic scattering Equation (1.46) is an integral equation of the second kind with a degenerate kernel. Hence, an ansatz of the following form can be made

$$f_1(\Omega) = \alpha_0 + \alpha_1\Omega_1 + \alpha_2\Omega_2 + \alpha_3\Omega_3,$$

which leads to the solution $f_1(\Omega) = \frac{3}{3-A_1}\Omega_1$ and hence

$$\alpha = \frac{4\pi}{3 - A_1}.$$

An analytical computation of C can only be carried out for simple cases, e.g the case of isotropic scattering in a $1D$ slab geometry, i.e. a plane-parallel plate, which extends into x and y direction until infinity, and where neither temperature nor radiative properties vary across planes parallel to the (x, y)-plane. These assumptions imply that the intensity only depends on the horizontal position, i.e. the z-coordinate, and the polar angle θ. Hence, the RTE simplifies to the following equation, where $\mu = \cos\theta$,

$$\mu\frac{\partial I}{\partial z}(z,\mu) + \gamma I(z,\mu) = \frac{\sigma}{2}\int_{-1}^{1} I(z,\mu')d\mu' + \kappa B(T(z)), \quad z \in (-z_0, z_0), \quad \mu \in [-1,1]. \tag{1.55}$$

As in the $3D$ case, the following operator notation is useful

$$H_\mu = \mu\frac{\partial}{\partial z}\nabla_x + \gamma Id, \qquad L = \frac{1}{2}\int_{-1}^{1} \cdot' d\mu', \tag{1.56}$$

where H_μ is defined on the function space

$$\{I(z,\mu) \,|\, z \in [-z_0, z_0],\ \mu \in [-1,1],\ I(-z_0,\mu) = 0\} \qquad \text{if } \mu > 0, \tag{1.57}$$

and

$$\{I(z,\mu) \,|\, z \in [-z_0, z_0],\ \mu \in [-1,1],\ I(z_0,\mu) = 0\} \qquad \text{if } \mu < 0. \tag{1.58}$$

In this case the constant C from the Milne problem is given by the constant of Chandrasekhar (see [5])

$$C = C_C = \int_0^1 \mu'^2 H(\mu') d\mu' \approx 0.7104, \tag{1.59}$$

where H denotes the Chandrasekhar function.

The boundary value problem (1.52), (1.53) plays an important role for the preconditioning of the discrete system resulting from the discretization of the integral equation discussed in Chapter 3. Since the integral equation is an equation for the energy G, it is helpful to rewrite the diffusion equation in terms of G. Note that the approximation $\hat{I}_\varepsilon$ is independent of Ω. Hence, the approximation for the energy is given by $G = 4\pi \hat{I}_\varepsilon$. Undoing the scaling results in the following diffusion equation in case of isotropic scattering

$$-\frac{1}{3\sigma}\Delta G + \kappa G = \kappa B, \quad x \in D, \tag{1.60}$$

$$G + \frac{C}{\sigma}\frac{\partial G}{\partial n} = 0, \quad x \in \partial D. \tag{1.61}$$

Equation (1.60) with σ replaced by γ can also be derived by plugging an ansatz of the form

$$I(x,\Omega) = \frac{1}{4\pi}\Big(G(x) + 3q(x)\cdot\Omega\Big) \tag{1.62}$$

into the RTE. This procedure is called P_1-approximation. The fact that σ is replaced by γ in the P_1-approximation does not mean a big difference since the equation anyway is valid only in the asymptotic diffusion limit and in this case σ and γ are almost identical. On the other hand, the ansatz (1.62) leaves some uncertain choices in the determination of the boundary conditions, and it does not contain any statement about when the approximation is valid.

For later use the diffusion operator is denoted by T_1

$$T_1 = -\frac{1}{3\gamma}\Delta + \kappa Id. \tag{1.63}$$

The operator is defined on the function space

$$\{G(x) \,|\, x \in \bar{D},\ G(x) + \frac{C}{\gamma}\frac{\partial}{\partial n}G(x) = 0 \text{ for } x \in \partial D\}, \tag{1.64}$$

i.e. the Robin boundary conditions (1.61) are included into the definition of the domain of the operator T_1.

Furthermore, we define the operator

$$\hat{T}_1 = T_1 + \sigma Id. \tag{1.65}$$

The main advantage of the numerical treatment of the diffusion approximation over treating the RTE directly, is the fact that the diffusion approximation does not depend on the angular variable. Therefore, the number of unknowns of the corresponding discrete problem is reduced tremendously. In the following chapter a formulation of the problem is presented, which is, like the diffusion approximation, independent of the angular variable, but does not depend on the validity of certain assumptions about the optical parameters.

Chapter 2

The Integral Formulation

The radiative transfer equation (1.1) describes the propagation of the intensity along the path of a ray. The intensity depends on five independent variables, three space coordinates and two coordinates describing the direction of the ray. The discretization of (1.1) e.g. by the discrete ordinate method or spherical harmonic expansion, therefore, leads to a very large number of unknowns. This is the major difficulty in the numerical treatment. However, in many applications one is not interested in the exact angular dependence of the intensity but rather in some angle integrated quantities like the energy G (see (1.4)). Hence, one might ask if it is possible to compute G without computing the angular dependent intensity I in a first step. Indeed, this is possible under certain conditions. In this chapter a reformulation of (1.1) into an integral equation for the energy is presented. The resulting system of equations is a well-posed formulation of the problem, i.e. a unique solution exists and depends smoothly on the input data. The reformulation is well known in case of isotropic scattering. In Section 2.4 the procedure is extended to the case of linear anisotropic scattering.

2.1 Derivation of the Integral Formulation

The radiative transfer equation (1.1) can be interpreted as a combination of ordinary differential equations along rays, coupled by the scattering integral and the boundary conditions. On this background, it is possible to integrate the equation formally along the rays with the coupling terms interpreted as source functions. This is the point of departure for the derivation of the integral formulation.

The aim is to get to a formulation which is independent of the angular variables. Since for specular reflecting boundaries, the intensity in an incoming direction depends on the intensity in one special outgoing direction, it is obvious that it is not possible to find an angle independent formulation in this case. Therefore, we restrict ourself to diffuse emitting and reflecting surfaces. Next, we have to consider the source term:

$$S(x,\Omega) = \frac{\sigma(x)}{4\pi} \int_{S^2} \Phi(\Omega,\Omega') I(x,\Omega') d\Omega' + \kappa(x) B(T(x)).$$

It consists of two parts: the emission term and the scattering integral. In case of thermal radiation the emission term B is given by the Planck function. It depends on the given temperature and the considered wavelength but not on the direction. On the other hand, it is obvious that arbitrary scattering phase functions Φ cannot be handled by an angle

independent formulation. The case of isotropic scattering leads to an isotropic source and can be handled easily. Since the case of linear anisotropic scattering will be considered in Section 2.4, the dependence on the angular variable is kept during the derivation of the integral formulation.

In the following, it is assumed that the boundary is smooth, i.e. $\partial D \in C^{1,\alpha}$. For non smooth boundaries, there are problems in the definition of reflecting boundary conditions (see Chapter 1). In Section 2.2 a weak formulation is introduced which is valid for arbitrary Lipschitz boundaries, $\partial D \in C^{0,1}$.

To simplify the notation the following abbreviations are used:

- $d(x,\Omega) := \max\{s \in \mathbb{R}^+ \,|\, x - t\Omega \in \bar{D} \quad \forall t \in [0,s]\}$,
- $x_b(x,\Omega) := x - d(x,\Omega)\Omega$, i.e. $x_b(x,\Omega)$ denotes the first point on the boundary crossed when going from $x \in D$ into direction $-\Omega$ (Note: for convex domains there is exactly one such intersection point.).

In order to integrate Equation (1.1), it is written as a set of linear ordinary differential equations:

$$\frac{dI}{ds}(x+s\Omega,\Omega) = -\gamma(x+s\Omega)I(x+s\Omega,\Omega) + S(x+s\Omega,\Omega), \qquad x \in \bar{D},\ \Omega \in S^2,\ s \in \mathbb{R} \quad \text{s.t.} \quad x+s\Omega \in \bar{D}. \tag{2.1}$$

Assuming that the source term S is given, this system can be solved analytically:

$$I(x,\Omega) = I_b(x_b(x,\Omega))e^{-\tau(x,x_b(x,\Omega))} + \int_0^{d(x,\Omega)} S(x-s\Omega,\Omega)e^{-\tau(x,x-s\Omega)}ds, \qquad x \in \bar{D},\ \Omega \in S^2, \tag{2.2}$$

where

$$\tau(x,y) = \|x-y\| \int_0^1 \gamma(tx+(1-t)y)dt. \tag{2.3}$$

To get rid of the dependency on the angular variable, we have to integrate over all directions $\Omega \in \mathcal{S}^2$. The first term on the right hand side in Equation (2.2) gives an integral over all boundary points which are visible from the point x. The line integral becomes an integral over a volume. A substitution of variables $y := x - s\Omega$ is used to account for that.

It holds $s = \|x-y\|$, $\Omega = \frac{x-y}{\|x-y\|}$ and

$$\begin{aligned} dy &= s^2 ds d\Omega \qquad \text{(volume element inside the domain } D\text{)}, \\ do(y) &= -\frac{s^2}{n(y)\cdot\Omega} d\Omega \qquad \text{(area element on the boundary } \partial D\text{)}. \end{aligned}$$

This substitution of variables yields

$$G(x) = \int_D v(x,y)\frac{e^{-\tau(x,y)}}{\|x-y\|^2} S(y, \frac{y-x}{\|y-x\|})dy + \int_{\partial D} v(x,y)\frac{n(y)\cdot(y-x)}{\|x-y\|^3} e^{-\tau(x,y)} I_b(y, \frac{y-x}{\|y-x\|})do(y) \quad \forall x \in D, \tag{2.4}$$

where the visibility function v is given by

$$v(x,y) = \begin{cases} 1 & \text{if } \{tx + (1-t)y \,|\, t \in [0,1]\} \subset \bar{D}, \\ 0 & \text{else.} \end{cases} \tag{2.5}$$

Remark 2.1. The following considerations give a physical interpretation: the square of the distance between the two points x and y appears in the dominator because isotropically emitted energy spreads out on concentric spherical surfaces. The measure of such a sphere grows quadratically with the radius. Hence, the strength of the influence between two points decays quadratically with their distance. This decay is the basis for the application of the matrix compression methods discussed in Chapter 4.

Assumption 2.1. We make the following assumptions.

1. The computational domain is convex. In this case the visibility function is equal to one: $v(x,y) = 1 \quad \forall x, y \in D$. This simplifies the numerical treatment of the equation since the discontinuous behavior of v for non convex domains leads to the following difficulties. The numerical approximation of the integrals described in Section 3.2 is much more difficult since the integrand is discontinuous. The matrix compression methods described in Chapter 4 are less efficient because they rely on the smoothness of the integral kernel. Hence, they can only be applied to regions where v is constant. In case of non convex domains we recommend to use some kind of domain decomposition.

2. In this section we assume further that the scattering is isotropic, i.e.

$$S(x,\Omega) = S(x) = \tfrac{\sigma(x)}{4\pi} G(x) + \kappa(x) B(T(x)).$$

The above assumptions lead to the following equation:

$$G(x) = \int_D \frac{e^{-\tau(x,y)}}{\|x-y\|^2} \Big(\frac{\sigma(y)}{4\pi} G(y) + \kappa(y) B(T(y)) \Big) dy + \int_{\partial D} \frac{n(y)\cdot(y-x)}{\|x-y\|^3} e^{-\tau(x,y)} I_b(y) do(y) \quad \forall x \in D. \tag{2.6}$$

For black boundaries, i.e. the reflection coefficient ρ is identically zero, $I_b(y)$ is a known quantity. Hence, Equation (2.6) is a closed system for the energy G, the so-called Peierls integral equation. The derivation of this equation can also be found for example in [56].

In case of homogeneous boundary conditions, i.e. $I_b \equiv 0$, and constant coefficients this equation can be written in the following form

$$(Id - \omega K) G = 4\pi (1-\omega) K B, \tag{2.7}$$

where the scattering albedo ω is defined as $\omega = \frac{\sigma}{\gamma}$. (In case of constant coefficients, we always assume that $\gamma > 0$. If this was not the case, the medium would be transparent, and the problem would reduce to radiative exchange between surfaces.). The integral operator K is given by

$$(KG)(x) = \frac{\gamma}{4\pi} \int_D \frac{e^{-\gamma\|x-y\|}}{\|x-y\|^2} G(y) dy. \tag{2.8}$$

For later use, an operator notation for the derivation of the integral equation is introduced. The formal solution can be interpreted as inverting the operator H_Ω (see (1.24)) for fixed Ω:

$$[H_\Omega^{-1}S](x,\Omega) = \int_0^{d(x,\Omega)} S(x - s\Omega)e^{-\tau(x,x-s\Omega)}ds. \tag{2.9}$$

The integration over all directions is denoted by the operator L

$$(LI)(x) = \frac{1}{4\pi}\int_{S^2} I(x,\Omega)d\Omega. \tag{2.10}$$

A right inverse L^+ ($LL^+ = Id$) of this operator is given by

$$(L^+G)(x,\Omega) = G(x) \qquad \forall\Omega. \tag{2.11}$$

By using these operators, the integral operator K can be decomposed as follows

$$K = \gamma LH^{-1}L^+. \tag{2.12}$$

The integral formulation for the one dimensional form (1.55) of the RTE with homogeneous Dirichlet boundary conditions is derived analogously to the $3D$ case: inverting the operator H_μ (see (1.56)) for fixed μ and integrating over all directions, i.e. applying the operator L of (1.56). Interchanging the order of integration, i.e. the integration w.r.t. μ becomes the inner integral, yields the equation

$$(Id - \omega K)G = 4\pi(1-\omega)KB, \tag{2.13}$$

for the energy $G = 4\pi LI$, where the integral operator $K = \gamma LH^{-1}L^+$ is given by

$$(KG)(x) = \frac{\gamma}{2}\int_0^l E_1(\gamma|x-y|)G(y)dy. \tag{2.14}$$

E_1 denotes the exponential integral of order one. The exponential integral of order n is defined as

$$E_n(x) = \int_1^\infty e^{-xt}\frac{1}{t^n}dt = \int_0^1 \mu^{n-2}e^{-\frac{x}{\mu}}d\mu, \quad x \geq 0 \quad n = 0,1,\ldots . \tag{2.15}$$

Differentiating Equation (2.15), a recurrence relationship is found as

$$\frac{dE_n}{dx}(x) = -E_{n-1}(x), \quad n = 1,2,\ldots, \tag{2.16}$$

especially

$$\frac{dE_1}{dx}(x) = -E_0(x) = -\frac{e^{-x}}{x}, \tag{2.17}$$

i.e. E_1 has a logarithmical singularity at the origin, and, hence, Equation (2.13) is a weakly singular Fredholm integral equation of the second kind.

Remark 2.2. The DOM of Section 1.2 is a discretization of the RTE. Since the integral equation (2.7) is derived from the RTE, the DOM can also be interpreted as a discretization of this equation. The integral operator K is replaced by an operator K_m, defined analogously to K, whereby the operator H_Ω is only inverted for a finite set of m ordinate directions Ω_i. Respectively, the integration over all directions is replaced by a weighted sum. Let $\mathbf{I} = (I_1, \ldots, I_m)$ denote the intensities in the chosen directions Ω_i. The discrete operators are defined as

$$(L_m \mathbf{I})(x) = \frac{1}{4\pi} \sum_{i=1}^{m} w_i I_i(x), \tag{2.18}$$

$$(L_m^+ G_m)(x, \Omega_i) = G_m(x) \qquad \forall i \in \{1, \ldots, m\}. \tag{2.19}$$

By using these operators, the operator K_m can be decomposed as follows

$$K_m = \gamma L_m \begin{pmatrix} H_{\Omega_1}^{-1} & & \\ & \ddots & \\ & & H_{\Omega_m}^{-1} \end{pmatrix} L_m^+ =: \gamma L_m H_m^{-1} L_m^+. \tag{2.20}$$

In Section 3.4 this method is compared with a numerical method which discretizes the integral equation directly. This method is introduced in Chapter 3.

Remark 2.3. The derivation of Equation (2.6) is valid only for $x \in D$. For $x \in \partial D$ the whole domain D is already covered by integrating over all directions with $n \cdot \Omega > 0$. Integration over the whole unit sphere gives:

$$G(x) = \int_{n\cdot\Omega<0} \ldots + \int_{n\cdot\Omega>0} = I_b(x) \underbrace{\int_{n\cdot\Omega<0} 1 d\Omega}_{=2\pi} + \int_D \ldots + \int_{\partial D} \ldots \quad . \tag{2.21}$$

That is, for x on the boundary an additional term appears in the equation. This is consistent with the following considerations. If the coefficients κ and σ are constant, it holds that $\tau(x, y) = \gamma \|x - y\|$. In this case the boundary integral can be written as a double layer potential: the vector valued function

$$k_{\text{vec}}(x, y) = -\frac{e^{-\gamma\|x-y\|}}{\|x - y\|^3}(y - x) \tag{2.22}$$

fulfills the integrability criterion for $y \in \mathbb{R}^3 \backslash \{x\}$, i.e. $\frac{\partial k_{\text{vec}\,i}}{\partial y_j} = \frac{\partial k_{\text{vec}\,j}}{\partial y_i}$. Since this domain is simply connected a primitive function k_{pot} w.r.t. y exists (i.e. $k_{\text{vec}}(x, y) = \nabla_y k_{\text{pot}}(x, y) \, \forall y \in \mathbb{R}^3 \backslash \{x\}$), e.g.:

$$k_{\text{pot}}(x, y) = \frac{e^{-\gamma\|x-y\|}}{\|x - y\|} - \gamma E_1(\gamma\|x - y\|). \tag{2.23}$$

Hence, Equation (2.6) can be written as

$$G(x) = \int_D \ldots - \int_{\partial D} \frac{\partial}{\partial n(y)} k_{\text{pot}}(x, y) I_b(y) do(y).$$

Since E_1 has a logarithmical singularity at the origin, k_{pot} has a singularity of first order. Hence, the above integral defines a double layer potential V. Let $\varphi : \partial D \to \mathbb{R}$:

$$(V\varphi)(x) = \frac{1}{4\pi} \int_{\partial D} \frac{\partial}{\partial n(y)} k_{\text{pot}}(x, y) \varphi(y) do(y), \quad x \in D.$$

The corresponding double layer operator is given by:

$$(D\varphi)(x) = \frac{1}{4\pi}\int_{\partial D}\frac{\partial}{\partial n(y)}k_{\text{pot}}(x,y)\varphi(y)do(y), \quad x \in \partial D.$$

It is well known that a double layer potential has a jump at the boundary (see [24]):

$$\lim_{x\to x_0}(V\varphi)(x) = -\frac{1}{2}\varphi(x_0) + (D\varphi)(x_0).$$

By using this jump relation, the limit $x \to \partial D$ in equation(2.6) is computed as

$$G(x) = \int_D \ldots + 4\pi\frac{1}{2}I_b(y) - \int_{\partial D}\frac{\partial}{\partial n(y)}k_{\text{pot}}(x,y)I_b(y)do(y).$$

This is the same relation as (2.21) derived directly from the RTE for $x \in \partial D$. The fact that the boundary integral is a double layer potential is used also in Section 4.3.

If the reflection coefficient ρ is greater than zero the intensities on the boundary for an incoming direction depend on the outgoing flux q_{out}.

$$\begin{aligned} G(x) = &\int_D \frac{e^{-\tau(x,y)}}{\|x-y\|^2}\Big(\frac{\sigma(y)}{4\pi}G(y) + \kappa(y)B(T(y))\Big)dy \\ &+ \int_{\partial D}\frac{n(y)\cdot(y-x)}{\|x-y\|^3}e^{-\tau(x,y)}\Big(\frac{\rho(y)}{\pi}q_{\text{out}}(y) + (1-\rho(y))B(T_{\text{out}}(y))\Big)do(y) \quad \forall x \in D. \quad (2.24)\end{aligned}$$

Hence, an additional equation is needed to close the system. This equation can be obtained by multiplying Equation (2.2) by $n\cdot\Omega$ and integrating over all directions with $n\cdot\Omega > 0$. For x on the smooth part of the boundary of a convex domain, integrating over $n\cdot\Omega > 0$, covers exactly the whole domain:

$$\begin{aligned} q_{\text{out}}(x) = &\int_D \frac{n(x)\cdot(x-y)}{\|x-y\|^3}e^{-\tau(x,y)}\Big(\frac{\sigma(y)}{4\pi}G(y) + \kappa B(T(y))\Big)dy \\ &+ \int_{\partial D}\frac{n(x)\cdot(x-y)}{\|x-y\|^2}\frac{n(y)\cdot(y-x)}{\|x-y\|^2}e^{-\tau(x,y)}\Big(\frac{\rho(y)}{\pi}q_{\text{out}}(y) + (1-\rho(y))B(T_{\text{out}}(y))\Big)do(y) \\ &\qquad \forall x \in \partial D. \quad (2.25)\end{aligned}$$

Remark 2.4. If x lies on an edge or vertex of the domain there are two difficulties:

1. The outer normal is not defined in such points.
2. There is no vector $\tilde{n}$ such that integrating over $\tilde{n}\cdot\Omega > 0$ covers exactly the whole domain.

Hence, there is no possibility to modify the equation in such a way that it is valid on the non smooth parts of the boundary.

Remark 2.5. The above derivation shows that, if the intensity I fulfills the RTE (1.1) with isotropic scattering and diffuse reflecting boundary conditions, then the corresponding energy G and flux q_{out}, defined by (1.4) and (1.7) respectively, satisfy the system of integral equations (2.24) and (2.25). On the other hand, it does not follow that for a solution (G, q_{out}) of the

integral equation, there exists an intensity I which solves the RTE, and such that G and q_{out} are given by

$$G(x) = \int_{S^2} I(x,\Omega)d\Omega, \qquad \text{and} \qquad q_{\text{out}}(x) = \int_{n(x)\cdot\Omega>0} n(x)\cdot\Omega I(x,\Omega)d\Omega.$$

This assertion is only true if the solution of the integral equation is unique, and if the existence of a solution of the RTE is guaranteed.

2.2 Existence and Uniqueness

The investigation of existence and uniqueness of the solution of the system (2.24) and (2.25) is based on an analysis of the mapping properties of the involved integral operators. These operators are defined as follows:

- $\forall x \in D$

$$(K_{11}u)(x) := \int_D k_{11}(x,y)u(y)dy := \int_D \frac{\sigma(y)}{4\pi}\frac{e^{-\tau(x,y)}}{\|x-y\|^2}u(y)dy,$$
$$(K_{12}u)(x) := \int_{\partial D} k_{12}(x,y)u(y)do(y) := \int_{\partial D} \frac{\rho(y)}{\pi}\frac{n(y)\cdot(y-x)}{\|x-y\|^3}e^{-\tau(x,y)}u(y)do(y),$$

- $\forall x \in \partial D$

$$(K_{21}u)(x) := \int_D k_{21}(x,y)u(y)dy := \int_D \frac{\sigma(y)}{4\pi}\frac{n(x)\cdot(x-y)}{\|x-y\|^3}e^{-\tau(x,y)}u(y)dy,$$
$$(K_{22}u)(x) := \int_{\partial D} k_{22}(x,y)u(y)do(y) := \int_{\partial D} \frac{\rho(y)}{\pi}\frac{n(x)\cdot(x-y)}{\|x-y\|^2}\frac{n(y)\cdot(y-x)}{\|x-y\|^2}e^{-\tau(x,y)}u(y)do(y).$$

The function $k_{ij}(x,y)$ is called kernel of the integral operator K_{ij}. Since D is convex, the kernels are always non-negative. They are continuous if $x \neq y$ and have a singularity on the diagonal $x = y$. In the following the notation $\bar{k}_{ij}(x,y)$ is used to denote the kernel without the factor $\sigma(y)$ in case of the volume integrals and $\rho(y)$ in case of the surface integrals.

By using the above definitions, the integral equation can be written in the following operator notation

$$\begin{pmatrix} G \\ q_{\text{out}} \end{pmatrix} - \underbrace{\begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}}_{:=K} \begin{pmatrix} G \\ q_{\text{out}} \end{pmatrix} = \underbrace{\begin{pmatrix} K_{11}^{(\kappa)} & K_{12}^{(1-\rho)} \\ K_{21}^{(\kappa)} & K_{22}^{(1-\rho)} \end{pmatrix} \begin{pmatrix} B \\ B_{\text{out}} \end{pmatrix}}_{=:f}, \tag{2.26}$$

where $K_{i1}^{(\kappa)}$ and $K_{i2}^{(1-\rho)}$ means that the factors $\frac{\sigma(y)}{4\pi}$ and $\frac{\rho(y)}{\pi}$ are replaced by $\kappa(y)$ and $(1-\rho(y))$ respectively.

Equation (2.26) is a Fredholm integral equation of the second kind. There are essentially two possibilities to show existence and uniqueness for this kind of equation. First, if the operator norm of K is smaller than one, a Neumann series argument can be used. Second, if the operator K is compact, Fredholm alternative can be used. Since both properties depend on the underlying function spaces, these spaces have to be defined first.

G is defined in the interior of the domain whereas q_{out} is defined on the boundary. Hence, function spaces on D and ∂D have to be defined. L^p-spaces can be defined on arbitrary measure spaces. Since ∂D has one dimension less than D, it is reasonable to scale the measure on D by σ, which has the unit of a reciprocal length. Furthermore, the surface measure is scaled by the dimensionless parameter ρ. This leads to some favorable properties, as will be shown later on.

Definition 2.1. Let $A_1 \subset D$ and $A_2 \subset \partial D$ be Borel sets. The measures μ_1 and μ_2 on these sets are defined via

$$\mu_1(A_1) = \int_{A_1} \sigma d\lambda^3, \quad A_1 \subset D, \qquad \mu_2(A_2) = \int_{A_2} \rho do(\lambda^2), \quad A_2 \subset \partial D, \tag{2.27}$$

where λ^n denotes the Lebesgue measure in $\mathbb{R}^n$.

The corresponding L^p-spaces, $p \geq 1$, are denoted by $L^p(D, \mu_1)$ and $L^p(\partial D, \mu_2)$. These two spaces have to be combined to yield one space. This can be done as follows (refer to [57]).

Let U, V be normed spaces, $1 \leq p \leq \infty$. Then

$$\|\binom{u}{v}\|_p := \begin{cases} (\|u\|_U^p + \|v\|_V^p)^{\frac{1}{p}} & \text{if } p < \infty, \\ \max\{(\|u\|_U, \|v\|_V\} & \text{if } p = \infty, \end{cases}$$

defines a norm on the direct sum $U \oplus V$. The normed space is denoted by $U \oplus_p V$. Furthermore, if U and V are complete, the same assertion holds for $U \oplus_p V$. If U, V are Hilbert spaces

$$\langle \binom{u_1}{v_1}, \binom{u_2}{v_2} \rangle := \langle u_1, u_2 \rangle_U + \langle v_1, v_2 \rangle_V$$

defines a scalar product on $U \oplus V$.

In the following the three spaces $L^p(\mu) := L^p(D, \mu_1) \oplus_p L^p(\partial D, \mu_2)$, $p = 1, 2, \infty$, are used. The norms and inner products on these spaces are given by

$$\|\binom{G}{q_{\text{out}}}\|_{L^1(\mu)} = \int_D \sigma(x)|G(x)|dx + \int_{\partial D} \rho(x)|q_{\text{out}}(x)|do(x), \tag{2.28}$$

$$\langle \binom{G^{(1)}}{q_{\text{out}}^{(1)}}, \binom{G^{(2)}}{q_{\text{out}}^{(2)}} \rangle_{L^2(\mu)} = \int_D \sigma(x) G^{(1)}(x) G^{(2)}(x) dx + \int_{\partial D} \rho(x) q_{\text{out}}^{(1)}(x) q_{\text{out}}^{(2)}(x) do(x), \tag{2.29}$$

$$\|\binom{G}{q_{\text{out}}}\|_{L^\infty(\mu)} = \max\{ \inf_{\substack{N \subset D, \\ \mu(N)=0}} \sup_{x \in D \setminus N} |G(x)|, \quad \inf_{\substack{N \subset \partial D, \\ \mu(N)=0}} \sup_{x \in \partial D \setminus N} |q_{\text{out}}(x)| \}. \tag{2.30}$$

L^p- spaces have the following properties (for a proof refer to [57]). Let ν be a measure on an arbitrary set A, and let $L^p(A, \nu)$, $1 \leq p \leq \infty$, denote the corresponding L^p-spaces. Then:

- $L^p(A, \nu)$, $p \geq 1$ is a Banach spaces, and $L^2(A, \nu)$ is a Hilbert space.
- Let $\frac{1}{p} + \frac{1}{q} = 1$. Let $f \in L^p(A, \nu)$, $g \in L^q(A, \nu)$. Then: $fg \in L^1(A, \nu)$ and

$$\|fg\|_{L^1} \leq \|f\|_{L^p} \|g\|_{L^q} \qquad \text{(Hölder inequality)}. \tag{2.31}$$

- If A is bounded, i.e. $\nu(A) < \infty$, there are positive constants c_1, c_2 such that

$$\|f\|_{L^1} \leq c_1 \|f\|_{L^2} \quad \forall f \in L^2(A, \nu), \tag{2.32}$$

$$\|f\|_{L^2} \leq c_2 \|f\|_{L^\infty} \quad \forall f \in L^\infty(A, \nu), \tag{2.33}$$

  i.e. $L^\infty(A, \nu) \subset L^2(A, \nu) \subset L^1(A, \nu)$.

- Let $\frac{1}{p} + \frac{1}{q} = 1$. The mapping $T : L^q(A, \nu) \to (L^p(A, \nu))'$ defined via

$$(Tg)(f) = \int_A f g d\nu \tag{2.34}$$

is an isometrical isomorphism. ($(L^p(A, \nu))'$ denotes the dual space of $L^p(A, \nu)$.)

Due to the definition of the norm on $U \oplus_p V$, the properties above also hold for $L^p(\mu)$, $1 \leq p \leq \infty$. The mapping T in (2.34) has to be replaced by

$$(Tg)(f) = \int_D f_1 g_1 d\mu_1 + \int_{\partial D} f_2 g_2 d\mu_2, \tag{2.35}$$

with $f = \binom{f_1}{f_2} \in L^p(\mu)$ and $g = \binom{g_1}{g_2} \in L^q(\mu)$.

Remark 2.6. Note that a subset $A_1 \subset D$ has measure zero, $\mu_1(A_1) = 0$, if $\sigma(x) = 0$ on A_1. The same holds for the boundary: a subset $A_2 \subset \partial D$ has measure zero, $\mu_2(A_2) = 0$, if $\rho(x) = 0$ on A_2. This means that functions which only differ on such sets are indistinguishable in the $L^p(\mu)$-spaces. This is in order since the solution on these sets does not influence the solution on other parts of the domain, hence, the problem decouples in the following sense: Let $N_1 \subset D$ and $N_2 \subset \partial D$ denote the largest set of measure zero, i.e.

$$N_1 = \bigcup_{\substack{A_1 \subset D, \\ \mu_1(A_1)=0}} A_1, \qquad N_2 = \bigcup_{\substack{A_2 \subset \partial D, \\ \mu_2(A_2)=0}} A_2.$$

Then the solution on $D \cup \partial D \backslash (N_1 \cup N_2)$ can be computed in a first step. If the solution on $D \cup \partial D \backslash (N_1 \cup N_2)$ is known, the solution on $N_1 \cup N_2$ can be computed by simply evaluating the integrals in Equation (2.26). Hence, it is sufficient to show existence and uniqueness in the $L^p(\mu)$-spaces.

The space of continuous functions $C(D) \oplus C(\partial D)$ equipped with the norm $\| \cdot \|_\infty$ is also a Banach space. In the following the two cases $C(D) \oplus C(\partial D)$ and $L^2(D, \mu_1) \oplus_2 L^2(\partial D, \mu_2)$ are treated since the two most popular methods for treating an integral equation numerically, namely the collocation method and the Galerkin method (refer to Chapter 3), are based on these two spaces respectively.

The Space $C(D) \oplus C(\partial D)$

To have a well defined equation it is necessary to show that K is a bounded operator from $C(D) \oplus C(\partial D)$ to $C(D) \oplus C(\partial D)$. If the boundary is smooth, the operator is even compact.

Theorem 2.1. *Let the boundary be smooth, i.e. $\partial D \in C^{1,\alpha}$ and assume that the functions $\kappa(x)$, $\sigma(x)$ and ρ are continuous.*

- *K_{11} is a compact operator from $C(D)$ to $C(D)$.*
- *K_{21} is a compact operator from $C(D)$ to $C(\partial D)$.*
- *K_{12} is a compact operator from $C(\partial D)$ to $C(D)$.*
- *K_{22} is a compact operator from $C(\partial D)$ to $C(\partial D)$.*

Proof. Consider the volume integrals. First of all, it has to be investigated if the functions $(K_{11}u)(x)$ and $(K_{21}u)(x)$ are continuous functions. Since both kernels are weakly singular and smooth outside the singularity at $x = y$, it is easy to show that $(K_{i1}u)(x)$ is bounded. The proof shows even more, namely that the family of functions $\{K_{i1}u \mid \|u\|_\infty \leq 1\}$ is uniformly bounded and equi-continuous. Hence, the lemma of Ascoli-Arzelà shows that the corresponding integral operator is compact. (For a detailed proof refer to [56] Lemma 2.2.)

For smooth surfaces $\partial D \in C^{1,\alpha}$, the following estimate is satisfied (see Appendix A.1)

$$|n(y) \cdot (y - x)| \leq C_{\partial D}\|x - y\|^{1+\alpha}. \tag{2.36}$$

Due to this fact the kernels k_{12} and k_{22} are integrable and compactness can be shown as in case of a volume integral. □

If the boundary is smooth, K is a compact operator on $C(D) \oplus C(\partial D)$ and Fredholm alternative can be applied. If ∂D is only piecewise smooth, the function space needs to be changed to account for the discontinuity of the outer normal $n(x)$ in non-smooth boundary points. We use the Banach space $L^\infty(\mu)$ to allow for discontinuities along edges and corners of ∂D.

The integrals $\int_D k_{11}(x, y)dy$ and $\int_{\partial D} k_{12}(x, y)do(y)$ exist for all $x \in D$. The integrals $\int_D k_{21}(x, y)dy$ and $\int_{\partial D} k_{22}(x, y)do(y)$ exist for all $x \in \partial D$ for which the surface normal is defined. Therefore, the mapping properties in case of piecewise smooth boundaries are very similar to the ones in case of smooth boundaries: the volume integral operators K_{11} and K_{21} are compact and the surface integral operators K_{12} and K_{22} are bounded in the $L^\infty(\mu)$-norm. On the other hand, as shown in [24], the compactness of the boundary integral operators is lost if the boundary has at least one vertex or edge. Hence, Fredholm alternative does not hold, instead a Neumann series argument has to be used to show existence and uniqueness.

Theorem 2.2. *Let V be a Banach space and $T \in L(V)$ satisfy $\|T\| < 1$. Then $Id - T$ is invertible with*

$$(Id - T)^{-1} = \sum_{n=0}^{\infty} T^n, \tag{2.37}$$

especially

$$\|(Id - T)^{-1}\| \leq (1 - \|T\|)^{-1}, \tag{2.38}$$

i.e. the corresponding operator equation of the second kind is uniquely solvable and the solution operator is bounded.

Proof. refer to [1] □

Remark 2.7. An operator which maps non-negative functions to non-negative functions is called a positive operator. If T is a positive operator with $\|T\| < 1$, then the operator $(Id - T)^{-1}$ is also positive due to the representation (2.37).

The proof of Theorem 2.3 (see below) shows that $\|K\|_{L^\infty(\mu)}$ is not smaller than one in general, but a simple scaling of q_{out} by factor four, yields an operator $\hat{K}$ which satisfies this property. Define: $\hat{q}_{\text{out}} = 4q_{\text{out}}$. This scaling can be interpreted as a kind of balancing of G and q_{out}, since the integrals involved in the definition of this quantities satisfy the following relation

$$\int_{S^2} 1 d\Omega = 4\pi = 4 \int_{n\cdot\Omega>0} n \cdot \Omega d\Omega.$$

The scaled equation is given by

$$\begin{pmatrix} G \\ \hat{q}_{\text{out}} \end{pmatrix} - \underbrace{\begin{pmatrix} K_{11} & \frac{1}{4}K_{12} \\ 4K_{21} & K_{22} \end{pmatrix}}_{:=\hat{K}} \begin{pmatrix} G \\ \hat{q}_{\text{out}} \end{pmatrix} = \hat{f}. \tag{2.39}$$

Note that the scaling can also be interpreted as a similarity transformation $\hat{K} = TKT^{-1}$ with $T = \left(\begin{smallmatrix} Id & 0 \\ 0 & 4Id \end{smallmatrix}\right)$. This will become important in the next subsection.

Theorem 2.3. *Let the optical parameters of the problem fulfill one of the following conditions:*

1. *There is a subset $A_1 \subset D$ with $\lambda^3(A_1) > 0$ such that $\kappa(y) > 0 \, \forall y \in A_1$.* *or*
2. *There is a $\varepsilon > 0$ such that for all $x \in \partial D$ it exist a subset $A_2 = A_2(x) \subset \partial D$ with $\lambda^2(A_2) \geq \varepsilon$ such that $n(x) \cdot (x - y) \geq \varepsilon$ and $\rho(y) \leq 1 - \varepsilon \, \forall y \in A_2$.*

Then $\hat{K}$ is a bounded operator from $L^\infty(\mu)$ to $L^\infty(\mu)$ with $\|\hat{K}\|_{L^\infty(\mu)} < 1$. In the special case of constant coefficients a sharper estimate can be given

$$\|\hat{K}\|_{L^\infty(\mu)} \begin{cases} < \max\{\omega, \rho\}, & \text{if } \rho \neq \omega \\ \leq \rho & \text{if } \rho = \omega. \end{cases} \tag{2.40}$$

where $\omega = \frac{\sigma}{\gamma}$ denotes the scattering albedo.

Proof. The L^∞-norm of an integral operator T

$$(Tu)(x) = \int_\Sigma k(x, y)u(y)dy \tag{2.41}$$

is estimated as follows:

$$\|T\|_\infty = \sup_{u \in L^\infty(\Sigma),\, \|u\|_\infty = 1} \sup_{x \in \Sigma} \left| \int_\Sigma k(x, y)u(y)dy \right| \leq \sup_{x \in \Sigma} \int_\Sigma |k(x, y)|dy. \tag{2.42}$$

To show the estimate (2.40), $\hat{K}$ is split into two parts: $\hat{K} = \left(\begin{smallmatrix} \hat{K}_1 \\ \hat{K}_2 \end{smallmatrix}\right)$, $\|\hat{K}\|_{L^\infty(\mu) \leftarrow L^\infty(\mu)} = \max\{\|\hat{K}_1\|_{L^\infty(D,\mu_1) \leftarrow L^\infty(\mu)}, \|\hat{K}_2\|_{L^\infty(\partial D,\mu_2) \leftarrow L^\infty(\mu)}\}$. According to (2.42), the norm of the combined volume and surface integral operators can be estimated as follows

$$\|\hat{K}_1\|_\infty \leq \sup_{x \in D}\{\int_D k_{11}(x, y)dy + \frac{1}{4}\int_{\partial D} k_{12}(x, y)do(y)\}, \tag{2.43}$$

and

$$\|\hat{K}_2\|_\infty \leq \sup_{x \in \partial D}\{4\int_D k_{21}(x, y)dy + \int_{\partial D} k_{22}(x, y)do(y)\}. \tag{2.44}$$

For constant coefficients, the values of the volume integrals are easily estimated by embedding the domain D into a larger standard domain, a sphere in case of K_{11} and a half sphere in case of K_{12}. On this standard domains the integral is easily computable. Since the kernels are always non-negative, the value of the integral can only be increased by enlarging the domain.

Estimating the values of the surface integrals directly is not possible because the surface cannot be embedded into a larger standard domain. The Theorem of Gauss can be used to

transform the boundary integrals into integrals over the volume. This gives an estimate for the integrals in (2.43) and (2.44) respectively. The transformation is rather complicated since the singularity of the kernels for $x = y$ has to be excluded before the Theorem of Gauss can be applied, i.e. a small ball $B_\varepsilon(x)$ is used and the limit $\varepsilon \to 0$ is contemplated. For more details refer to [3], where the case of radiative transfer between diffuse reflecting surfaces is treated.

A much simpler approach to obtain an estimate for the integrals in (2.43) and (2.44) respectively, reverses the variable transformation used in the derivation of the integral equation. By this approach, the fact that $\|\hat{K}\|_{L^\infty(\mu)} \leq 1$, is simply a consequence of the conservation of energy described by the RTE.

The volume integral in (2.43) is computed as follows

$$I_{11} := \int_D k_{11}(x,y)dy = \frac{1}{4\pi}\int_{S^2}\int_0^{d(x,\Omega)} \sigma(x - s\Omega)e^{-\tau(x,x-s\Omega)}ds d\Omega.$$

Using the fact that $\gamma = \sigma + \kappa$ and that

$$\frac{d}{ds}\tau(x, x - s\Omega) = \frac{d}{ds}\int_0^s \gamma(x - t\Omega) = \gamma(x - s\Omega),$$

yields

$$I_{11} = \frac{1}{4\pi}\int_{S^2} \underbrace{[e^{-\tau(x,x-s\Omega)}]_0^{d(x,\Omega)}}_{=1-e^{-\tau(x,x_b(x,\Omega))}} d\Omega - \int_D \kappa(y)\bar{k}_{11}(x,y)dy. \tag{2.45}$$

(Please, remember $\bar{k}_{ij}(x,y)$ is used to denote the kernel without the factor $\sigma(y)$ in case of the volume integrals and $\rho(y)$ in case of the surface integrals.)

The surface integral in (2.43) is given by

$$I_{12} := \int_{\partial D} k_{12}(x,y)do(y) = \frac{1}{\pi}\int_{S^2} \rho(x_b(x,\Omega))e^{-\tau(x,x_b(x,\Omega))}d\Omega.$$

Since the scaling of q_{out} gives a factor $\frac{1}{4}$ in front of the boundary integral, the following estimate holds

$$\|\hat{K}_1\|_\infty \leq \sup_{x\in D}\Big(1 - \frac{1}{4\pi}\int_{S^2}[1 - \rho(x_b(x,\Omega))]e^{-\tau(x,x_b(x,\Omega))}d\Omega - \int_D \kappa(y)\bar{k}_{11}(x,y)dy\Big).$$

An analog transformation of the integrals in (2.44) yields an estimate for $\|\hat{K}_2\|$

$$\|\hat{K}_2\|_\infty \leq \sup_{x\in\partial D}\Big(1 - \frac{1}{\pi}\int_{n(x)\cdot\Omega>0}[1 - \rho(x_b(x,\Omega))]n(x)\cdot\Omega e^{-\tau(x,x_b(x,\Omega))}d\Omega - 4\int_D \kappa(y)\bar{k}_{21}(x,y)dy\Big).$$

These estimates give the first assertion of the theorem. In case of constant coefficients the estimates simplify as follows:

$$\|\hat{K}_1\|_\infty \leq \sup_{x\in D}\Big[\frac{\sigma}{\gamma} + \underbrace{\frac{1}{4\pi}\int\limits_{S^2} \underbrace{e^{-\gamma\|x-x_b(x,\Omega)\|}}_{<1} d\Omega}_{\in(0,1)}(\rho - \frac{\sigma}{\gamma})\Big], \tag{2.46}$$

and

$$\|\hat{K}_2\|_\infty \leq \sup_{x\in\partial D}\Big[\frac{\sigma}{\gamma} + 4\underbrace{\frac{1}{\pi}\int\limits_{n(x)\cdot\Omega>0} n(x)\cdot\Omega\,\underbrace{e^{-\gamma\|x-x_b(x,\Omega)\|}}_{<1} d\Omega}_{\in(0,\frac{1}{4})}(\rho - \frac{\sigma}{\gamma})\Big]. \tag{2.47}$$

This gives the estimate (2.40). □

Remark 2.8. The optical thickness τ of the problem is defined as

$$\tau = \max_{x,y\in D} \tau(x,y). \tag{2.48}$$

For the following considerations it is assumed that the coefficients κ, σ and ρ are constant in space. In this case τ is given by $\tau = \gamma\cdot\mathrm{diam}(D)$. To see how the optical thickness influences the operator norm, two extreme cases are considered: the case of black boundaries, i.e. $\rho = 0$ and the case of a non-scattering medium, i.e. $\sigma = 0$. In both cases the system (2.26) decouples. In the first case one only has to consider the operator K_{11} in the equation for G and in the second case the operator K_{22} in the equation for $\hat{q}_{\mathrm{out}}$.

In the first case, the right hand side in (2.46) becomes maximal for x in the center of the domain. For this x the following estimate holds

$$\tau(x, x_b(x,\Omega)) \leq \frac{1}{2}\tau \quad \forall\Omega,$$

and, hence,

$$\|\hat{K}_{11}\|_{L^\infty} \leq \omega(1 - e^{-\frac{1}{2}\tau}). \tag{2.49}$$

That is, $\|K_{11}\|_\infty$ gets close to one if the scattering albedo, $\omega = \frac{\sigma}{\gamma}$, is close to one and τ is large, i.e. in the diffusion limit (see Section 1.2). This becomes important in the discussion of the convergence of iterative solution methods in Section 3.3.

In the second case, $\sigma = 0$, the right hand side in (2.47) becomes large when $e^{-\gamma\|x-x_b(x,\Omega)\|}$ is close to one for some directions Ω, i.e. for x near non-smooth boundary points or if the optical thickness of the problem is small.

If $\rho \approx \omega$, the optical thickness has little influence on the operator norms.

Using the estimates in Theorem 2.3 together with Theorem 2.2 gives

Theorem 2.4. *Let the prerequisite of Theorem 2.3 be fulfilled. Then the integral equation has a unique solution in $L^\infty(\mu)$ and the solution operator is continuous. Furthermore, the solution is non-negative.*

Proof. The existence of a unique solution follows from the Theorems 2.3 and 2.2. The integral operator $\hat{K}$ is a positive operator since all integral kernels k_{ij} are non-negative. Therefore, due to Remark 2.7, the operator $Id - \hat{K}$ is also positive. □

Remark 2.9. The non-negativity of the solution is a desirable result since the energy G and the flux q_{out} are non-negative physical quantities.

The Space $L^2(D,\mu_1) \oplus_2 L^2(\partial D,\mu_2)$

The space $L^2(D,\mu_1) \oplus_2 L^2(\partial D,\mu_2)$ is a Hilbert space. The weak formulation of the integral Equation (2.26) in this space reads as follows:

For given $f \in L^2(\mu)$, find $u \in L^2(\mu)$ such that

$$a(u,v) := \langle Au, v\rangle_{L^2(\mu)} = \langle f, v\rangle_{L^2(\mu)} \qquad \forall v \in L^2(\mu), \tag{2.50}$$

where the operator A is defined as $A := Id - K$. (Id denotes the identity operator.)

The original Equation (2.26) is only defined for a smooth boundary, but the weak form is valid for domains with Lipschitz boundary, i.e. $\partial D \in C^{1,0}$. An additional advantage is that $L^2(\mu)$ is a Hilbert space. This offers the possibility to exploit such powerful theorems as the Lax-Milgram lemma.

To have a well defined equation, the image of the integral operator has to be an L^2-function, e.g. for $u \in L^2(D)$, $K_{11}u$ has to belong to $L^2(D)$. A rather obvious approach to show this, is to use Cauchy-Schwarz inequality, but this attempt does not work since $k_{11} \notin L^2(D \times D)$. A more profound result from functional analysis, namely the Calderon-Zygmund lemma, has to be employed.

Theorem 2.5 (Calderon-Zygmund inequality). *Let $1 < p < \infty$. Let $\omega : \mathbb{R}^n \backslash \{0\} \mapsto \mathbb{R}$ be measurable on $\partial B_1(0)$ ($B_r(x)$ = ball around x with radius r) and bounded and satisfy the following two conditions*

- $\int\limits_{\partial B_1(0)} \omega(\xi) do(\xi) = 0,$
- $w(x) = w(\frac{x}{\|x\|}).$

Then for $0 < \varepsilon \leq 1$ and $f \in L^p(\mathbb{R}^n)$

$$(T_\varepsilon f)(x) := \int\limits_{\mathbb{R}^n \backslash B_\varepsilon(x)} \frac{w(x-y)}{\|x-y\|^n} f(y) dy$$

is a well defined function in $L^p(\mathbb{R}^n)$. Furthermore, the operator T_ε is a bounded linear operator from $L^p(\mathbb{R}^n)$ to $L^p(\mathbb{R}^n)$ and the limit $Tf := \lim\limits_{\varepsilon \to 0} T_\varepsilon f$ exists with

$$\|Tf\|_{L^p(\mathbb{R}^n)} \leq C(n,p) \|\omega\|_{L^\infty(\partial B_1(0))} \|f\|_{L^p(\mathbb{R}^n)}.$$

Proof. refer to [1], Chapter 8.20 □

Since the functions σ and ρ are in $L^\infty(D)$ and $L^\infty(\partial D)$ respectively, the mapping properties of K in L^2 do not depend on the special choice of the measure. Hence, in the following the spaces $L^2(D) := L^2(D,\lambda^3)$ and $L^2(\partial D) := L^2(\partial D,\lambda^2)$ equipped with the Lebesgue measure are used.

For the proofs of the mapping properties of K some additional theorems are needed.

The adjoint of a compact operator is compact.

Theorem 2.6 (Schauder). *Let U, V be Hilbert spaces $T \in L(U,V)$. Then*

$$T \in K(U,V) \qquad \Leftrightarrow \qquad T^* \in K(V,U).$$

Proof. refer to [1]. □

The embedding of H^1 in L^2 is compact.

Theorem 2.7 (Rellich). *Let $D \subset \mathbb{R}^n$ be open and bounded with Lipschitz boundary, i.e. $\partial D \in C^{0,1}$. Then the embedding*

$$H^s(D) \hookrightarrow L^2(D)$$

is compact for $s > 0$.

Proof. refer to [23], Theorem 6.4.8 c) □

These theorems are used to proof the following mapping properties of the integral operators in the L^2-spaces.

Theorem 2.8. *Let $\partial D \in C^{0,1}$ and assume that the functions $\kappa(x)$ and $\sigma(x)$ are continuous. Then*

1. *K_{11} is a compact operator from $L^2(D)$ to $L^2(D)$.*
2. *K_{12} is a compact operator from $L^2(\partial D)$ to $L^2(D)$.*
3. *K_{21} is a compact operator from $L^2(D)$ to $L^2(\partial D)$.*
4. *K_{22} is a bounded operator from $L^2(D)$ to $L^2(\partial D)$.*

Proof.

1. In order to apply Theorem 2.5 the kernel k_{11} is split up as follows

$$k_{11}(x,y) = \frac{\sigma(y)}{4\pi} e^{-\tau(x,y)} \underbrace{\frac{1}{\|x-y\|^2}}_{:=k_0(x,y)}.$$

Since the function k_0 contains the singularity, this part is most difficult to handle. By using the Calderon-Zygmund lemma, it can be shown that the operator K_0 defined by

$$(K_0 u)(x) := \int_{\mathbb{R}^3} k_0(x,y)u(y)dy = \int_{\mathbb{R}^3} \frac{1}{\|x-y\|^2} u(y)dy$$

maps $L^2(\mathbb{R}^3)$ to $H^1(\mathbb{R}^3)$. To this end an operator $K^{(\varepsilon)}$ ($\varepsilon > 0$ arbitrary, fixed) is defined by

$$(K^{(\varepsilon)}u)(x) := \int_{\mathbb{R}^3 \setminus B_\varepsilon(x)} \frac{1}{\|x-y\|^2} u(y)dy.$$

For $u \in C_0^0(\mathbb{R}^3)$, it holds $(K^{(\varepsilon)}u)(x) \in H^1(\mathbb{R}^3)$ with

$$\frac{d}{dx_i}(K^{(\varepsilon)}u)(x) = \int_{\mathbb{R}^3 \setminus B_\varepsilon(x)} \frac{-2\frac{x_i - y_i}{\|x-y\|}}{\|x-y\|^3} u(y)dy + r_i^\varepsilon(x;u). \tag{2.51}$$

The term $r_i^\varepsilon(x;u)$ stems from the fact that the domain of integration depends on x. It is given by (see Appendix A.3)

$$r_i^\varepsilon(x;u) := \int\limits_{\partial B_\varepsilon(x)} \frac{1}{\|x-y\|^2} u(y) \frac{x_i - y_i}{\varepsilon} do(y) = \int\limits_{\partial B_1(0)} \frac{1}{\varepsilon^2} u(x+\varepsilon y) \frac{\varepsilon y_i}{\varepsilon} \varepsilon^2 do(y)$$

$$= \int\limits_{\partial B_1(0)} y_i u(x+\varepsilon y) do(y) \overset{\varepsilon \to 0}{\to} \int\limits_{\partial B_1(0)} y_i do(y) u(x) = 0. \quad (2.52)$$

For $u \in L^2(\mathbb{R}^3)$ define

$$(T_\varepsilon^{(i)} u)(x) := \int\limits_{\mathbb{R}^3 \setminus B_\varepsilon(x)} \frac{-2\frac{x_i - y_i}{\|x-y\|}}{\|x-y\|^3} u(y) dy.$$

Since the requirements of Theorem 2.5 are fulfilled, $T_\varepsilon^{(i)}(u)$ is well defined in $L^2(\mathbb{R}^3)$, and the operator $T^{(i)}$ given by

$$(T^{(i)} u)(x) := \int\limits_{\mathbb{R}^3} \frac{-2\frac{x_i - y_i}{\|x-y\|}}{\|x-y\|^3} u(y) dy$$

is in $L(L^2(\mathbb{R}^3), L^2(\mathbb{R}^3))$.

It remains to show that $\forall u \in L^2(\mathbb{R}^3) : T^{(i)} u = \frac{d}{dx_i}(K_0 u)$ in $L^2(\mathbb{R}^3)$, i.e.

$$\langle T^{(i)} u, \phi \rangle = -\langle K_0 u, \frac{d}{dx_i} \phi \rangle \quad \forall \phi \in C_0^\infty(\mathbb{R}^3). \quad (2.53)$$

Let $u \in C_0^0(\mathbb{R}^3)$. Then

$$\langle T^{(i)} u, \phi \rangle = \langle \lim_{\varepsilon \to 0} T_\varepsilon^{(i)} u, \phi \rangle \overset{(2.51)}{=} \lim_{\varepsilon \to 0} \langle \frac{d}{dx_i} K^{(\varepsilon)} u, \phi \rangle - \underbrace{\langle \lim_{\varepsilon \to 0} r_i^\varepsilon(\cdot; u), \phi \rangle}_{\overset{(2.52)}{=} 0}$$

$$= -\lim_{\varepsilon \to 0} \langle K^{(\varepsilon)} u, \frac{d}{dx_i} \phi \rangle = -\langle K_0 u, \frac{d}{dx_i} \phi \rangle.$$

Since $C_0^0(\mathbb{R}^3)$ is dense in $L^2(\mathbb{R}^3)$, (2.53) holds for all $u \in L^2(\mathbb{R}^3)$. Hence, the assertion that $K_0 \in L(L^2(\mathbb{R}^3), H^1(\mathbb{R}^3))$ is shown. Restricting this result to the domain D and using the fact that the embedding from $H^1(D)$ to $L^2(D)$ is compact (refer to Theorem 2.7) gives that $K_0 \in K(L^2(D), L^2(D))$.

To show that the kernel $k_{11}(x,y) = \frac{\sigma(y)}{4\pi} e^{-\tau(x,y)} k_0(x,y)$ also yields a compact operator, a sequence of compact operators is constructed, which converges to K_{11} in the operator norm. Then the limit operator, i.e. K_{11}, is also compact (refer to [57]).

Let $\varepsilon > 0$ be arbitrary, fixed. Since $f(x,y) = e^{-\tau(x,y)}$ is continuous on the compact domain $\bar{D} \times \bar{D}$ the approximation Theorem of Weierstraß gives a polynomial $p_\varepsilon(x,y) = \sum_{\nu,\mu} \alpha_{\nu,\mu} x^\nu y^\mu$ such that $\|f - p_\varepsilon\|_{L^\infty(\bar{D} \times \bar{D})} \leq \varepsilon$. Define the kernel $k_\varepsilon(x,y) = \frac{\sigma(y)}{4\pi} p_\varepsilon(x,y) k_0(x,y)$. Then:

$$\|(K - K_\varepsilon)u\|_{L^2} = \Big(\int_D \Big[\int_D \frac{\sigma(y)}{4\pi} \underbrace{|f(x,y) - p_\varepsilon(x,y)|}_{\leq \varepsilon} k_0(x,y) |u(y)| dy \Big]^2 dx \Big)^{\frac{1}{2}}$$

$$\leq \varepsilon \frac{\|\sigma\|_\infty}{4\pi} \|K_0\|_{L^2} \|u\|_{L^2},$$

i.e.

$$\|K - K_\varepsilon\|_{L^2} \leq \varepsilon \frac{\|\sigma\|_\infty}{4\pi} \|K_0\|_{L^2}.$$

It remains to show that the operator K_ε is compact. K_ε can be written in the following form: $K_\varepsilon = \sum_{\nu,\mu} \alpha_{\nu,\mu} M_{x^\nu} \circ K_0 \circ M_{y^\mu} \circ M_{\frac{\sigma(y)}{4\pi}}$, where the multiplication operator M_g is defined as follows

$$M_g : \begin{cases} L^2 \to L^2, \\ f \to gf. \end{cases}$$

M_g is bounded if $g \in L^\infty$. Hence, M_{x^ν}, M_{y^μ} and $M_{\frac{\sigma(y)}{4\pi}}$ are bounded and $M_{x^\nu} \circ K_0 \circ M_{y^\mu} \circ M_{\frac{\sigma(y)}{4\pi}}$ is compact as a concatenation of continuous operators, where one of the operators is compact. Therefore, K_ε is compact as a sum of compact operators and the compactness of K_{11} is proven.

2. To proof that K_{12} is compact, again a multiplicative splitting of the kernel is used

$$k_{12}(x,y) = \frac{\rho(y)}{\pi} e^{-\tau(x,y)} \sum_{i=1}^{3} n_i(y) \underbrace{\frac{y_i - x_i}{\|x-y\|^3}}_{:=k_i(x,y)}.$$

$k_i(x,y) = -\frac{d}{dy_i} \frac{1}{\|x-y\|}$ is the derivative of the Newton volume potential. Therefore, the corresponding operator is a bounded operator from $L^2(\partial D)$ to $H^{\frac{1}{2}}(D)$ (for a proof refer to [39]). Hence, it is compact from $L^2(\partial D)$ to $L^2(D)$ due to Theorem 2.7. Using the same argument as in case of the operator K_{11} yields the compactness of K_{12}.

3. As shown below the operators K_{12} and K_{21} are adjoint except for a constant factor. Hence, the compactness of K_{21} follows by Theorem 2.6 from the compactness of K_{12}.

4. Since $\rho(y) \leq 1$ and $e^{-\tau(x,y)} < 1$ and $n(x) \cdot (x-y) \leq \|x-y\|$ the kernel k_{22} is bounded by the kernel of the double layer potential of the Laplace equation. Since the corresponding operator is a bounded operator from $L^2(\partial D)$ to $L^2(\partial D)$, the same statement holds for K_{22}.

□

The operator K_{22} is not compact for general surfaces and, hence, as in case of L^∞-functions, Fredholm alternative cannot be used. The fact that $L^2(\mu)$ is a Hilbert space offers the opportunity to employ the Lax-Milgram lemma to show existence and uniqueness. To this end it has to be shown that the bilinear form $a(\cdot,\cdot)$ is coercive.

Theorem 2.9 (Lax-Milgram). *Let V be a Hilbert space and $a(\cdot,\cdot)$ a continuous and coercive bilinear form, i.e. there exist positive constants C_s and α*

$$|a(u,v)| \leq C_s \|u\| \|v\| \quad \forall u, v \in V, \tag{2.54}$$

$$|a(u,u)| \geq \alpha \|u\|^2 \quad \forall u \in V. \tag{2.55}$$

and let $F : V \to \mathbb{R}$ be a bounded linear functional. Then it exists exactly one $u \in V$ such that

$$a(u,v) = F(v) \quad \forall v \in V.$$

Or in other words, there exists a unique continuous linear operator A on V such that

$$a(u,v) = \langle Au, v \rangle_V \quad \forall u, v \in V.$$

Furthermore, A is bijective and the following estimates hold: $\|A\| \leq C_s$ and $\|A^{-1}\| \leq \frac{1}{\alpha}$.

proof. refer to [1] □

For the following theoretical considerations as well as for the numerical treatment of the weak formulation, it would be advantageous if the operator K was self-adjoint. Due to the choice of the scalar products in $L^2(D,\mu_1)$ and $L^2(\partial D,\mu_2)$ the integral operators K_{11} and K_{22} are self-adjoint. Furthermore, it holds

$$\bar{k}_{12}(x,y) = 4\bar{k}_{21}(y,x), \tag{2.56}$$

and, therefore, $\forall v \in L^2(D,\mu_1)$, $\forall u \in L^2(\partial D,\mu_2)$

$$\begin{aligned}\langle K_{12}u, v\rangle_{L^2(D,\mu_1)} &= \int_D \sigma(x)v(x) \int_{\partial D} \rho(y)\bar{k}_{12}(x,y)u(y)do(y)dx \\ &= \int_{\partial D} \rho(y)u(y) \int_D \sigma(x)4\bar{k}_{21}(y,x)v(x)dxdo(y) = 4\langle u, K_{21}v\rangle_{L^2(\partial D,\mu_2)}.\end{aligned} \tag{2.57}$$

Hence, a similarity transformation $T = \left(\begin{smallmatrix} Id & 0 \\ 0 & 2Id \end{smallmatrix}\right)$ can be used to get a self-adjoint operator

$$K_{\text{sym}} = TKT^{-1} = \begin{pmatrix} K_{11} & \frac{1}{2}K_{12} \\ 2K_{21} & K_{22} \end{pmatrix}. \tag{2.58}$$

Since K is a bounded operator on $L^2(\mu)$, K_{sym} is also bounded. The corresponding bilinear form

$$a_{\text{sym}}(u,v) = \langle (Id - K_{\text{sym}})u, v\rangle_{L^2(\mu)} \tag{2.59}$$

satisfies the estimate

$$a_{\text{sym}}(u,u) \geq (1 - \|K_{\text{sym}}\|_{L^2(\mu)})\|u\|^2_{L^2(\mu)} \quad \forall u \in L^2(\mu). \tag{2.60}$$

Hence, $a_{\text{sym}}(\cdot,\cdot)$ is coercive if $\|K_{\text{sym}}\|_{L^2(\mu)} < 1$. As mentioned above, an estimate in this norm cannot be derived directly since the kernels are no L^2-functions. On the other hand, an estimate for the $L^1(\mu)$-norm is easily derived. Having an estimate in the stronger $L^\infty(\mu)$-norm and in the weaker $L^1(\mu)$-norm, yields an estimate in the required $L^2(\mu)$-norm by using interpolation theory on operator spaces. First of all, the estimate in the $L^1(\mu)$ norm has to be shown.

Theorem 2.10. *K is a bounded operator from $L^1(\mu)$ to $L^1(\mu)$. Furthermore, $\|K\|_{L^1(\mu)}$ satisfies the same estimates as $\|\hat{K}\|_{L^\infty(\mu)}$ in Theorem 2.3.*

Proof. The L^1-norm of an integral operator T

$$(Tu)(x) = \int_\Sigma k(x,y)u(y)dy \tag{2.61}$$

is estimated by using the Theorem of Fubini:

$$\begin{aligned}\|T\|_1 &= \sup_{u\in L^1(\Sigma),\,\|u\|_1=1} \int_\Sigma |\int_\Sigma k(x,y)u(y)dy|dx \\ &\leq \sup_{u\in L^1(\Sigma),\,\|u\|_1=1} \int_\Sigma \int_\Sigma |k(x,y)|dx|u(y)|dy \leq \|\int_\Sigma |k(x,\cdot)|dx\|_\infty \underbrace{\|u\|_1}_{=1}.\end{aligned}$$

This shows that K is a bounded operator from $L^1(\mu)$ to $L^1(\mu)$ since all kernels $k_{ij}(x,\cdot)$ are in L^∞. A more detailed investigation, regarding the fact that the measure μ is used instead of the Lebesgue measure, yields:

$$\begin{aligned}
\|K\binom{G}{q_{\text{out}}}\|_{L^1(\mu)} \leq & \int_D \sigma(x) \int_D \sigma(y)\bar{k}_{11}(x,y)|G(y)|dydx \\
& + \int_D \sigma(x) \int_{\partial D} \rho(y)\bar{k}_{12}(x,y)|q_{\text{out}}(y)|do(y)dx \\
& + \int_{\partial D} \rho(x) \int_D \sigma(y)\bar{k}_{21}(x,y)|G(y)|dydx \\
& + \int_{\partial D} \rho(x) \int_{\partial D} \rho(y)\bar{k}_{22}(x,y)|q_{\text{out}}(y)|do(y)dx \\
\leq & \int_D \underbrace{\Big[\int_D \sigma(x)\bar{k}_{11}(x,y)dx + \int_{\partial D} \rho(x)\bar{k}_{21}(x,y)do(x)\Big]}_{=:c_1(y)} \sigma(y)|G(y)|dy \\
& + \int_{\partial D} \underbrace{\Big[\int_D \sigma(x)\bar{k}_{12}(x,y)dx + \int_{\partial D} \rho(x)\bar{k}_{22}(x,y)do(x)\Big]}_{=:c_2(y)} \rho(y)|q_{\text{out}}(y)|dy \\
& \leq \max\{\sup_{y\in D} c_1(y), \sup_{y\in\partial D} c_2(y)\}\|\binom{G}{q_{\text{out}}}\|_{L^1(\mu)}.
\end{aligned}$$

Since $\bar{k}_{11}$ and $\bar{k}_{22}$ are symmetric in x and y, and since $\bar{k}_{12}(x,y) = 4\bar{k}_{21}(y,x)$, the quantities $c_1(y)$ and $c_2(y)$ are the same as the expressions in Equation (2.43) and (2.44) respectively. Hence, $\|K\|_{L^1(\mu)}$ satisfies the same estimates as $\|\hat{K}\|_{L^\infty(\mu)}$ □

Having an estimate for the $L^\infty(\mu)$-norm and the $L^1(\mu)$-norm the interpolation Theorem of Riesz-Thorin can be used to get an estimate for the $L^2(\mu)$-norm.

Theorem 2.11 (Riesz-Thorin interpolation theorem).
Let $1 \leq p_0, p_1, q_0, q_1 \leq \infty$, $0 < \theta < 1$ and let the indices p and q are given by

$$\frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}, \qquad \frac{1}{q} = \frac{1-\theta}{q_0} + \frac{\theta}{q_1}.$$

Furthermore, let ν_1 and ν_2 be measures and assume that

$$T : L^{p_0}(\nu_1) \to L^{q_0}(\nu_2) \quad \textit{with norm } M_0, \tag{2.62}$$
$$T : L^{p_1}(\nu_1) \to L^{q_1}(\nu_2) \quad \textit{with norm } M_1. \tag{2.63}$$

Then

$$\|Tf\|_{L^q} \leq cM_0^{1-\theta}M_1^\theta\|f\|_{L^q} \qquad \forall f \in L^{p_0}(\nu_1) \cap L^{p_1}(\nu_1),$$

with $c = 1$ in the complex case and $c = 2$ in the real case. Hence, the operator can be continued to a continuous linear operator

$$T : L^p(\nu_1) \to L^q(\nu_2) \quad \textit{with norm } M \leq cM_0^{1-\theta}M_1^\theta.$$

Proof. refer to [57] □

The spaces $L^p(D,\mu_1)\oplus_p L^p(\partial D,\mu_2)$, $p=1,2,\infty$, are sums of L^p spaces, therefore, the theorem above cannot be applied directly to these spaces. On the other hand, when looking at the proof of the theorem, it can be seen that the two facts needed about L^p-spaces, are the validity of the Hölder inequality (2.31), and the fact that the mapping T in (2.34) is an isometrical isomorphism. Both properties are satisfied for the $L^p(\mu)$-spaces. Therefore, the theorem is also valid for these spaces.

To exploit the estimates for $\|\hat{K}\|_{L^\infty(\mu)}$ and $\|K\|_{L^1(\mu)}$, Theorem 2.11 is applied with the indices $p_0=q_0=\infty$ and $p_1=q_1=1$. Setting $\theta=\frac{1}{2}$, yields $p=q=2$.

Remark 2.10. Theorem 2.11 gives an alternative proof of the boundedness of the operator K in $L^2(D,\mu_1)\oplus_2 L^2(\partial D,\mu_2)$, but Theorem 2.8 contains stronger assertions, namely that some parts of the operator are even compact.

Unfortunately Theorem 2.11 cannot be applied directly to yield the estimate $\|K\|_{L^2(\mu)}<1$ because of two facts

- The $L^\infty(\mu)$ estimate only holds for the scaled operator $\hat{K}$ but not for K itself.
- Because of the constant $c=2$ in the real case, only $\|K\|_{L^2(\mu)}<2$ could be shown, which is not enough for a Neumann series argument. For an example that $c=1$ does not hold in general in the real case refer to [57].

Nevertheless Theorem 2.11 is useful when examining the spectrum of K as a bounded operator in the different spaces. To this end some results from functional analysis are needed. The proofs can be found in [57].

Definition 2.2. Let V be a Banach space and $T\in L(V)$. The resolvent set $\rho(T)$ is defined as

$$\rho(T)=\{\lambda\in\mathbb{C}\,|\,(\lambda Id-T)^{-1}\text{ exists in }L(V)\},$$

and the spectrum is defined as $\sigma(T)=\mathbb{C}\backslash\rho(T)$.

The spectrum of a bounded operator is bounded by its norm. For normal operators in Hilbert spaces the converse assertion is also true.

Theorem 2.12. *Let V be a Banach space, $T\in L(V)$. Then*

$$\sup_{\lambda\in\sigma(T)}|\lambda|\le\|T\|. \tag{2.64}$$

If V is a Hilbert space and T a normal operator. Then

$$\sup_{\lambda\in\sigma(T)}|\lambda|=\|T\|. \tag{2.65}$$

Proof. refer to [57] □

Due to the assertions in Theorem 2.3 and Theorem 2.10, there is a positive constant $c\le 1$ such that $\|\hat{K}\|_{L^\infty(\mu)}\le c$ and $\|K\|_{L^1(\mu)}\le c$.

Let $\lambda\in\mathbb{C}$ with $|\lambda|>c$. Inequality (2.64) implies that $(\lambda Id-\hat{K})^{-1}$ and $(\lambda Id-K)^{-1}$ exists in $L(L^\infty(\mu))$ and $L(L^1(\mu))$ respectively. Since $\hat{K}$ can be transformed to K by a similarity transformation, $\hat{K}$ and K have the same spectrum, i.e. $(\lambda Id-K)^{-1}$ exists in $L(L^\infty(\mu))$.

Applying Theorem 2.11 to $(\lambda Id - K)^{-1}$, shows that $(\lambda Id - K)^{-1}$ exists in $L(L^2(\mu))$. Since $|\lambda| > c$ is arbitrary, it is shown that

$$\sup_{\lambda \in \sigma(K_{L^2(\mu) \leftarrow L^2(\mu)})} |\lambda| \leq c. \tag{2.66}$$

Since K_{sym} can be transformed to K by a similarity transformation, the same statement holds for the spectrum of K_{sym}. Furthermore, K_{sym} is self-adjoint. Hence, Equation (2.65) can be applied to get

$$\|K_{\text{sym}}\|_{L^2(\mu)} \leq c. \tag{2.67}$$

Remark 2.11. The estimate above is in general not sharp. For the $1D$ integral equation in case of homogeneous Dirichlet boundaries (see Equation (2.13)). The following estimate for $\|K\|_{L^2}$ is derived in [38]

$$\|K\|_{L^2} \leq 1 - \frac{3\pi^2 + 4}{12} \frac{1}{\tau^2}, \quad \text{if } \tau \gg 1, \tag{2.68}$$

where τ denotes the optical thickness of the slab. For large τ, this is much smaller than the L^∞-norm of the operator, which is given by

$$\|K\|_{L^\infty} = 1 - E_2(\frac{\tau}{2}).$$

As mentioned above, the estimate for $\|\hat{K_{\text{sym}}}\|_{L^2(\mu)}$ yields the coercivity of the bilinear form $a_{\text{sym}}(\cdot, \cdot)$. Applying the Lax-Milgram lemma(2.9) proofs the following theorem.

Theorem 2.13. *Under the same assumptions as in Theorem 2.4 the bilinear form $a_{sym}(\cdot, \cdot)$ is symmetric and coercive in $L^2(\mu)$ with coercivity constant $\alpha = (1 - \|\hat{K_{sym}}\|_{L^2(\mu)}) > 0$. Therefore, the solution of the weak formulation is unique and the solution operator $(Id - K_{sym})^{-1}$ is bounded.*

Remark 2.12. The article [35] deals with an integral equation which is very similar to Equation (2.26). Existence and uniqueness is shown under the assumption that $\kappa(y) \geq \kappa_0 > 0 \, \forall y \in D$ and $\rho(y) \leq \rho_0 < 1 \, \forall y \in \partial D$. This assumption is much stronger than the prerequisite in Theorem 2.3.

2.3 Regularity of the Solution

Consider the case of homogeneous Dirichlet boundaries and constant coefficients, then the integral equation (2.26) simplifies to (2.7)

$$G - \omega K G = f, \tag{2.69}$$

where K is a weakly singular integral operator. The following one dimensional example should illustrate the behavior of the solution of such an equation.

Example 2.1. Since the considered integral equation is of the second kind the regularity of the solution can be at most as high as the regularity of the function which results when the integral operator is applied to a C^∞-function. Consider the following weakly singular integral operator in one dimension

$$(Ku)(x) = \int_0^1 |x - y|^{\alpha - 1} u(y) dy, \quad \alpha \in (0, 1). \tag{2.70}$$

Let the function $u \in C^1((0,1))$. Splitting the integration domain into the intervals $(0,x)$ and $(x,1)$ to get rid of the modulus, and integration by parts yields

$$(Ku)(x) = \frac{1}{\alpha}\Big(x^\alpha u(0) + (1-x)^\alpha u(1) + \int_0^x (x-y)^\alpha [\frac{d}{dy}u(y)]dy + \int_x^1 (y-x)^\alpha [\frac{d}{dy}u(y)]dy\Big).$$

Hence, the derivative is computed as

$$[\frac{d}{dx}(Ku)(x) = x^{\alpha-1}u(0) + (1-x)^{\alpha-1}u(1) + \int_0^1 |x-y|^{\alpha-1}[\frac{d}{dy}u(y)]dy.$$

Since the kernel in the last integral is weakly singular, the integral as a function of x is bounded. On the other hand, the first two terms have a singularity for $x = 0$ and $x = 1$ respectively.

The example above shows that the derivatives of a solution of a weakly singular integral equation might be singular near the boundary. This is indeed true as the investigations in the article [44] and in the monograph [56] show. The regularity of the solution in the interior of the domain is determined by the regularity of the right hand side and the singular behavior of the derivatives at the boundary is determined by the order of the singularity of the kernel. The main result is the following:

Lemma 2.1. *Let the function space $C_1^m(D)$ be defined as the functions which are m times differentiable in D and whose derivatives satisfy the estimate*

$$|D^\alpha u(x)| \le c \begin{cases} 1 & \text{if } \alpha = 0, \\ |\log \rho(x)| & \text{if } |\alpha| = 1, \\ \rho(x)^{1-|\alpha|} & \text{if } |\alpha| > 1, \end{cases}$$

where $\rho(x) = \inf_{y\in\partial D} \|x-y\|$ denotes the distance to the boundary. Let $f \in C_1^m(D)$. Then the solution of the integral equation (2.69) is in $C_1^m(D)$.

Proof. refer to [56], Theorem 3.1 □

A more thorough investigation shows even more, namely, that for a smooth surface only the normal derivatives are singular. The tangential derivatives are bounded.

Concerning the boundary integral operators in (2.26), the following assertions hold. If the boundary is smooth, the solution is also smooth. If the solution is only piecewise smooth, the behavior of the solution when approaching the non-smooth points of ∂D is similar to that of the solution of a volume integral equation when approaching the boundary. This can be motivated as follows: the boundary integral may be interpreted as a sum over integrals over parameter domains, where each parameter domain corresponds to a smooth part of the surface. The non-smooth boundary points correspond to the boundaries of the parameter domains.

The discretization of (2.26) introduced in Chapter 3 uses a triangulation of the domain. Due to the above considerations about the regularity of the solution a higher resolution is needed near the boundaries. Hence, the mesh should be graded towards the boundary. Such a mesh implies a mesh on the boundary which is anisotropically refined towards edges and vertices. This is exactly what is needed for a good resolution of the singularities on the surface.

2.4 The Case of Linear Anisotropic Scattering

In this section we assume that the coefficients κ and σ are constant in space and that the domain D is convex. These assumptions are essential for the considerations in this section since we have to compute a primitive function of an integral kernel, which is only possible for constant coefficients.

As pointed out in Section 1.1, in many practical applications it is sufficient to handle the case of linear anisotropic scattering. Therefore, it would be advantageous if an integral formulation for this case could be derived. The linear anisotropic part of the scattering phase function gives rise to an additional term in the integral equation for the energy G, which does not appear for isotropic scattering:

$$G(x) = \int_D \ldots + \frac{A_1\sigma}{4\pi} \int_D \underbrace{\frac{e^{-\gamma\|x-y\|}}{\|x-y\|^3}(x-y)}_{=k_{\mathrm{vec}}(x,y)(\mathrm{see}(2.22))} \cdot q(y)dy. \tag{2.71}$$

In the above equation the flux q occurs as an additional unknown. Hence, an additional equation for q is needed to close the system. This can be done by multiplying (2.2) by Ω and integrating over all directions. By this, the system becomes much more complex and a numerical treatment is much more costly in comparison with the isotropic case. Therefore, we look for an approach which does not introduce additional equations. To this end we exploit the fact that a relationship between the energy and the divergence of the flux is known. Please remember the momentum equation in Section 1.1

$$\nabla \cdot q(x) + \kappa G(x) = 4\pi\kappa B(T(x)). \tag{2.72}$$

Since q and not $\nabla \cdot q$ appears in Equation (2.71), integration by parts has to be used. The vector valued kernel $k_{\mathrm{vec}}(x,y)$ in Equation (2.71) already appeared in Remark 2.3. Hence, a primitive function is given by

$$k_{\mathrm{pot}}(x,y) = \frac{e^{-\gamma\|x-y\|}}{\|x-y\|} - \gamma E_1(\gamma\|x-y\|). \tag{2.73}$$

Applying the Theorem of Gauss yields:

$$\begin{aligned} G(x) = \ldots + \frac{A_1\sigma}{4\pi} \lim_{\varepsilon\to 0} \Big[&- \int_{D\setminus B_\varepsilon(x)} k_{\mathrm{pot}}(x,y)\nabla \cdot q(y)dy \\ &+ \int_{\partial B_\varepsilon(x)} k_{\mathrm{pot}}(x,y)q(y)\cdot n(y)do(y) + \int_{\partial D} k_{\mathrm{pot}}(x,y)q(y)\cdot n(y))do(y)\Big]. \end{aligned} \tag{2.74}$$

It holds

$$\lim_{\varepsilon\to 0} \int_{\partial B_\varepsilon(x)} k_{\mathrm{pot}}(x,y)q(y)\cdot n(y)do(y) = 0, \tag{2.75}$$

since the highest singularity is of order one.

The flux at the boundary is split into the incoming and the outgoing part:

$$q\cdot n = q_{\mathrm{out}} - q_{in} = (1-\rho)q_{\mathrm{out}} - \pi(1-\rho)B(T_{\mathrm{out}}). \tag{2.76}$$

Using the momentum equation to replace $\nabla \cdot q$ in (2.74) yields

$$\begin{aligned} G(x) = \int_D \Big[\frac{e^{-\gamma\|x-y\|}}{\|x-y\|^2} - A_1\sigma k_{\text{pot}}(x,y)\Big]\kappa B(T(y)) + \\ \Big[\frac{e^{-\gamma\|x-y\|}}{\|x-y\|^2} + A_1\kappa k_{\text{pot}}(x,y)\Big]\frac{\sigma}{4\pi}G(y)dy \\ + \int_{\partial D}\Big[\frac{n(y)\cdot(y-x)}{\|x-y\|^3}e^{-\gamma\|x-y\|}\rho(y) + k_{\text{pot}}(x,y)\frac{A_1\sigma}{4}(1-\rho(y))\Big]\frac{1}{\pi}q_{\text{out}}(y) \\ + \Big[\frac{n(y)\cdot(y-x)}{\|x-y\|^3}e^{-\gamma\|x-y\|} + k_{\text{pot}}(x,y)\frac{A_1\sigma}{4}\Big](1-\rho(y))B(T_{\text{out}}(y))do(y). \quad (2.77) \end{aligned}$$

Remark 2.13. Equation (2.77) is very similar to (2.24). The additional terms of the integral kernel are weakly singular. Hence, the mapping properties described in Theorem 2.8 remain valid. When solving this equation numerically the only additional costs in comparison with the isotropic case are due to the fact that an evaluation of the kernel involves more operations. Therefore, the assembly of the system matrix is a little more costly, but the number of unknowns stays the same.

Up to now, we have only considered the integral equation for the energy G but in the equation for q_{out}, there is also an additional term due to the anisotropic scattering:

$$q_{out}(x) = \ldots + \frac{A_1\sigma}{4\pi}\int_D \underbrace{\frac{n(x)\cdot(x-y)}{\|x-y\|^4}e^{-\gamma\|x-y\|}(x-y)}_{=:k_{\text{vec}}(x,y)}\cdot q(y)dy. \quad (2.78)$$

Due to the dependence on $n(x)$, the integrability criterion cannot be fulfilled for arbitrary $y \neq x$. Hence, a primitive function w.r.t. y does not exist and the trick with the momentum equation cannot be applied. Therefore, we propose to approximate the anisotropic scattering phase function by isotropic scattering plus a peak in forward direction (see also Section 1.1)

$$\frac{1}{4\pi}\int_{S^2}(1 + A_1\Omega\cdot\Omega')I(x,\Omega')d\Omega' \simeq fI(x,\Omega) + (1-f)\frac{1}{4\pi}\int_{S^2} I(x,\Omega')d\Omega'. \quad (2.79)$$

This time f is not determined heuristically but by the requirement that certain moment conditions are fulfilled (see [40], page 416). This conditions can be interpreted by looking at the $1D$ situation

$$\frac{1}{2}\int_{-1}^{1}(1 + A_1\mu\mu')I(x,\mu')d\mu' \simeq fI(x,\mu) + (1-f)\frac{1}{2}\int_{-1}^{1} I(x,\mu')d\mu'.$$

Due to construction, equality holds in (2.4) for arbitrary f if $I(x,\mu)$ is independent of μ. Requiring that equality holds also for the case that $I(x,\mu)$ depends linearly on the angular variable, i.e. $I(x,\mu) = I_0(x) + \mu I_1(x)$, leads to $f = \frac{A_1}{3}$. Thus, the isotropic approximation is obtained by replacing the scattering cross section σ by $(1 - \frac{A_1}{3})\sigma$ (see (2.79)).

Chapter 3

Numerical Solution

There are different methods for the numerical solution of integral equations. The most popular ones are the collocation method and the Galerkin method. In Section 3.1 these methods are briefly introduced and their advantages and disadvantages are discussed.

Both methods are only semi-discrete methods because the assembly of the system matrices involves the evaluation of integrals that cannot be computed analytically. Instead, cubature formulae are used to get a fully discrete method. Because of the weak singularity of the kernel, some of the integrals have singular or near-singular behavior. In this case, standard cubature formulae cannot be applied directly. Very similar integrals occur in context of the boundary element method (BEM). In [49] regularizing coordinate transformations have been introduced, which transform the singular integrals such that the integrands are analytic and standard cubature formulae can be used. The case of surface elements is the only case that occurs in context of the BEM. The transformations are extended to the case of volume elements in Section 3.2

The discretization leads to a linear system of algebraic equations with a full matrix. The matrix compression methods discussed in Chapter 4 make iterative methods an attractive approach for the solution of such systems. The number of iterations needed to achieve a certain accuracy strongly depends on the spectrum of the system matrix. Since the equation is of second kind, it is reasonable to expect that the spectrum is well suited for iterative methods. This proves to be true for a wide range of parameters describing the physical problem, but for a certain regime, the diffusion limit of Section 1.2, the matrix is ill-conditioned. In Section 3.3 a preconditioner is introduced which leads to very fast convergence in this case. A comparison between the solution of the integral equation by the Galerkin method and the solution of the RTE by the discrete ordinate method is given in Section 3.4.

3.1 The Galerkin Method

In case of the collocation method as well as in case of the Galerkin method an approximate solution of the integral equation is sought in a finite dimensional ansatz space. Usually, ansatz functions with local support are used, e.g. piecewise constant or piecewise linear functions on a triangulation of the domain. The advantage of these ansatz spaces is that they are very flexible. Usually, the solution requires a higher resolution in some parts of the domain, e.g. near the boundary because of the possible singularity of the normal derivatives (see Chapter 2). This could be achieved by a local refinement of the grid.

A triangulation $\mathcal{G}^{(v)} = \{\tau_1, \ldots, \tau_{m^{(v)}}\}$ of the volume D consists of disjoint, open subsets of D. It should have the following properties:

- $\mathcal{G}^{(v)}$ is a partition of D, i.e. $\bar{D} = \bigcup_{i=1}^{m^{(v)}} \bar{\tau}_i$
- For every $\tau \in \mathcal{G}^{(v)}$ there is an affine mapping χ_τ of the volume reference element $T^{(3)} = \{(x_1, x_2, x_3) : 0 < x_1 < 1, 0 < x_2 < x_1, 0 < x_3 < x_2\}$ onto τ, i.e. the volume elements are tetrahedrons.
- The triangulation is admissible, i.e. $\forall\, \tau_1,\ \tau_2 \in \mathcal{G}$ with $\tau_1 \neq \tau_2$ holds:

$$\bar{\tau}_1 \cap \bar{\tau}_2 = \begin{cases} \emptyset & \text{or} \\ \text{common vertex} & \text{or} \\ \text{common edge} & \text{or} \\ \text{common face.} & \end{cases}$$

The grid size h is defined as

$$h := \max\{\operatorname{diam}(\tau) \,|\, \tau \in \mathcal{G}^{(v)}\}.$$

The uniformity of the triangulation $\mathcal{G}^{(v)}$ is characterized by the smallest constant C_u satisfying

$$h \leq C_u \operatorname{diam}(\tau) \quad \forall \tau \in \mathcal{G}^{(v)}, \tag{3.1}$$

and the shape regularity of the elements is characterized by the smallest constant C_q satisfying

$$\operatorname{diam}(\tau)^3 \leq C_q |\tau| \quad \forall \tau \in \mathcal{G}^{(v)}. \tag{3.2}$$

Since $\mathcal{G}^{(v)}$ contains only finitely many elements, the constants C_u and C_q are always bounded. For asymptotic results a family of triangulations with an increasing number $m^{(v)}$ of elements is used. We require that the values of C_u and C_q do not depend on $m^{(v)}$.

The set of vertices of $\mathcal{G}^{(v)}$ is given by

$$\Xi^{(v)} = \{\xi \in D \,|\, \xi \text{ is vertex of an element } \tau \in \mathcal{G}^{(v)}\} = \{\xi_1, \ldots, \xi_{n^{(v)}}\}. \tag{3.3}$$

The triangulation $\mathcal{G}^{(v)}$ of the domain D implies in a canonical way a triangulation $\mathcal{G}^{(s)} = \{\tau_{m^{(v)}+1}, \ldots, \tau_m\}$, $(m = m^{(v)} + m^{(s)})$ of the boundary ∂D. This triangulation consists of flat triangles, i.e. for every $\tau \in \mathcal{G}^{(s)}$ there is an affine mapping χ_τ of the surface reference element $T^{(2)} = \{(x_1, x_2) : 0 < x_1 < 1, 0 < x_2 < x_1\}$ onto τ. The triangulation of the complete computational domain is given by $\mathcal{G} = \mathcal{G}^{(v)} \cup \mathcal{G}^{(s)}$.

Remark 3.1. The condition $\bar{D} = \bigcup_{i=1}^{m^{(v)}} \bar{\tau}_i$ implies that $\bar{D}$ is a polyhedron. In the general case, i.e. ∂D is a piecewise smooth surface, the union $\bigcup_{i=m^{(v)}+1}^{m^{(v)}+m^{(s)}} \bar{\tau}_i$ has to interpolate the nodal points on ∂D. This procedure guarantees that for a convex domain D the approximate domain $\tilde{D}$ is also convex. This is for example not possible if parallelepiped elements are used instead of tetrahedrons to approximate the domain. For simplicity, we assume in the following that D is a polyhedron.

The Ansatz Space

For the energy G we choose the finite element space

$$V_h^{(v)} := \{\psi \in C(D) \,|\, \psi_{|\tau} \in \mathcal{P}_k(\tau)\, \forall \tau \in \mathcal{G}^{(v)}\} \tag{3.4}$$

as ansatz space. (For $k = 0$ we do not assume continuity.). $\mathcal{P}_k$ denotes the set of polynomials up to degree k in $\mathbb{R}^3$

$$\mathcal{P}_k := \{\sum_{|\alpha| \le k} \lambda_\alpha \prod_{i=1}^{3} x_i^{\alpha_i},\ \lambda_\alpha \in \mathbb{R}\}, \quad \alpha \text{ a multi-index.} \tag{3.5}$$

The ansatz space for the fluxes q_{out} is constructed by lifting a finite element space on a parameter plane onto the surface ∂D via the mapping χ_τ:

$$V_h^{(s)} := \{\psi \in C(\partial D) \,|\, \psi \circ \chi_\tau \in \mathcal{P}_k(T^{(2)})\, \forall \tau \in \mathcal{G}^{(s)}\}. \tag{3.6}$$

(For $k = 0$ we again do not assume continuity.)

These two spaces together yield the ansatz space for the whole system of equations

$$V_h = V_h^{(v)} \oplus V_h^{(s)}. \tag{3.7}$$

For piecewise linear elements, i.e. $k = 1$ the Lagrange or nodal basis $\{b_1, \ldots, b_{n^{(v)}}\}$ of $V_h^{(v)}$ is defined via the condition $b_k(\xi_l) = \delta_{kl}$. Analogously for $V_h^{(s)}$.

The Collocation Method

The collocation and the Galerkin method differ by the fact which equation the approximate solution has to fulfill. The collocation method requires that the integral equation is fulfilled in certain points, the so-called collocation points. This only makes sense for functions which allow the evaluation of point values, e.g. for continuous functions but not for L^∞-functions. The position and number of the collocation points has to be chosen in dependence on the used ansatz space to get a well defined discrete problem. For continuous piecewise linear ansatz functions these points are the vertices of the elements. For a non smooth boundary some of the vertices of the boundary elements lie on edges or vertices of the boundary. As mentioned in Section 2.4 the integral equation is not defined in non smooth boundary points because the outer normal is not defined. Furthermore, the volume integral equation is only valid inside the open domain D (see Remark 2.3). This also makes problems when piececewise linear elements are used. Hence, piecewise constant elements or discontinuous piecewise linear elements have to be used (see [3]).

Another drawback of the collocation method is the following. The discrete system is not symmetric even though the integral operator K_{sym} is self-adjoint. Therefore, the amount of storage needed is nearly twice as high as that of the Galerkin method.

Concerning the stability of the collocation method, the following can be said. Since the compactness of the boundary integral operator is lost for non-smooth surfaces, Fredholm alternative cannot be applied to prove stability. This is the reason why the collocation method lacks a proof of stability when applied to the integral equations stemming from the BEM. On the other hand, in Section 2.2 it is shown that the L^∞-norm of the integral operator $\hat{K}$ is less than one. Hence, a Neumann series argument can be used to show stability of the collocation method when applied to the integral equation as derived from the RTE (see Equation (2.39)).

The Galerkin Method

The Galerkin method is based on the variational formulation (2.50) of the problem, where the same finite dimensional subspace $V_h \subset V := L^2(D,\mu_1) \oplus_2 L^2(\partial D,\mu_2)$ is used as ansatz and test space. Due to the considerations in Section 2.2, it is advantageous to use the symmetric bilinear form $a_{\mathrm{sym}}(\cdot,\cdot)$ instead of $a(\cdot,\cdot)$. Since the non-symmetric form is not used any further, the symmetric form is denoted by $a(\cdot,\cdot)$ in this chapter. This leads to the following Galerkin problem:

For given $f \in V$, find $u_h \in V_h$ such that

$$a(u_h, v_h) = \langle f, v_h \rangle_V \qquad \forall v_h \in V_h. \tag{3.8}$$

By choosing a basis $\{b_1, \ldots, b_n\}$ of the subspace V_h, the Galerkin Equation (3.8) is transformed into a linear system of algebraic equations. Let $\{u_l\}$ be the set of coefficients of the Galerkin solution u_h with respect to the chosen basis, i.e.

$$u_h = \sum_{l=1}^{n} u_l b_l.$$

Then these coefficients have to fulfill the following system of equations

$$a(u_h, b_k) = \sum_{l=1}^{n} u_l a(b_l, b_k) = \langle f, b_k \rangle, \qquad k = 1, \ldots, n,$$

or in matrix notation

$$\mathbf{A}\mathbf{u} = \mathbf{f},$$

with $\mathbf{A} = a(b_l, b_k))_{k,l=1}^{n}$, $\mathbf{f} = (\langle f, b_1 \rangle, \ldots, \langle f, b_n \rangle)^T$ and $\mathbf{u} = (u_1, \ldots, u_n)^T$.

In Section 2.2 it is shown that, if the prerequisite of Theorem 2.3 is fulfilled, $a(\cdot,\cdot)$ is coercive in V and, hence, also in $V_h \subset V$

$$a(u,u) \geq \alpha \|u\|^2_{L^2(\mu)} \quad \forall u \in V_h. \tag{3.9}$$

Applying the Lax-Milgram lemma, yields existence and uniqueness of the solution of the Galerkin Equation (3.8) and that the inverse operator A_h^{-1} is bounded by the constant $\frac{1}{\alpha}$ independent of h. Furthermore, the coercivity condition guarantees that the Galerkin solution is quasi-optimal, i.e. the error is bounded by a constant times the error of the best approximation in the space V_h.

Theorem 3.1 (Céa Lemma). *Let $a(\cdot,\cdot)$ be a bilinear form, which is continuous with continuity constant C_s and coercive with constant $\alpha > 0$. Let $u \in V$ be the solution of (2.50) and $u_h \in V_h$ be the Galerkin solution. Then the following estimate holds:*

$$\|u - u_h\| \leq \frac{C_s}{\alpha} \inf_{w_h \in V_h} \|u - w_h\|.$$

If $a(\cdot,\cdot)$ is symmetric the constant $\frac{C_s}{\alpha}$ in the estimate above can be replaced by $\sqrt{\frac{C_s}{\alpha}}$.

Proof. refer to [23] □

This lemma together with the following approximation property (3.11) of the space V_h gives the convergence of the Galerkin method.

Lemma 3.1. *Let the Grid $\mathcal{G}$ be quasi-uniform and shape regular. Then there is a constant C_{approx} such that $\forall u \in H^s(D,\mu_1) \oplus H^s(\partial D,\mu_2)$, $s < k+1$, exists a function $u_h \in V_h$ (V_h as defined in (3.7)) such that*

$$\|u - u_h\|_{L^2} \leq C_{approx} h^s \|u\|_{H^s}, \tag{3.10}$$

especially

$$\lim_{h\to 0} \inf_{u_h \in V_h} \|u - u_h\|_V = 0 \qquad \forall u \in V. \tag{3.11}$$

Combining the two lemmas gives the following:

Theorem 3.2. *Let the Grid $\mathcal{G}$ be quasi-uniform and shape regular. Then the Galerkin problem (3.8) has a unique solution u_h and the quasi optimal error estimate*

$$\|u - u_h\|_{L^2} \leq \frac{C_s}{\alpha} C_{approx} h^s \|u\|_{H^s}, \tag{3.12}$$

holds if the solution u of the continuous problem is in H^s, $s \leq k+1$.

In Section 2.3 it is mentioned that the solution u of (2.50) is not in $H^2(D) \oplus H^2(\partial D)$. For the interior part the derivatives in normal direction have a logarithmical singularity when approaching the boundary. The boundary part is non smooth near edges and vertices. To equi-distribute the approximation error over the whole grid, it is advantageous to use a grid in the interior which is anisotropically refined near the boundary and a grid on the boundary which is anisotropically refined towards edges and vertices. This is somehow difficult to obtain by an adaptive grid refinement based on an error estimator. On the other hand, since it is known a priori where the refinement has to take place, it could be done in a preprocessing step.

As has been seen in Section 2.2 the bilinear form $a_{\mathrm{sym}}(\cdot,\cdot)$ is symmetric. (Remember $a_{\mathrm{sym}}(\cdot,\cdot)$ is denoted by $a(\cdot,\cdot)$ in this chapter.) Therefore, the Galerkin matrix is also symmetric

$$a_{kl} = a(b_l, b_k) = a(b_k, b_l) = a_{lk}.$$

The major drawback of the Galerkin method is the fact that the computation of the matrix entries is very costly in comparison to the collocation method. This is due to the fact that e.g. for the volume integrals one has to compute $6D$ integrals instead of $3D$ integrals in case of the collocation method. If one of the matrix compression methods of Chapter 4 is used, this drawback is relaxed since these methods avoid the assembly of the whole matrix.

3.2 Numerical Integration

To compute the entries of the system matrix $\mathbf{A}$ the bilinear form $a(\cdot,\cdot)$ has to be evaluated for the ansatz functions. An analytic integration of the involved integrals is in general not possible. Hence, numerical cubature techniques have to be used to get a fully discrete method. The corresponding bilinear form is denoted by $a_h(\cdot,\cdot)$.

An approximation of the double integrals in x and y direction is calculated as a sum of approximations on pairs of elements. Depending on which part of the matrix is considered these elements are either volume or surface elements. To treat all cases at once, we use the following notation

$$I_{\tau,t}(u,v) = \int_{\tau\times t} v(x)k(x,y)u(y)dydx, \tag{3.13}$$

where the kernel k is one of the kernels k_{ij} described in Chapter 2 and τ and t is either a volume or a surface element. With this notation the bilinear form $a(\cdot,\cdot)$ can be written as

$$a(u,v) = \sum_{(\tau,t)\in\mathcal{G}\times\mathcal{G}} I_{\tau,t}(u,v), \tag{3.14}$$

and the approximation $a_h(\cdot,\cdot)$ is given by

$$a_h(u,v) = \sum_{(\tau,t)\in\mathcal{G}\times\mathcal{G}} Q_{\tau,t}(u,v), \tag{3.15}$$

where $Q_{\tau,t}$ denotes some cubature formula used to approximate the integral $I_{\tau,t}$.

As usual the elements are first mapped onto the master elements $T^{(3)}$ or $T^{(2)}$ via the affine mappings χ and then the cubature techniques are applied to this reference situation. The integral (3.13) in local coordinates is given as

$$I_{\tau,t}(u,v) = \int_{T_\tau\times T_t} \hat{v}(\hat{x})\hat{k}(\hat{x},\hat{y})\hat{u}(\hat{y})d\hat{y}d\hat{x} \tag{3.16}$$

where

$$T_\tau = \begin{cases} T^{(3)} & \text{if } \tau \text{ is a volume element,} \\ T^{(2)} & \text{if } \tau \text{ is a surface element,} \end{cases}$$

and

$$\hat{k}(\hat{x},\hat{y}) := k(\chi_\tau\hat{x},\chi_t\hat{y})\, g_\tau g_t, \qquad \hat{v} = v\circ\chi_\tau, \qquad \hat{u} = u\circ\chi_t.$$

g_τ denotes the determinant of the Jacobian D_{χ_τ} of χ_τ if τ is a volume element and $\sqrt{\det(D_{\chi_\tau}^T D_{\chi_\tau})}$ if τ is a surface element. Since χ_τ is an affine mapping, $g_\tau = p|\tau|$ with $p = 2$ or 6 if τ is a volume or surface element respectively. By construction of the ansatz spaces, the functions $\hat{u}$ and $\hat{v}$ are polynomials of degree zero in case of piecewise constant elements and of degree one in case of piecewise linear elements.

The major difficulty for the integration is the weak singularity of the kernels: if $\bar{\tau}\cap\bar{t}\neq\emptyset$, the integrand is singular. In this situation, a transformation is applied to remove the singularity, before standard cubature methods are used. These transformations depend on the kind of the two involved elements (volume or surface element) and their relative position. Hence, the following 13 cases have to be distinguished:

1. τ and t are both volume elements:
   (a) 'Identical Elements': $\tau = t$.
   (b) 'Common Face': τ and t share exactly one face.
   (c) 'Common Edge': τ and t share exactly one edge.
   (d) 'Common Vertex': τ and t share exactly one vertex.
   (e) 'Regular Case': $\bar{\tau}\cap\bar{t} = \emptyset$.
2. τ is a volume element and t is a surface element:
   (a) 'Common Face': τ and t share exactly one face.
   (b) 'Common Edge': τ and t share exactly one edge.
   (c) 'Common Vertex': τ and t share exactly one vertex.

(d) 'Regular Case': $\bar{\tau} \cap \bar{t} = \emptyset$.

3. τ and t are both surface elements:

   (a) 'Identical Elements': $\tau = t$.

   (b) 'Common Edge': τ and t share exactly one edge.

   (c) 'Common Vertex': τ and t share exactly one vertex.

   (d) 'Regular Case': $\bar{\tau} \cap \bar{t} = \emptyset$.

Assumption 3.1. In the singular case $\bar{\tau} \cap \bar{t} \neq \emptyset$, we assume that the transformations χ_τ, χ_t on the reference elements are chosen such that

$$x = y \quad \Rightarrow \quad \chi_\tau^{-1}(x) = \chi_t^{-1}(y).$$

Furthermore, if $\bar{\tau}$ and $\bar{t}$ share exactly one vertex, it should be mapped to the origin. If $\bar{\tau}$ and $\bar{t}$ share exactly one edge, this edge should be mapped onto the x_1 axis. If $\bar{\tau}$ and $\bar{t}$ share exactly one face, this face should be mapped into the (x_1, x_2) plane. (Refer to Figure 3.1 for the case of common face.)

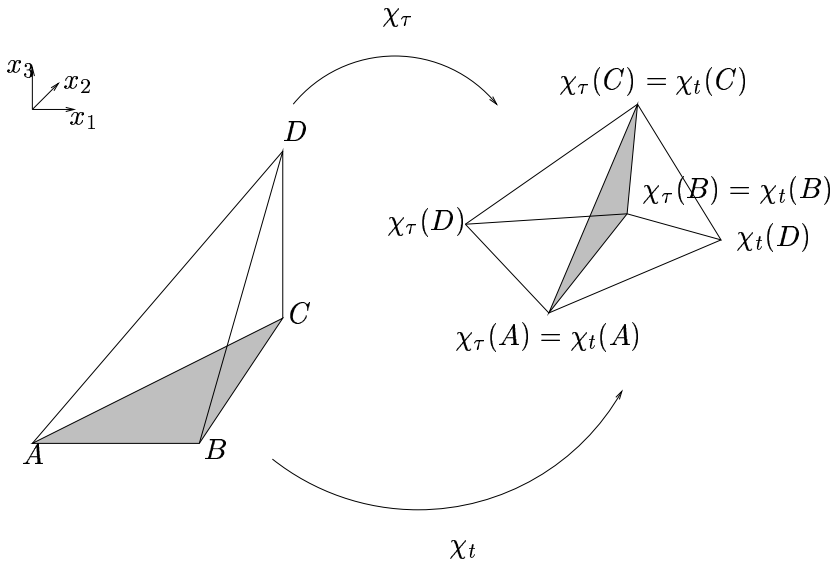

Figure 3.1: Mapping onto reference element in case of common face

We use the following notation:
Let $S^{(d)}$ denote the d-dimensional unit-simplex

$$S^{(d)} = \left\{ x \in \mathbb{R}^d_+ : \sum_{i=1}^{d} x_i < 1 \right\}, \tag{3.17}$$

$T^{(d)}$ denotes also an open d-dimensional simplex

$$T^{(d)} = \{0 < x_1 < 1, 0 < x_2 < x_1, \ldots, 0 < x_d < x_{d-1}\}, \tag{3.18}$$

and $C^{(d)}$ denotes the d-dimensional unit-cube

$$C^{(d)} = (0,1)^d. \tag{3.19}$$

The example below shows the simplest example of a transformation which makes a weakly singular integrand regular.

Example 3.1. Consider the integral of a function $f(z) = \frac{1}{\|z\|} g(z)$ over $T^{(2)}$, where g is analytic in the region of integration. Then the simple transformation $F : C^{(2)} \to T^{(2)}, \quad F(\omega, \eta) = (\omega, \omega\eta)^T$ renders the integrand analytic:

$$If = \int_0^1 \int_0^{z_1} f(z_1, z_2) dz_2 dz_1 = \int_0^1 \int_0^1 \omega f(\omega, \omega\eta) d\omega d\eta = \int_0^1 \int_0^1 \frac{1}{\sqrt{1+\eta^2}} g(\omega, \omega\eta) d\omega d\eta$$

The above transformation is called Duffy transformation (see [17]).

A similar procedure also works for weak singularities in higher dimensions. It is not restricted to simplices, but can be applied to more general d-dimensional polyhedrons. Let $\mathrm{col_d}$ denote a mapping $\mathrm{col_d} : \{2, \ldots, d\} \to \{1, \ldots, d-1\}$ with $\mathrm{col_d}(i) \leq i - 1$ (especially: $\mathrm{col_d}(2) = 1$, i.e. there are $d-2$ degrees of freedom). The corresponding d-dimensional polyhedron $P_{\mathrm{col_d}}$ is defined by

$$P_{\mathrm{col_d}} := \{0 < z_1 < 1,\ 0 < z_i < z_{\mathrm{col_d}(i)},\ i = 2, \ldots, d\}. \tag{3.20}$$

$P_{\mathrm{col_d}}$ is a d-dimensional convex polyhedron with the origin as a distinguished vertex, i.e. the remaining vertices of $P_{\mathrm{col_d}}$ lie in a $(d-1)$-dimensional affine subspace. One can say even more about these vertices: they are of the form $(1, \tilde{V})^T$ where $\tilde{V}$ is a vertex of the $(d-1)$-dimensional unit cube.

To illustrate the abstract definition of $P_{\mathrm{col_d}}$ the case $d = 3$ is considered. As mentioned above there is only one degree of freedom in this case: $\mathrm{col_3}(3) \in \{1, 2\}$. The corresponding polyhedrons are the tetrahedron $P(0 < z_3 < z_2) := \{0 < z_1 < 1, 0 < z_2 < z_1, 0 < z_3 < z_2\}$ and the pyramid $P(0 < z_3 < z_1) := \{0 < z_1 < 1, 0 < z_2 < z_1, 0 < z_3 < z_1\}$.

We are looking for a transformation which removes a weak singularity at the origin. To this end, we define the non-linear mapping

$$F_{\mathrm{col_d}} : \begin{cases} \mathbb{R}^d & \to \quad \mathbb{R}^d, \\ (\omega_1, \eta_1, \ldots, \eta_{d-1}) & \to \quad (z_1, \ldots, z_d), \end{cases}$$

by

$$\begin{aligned} z_1 &:= \omega_1, \\ z_i &:= \omega_1 \prod_{j=1}^{i-2} \eta_j^{e_{ij}} \eta_{i-1}, \quad i = 2, \ldots, d, \end{aligned}$$

where the numbers $e_{i,j}$, $i = 3, \ldots, d$, $j = 1, \ldots, i-2$ are recursively defined via

$$\begin{aligned} e_{i,i-2} &:= c_{i,i-1}, \\ e_{i,j} &:= c_{i,j+1} + \sum_{k=j+1}^{i-2} c_{k+1,j+1} e_{i,k}, \quad j = i-3, \ldots, 1. \end{aligned} \tag{3.21}$$

The numbers $c_{i,j}$, $i = 2, \ldots, d$, $j = 1, \ldots, i-1$ are given by

$$c_{i,j} := \begin{cases} 1 & \text{if } j = \mathrm{col_d}(i), \\ 0 & \text{else.} \end{cases} \tag{3.22}$$

$F_{\mathrm{col_d}}$ fulfills the prerequisite of the substitution rule as a mapping from $S^{(1)} \times C^{(d-1)}$ to $P_{\mathrm{col_d}}$, i.e. it is bijective and the determinant of its Jacobian is always positive. The rather technical proof can be found in Appendix A.4.

$F_{\mathrm{col_d}}$ has the additional properties

$$\|F_{\mathrm{col_d}}(\omega_1, \eta_1, \ldots, \eta_{d-1})\|_2 \geq |\omega_1|, \tag{3.23}$$

$$F_{\mathrm{col_d}}(\omega_1, \eta_1, \ldots, \eta_{d-1}) = \omega_1 F_{\mathrm{col_d}}(1, \eta_1, \ldots, \eta_{d-1}), \tag{3.24}$$

$$\Big(\Longrightarrow \det(F'_{\mathrm{col_d}}) = \omega_1^{d-1} f(\eta_1, \ldots, \eta_{d-1}) \Big). \tag{3.25}$$

Let $L : \mathbb{R}^d \to \mathbb{R}^3$ denote a linear mapping with $Lz \neq 0 \forall z \in \bar{P_{\mathrm{col_d}}} \backslash \{0\}$. Applying the transformation $F_{\mathrm{col_d}}$ to a singular integral of the form

$$\int\limits_{P_{\mathrm{col_d}}} \frac{1}{\|Lz\|^\alpha} dz, \quad \alpha > 0, \tag{3.26}$$

and exploiting the properties of $F_{\mathrm{col_d}}$ results in

$$\int\limits_{S^{(1)} \times C^{(d-1)}} \frac{1}{\|LF_{\mathrm{col_d}}(1, \eta_1, \ldots, \eta_{d-1})\|^\alpha} \omega_1^{d-1-\alpha} f(\eta_1, \ldots, \eta_{d-1}) d\omega_1 d\eta. \tag{3.27}$$

Since $LF_{\mathrm{col_d}}(1, \eta) \neq 0$, the integrand is regular if $d \geq \alpha + 1$.

The properties of $F_{\mathrm{col_d}}$ are conserved if $F_{\mathrm{col_d}}$ is concatenated with an injective linear mapping. Hence, a transformation to regularize the integrand exists for every d-dimensional polyhedron which could be mapped to a polyhedron of the form (3.20) by a linear transformation.

If some of the variables are not involved in the singularity, a Duffy transformation is only applied to the remaining coordinates. The number of these coordinates is denoted by d_s. Let $\mathrm{col_{d,d_s}}$ denote a mapping $\mathrm{col_{d,d_s}} : \{d - d_s + 2, \ldots, d\} \to \{d - d_s + 1, \ldots, d - 1\}$ with $\mathrm{col_{d,d_s}}(i) \leq i - 1$ (especially: $\mathrm{col_{d,d_s}}(d - d_s + 2) = d - d_s + 1$, i.e. there are $d_s - 2$ degrees of freedom). The corresponding d-dimensional polyhedron $P_{\mathrm{col_{d,d_s}}}$ is defined by

$$P_{\mathrm{col_{d,d_s}}} := \{0 < z_1 < 1, \ldots, 0 < z_{d-d_s+1} < z_{d-d_s}, \\ 0 < z_i < z_{\mathrm{col_{d,d_s}}(i)},\ i = d - d_s + 2, \ldots, d\}. \tag{3.28}$$

The definition of $P_{\mathrm{col_{d,d_s}}}$ is illustrated by some examples for $d = 6$.

Example 3.2.

- $d_s = 3$: one degree of freedom: $\mathrm{col}_{6,3}(6) \in \{4, 5\}$
  - $P(0 < z_6 < z_5) := T^{(6)}$, simplex,
  - $P(0 < z_6 < z_4) := \{0 < z_1 < 1, 0 < z_2 < z_1, 0 < z_3 < z_2, 0 < z_4 < z_3, 0 < z_5 < z_4, 0 < z_6 < z_4\}$, union of 2 simplices,
- $d_s = 4$: two degrees of freedom: $\mathrm{col}_{6,4}(5) \in \{3, 4\}$, $\mathrm{col}_{6,4}(6) \in \{3, 4, 5\}$
  - $P(0 < z_5 < z_4, 0 < z_6 < z_5) := T^{(6)}$, simplex,
  - $P(0 < z_5 < z_3, 0 < z_6 < z_5) := \{0 < z_1 < 1, 0 < z_2 < z_1, 0 < z_3 < z_2, 0 < z_4 < z_3, 0 < z_5 < z_3, 0 < z_6 < z_5\}$, union of 3 simplices.

Define the bijective non-linear mapping

$$F_{\mathrm{col_{d,d_s}}} : \left\{ \begin{array}{ccc} S^{(d-d_s+1)} \times C^{(d_s-1)} & \to & P_{\mathrm{col_{d,d_s}}}, \\ (\omega_1, \ldots, \omega_{d-d_s+1}, \eta_1, \ldots, \eta_{d_s-1}) & \to & (z_1, \ldots, z_d), \end{array} \right.$$

by

$$\begin{aligned} z_i &:= \sum_{j=1}^{d-d_s+2-i} \omega_j, \quad i \in \{1, \ldots, d-d_s+1\}, \\ z_i &:= \omega_1 \prod_{j=d-d_s+1}^{i-2} \eta_{j-(d-d_s)}^{e_{ij}} \eta_{i-(d-d_s)-1}, \quad i \in \{d-d_s+2, \ldots, d\}, \end{aligned}$$

where the numbers $e_{i,j}$, $i = d-d_s+3, \ldots, d$, $j = d-d_s+1, \ldots, i-2$ are defined as in (3.21). The proof of the bijectivity of $F_{\mathrm{col_{d,d_s}}}$ is analog to that of $F_{\mathrm{col_d}}$. This transformation renders the integrand analytic.

The transformations above cannot be applied to the integral (3.16) directly, since neither the integration domain nor the integrand satisfy the requirements. First, the singularity has to be fixed at the origin. This is done by introducing relative coordinates $z_i = x_i - y_i$ for all coordinate directions i for which the integrand is singular on the set $\{x_i = y_i\}$. Let I be the set of these indices i. Due to Assumption 3.1, it holds:

- $I = \{1, 2, 3\}$ in case of identical volume elements,
- $I = \{1, 2\}$ in case of a common face,
- $I = \{1\}$ in case of a common edge,
- $I = \emptyset$ in case of a common vertex.

From this point on, we can distinguish to kinds of variables: the singular and the regular variables. The integrand is singular if all the variables z_i, $i \in I$, and x_j, y_j, $j \notin I$ are equal to zero. These variables are called singular variables in the following. The remaining variables x_i, $i \in I$ have no influence on the singularity. They are called regular variables. For the applicability of a Duffy transformation, the geometry of the integration domain w.r.t. the singular variables is important: it has to be a polyhedron of the form $P_{\mathrm{col_d}}$. Therefore, it is advantageous to interchange the order of integration such that the regular coordinates are the innermost variables, because then the singular variables can be considered independently of

the regular variables. After interchanging the variables the domain has to be subdivided into parts to which a Duffy transformation can be applied. This could be done by first splitting the whole domain into simplices containing the origin as a vertex and then merging some of these simplices. For efficiency reasons it is advantageous to have as few parts as possible.

Remark 3.2. To keep the number of different cases which have to be distinguished as small as possible, it is advantageous to first of all split up the integration domain into two parts:

$$I_{\tau,t}(u,v) = \int\limits_{\hat{y}_1<\hat{x}_1} \ldots d\hat{y}d\hat{x} + \int\limits_{\hat{x}_1<\hat{y}_1} \ldots d\hat{x}d\hat{y} =: I^{(1)}_{\tau,t}(u,v) + I^{(2)}_{\tau,t}(u,v). \tag{3.29}$$

These two cases can then be treated in an analog way since one of them can be transformed into the other by interchanging the roles of the variables $\hat{x}$ and $\hat{y}$. Since all polyhedrons of the form (3.20) have the origin as a distinguished vertex, such a splitting has to be done anyway sooner or later.

After these rather theoretical considerations, the results for some of the cases mentioned above are presented.

(a) The Case of Identical Volume Elements

Let the affine mapping χ_τ be given by:

$$\chi_\tau(\hat{\mathbf{x}}) = \mathbf{x}_0 + \mathbf{D}_{\chi_\tau}\hat{\mathbf{x}}, \quad \mathbf{x}_0 \in \mathbb{R}^3,\ \mathbf{D}_{\chi_\tau} \in \mathbb{R}^{3\times 3} \text{ (Jacobi matrix)}.$$

Then the integrand in (3.16) has a singularity of the form

$$\|\chi_\tau(\hat{\mathbf{x}}) - \chi_t(\hat{\mathbf{y}})\|^{-2} = \|\mathbf{D}_{\chi_\tau}(\hat{\mathbf{x}} - \hat{\mathbf{y}})\|^{-2} \overset{\hat{\mathbf{z}}=\hat{\mathbf{x}}-\hat{\mathbf{y}}}{=} \|\mathbf{D}_{\chi_\tau}\hat{\mathbf{z}}\|^{-2}.$$

Since $\mathbf{D}_{\chi_\tau}$ is non singular, the integrand is of the required form (3.26). It remains to split up the integration domain into polyhedrons of the form $P_{\mathrm{col}_{6,3}}$. To this end, we first apply the splitting into the two parts $\hat{y}_1 < \hat{x}_1$ and $\hat{y}_1 > \hat{x}_1$. The first part can be described by the following system of inequalities (For the second part one just has to interchange the roles of $\hat{x}$ and $\hat{y}$.):

$$\left\{\begin{array}{r} 0 < \hat{x}_1 < 1 \\ 0 < \hat{x}_2 < \hat{x}_1 \\ 0 < \hat{x}_3 < \hat{x}_2 \\ 0 < \hat{y}_1 < \hat{x}_1 \\ 0 < \hat{y}_2 < \hat{y}_1 \\ 0 < \hat{y}_3 < \hat{y}_2 \end{array}\right\} \quad \overset{\hat{z}=\hat{x}-\hat{y}}{\longrightarrow} \quad \left\{\begin{array}{r} 0 < \hat{x}_1 < 1 \\ 0 < \hat{x}_2 < \hat{x}_1 \\ 0 < \hat{x}_3 < \hat{x}_2 \\ 0 < \hat{z}_1 < \hat{x}_1 \\ \hat{x}_2 - \hat{x}_1 + \hat{z}_1 < \hat{z}_2 < \hat{x}_2 \\ \hat{x}_3 - \hat{x}_2 + \hat{z}_2 < \hat{z}_3 < \hat{x}_3 \end{array}\right\}.$$

Interchanging the ordering of integration results in:

$$\left\{\begin{array}{r} 0 < \hat{z}_1 < 1 \\ \hat{z}_1 - 1 < \hat{z}_2 < 1 \\ \max\{\hat{z}_1, \hat{z}_2, \hat{z}_2 - \hat{z}_1\} - 1 < \hat{z}_3 < 1 + \min\{0, \hat{z}_2, \hat{z}_2 - \hat{z}_1\} \\ \max\{\hat{z}_1, \hat{z}_2, \hat{z}_3, \hat{z}_1 - \hat{z}_2, \hat{z}_1 - \hat{z}_3, \hat{z}_2 - \hat{z}_3, \hat{z}_1 - \hat{z}_2 + \hat{z}_3\} < \hat{x}_1 < 1 \\ \max\{0, \hat{z}_2, \hat{z}_3, \hat{z}_2 - \hat{z}_3\} < \hat{x}_2 < \hat{x}_1 + \min\{0, \hat{z}_2 - \hat{z}_1\} \\ \max\{0, \hat{z}_3\} < \hat{x}_3 < \hat{x}_2 + \min\{0, \hat{z}_3 - \hat{z}_2\} \end{array}\right\}$$

The min/max conditions can be removed by splitting the integration domain resulting in ten 6d simplices having the origin as a common vertex. These simplices could be mapped on $T^{(6)}$ by a linear transformation. Since only three of the six variables, namely the z_i are involved in the singularity, a $3D$ Duffy transformation can be applied to render the integrand analytic, $T^{(6)}$ is mapped onto $C^{(2)} \times S^{(4)}$.

As has been seen in Example 3.2, a $3D$ Duffy transformation can also be applied to $P(0 < z_6 < z_4)$, which is a union of two simplices. Therefore, the computational complexity could be reduced if some of the simplices can be merged to build a polyhedron of the form $P(0 < z_6 < z_4)$. Indeed, this is possible for six out of these ten.

Putting together all the transformations, gives the following result:

$$I^{(1)}_{\tau,t}(u,v) = \sum_{j=1}^{7} \int_{C^{(2)}} \int_{S^{(4)}} \omega_1^2 f^{(j)}(\eta) \hat{v}(\hat{x}^{(j)}) \hat{k}(\hat{x}^{(j)}, \hat{y}^{(j)}) \hat{u}(\hat{y}^{(j)}) d\omega d\eta, \tag{3.30}$$

where the transformations $(\omega, \eta) \to (\hat{x}^{(j)}, \hat{y}^{(j)})$ are given by:

$$\begin{pmatrix} \hat{x}_1^{(1)} \\ \hat{x}_2^{(1)} \\ \hat{x}_3^{(1)} \\ \hat{y}_1^{(1)} \\ \hat{y}_2^{(1)} \\ \hat{y}_3^{(1)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \eta_2\omega_1 + \omega_2 + \omega_3 \\ \omega_2 \\ (1-\eta_1)\omega_1 + \omega_2 + \omega_3 + \omega_4 \\ (1-\eta_1)\omega_1 + \omega_2 + \omega_3 \\ (1-\eta_1)\omega_1 + \omega_2 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(2)} \\ \hat{x}_2^{(2)} \\ \hat{x}_3^{(2)} \\ \hat{y}_1^{(2)} \\ \hat{y}_2^{(2)} \\ \hat{y}_3^{(2)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \eta_2\omega_1 + \omega_2 + \omega_3 \\ \eta_2\omega_1 + \omega_2 \\ (1-\eta_1)\omega_1 + \omega_2 + \omega_3 + \omega_4 \\ (1-\eta_1)\omega_1 + \omega_2 + \omega_3 \\ \omega_2 \end{pmatrix},$$

$$\begin{pmatrix} \hat{x}_1^{(3)} \\ \hat{x}_2^{(3)} \\ \hat{x}_3^{(3)} \\ \hat{y}_1^{(3)} \\ \hat{y}_2^{(3)} \\ \hat{y}_3^{(3)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \omega_1 + \omega_2 + \omega_3 \\ \eta_2\omega_1 + \omega_2 \\ (1-\eta_1)\omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \omega_2 + \omega_3 \\ \omega_2 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(4)} \\ \hat{x}_2^{(4)} \\ \hat{x}_3^{(4)} \\ \hat{y}_1^{(4)} \\ \hat{y}_2^{(4)} \\ \hat{y}_3^{(4)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \omega_2 + \omega_3 \\ \omega_2 \\ \eta_1\omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \eta_1\omega_1 + \omega_2 + \omega_3 \\ \eta_1\eta_2\omega_1 + \omega_2 \end{pmatrix},$$

$$\begin{pmatrix} \hat{x}_1^{(5)} \\ \hat{x}_2^{(5)} \\ \hat{x}_3^{(5)} \\ \hat{y}_1^{(5)} \\ \hat{y}_2^{(5)} \\ \hat{y}_3^{(5)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \eta_1\omega_1 + \omega_2 + \omega_3 \\ \eta_1\eta_2\omega_1 + \omega_2 \\ \omega_2 + \omega_3 + \omega_4 \\ \omega_2 + \omega_3 \\ \omega_2 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(6)} \\ \hat{x}_2^{(6)} \\ \hat{x}_3^{(6)} \\ \hat{y}_1^{(6)} \\ \hat{y}_2^{(6)} \\ \hat{y}_3^{(6)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \omega_1 + \omega_2 + \omega_3 \\ \omega_1 + \omega_2 \\ (1-\eta_1\eta_2)\omega_1 + \omega_2 + \omega_3 + \omega_4 \\ (1-\eta_1)\omega_1 + \omega_2 + \omega_3 \\ \omega_2 \end{pmatrix},$$

$$\begin{pmatrix} \hat{x}_1^{(7)} \\ \hat{x}_2^{(7)} \\ \hat{x}_3^{(7)} \\ \hat{y}_1^{(7)} \\ \hat{y}_2^{(7)} \\ \hat{y}_3^{(7)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \omega_1 + \omega_2 + \omega_3 \\ \omega_2 \\ (1-\eta_1+\eta_1\eta_2)\omega_1 + \omega_2 + \omega_3 + \omega_4 \\ \eta_1\eta_2\omega_1 + \omega_2 + \omega_3 \\ \eta_1\eta_2\omega_1 + \omega_2 \end{pmatrix},$$

and $f^{(j)}(\eta)$ is given by

$$f^{(1)}(\eta) = f^{(2)}(\eta) = f^{(3)}(\eta) = 1, \quad f^{(4)}(\eta) = f^{(5)}(\eta) = f^{(6)}(\eta) = f^{(7)}(\eta) = \eta_1.$$

The integrand in (3.30) is analytic and can be approximated by standard cubature techniques.

Remark 3.3. The techniques above are fully implicit in the sense that the form of the kernel function, the trial and the test space, and the order of the singularity of the kernel function are not used explicitly. This is very useful when testing the implementation. Simple test kernels, where the integrals can be computed analytically, e.g. polynomials, can be used to this end.

Remark 3.4. Due to the fact that some of the variables are not involved in the singularity, the variable ω_i, $i \geq 2$ do not appear in the kernel $\hat{k}(\hat{x}^{(j)}, \hat{y}^{(j)})$. They only come into play through the functions $\hat{v}(\hat{x}^{(j)})$ and $\hat{u}(\hat{y}^{(j)})$. Due to the choice of the ansatz functions, these functions are polynomials. Hence, the integration w.r.t. the variables ω_i, $i \geq 2$ can be carried out analytically.

(b) The Case of a Common Face

Let the affine mappings χ_τ and χ_t be given by:

$$\chi_\tau(\hat{\mathbf{x}}) = \mathbf{x}_0 + \mathbf{D}_{\chi_\tau}\hat{\mathbf{x}}, \quad \mathbf{x}_0 \in \mathbb{R}^3, \mathbf{D}_{\chi_\tau} = (\mathbf{t}_1 \,|\, \mathbf{t}_2 \,|\, \mathbf{t}_{3_x}),$$

and

$$\chi_t(\hat{\mathbf{y}}) = \mathbf{x}_0 + \mathbf{D}_{\chi_t}\hat{\mathbf{y}}, \quad \mathbf{D}_{\chi_t} = (\mathbf{t}_1 \,|\, \mathbf{t}_2 \,|\, \mathbf{t}_{3_y}).$$

The special form of the affine mappings results from the Assumption 3.1.

The integrand in (3.16) has a singularity of the form

$$\|\chi_\tau(\hat{\mathbf{x}}) - \chi_t(\hat{\mathbf{y}})\|^{-2} = \|\mathbf{D}_{\chi_\tau}\hat{\mathbf{x}} - \mathbf{D}_{\chi_t}\hat{\mathbf{y}}\|^{-2} \overset{\hat{z}_i = \hat{x}_i - \hat{y}_i,\, i=1,2}{=} \|\hat{z}_1\mathbf{t}_1 + \hat{z}_2\mathbf{t}_2 + \hat{x}_3\mathbf{t}_{3_x} - \hat{y}_3\mathbf{t}_{3_y}\|^{-2}.$$

The integrand in (3.13) is only singular for $\mathbf{x} = \mathbf{y}$. Due to the Assumption 3.1 the integrand above is only singular for $\hat{x}_i = \hat{y}_i$, $i = 1, 2$, and $\hat{x}_3 = \hat{y}_3 = 0$, i.e. for $\hat{z}_1 = \hat{z}_2 = \hat{x}_3 = \hat{y}_3 = 0$. Hence, the linear mapping L given by the matrix $(\mathbf{t}_1 \,|\, \mathbf{t}_2 \,|\, \mathbf{t}_{3_x} \,|\, -\mathbf{t}_{3_y})$ fulfills the required condition (see (3.26)).

The splitting of the integration domain is done analogously to the case (a) and results again in ten $6D$ simplices having the origin as a common vertex. This time four variables, namely z_1, z_2, x_3 and y_3 are involved in the singularity. Therefore, a $4D$ Duffy transformation renders the integrand analytic, $T^{(6)}$ is mapped onto $C^{(3)} \times S^{(3)}$.

As has been seen in Example 3.2, a $4D$ Duffy transformation can be also applied to $P(0 < z_5 < z_3, 0 < z_6 < z_5)$, which is a union of three simplices. Therefore, the computational complexity could be reduced by merging some of the simplices. Indeed, this is possible for nine out of these ten.

The result of all the transformations is given by

$$I^{(1)}_{\tau,t}(u,v) = \sum_{j=1}^{4} \int_{C^{(3)}} \int_{S^{(3)}} \omega_1^3 f^{(j)}(\eta)\hat{v}(\hat{x}^{(j)})\hat{k}(\hat{x}^{(j)}, \hat{y}^{(j)})\hat{u}(\hat{y}^{(j)}) d\omega d\eta, \tag{3.31}$$

where the transformations $(\omega, \eta) \to (\hat{x}^{(j)}, \hat{y}^{(j)})$ are given by:

$$\begin{pmatrix} \hat{x}_1^{(1)} \\ \hat{x}_2^{(1)} \\ \hat{x}_3^{(1)} \\ \hat{y}_1^{(1)} \\ \hat{y}_2^{(1)} \\ \hat{y}_3^{(1)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 \\ \eta_1\omega_1 + \omega_2 \\ \eta_1\omega_1 \\ (1 - \eta_2\eta_3)\omega_1 + \omega_2 + \omega_3 \\ (1 - \eta_2\eta_3)\omega_1 + \omega_2 \\ \eta_2(1 - \eta_3)\omega_1 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(2)} \\ \hat{x}_2^{(2)} \\ \hat{x}_3^{(2)} \\ \hat{y}_1^{(2)} \\ \hat{y}_2^{(2)} \\ \hat{y}_3^{(2)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 \\ (1 - \eta_2(1 - \eta_3))\omega_1 + \omega_2 \\ (1 - \eta_2)\omega_1 \\ \eta_1\omega_1 + \omega_2 + \omega_3 \\ \eta_1\omega_1 + \omega_2 \\ \eta_1\omega_1 \end{pmatrix},$$

$$\begin{pmatrix} \hat{x}_1^{(3)} \\ \hat{x}_2^{(3)} \\ \hat{x}_3^{(3)} \\ \hat{y}_1^{(3)} \\ \hat{y}_2^{(3)} \\ \hat{y}_3^{(3)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 \\ \omega_1 + \omega_2 \\ \eta_1\omega_1 \\ (1 - \eta_2\eta_3)\omega_1 + \omega_2 + \omega_3 \\ (1 - \eta_2)\omega_1 + \omega_2 \\ (1 - \eta_2)\omega_1 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(4)} \\ \hat{x}_2^{(4)} \\ \hat{x}_3^{(4)} \\ \hat{y}_1^{(4)} \\ \hat{y}_2^{(4)} \\ \hat{y}_3^{(4)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 \\ \omega_1 + \omega_2 \\ \omega_1 \\ (1 - \eta_1\eta_2\eta_3)\omega_1 + \omega_2 + \omega_3 \\ (1 - \eta_1\eta_2)\omega_1 + \omega_2 \\ \eta_1(1 - \eta_2)\omega_1 \end{pmatrix},$$

and $f^{(j)}(\eta)$ is given by

$$f^{(1)}(\eta) = f^{(2)}(\eta) = f^{(3)}(\eta) = \eta_2, \quad f^{(4)}(\eta) = \eta_1^2\eta_2.$$

The integrand in (3.31) is analytic and can be approximated by standard cubature techniques.

The transformation for the cases 1.(c) and (d) as well as the transformations in case of one volume and one surface element can be found in Appendix A.5 . The case of pairs of surface elements occurs also in the BEM and is treated in [51].

Cubature Orders

The accuracy of the cubature methods has to be chosen such that the error resulting from this further discretization does not increase the asymptotic discretization error of the Galerkin method. The impact of the cubature error is estimated by the following lemma.

Lemma 3.2 (first Strang lemma). *Let the bilinear form a_h be continuous and satisfy the coercivity condition (3.9) with α independent of h. Then the following error estimate holds for the solution $\tilde{u}_h$ of the fully discrete Galerkin system:*

$$\|u - \tilde{u}_h\| \leq \inf_{v_h \in S_h} \Big((1 + \frac{C_S}{\alpha}) \|u - v_h\| + \frac{1}{\alpha} \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|} \Big) + \frac{1}{\alpha} \sup_{w_h \in V_h} \frac{\langle f, w_h \rangle - \langle f_h, w_h \rangle}{\|w_h\|}. \quad (3.32)$$

where C_s denotes the continuity constant and α the coercivity constant.

Proof. refer to [10] □

The coercivity of a_h follows from that of a if the error between a and a_h is small

$$|a(u, v) - a_h(u, v)| \leq \varepsilon \|u\| \|v\| \quad \forall u, v \in V_h,$$

for a suitable $\varepsilon < \alpha$. In this case a_h is coercive with coercivity constant $\tilde{\alpha} \geq \alpha - \varepsilon$.

As shown in Section 3.1, the Galerkin solution has an error of order $O(h^{k+1})$ in regions where the solution u of the continuous problem is smooth enough (see Theorem 3.2). To guaranty the same estimate for the fully discrete method, the following consistency condition has to be required (due to Lemma 3.2):

$$|a(u,v) - a_h(u,v)| \leq Ch^{k+1}\|v\|_{L^2}\|u\|_{H^k} \quad \forall u, v \in V_h, \tag{3.33}$$

$$|\langle f - f_h, v\rangle| \leq Ch^{k+1}\|v\|_{L^2} \quad \forall v \in V_h, \tag{3.34}$$

where $C = C_s C_{\text{approx}}$ (see (3.12))

This global condition could be split up into a local condition for the error on pairs of elements:

$$E_{\tau,t} = |Q_{\tau,t} - I_{\tau,t}|,$$

where $Q_{\tau,t}$ denotes the cubature formula used to approximate the integral $I_{\tau,t}$ in (3.16).

Lemma 3.3. *Let the Grid $\mathcal{G}$ be quasi-uniform and shape regular and the cubature formulae be chosen such that for all pairs (τ, t) of elements in $\mathcal{G} \times \mathcal{G}$*

$$E_{\tau,t} \leq \varepsilon_{\tau,t}\|v\|_{L^2(\tau)}\|u\|_{H^k(t)}, \tag{3.35}$$

with

$$\varepsilon_{\tau,t} \leq \begin{cases} Ch^{3+k+1} & \text{if } \tau, t \in \mathcal{G}^{(v)}, \\ Ch^{2\frac{1}{2}+k+1} & \text{if } \tau \in \mathcal{G}^{(v)} \text{ and } t \in \mathcal{G}^{(s)} \text{ or vice versa}, \\ Ch^{2+k+1} & \text{if } \tau, t \in \mathcal{G}^{(s)}. \end{cases} \tag{3.36}$$

Then the global condition (3.33) is fulfilled.

Proof. It holds

$$|a(u,v) - a_h(u,v)| \overset{(3.35)}{\leq} \sum_{(\tau,t)\in\mathcal{G}\times\mathcal{G}} \varepsilon_{\tau,t}\|v\|_{L^2(\tau)}\|u\|_{H^k(t)}.$$

Splitting up the grid into its volume and surface part:

$$\mathcal{G} \times \mathcal{G} = \mathcal{G}^{(v)} \times \mathcal{G}^{(v)} \cup \mathcal{G}^{(v)} \times \mathcal{G}^{(s)} \cup \mathcal{G}^{(s)} \times \mathcal{G}^{(v)} \cup \mathcal{G}^{(s)} \times \mathcal{G}^{(s)}$$

and using the notation $u = \binom{u_1}{u_2}$, $u_1 \in V_h^{(v)}$ and $u_2 \in V_h^{(s)}$, one obtains for the volume part $\mathcal{G}^{(v)} \times \mathcal{G}^{(v)}$:

$$\sum_{(\tau,t)\in\mathcal{G}^{(v)}\times\mathcal{G}^{(v)}} \varepsilon_{\tau,t}\|v_1\|_{L^2(\tau)}\|u_1\|_{H^k(t)}$$

$$\overset{(3.36)}{\leq} Ch^{3+k+1}\Big(\sum_{\tau\in\mathcal{G}^{(v)}} \|v_1\|^2_{L^2(\tau)} \sum_{t\in\mathcal{G}^{(v)}} 1\Big)^{\frac{1}{2}}\Big(\sum_{t\in\mathcal{G}^{(v)}} \|u_1\|^2_{H^k(t)} \sum_{\tau\in\mathcal{G}^{(v)}} 1\Big)^{\frac{1}{2}}$$

$$= Ch^{k+1}h^3\,\sharp\,\mathcal{G}^{(v)}\|v_1\|_{L^2(D)}\|u_1\|_{H^k(D)}$$

Since $\mathcal{G}$ is quasi-uniform, it holds $\sharp\,\mathcal{G}^{(v)} \sim \frac{1}{h^3}$, where $\sharp\,\mathcal{G}^{(v)}$ denotes the number of nodes in $\mathcal{G}^{(v)}$ when piecewise linear ansatz functions are used and the number of elements in case of piecewise constant ansatz functions. Analogous considerations for the remaining parts, using the fact that $\sharp\,\mathcal{G}^{(s)} \sim \frac{1}{h^2}$ yields altogether

$$|a(u,v) - a_h(u,v)| \leq Ch^{k+1}\Big(\|v_1\|_{L^2(D)}\|u_1\|_{H^k(D)} + \|v_1\|_{L^2(D)}\|u_2\|_{H^k(\partial D)}$$

$$+ \|v_2\|_{L^2(\partial D)}\|u_1\|_{H^k(D)} + \|v_2\|_{L^2(\partial D)}\|u_2\|_{H^k(\partial D)}\Big)$$

$$\leq 2Ch^{k+1}\|v\|_{L^2}\|u\|_{H^k}.$$

□

In the integral equation stemming from the RTE the right hand side f is computed by applying an integral operator similar to that on the left hand side. Therefore, the condition (3.34) can be shown analogously to (3.33). This gives the following theorem.

Theorem 3.3. *Let the Grid $\mathcal{G}$ be quasi-uniform and shape regular and the cubature formulae be chosen such that the estimate (3.35) is fulfilled for all pairs of elements. Then for sufficiently small $h \leq h_0$ the fully discrete problem is uniquely solvable and the quasi optimal error estimate*

$$\|u - \tilde{u}_h\| \leq (1 + \frac{C_s}{\alpha}) C_{approx} h^s \|u\|_{H^s}, \tag{3.37}$$

holds if the solution u of the continuous problem is in H^s, $s \leq k+1$.

In the following the methods for calculating the multiple integrals are presented and the number of quadrature points necessary to meet the local condition (3.35) is determined. As shown above, the singular integrals are transformed into integrals over domains of the form $C^{(d_1)} \times S^{(d_2)}$. The integrals over the simplices $S^{(d_2)}$ are pulled back onto $C^{(d_2)}$ by using simplex coordinates. The same applies to the regular integrals over the product domains $T^{(d_1)} \times T^{(d_2)}$. Hence, all integrals are of the form

$$\int_{C^{(d)}} f(x) dx,$$

with $d = 6$ in case of pairs of volume elements, $d = 5$ in case of one volume and one surface element and $d = 4$ in case of surface elements. For the integrals over $C^{(d)}$ d-dimensional tensor product versions of Gauss quadrature are used.

Let $(\xi_{i,q}, w_{i,q})$ denotes the nodes and weights of a q-point Gauss quadrature formula and let $q = (q_1, \ldots, q_d)^T \in \mathbb{N}^d$ then the integral over a function $f \in C^0(C^{(d)})$ could be approximated by the tensor product Gauss formula

$$Q_q(f) = \sum_{i_1=1}^{q_1} \cdots \sum_{i_d=1}^{q_d} w_{i_1,q_1} \cdots w_{i_d,q_d} f(\xi_{i_1,q_1}, \ldots \xi_{i_d,q_d}).$$

The error of such a tensor product quadrature can be estimated by transforming the problem into a sequence of $1D$ integrations, e.g. in $2D$

$$(I - Q)f = I_1[(I_2 - Q_2)f + Q_2[(I_1 - Q_1)f].$$

Hence, the error can be estimated by the sum of the $1D$ errors times the size of the domain in the remaining directions. The problem of estimating the error is reduced to estimating the error of a $1D$ quadrature formula.

When estimating the error for a given number of Gauss points, again the singularity of the kernels is the major difficulty. A standard procedure to get an error estimate is using Taylor expansion of the integrand and exploiting the degree of exactness of the method ($2q - 1$ for a q-point Gauss quadrature). Since the variables in the integrand are scaled with the mesh size h, a factor h^{2q} for the error term is obtained through the chain rule. On the other hand, the $2q$-th derivative of the integrand is of the order $O(\delta^{-(2q+\alpha)})$ where $\delta = \text{dist}(x, y)$ for some $x \in \tau$, $y \in t$, and α is the order of the singularity of the kernel. Hence, standard quadrature formulae with fixed degree to obtain the required accuracy can only be applied if $\text{dist}(\tau, t) \geq C$ with C independent of h. This region is called farfield (w.r.t. integration). For the nearly singular case, $\text{dist}(\tau, t) = ch, c > 1$, the error is proportional to

$$h^{2q}(ch)^{-(2q+\alpha)} = c^{-(2q+\alpha)} h^{-\alpha}.$$

Since $c > 1$, convergence can be achieved by choosing $q \sim |\log h|$. The minimal number of Gauss points in x and y direction for every element pair needed to meet the local consistency requirement (3.36) can be found in [52].

It remains to consider the case where the integrand is singular. In this case some kind of Duffy transform is used to render the integrand analytic. In the $2D$ Example 3.1 the integrand is of the form $\frac{1}{\sqrt{1+\eta^2}} g(\omega, \omega\eta) =: \tilde{g}(\omega, \eta)$. A simple scaling yields

$$I_h = \int_0^h \int_0^1 \tilde{g}(\omega, \eta) d\eta d\omega = \int_0^1 \int_0^1 h\tilde{g}(h\zeta, \eta) d\eta d\zeta. \tag{3.38}$$

To estimate the error of the integration w.r.t. ζ, Taylor expansion exploiting the degree of exactness can be used to show that a fixed number of Gauss points is sufficient to obtain the required accuracy. On the other hand, for η no powers of h are obtained through the chain-rule since the integrand only depends on η rather than on $h\eta$. Hence, another method, which does not involve derivatives of the integrand, is needed to estimate the error.

Theorem 3.4. *Let f be analytic in $[-1, 1]$ and admit an analytic continuation into the closed ellipse $\mathcal{E}_\rho \subset \mathbb{C}$ with foci at ± 1 and semi-axis sum $\rho > 1$. Then*

$$|E_q f| = |If - Q_q f| \leq C\rho^{-2q} \max_{z \in \partial\mathcal{E}_\rho} |f(z)|, \tag{3.39}$$

where Q_q denotes the q-point Gauss-Legendre quadrature formula on $[-1, 1]$. *Proof.* refer to [15] □

Since the integrand in (3.38) has a singularity for the complex values $\eta = \pm i$ the size of the ellipse where the integrand is analytic w.r.t. η is of size $O(1)$. The estimate in (3.39) meets the local consistency condition (3.36) if $O(|\log h|)$ Gauss points in direction η are used. Similar results hold for the more general Duffy transformations applied to compute the matrix entries for the integral equation. More details can be found in [51] and [43].

Asymptotic results are only of questionable value for practical applications since the constants are only known with uncertainty. The numerical results in [51] for the double layer potential of the Laplace equation show that in this case the following number of quadrature points is sufficient to preserve the convergence rate of the Galerkin method: when the elements are close to each other, in each direction, in which asymptotically $O(|\log h|)$ quadrature points are needed, two quadrature points are sufficient, in the other directions only one point is needed. In the farfield the midpoint rule is sufficient.

The kernels in Equation (2.26) contain the exponential damping factor $e^{-\gamma\|x-y\|}$. The decay due to this factor is very strong when the optical thickness of the considered elements is large. This complicates the approximation of the integrals substantially as the following example shows.

Example 3.3. Let the function g in Example 3.1 be given by $g(z) = e^{-\gamma\|z\|}$. Applying Duffy transformation results in

$$If = \int_0^1 \int_0^1 \frac{1}{\sqrt{1+\eta^2}} e^{-\gamma\omega\sqrt{1+\eta^2}} d\omega d\eta. \tag{3.40}$$

Numerical experiments show that with increasing γ the number q_ω of Gauss points in direction ω, needed to obtain a given relative error tolerance, increases. While the number q_η of Gauss

points, needed for the direction η, is nearly independent of γ. For small γ, q_ω is smaller than q_η as predicted by the asymptotic error analysis, but for large γ the reverse statement is true. Table 3.1 shows the number of Gauss points for the variables ω and η needed to meet the error tolerance $tol = 5 \cdot 10^{-4}$.

γ	0.01	0.1	1.0	2.0	5.0	10.0
q_ω	1	1	2	3	4	6
q_η	2	2	2	3	3	3

Table 3.1: Number of Gauss points for different optical thicknesses

The integration with respect to ω in the integral (3.40) can be carried out analytically

$$If = \int_0^1 \frac{1}{\gamma\sqrt{1+\eta^2}}\left(1 - e^{-\gamma\sqrt{1+\eta^2}}\right)d\eta. \tag{3.41}$$

The problem with the exponential decay is relaxed since the exponent in (3.40) varies in the interval $[0, \gamma\sqrt{2}]$ whereas the exponent in (3.41) is in $[\gamma, \gamma\sqrt{2}]$. Thus, the number of Gauss points needed for (3.41) is nearly independent of γ.

The example above shows that the number of Gauss points needed for the variable ω_1 becomes large for large γ. The remedy is to use analytical integration for this variable. Unfortunately, due to the inner integration w.r.t. ω_i, $i \geq 2$ and due to the use of piecewise linear trial functions, the integrand resulting from the discretization of the integral equation as derived from the RTE (see Equation (2.26)) is of the form

$$\int_0^1 \cdots \int_0^1 p_1(\omega_1)\frac{1}{p_2(\eta)}e^{-\gamma\omega_1\sqrt{p_2(\eta)}}d\omega d\eta_{d_s-1}\cdots d\eta_1,$$

where p_1 and p_2 are polynomials. The integration w.r.t. ω_1 can still be carried out analytically but the resulting term is rather complicated. Therefore, it is not recommendable to do that in praxis. Instead, a large number of Gauss points for the variable ω_1 has to be used. Furthermore, the problem due to the exponential decay occurs also when the integrand is regular, and in this case it is not possible to carry out the integration analytically with respect to some of the variables.

3.3 Solution of the Linear System

The discretization of an integral equation leads to a linear system of algebraic equations with a full matrix. A direct solver, like the Gauss elimination method, needs $O(n^3)$ operations and is therefore not suitable for this kind of problem. The costs could be reduced tremendously by using an iterative method. The only knowledge about the matrix which is required for an iterative method is the possibility to perform a matrix-vector multiplication. By applying the matrix compression methods discussed in Chapter 4, the approximate computation of such a matrix-vector multiplication is possible with almost linear complexity although the matrix is fully populated.

Using the collocation method for the integral equation (2.39) leads to a linear system of the form

$$\mathbf{Au} := (\mathbf{I} - \mathbf{K})\mathbf{u} = \mathbf{f}, \tag{3.42}$$

where $\mathbf{K}$ is the discretization of the integral operator $\hat{K}$ of Chapter 2.

The classical iterative method for such an equation of the second kind is the Picard iteration based on the Neumann series $(Id - K)^{-1} = \sum_{i=1}^{\infty} K^i$, if $\|K\| < 1$

$$\mathbf{u}_{i+1} = \mathbf{K}\mathbf{u}_i + \mathbf{f}, \qquad i = 0, 1, \ldots,$$

with initial guess $\mathbf{u}_0$.

The method converges if $\|\mathbf{K}\| < 1$ (for some norm). In Chapter 2 it is shown that, except for the case of total reflection and zero absorption, $\|\hat{K}\|_\infty < 1$. The sum of row i of the matrix $\mathbf{K}$ (assuming that the integrals are computed exactly) can be estimated as follows

$$\sum_j |k_{ij}| = \sum_j k_{ij} = \sum_j (\hat{K} b_j)(x_i) = (\hat{K} \underbrace{\sum_j b_j}_{\equiv 1})(x_i) \leq \|\hat{K}1\|_\infty \leq \|\hat{K}\|_\infty,$$

where it has been exploited that the basis functions form a partition of unity. The above estimate shows that $\|\mathbf{K}\|_\infty \leq \|\hat{K}\|_\infty < 1$ and, hence, the Picard iteration converges with convergence rate $\|\hat{K}\|_\infty$.

Remark 3.5. The above considerations show a further favorable property, namely that the matrix A in (3.42) is a M-matrix, i.e. it is strictly diagonal dominant and $a_{ij} \leq 0$ for $i \neq j$. The inverse $\mathbf{A}^{-1}$ of such a matrix is element wise non negative (refer to [23]). This property guarantees that the solution $\mathbf{u}$ of (3.42) is non-negative if the right hand side is non-negative. That is, the non-negativity of the solution of the continuous problem (see Theorem 2.4) is inherited by the discrete solution. Note that, when the RTE is discretized directly, e.g. by the DOM, one often encounters the problem that the computed intensities are negative (see [12]).

The convergence rate of the Picard iteration can be very slow. Other methods, like the Krylov subspace methods, yield better convergence rates. They minimize the iteration error over the affine spaces

$$X_i = \mathbf{u}_0 + \mathcal{K}_i, \quad \mathcal{K}_i = \text{span}\{\mathbf{r}_0, \mathbf{A}r_0, \ldots, \mathbf{A}^{i-1}\mathbf{r}_0\},$$

with initial residual $\mathbf{r}_0 = \mathbf{f} - \mathbf{A}\mathbf{u}_0 = \mathbf{A}\mathbf{e}_0$, where $\mathbf{e}_0 = \mathbf{A}^{-1}\mathbf{f} - \mathbf{u}_0$ denotes the initial error. The different methods are distinguished by the fact in which norm the error is minimized. For symmetric matrices (Please remember that the Galerkin discretization of the self-adjoint operator K_{sym} results in a symmetric matrix.) the MINRES method minimizes the Euclidean norm of the residual $\mathbf{r}_i$, the conjugate gradient (CG) method minimizes the energy norm of the error $\mathbf{e}_i$. The energy norm is only defined for positive definite matrices. Hence, the CG method can only be applied to such matrices.

The discrete Galerkin system for the weak formulation is of the form

$$\mathbf{A}\mathbf{u} := (\mathbf{M} - \mathbf{K})\mathbf{u} = \mathbf{f}, \tag{3.43}$$

where $\mathbf{M}$ is the mass matrix and $\mathbf{K}$ is the discretization of the integral operator K_{sym}.

Since the operator A and therefore its restriction to the finite dimensional ansatz space is coercive, the matrix $\mathbf{A}$ is positive definite:

$$\mathbf{u}^T\mathbf{A}\mathbf{u} = \sum_{i,j} u_i u_j \langle b_i, Ab_j \rangle_{L^2(\mu)} = \langle u, Au \rangle_{L^2(\mu)} \geq \alpha \langle u, u \rangle_{L^2(\mu)} > 0.$$

Due to the structure of the Krylov spaces the minimization property can be written as follows

$$\|e_i\| \leq \min_{\substack{p\in\mathcal{P}_i,\\ p(0)=1}} \max_{\lambda\in\sigma(A)} |p(\lambda)|\|e_0\|, \tag{3.44}$$

where $\|\cdot\| = \|\cdot\|_{\mathbf{A}}$ in case of CG. (Remember $\mathcal{P}_i$ is the space of polynomials up to degree i.) This minimum is small if the eigenvalues are clustered in a few regions. Using the approximation property of Chebychev polynomials one can derive an estimate for the case that only the condition number of the matrix is known

$$\|e_i\| \leq 2\Big(\frac{\sqrt{\kappa(\mathbf{A})}-1}{\sqrt{\kappa(\mathbf{A})}+1}\Big)^i \|e_0\|, \tag{3.45}$$

where $\kappa(\mathbf{A})$ denotes the condition number of $\mathbf{A}$. A small condition number leads to fast convergence. The convergence rate is much faster than in case of the Picard iteration. Furthermore, as shown above, the Krylov methods converge even faster if the eigenvalues are clustered.

According to Chapter 2, the eigenvalues of the symmetric operator K_{sym} lie in the interval $[-c, c]$ with $c = \|\hat{K}\|_{L^\infty(\mu)} < 1$. Since the considered equation is of the second kind the critical eigenvalues are the ones close to one. Such eigenvalues can only occur if c is close to one. Due to Remark 2.8 this is the case if the optical parameters satisfy one of the following conditions

1. $\omega \approx 1$ and $\tau \gg 1$,
2. $\rho \approx 1$ and τ small.

Since c is only an upper bound for the eigenvalues, c close to one, does not necessarily imply that there exist eigenvalues close to one. Furthermore, according to Chapter 2, K_{11}, K_{12} and K_{21} are compact operators and K_{22} is compact if the surface is smooth. Hence, it is reasonable to assume that the spectrum of K resembles that of a compact operator, i.e. the eigenvalues λ_k tend to zero. If this convergence is fast, most of the eigenvalues are clustered near zero and Krylov methods yield fast convergence even if the largest eigenvalues are close to one. These considerations show that the convergence of Krylov methods is not necessarily slow in the two cases above.

Remark 3.6. Since the Galerkin method is used for the discretization, the eigenvalues of the discrete operator correspond to generalized eigenvalues of the system matrix. (This relationship is examined in more detail later in this section (see (3.68) below)). The generalized eigenvalues resemble the eigenvalues of the continuous operator, e.g. the values should lie in the interval $(0, 2)$ and should be clustered near one if the convergence of the eigenvalues of the continuous integral operator to zero is sufficiently fast. Therefore, the figures in this section show the generalized eigenvalues instead of the standard eigenvalues to facilitate a better interpretation of the results. We have to keep in mind that the convergence of the Krylov methods is determined by the standard eigenvalues rather than the generalized eigenvalues. Since the mass matrix is diagonal dominant, the qualitative behavior of the distribution of the standard eigenvalues is rather similar to that of the generalized eigenvalues. That is, if the distribution of the generalized eigenvalues is well suited for a Krylov method, the same statement holds for the standard eigenvalues. On the other hand, it has to be said that in cases where the convergence of the eigenvalues to zero is very fast, e.g. in the optically thin case. The convergence of Krylov methods is even faster if the collocation method is used since the eigenvalue distribution of the integral operator is not smeared out by the convolution with the mass matrix as it is the case if the Galerkin method is used.

We begin with the investigation of the second of the two critical cases above. Since we have not managed to derive analytical statements about the spectrum of the integral operator that are more detailed than the upper bound derived above, we examine one special situation by computing the eigenvalues of the discretized operator numerically.

Remark 3.7. Before describing the setup of the considered problem, we want to mention at this point that all numerical experiments presented in this thesis handle situations where the coefficients ρ, κ and σ are constant in space since the numerical methods have only been implemented for this case.

The computational domain is the unit cube, and the optical parameters are chosen as follows: $\rho = 1.0$, $\gamma = 0.1$ and $\omega = 0.5$. For the discretization piecewise linear ansatz functions on a regular decomposition of the domain into tetrahedral elements are used. The number of vertices is $9^3 = 729$, 386 of which lie on the surface of the domain, i.e. the number of unknowns is 1115. Figure 3.2 shows the distribution of the generalized eigenvalues of the system matrix $\mathbf{A}$ in the complex plane. Since the matrix is symmetric, all eigenvalues lie on the real axis.

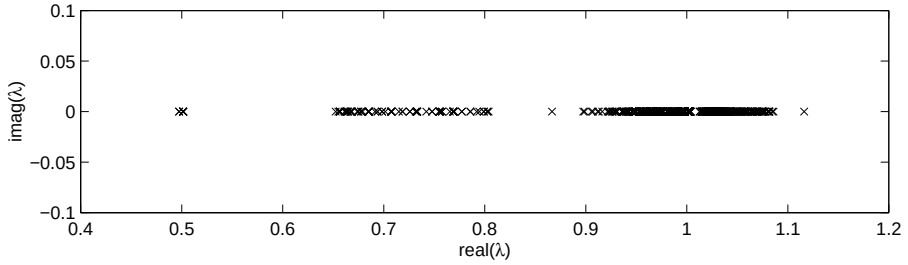

Figure 3.2: Spectrum of the system matrix in the strong reflecting case

The smallest eigenvalue is approximately 0.5. That is, the largest eigenvalue of the discrete integral operator is much smaller than its $\|\cdot\|_\infty$ norm, which is computed numerically as 0.9651. Furthermore, many of the eigenvalues are clustered in the interval $[0.9, 1.1]$. This interval contains 1019 of the 1115 eigenvalues. That is, the convergence of the eigenvalues of the integral operator to zero is fast. The condition number of the system matrix is $\kappa(\mathbf{A}) = 2.3$. Altogether, it can be said that the eigenvalue distribution is well suited for Krylov methods and no preconditioning is necessary.

In the first case, the critical eigenvalue distribution is due to the operator K_{11}. Hence, it is reasonable to consider this operator on its own in a first step. This corresponds to the case of non reflecting boundaries (2.6). In a second step, it is examined, what conclusions can be drawn from this contemplations, when the whole system is considered.

To accelerate the convergence of an iterative method a preconditioner can be used, i.e. the system matrix $\mathbf{A}$ is multiplied by a matrix $\mathbf{P}$, which has the property that the eigenvalue distribution of $\mathbf{PA}$ is better suited for a Krylov method. That is, the eigenvalues are either clustered in a small number of small regions or the condition number is small. One way for the construction of a preconditioner is to find a matrix, which has eigenvalues and eigenfunctions resembling that of the system matrix and whose inverse is easy to compute.

The critical case corresponds to the diffusion limit considered in Section 1.2. In this case the diffusion equation (1.60) with Robin boundary conditions (1.61) is a good approximation of the RTE. Since the diffusion equation is a differential equation, its numerical treatment is much less costly compared with that of the integral equation. Hence, a discretization of this

problem seems to yield a reasonable preconditioner. This preconditioning is called diffusion-synthetic acceleration (DSA). It has been introduced by Larsen [36] in context of the discrete ordinate method.

In the diffusion limit the solution operator (refer to (1.65))

$$(\hat{T}_1 - \sigma Id)^{-1}\kappa$$

corresponding to the diffusion equation is a good approximation to the solution operator

$$(Id - \omega K)^{-1}(1-\omega)K$$

of the integral equation. A simple transformation

$$(Id - \omega K)^{-1}(1-\omega)K = [K^{-1}(Id - \omega K)]^{-1}(1-\omega) = [\gamma K^{-1} - \sigma Id]^{-1}\kappa,$$

shows that the integral operator K and the inverse of the diffusion operator $\hat{T}_1^{-1}$ should be alike in the diffusion limit. In the following, the spectral properties of the integral operator and the diffusion operator are examined in more detail. Since the exact spectrum depends on the computational domain, it is only possible to investigate a special reference situation. The general characteristics of the spectrum should be similar in other situations. To be able to obtain some theoretical results about the spectrum of the integral operator, the $1D$ problem (1.55) is considered in an article by Faber and Manteuffel [38].

It is convenient to rewrite the equation in optical coordinates $x = \gamma z$:

$$\underbrace{\mu \frac{\partial}{\partial x} I(x,\mu) + I(x,\mu)}_{=:H_\mu I} = \omega \underbrace{\frac{1}{2}\int_{-1}^{1} I(x,\mu')d\mu'}_{=LI=\frac{1}{4\pi}G} + (1-\omega)B(x), \quad x \in (-a,a), \mu \in [-1,1], \tag{3.46}$$

$$I(-a,\mu) = 0,\ \mu > 0, \qquad I(a,\mu) = 0,\ \mu < 0. \tag{3.47}$$

where $\omega = \frac{\sigma}{\gamma} \leq 1$. The asymptotic diffusion limit in this coordinates is characterized by $a \gg 1$ and $\omega \approx 1$. The integral equation in this coordinates reads as

$$G - \omega KG = 4\pi(1-\omega)KB, \tag{3.48}$$

where

$$(KG)(x) = \frac{1}{2}\int_{-a}^{a} E_1(|x-y|)G(y)dy. \tag{3.49}$$

Since E_1 is weakly singular, K is a self-adjoint, positive definite and compact operator. Furthermore, it holds $\|K\|_\infty = 1 - E_2(a) \overset{a\to\infty}{\to} 1$. Hence, the eigenvalues of $Id - \omega K$ lie in the interval $[\varepsilon, 1]$, where ε is small if $a \gg 1$ and $\omega \approx 1$, i.e. in the asymptotic diffusion limit. The eigenvalues near ε are responsible for the slow convergence of the Krylov methods. Hence, a preconditioner has to consider these eigenvalues. The main part of the eigenvalues is clustered near one and does not have a bad influence on the convergence rate.

The diffusion operator $\hat{T}_1$ corresponding to the diffusion Equation (1.60) for the $1D$ problem in optical coordinates is given by

$$\hat{T}_1 G = -\frac{1}{3}\frac{d^2}{dx}G(x) + G(x), \tag{3.50}$$

subjected to the Robin boundary conditions

$$G(\pm a) \pm \beta G'(\pm a) = 0, \tag{3.51}$$

where $\beta \approx 0.7104$ in Section 1.2.

Since $\hat{T}_1$ is a linear self-adjoint differential operator of second order on a $1D$ interval, its eigenvectors are sine or cosine functions: $\sin(\frac{\eta_k}{a}x)$ or $\cos(\frac{\eta_k}{a}x)$. The constants η_k have to be chosen such that the boundary conditions are fulfilled. The eigenfunctions of $\hat{T}_1$ are given by

$$e_k(x) = \begin{cases} \cos(\frac{\eta_k}{a}x) & \text{for } k \text{ odd}, \\ \sin(\frac{\eta_k}{a}x) & \text{for } k \text{ even}, \end{cases} \tag{3.52}$$

where $\eta_k(x)$ is the solution of the equation:

$$\begin{cases} \frac{\beta}{a}\eta_k = \cot \eta_k & \text{for } k \text{ odd}, \\ \frac{\beta}{a}\eta_k = -\tan \eta_k & \text{for } k \text{ even}. \end{cases} \tag{3.53}$$

The corresponding eigenvalues of $\hat{T}_1^{-1}$ (the Green's function of $\hat{T}_1$) are given by

$$\lambda_k = \frac{1}{1 + \frac{1}{3}(\frac{\eta_k}{a})^2}, \tag{3.54}$$

which yields the following estimate for $\frac{\beta k}{a} \ll 1$

$$\lambda_k \approx 1 - \frac{1}{3}(\frac{k\pi}{2a})^2. \tag{3.55}$$

That is, for $\frac{\beta k}{a} \ll 1$, λ_k is close to one. That is, the frequencies of interest correspond to a small value of k. For these k, η_k can be approximated as follows

$$\eta_k = \frac{k\pi}{2} - \frac{\beta k\pi}{2a} + O((\frac{\beta k}{a})^3). \tag{3.56}$$

Remark 3.8. The eigenfunctions e_k are highly oscillatory if k is large. Furthermore, the eigenvalues converge to zero. This is not surprising since $\hat{T}_1^{-1}$ should approximate the compact operator K. The eigenvalues for small k are close to one. These eigenvalues and the corresponding eigenvectors should resemble that of K to yield a good preconditioner.

An explicit computation of the spectrum of the integral operator K is not possible, even for this simplified $1D$ model problem. But some estimates can be derived by considering the factorization $K = LH^{-1}L^+$ (see (2.12)). Since H_μ is a linear differential operator of first order, the eigenvalue decomposition of $H_\mu^* H_\mu$ can be computed analytically. This corresponds to a singular value decomposition of H_μ (see Appendix A.6 for more details)

$$H_\mu^{-1} f = \sum_{k=1}^{\infty} \frac{\cos(2\xi_k)}{a\delta_0(\xi_k)} \begin{cases} \int_{-a}^{a} \sin(\frac{\xi_k}{a}(x+a)) \sin(\frac{\xi_k}{a}(s-a)) f(s) ds, \ \mu > 0, \\ \int_{-a}^{a} \sin(\frac{\xi_k}{a}(x-a)) \sin(\frac{\xi_k}{a}(s+a)) f(s) ds, \ \mu > 0, \end{cases} \tag{3.57}$$

where $a\delta_0(\xi_k)$ is a factor to normalize the singular vectors. The frequencies ξ_k are rather similar to the solutions of (3.53). For $\frac{\mu k}{2a} \ll 1$, it holds

$$\xi_k(\mu) = \frac{k\pi}{2} - \frac{\mu k\pi}{4a} + O((\frac{\mu k}{4a})^3) \quad \text{for } \mu > 0, \tag{3.58}$$

$$\xi_k(\mu) = \xi_k(-\mu) \quad \text{for } \mu < 0. \tag{3.59}$$

To find K, L has to be applied to (3.57). Considering a single term of (3.57), using (3.59) and expanding the sin terms yields

$$\int_0^1 \frac{\cos(2\xi_k)}{a\delta_0(\xi_k)} \Big(\cos^2 \xi_k \sin(\frac{\xi_k}{a}x) \sin(\frac{\xi_k}{a}s) - \sin^2 \xi_k \cos(\frac{\xi_k}{a}x) \cos(\frac{\xi_k}{a}s) \Big) d\mu. \tag{3.60}$$

This does not constitute a singular value decomposition of K since the functions of x and s are not separable (note: $\xi_k = \xi_k(\mu)$; but it does yield useful bounds for $\frac{k}{a}$ small. In this case one of the terms in (3.60) is very small. To see this, let $\xi_k = \frac{k\pi}{2} - \Theta_k$, with $0 < \Theta_k = \Theta_k(\mu) < \frac{\mu k\pi}{4a} \ll 1$. Hence, $\cos\Theta_k \approx 1$. Using the addition theorem for sine and cosine functions gives

$$\cos^2 \xi_k = \frac{1}{2}(1 + (-1)^k \cos\Theta_k) \approx 0 \text{ for } k \text{ odd},$$

$$\sin^2 \xi_k = \frac{1}{2}(1 - (-1)^k \cos\Theta_k) \approx 0 \text{ for } k \text{ even}.$$

Since $\cos(2\xi_k)$ is negative for k odd and positive for k even, each term in (3.60) includes a dominant positive part and a small negative part. Separating these parts results in a splitting of K: $K = K_1 - K_2$. Both K_1 and K_2 are self-adjoint and positive. The odd terms of K_1 are given by (see (3.60))

$$\int_0^1 \frac{\cos\Theta_k(1 + \cos\Theta_k)\delta_1(\xi_k)}{2\delta_0(\xi_k)} \frac{\cos(\frac{\xi_k}{a}x)\cos(\frac{\xi_k}{a}s)}{a\delta_1(\xi_k)} d\mu, \tag{3.61}$$

where $a\delta_1(\xi_k)$ is again a normalization factor. Using the mean value theorem yields

$$\frac{\cos(\frac{\hat{\xi}_k}{a}x)\cos(\frac{\hat{\xi}_k}{a}s)}{a\delta_1(\xi_k)} \int_0^1 \frac{\cos\Theta_k(1 + \cos\Theta_k)\delta_1(\xi_k)}{2\delta_0(\xi_k)} d\mu, \tag{3.62}$$

where $\hat{\xi}_k \in (\frac{k\pi}{2} - \frac{k\pi}{4a}, \frac{k\pi}{2})$, (refer to (3.58)). (3.62) shows that the eigenvectors of K_1 for odd k are given by $\cos(\frac{\hat{\xi}_k}{a}x)$. Compare this with the eigenvectors of $\hat{T}_1^{-1}$ given in (3.52) and note that η_k, for $\frac{k}{a}$ small, is very similar to $\hat{\xi}_k$ (see (3.56)).

An approximation of the eigenvalues of K_1 given by the integral in (3.62) is computed in [38]. The result is also very similar to the eigenvalues of $\hat{T}_1^{-1}$ given in (3.55). The same holds for even k.

Since K_2 is small for a small value of $\frac{k}{a}$, the low frequency part of the spectrum of K resembles that of K_1 for large a. Hence, the low frequency part of the spectrum of K is very similar to that of $\hat{T}_1^{-1}$. Exact upper and lower bounds for the eigenvalues of K can be found in [38]. The numerical results in this article show that a value of $\beta \in [0.68, 0.76]$ yields the best results. This matches the observations of Section 1.2.

Remark 3.9. As mentioned in Section 2.1, the discrete ordinate method can be interpreted as a discretization of the integral equation. If $\frac{k}{a}$ is small, ξ_k in Equation (3.58) varies only little as μ integrates from zero to one. Hence, a quadrature rule with a small number of quadrature points can be used to approximate the integral in (3.60) in this case. Approximating the integral by a quadrature rule, is exactly what is done when the DOM is used.
If a is large, the condition, $\frac{k}{a}$ is small, is satisfied for relatively large k, i.e. only a few discrete ordinates for the DOM are needed in this case. When the medium becomes optically thin more and more directions have to be used to yield accurate results. Furthermore, the eigenvalues corresponding to a small k are better approximated than the ones corresponding to a large k. The numerical results in Section 3.4 confirm these contemplations.

As shown above, the low frequency part of the spectrum is the important part when preconditioning is considered. On the other hand, the fact that this part is well approximated by $\hat{T}_1^{-1}$, does not automatically say that $\hat{T}_1^{-1}$ is a good preconditioner since it could happen that the small eigenvalues of the high frequency part are redistributed in an unfavorable way by the preconditioning. In [38] it is shown that this is not the case if the two parts of the spectrum are orthogonal. Since K and $\hat{T}_1^{-1}$ are self-adjoint, this condition is fulfilled.

Since Equation (3.48) is of the second kind, the following operator $P = (Id - \omega\hat{T}_1^{-1})^{-1}$ yields a good preconditioner. This expression can be simplified as follows (see Equation (1.63) and (1.65) for the definition of T_1 and $\hat{T}_1$)

$$P = (Id - \omega\hat{T}_1^{-1})^{-1}\hat{T}_1^{-1}\hat{T}_1 = (\hat{T}_1 - \omega Id)^{-1}(\hat{T}_1 - \omega Id + \omega Id) = Id + \omega T_1^{-1}. \tag{3.63}$$

Remark 3.10. In [2] an alternative way for the construction of a preconditioner in context of the DOM is described. Remember the definition of the operators in case of isotropic scattering in a $1D$ slab geometry using optical coordinates

$$A = I - \omega K, \quad K = LH^{-1}L^{+}, \quad T = H - \omega L^{+}L.$$

By simple algebraic transformations, using the fact that $LL^{+} = Id$, it follows

$$A^{-1} = Id + \omega LT^{-1}L^{+}. \tag{3.64}$$

Please note: since the inverse of the operator T appears in this formula, it is only of interest when a method for discretizing the RTE is interpreted as a method for discretizing the integral equation, but not for methods which discretize the integral equation directly.

As shown in Section 1.2, the operator T of the RTE can be approximated by the diffusion operator T_1 if the parameters of the problem are in the asymptotic diffusion limit. Since T_1 is defined for functions independent of the direction variable Ω, suitable prolongations and restrictions have to be used: $T^{-1} \approx L^{+}T_1^{-1}L$. Using this approximation in Equation (3.64) yields

$$\tilde{A}^{-1} = Id + \omega LL^{+}T_1^{-1}LL^{+} = Id + \omega T_1^{-1}. \tag{3.65}$$

This is exactly the same operator as in (3.63). That is, the two methods for constructing the preconditioner are equivalent.

Up to now, only the continuous operators have been considered. The following investigations show, what these considerations imply for the discretized operators. Consider the continuous eigenvalue problem

$$Ae = \lambda e, \tag{3.66}$$

where A is a self-adjoint operator. The discretization by a Petrov-Galerkin method yields the problem: find $e_h \in V_h$ such that

$$a(e_h, v) = \lambda_h \langle e_h, v \rangle \quad \forall v \in W_h. \tag{3.67}$$

Note: the collocation method can be interpreted as a Petrov-Galerkin method, where the test functions are chosen to be shifted Dirac delta functions $\delta(\cdot - x_i)$, where x_i is a collocation point.

Choosing a basis of V_h leads to a generalized algebraic eigenvalue problem for the coefficient vector $\mathbf{e}_h$ of e_h

$$\mathbf{A}\mathbf{e}_h = \lambda \mathbf{M}\mathbf{e}_h, \tag{3.68}$$

where $\mathbf{A}$ is the matrix corresponding to the discrete operator A_h and $\mathbf{M}$ is the mass matrix. In case of the collocation method, the mass matrix is the identity matrix and the eigenvalue problem is a standard eigenvalue problem.

The following result can be found in [23]: let λ_0 be an eigenvalue of (3.66) with eigenvector e. Then there exists a discrete eigenvalue λ_h with eigenvector e_h such that

$$|\lambda_0 - \lambda_h| \leq C \operatorname{dist}(e, V_h)^2, \tag{3.69}$$

$$\|e - e_h\|_V \leq C \operatorname{dist}(e, V_h). \tag{3.70}$$

Hence, if two continuous operators R and Q with similar eigenvalues and corresponding eigenfunctions are discretized with two methods using the same ansatz space and basis functions, then the generalized algebraic eigenvalues and corresponding eigenvectors in $\mathbb{R}^n$ are also similar.

The convergence rate of the Krylov methods does not depend on the generalized eigenvalues but on the standard eigenvalues. This has to be regarded when constructing a preconditioner for the discrete system. To simplify the considerations, we make the following assumptions:

- The function e is an eigenfunction of the two continuous operators R and Q, i.e. $Re = \lambda e$ and $Qe = \mu e$, where $\mu = \lambda$ or $\mu = \frac{1}{\lambda}$.
- The eigenfunction e lies in the finite dimensional ansatz space V_h. The corresponding coefficient vector in $\mathbb{R}^n$ is denoted by $\mathbf{e}$.

These assumptions imply that

$$\mathbf{R}\mathbf{e} = \lambda \mathbf{M}_1 \mathbf{e}, \qquad \mathbf{Q}\mathbf{e} = \mu \mathbf{M}_2 \mathbf{e},$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are different if different test functions are used to discretize the two operators. A good preconditioner $\mathbf{P}$ should yield a system matrix which has the eigenvalue one with eigenvector $\mathbf{e}$. The assumptions above are in general only fulfilled approximately (see the considerations about K and $\hat{T}_1$). In this case the eigenvalues are not exactly one, but lie in a small neighborhood of the point one.

In case $\mu = \lambda$, the operator Q resembles the spectrum of R. For $\mu = \frac{1}{\lambda}$, the operator Q^{-1} resembles the spectrum of R. In either case the operator Q and not Q^{-1} is discretized to yield the matrix $\mathbf{Q}$. This becomes very important when considering the spectra of the matrices. The following cases have to be distinguished:

1. R and Q^{-1} have a similar spectrum and the equation is of the first kind:
   (This situation occurs in the BEM, where R is the hyper-singular operator and Q is the single layer operator (refer to [55]).)
   Using the preconditioner
   $$\mathbf{P} = \mathbf{M}_2^{-1}\mathbf{Q}\mathbf{M}_1^{-1},$$
   yields $\mathbf{PRe} = \mathbf{M}_2^{-1}\mathbf{Q}(\lambda\mathbf{e}) = \mathbf{e}$. Note: the inversion of the two diagonal dominant mass matrices $\mathbf{M}_i$ can be approximated by a few steps of a Jacobi iteration. This is much cheaper than applying the dense matrices $\mathbf{R}$ or $\mathbf{Q}$.

2. R and Q have a similar spectrum and the equation is of the first kind:
   This corresponds to the situation in Remark 3.10, R is the operator A and Q is the approximation $\tilde{A}$ using the diffusion operator T_1 instead of the radiative transfer operator T.
   In this case
   $$\mathbf{P} = \mathbf{Q}^{-1}\mathbf{M}_2\mathbf{M}_1^{-1}$$
   is an optimal preconditioner. Using the same discretization method yields $\mathbf{M}_2\mathbf{M}_1^{-1} = \mathbf{I}$, i.e. no inversion of a mass matrix is needed. Since the diffusion equation is an elliptic equation and the RTE is a hyperbolic equation, it is complicated to find a stable discretization, which works for both. A special transport like discretization of the diffusion equation is derived in the two articles [2] and [11].

3. R and Q^{-1} have a similar spectrum and the equation is of the second kind:
   This corresponds to the situation considered in this section. R is the integral operator K and Q is the diffusion operator $\hat{T}_1$.
   In this case
   $$\mathbf{P} = (\mathbf{I} - (\mathbf{M}_2^{-1}\mathbf{Q})^{-1})^{-1}\mathbf{M}_1^{-1}$$
   is an optimal preconditioner. This expression can be simplified to avoid the nested inversion of two matrices:
   $$\mathbf{P} = (\mathbf{Q} - \mathbf{M}_2)^{-1}\mathbf{Q}\mathbf{M}_1^{-1}. \tag{3.71}$$
   That is, the mass matrix $\mathbf{M}_1$ has to be inverted when the Galerkin method is used to discretize the operator R. When the collocation method is used, it holds $\mathbf{M}_1 = \mathbf{I}$ and no inversion of a mass matrix is needed. The same statement holds when the DOM is interpreted as a discretization of the integral equation since then the operators H_{Ω_j} are inverted explicitly yielding the identity matrix (for more details see (3.82) below). In this case the preconditioner can be simplified further to yield:
   $$\mathbf{P} = \mathbf{I} + (\mathbf{Q} - \mathbf{M}_2)^{-1}\mathbf{M}_2. \tag{3.72}$$
   Thus, using the collocation method to discretize the integral equation, is advantageous when preconditioning is concerned.

Remark 3.11. When the DOM is used to discretize the integral equation, there are two possibilities to construct a preconditioner for the asymptotic diffusion limit. Regarding to Remark 3.10, these two preconditioners are equivalent for the continuous operators. On the other hand, in the case of the discrete operators one has to use the same discretization methods for the diffusion equation and the transport equation when the preconditioner in the second item of the enumeration above is applied, while an arbitrary discretization of the diffusion equation can be used when the preconditioner in the third item is applied. The numerical results below confirm this statement.

As a first example we consider a $1D$ slab geometry of optical thickness $\tau = 100$, with scattering albedo $\omega = 1$ and with black boundaries. For the discretization piecewise linear ansatz functions on a uniform mesh with $n = 100$ elements are used. The same ansatz functions are used for a standard Galerkin discretization of the diffusion operator. Figure 3.3 shows the generalized eigenvalues of the system matrix and the eigenvalues after DSA-preconditioning.

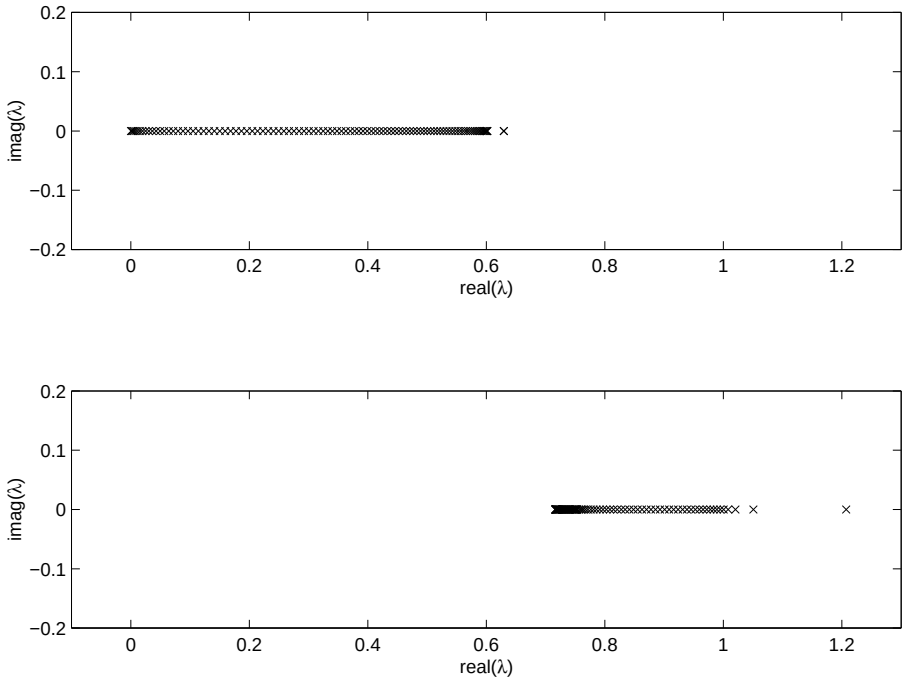

Figure 3.3: Spectrum of the system matrix before (top) and after DSA-preconditioning (bottom) in case of black boundaries

The smallest eigenvalue of the system matrix without preconditioning is very close to zero, i.e. the largest eigenvalue of the integral operator is close to one (numerical value: $\lambda_{\max} = 0.9996$). Furthermore, the convergence of the eigenvalues of the integral operator to zero is rather slow. After DSA-preconditioning the eigenvalues are well separated from zero. In fact approximately 70% percent of the eigenvalues lie in the interval $[0.7, 0.8]$. The condition numbers are $\kappa(\mathbf{A}) = 1630$ and $\kappa(\mathbf{P}^{-1}\mathbf{A}) = 1.7$. That is the preconditioning works very well.

Besides the Galerkin method, two other methods are used for the discretization: the collocation method and the DOM based on the even parity formulation described in the next section. All three methods use the same spatial ansatz functions. Since the collocation method as well as the DOM do not lead to symmetric system matrices, a Krylov method for non-symmetric matrices, e.g. GMRES, instead of CG has to be used to solve the resulting linear system. GMRES is applied with and without DSA-preconditioning, where the DSA-preconditioning is implemented as described above: that is (3.71) is used in case of the Galerkin method and (3.72) in case of the collocation method and the DOM. Table 3.2 shows the number of iterations needed by the GMRES-method to obtain an accuracy of 10^{-7}.

We observe that the DSA-preconditioning works well in either of the three cases. That is, there is no need to use a discretization of the diffusion operator which is adapted to the discretization of the integral operator, except that the ansatz functions have to be the same. On the other hand, the inversion of the mass matrix in case of the Galerkin method is essential. Numerical results show that, when (3.72) is applied, the preconditioning fails

	without preconditioning	with preconditioning
Galerkin	25	7
collocation	77	8
DOM	54	6

Table 3.2: Number of GMRES-iteration steps with and without preconditioning

completely, especially on non-uniform grids.

Further numerical experiments show that the number of iterations needed to obtain a certain accuracy is nearly independent of the grid size used. This is opposed to a matrix resulting from the discretization of a partial differential equation where the condition number grows when a smaller grid size is used.

The numerical results for the $3D$ case with black boundaries look almost the same as in the $1D$ case. Hence, we consider the case of reflecting boundaries $\rho > 0$. In this case the complete system of integral equations has to be considered. That is, the system matrix is given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{M}_1 & \\ & \mathbf{M}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}.$$

As shown above, in the optically thick and strongly scattering case, i.e. $\tau \gg 1$ and $\omega \approx 1$, the matrix $\mathbf{K}_{11}$ is responsible for the bad condition number of the matrix $\mathbf{A}$. Hence, it seems to be reasonable to apply the DSA-preconditioning only to the upper part of $\mathbf{A}$, i.e. to use the matrix

$$\hat{\mathbf{P}} = \begin{bmatrix} \mathbf{P} & \\ & \mathbf{I} \end{bmatrix}$$

as a preconditioner. Let n_1 and n_2 denote the number of degrees of freedom in the volume and on the boundary respectively, and $n = n_1 + n_2$.

If a matrix $\mathbf{A}_{11}$ of dimension n_1 is extended to a matrix $\mathbf{A} = \left(\begin{smallmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{smallmatrix}\right)$ of dimension $n = n_1 + n_2$ all its eigenvalues may be changed. On the other hand, a simple consideration yields that, if $\mathbf{A}_{11}$ has an eigenspace of dimension $\tilde{n}_1 > n_2$, the eigenspace of the matrix $\mathbf{A}$ corresponding to the same eigenvalue has the dimension $\tilde{n}_1 - n_2$. That is, preconditioning of only a part of the matrix might yield good results.

After the preconditioning above the system matrix is given by

$$\hat{\mathbf{P}}\mathbf{A} = \begin{bmatrix} \mathbf{P}(\mathbf{M}_1 - \mathbf{K}_{11}) & -\mathbf{P}\mathbf{K}_{12} \\ -\mathbf{K}_{21} & \mathbf{M}_2 - \mathbf{K}_{22} \end{bmatrix},$$

where all the eigenvalues of $\mathbf{P}(\mathbf{M}_1 - \mathbf{K}_{11})$ are close to one. Hence, we can hope that $n_1 - n_2$ eigenvalues of $\hat{\mathbf{P}}\mathbf{A}$ are close to one. To investigate this numerically we choose the optical parameters $\rho = 0.5$, $\gamma = 100.0$ and $\omega = 1.0$. The computational domain is again the unit cube and the grid used is the same as in the optically thin and strong reflecting case considered above. Figure 3.4 shows the eigenvalues of the system matrix before and after DSA-preconditioning.

The spectrum of $\mathbf{A}$ consists of two clusters: the interval $[0, 0.12]$ contains $n_1 = 729$ eigenvalues. They correspond to the volume part of the equation. The interval $[1, 1.05]$ contains $n_2 = 386$ eigenvalues. They correspond to the surface part of the equation. CG can exploit the fact that the eigenvalues are clustered, but for the first cluster it holds $\frac{\lambda^{(1)}_{\max}}{\lambda^{(1)}_{\min}} = 127$ (This value is called "local condition number" in the following.), which is already very large.

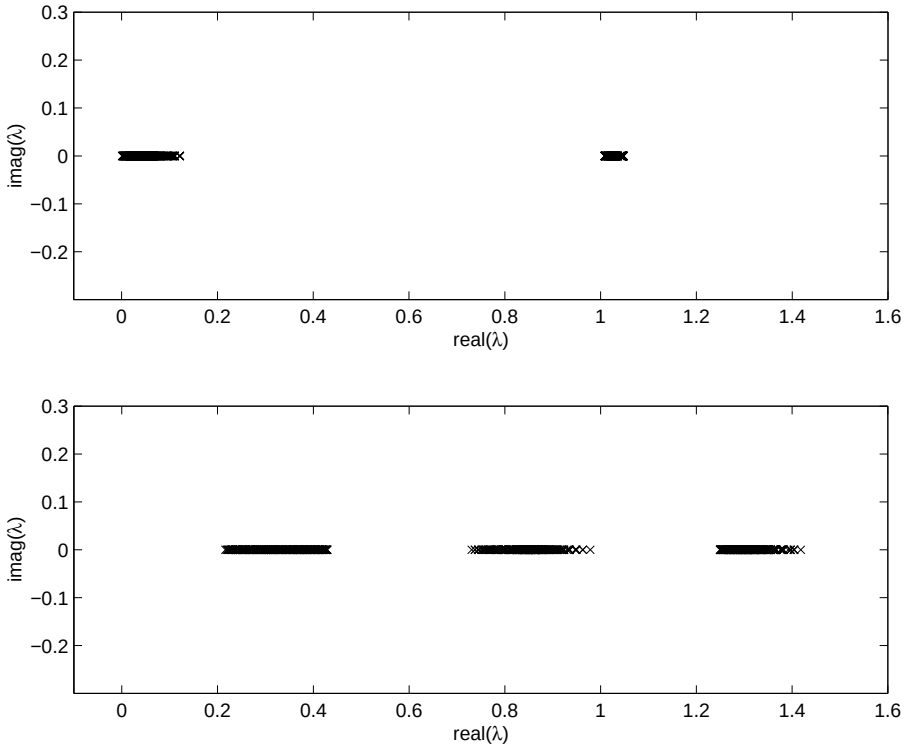

Figure 3.4: Spectrum of the system matrix before (top) and after DSA-preconditioning (bottom) in case of reflecting boundaries

The spectrum $\hat{\mathbf{P}}\mathbf{A}$ consists of three clusters: The interval $[0.73, 0.98]$ contains $n_1 - n_2 = 343$ eigenvalues. This interval is approximately the same as in the $1D$ case when preconditioning of the system with black boundaries is considered. The interval $[0.22, 0.43]$ contains $n_2 = 386$ eigenvalues. These are the eigenvalues that should have been moved to a region close to one by the DSA-preconditioning, but have been redistributed by the augmentation of the system matrix by the surface part. The local condition $\frac{0.43}{0.22} \approx 1.95$ is slightly larger than that of the cluster near one, but it is still small. This is maybe due to the fact that the eigenvalues of the original matrix $\mathbf{A}$ corresponding to the surface part are strongly clustered near one. The interval $[1.25, 1.42]$ contains $n_2 = 386$ eigenvalues. They correspond to the surface part of the equation. That is, these eigenvalues are shifted and slightly spread out by the preconditioning, but the local condition number of this cluster is still small. That is, the preconditioning yields better results as one might expect from the theoretical considerations above. Furthermore, the situation should become even better if a larger system is considered. Since $n_1 = O(h^{-3})$ whereas $n_2 = O(h^{-2})$, the influence of the boundary part becomes smaller for a smaller grid size h. Further numerical experiments show that the situation is very similar for other values of ρ, but for larger ρ the first cluster gets closer to zero.

3.4 Comparison With DOM: Accuracy

In Remark 2.2 it is explained that the DOM can be interpreted as a discretization of the integral equation. In this section this discretization is compared with the Galerkin discretization described in the previous sections. For simplicity we restrict ourselves to the case of a $1D$ slab geometry with isotropic scattering and homogeneous Dirichlet boundaries $I_b \equiv 0$. In this

case operator K_m of Remark 2.2 is an integral operator, and its kernel can be specified explicitly. Suppose that the chosen quadrature rule for approximating $\int_{-1}^{1} f(\mu)d\mu$ is symmetric, i.e. $m = 2M$ and

$$0 < \mu_1 < \cdots < \mu_M \leq 1, \qquad \mu_{-i} = -\mu_i, \qquad w_{-i} = w_i \geq 0, \quad \text{for } i = 1, \ldots, M.$$

Proceeding as in the derivation of the integral equation in the $1D$ case, using this quadrature rule instead of the integration over $[-1, 1]$ shows that the operator K_m is the integral operator given by

$$(K_m G_m)(x) = \frac{1}{2} \int_{-a}^{a} k_m(|x - y|) G_m(y) dy, \tag{3.73}$$

where

$$k_m(x) = \sum_{i=1}^{M} w_i \frac{1}{\mu_i} e^{-\frac{x}{\mu_i}}, \tag{3.74}$$

i.e. the integral in the definition of the exponential integral (see (2.15)) is replaced by an approximation using the quadrature rule above. Note that the kernel k_m does not have a singularity for $x = 0$ in contrast to the kernel E_1.

Instead of discretizing the RTE (3.46) directly by the DOM, the RTE is transformed into the so-called even parity formulation. This transformation is possible if the scattering is isotropic. The even and odd part of the intensity are given by

$$I^{\text{even}}(x, \Omega) := \frac{1}{2}\Big(I(x, \Omega) + I(x, -\Omega)\Big), \tag{3.75}$$

$$I^{\text{odd}}(x, \Omega) := \frac{1}{2}\Big(I(x, \Omega) - I(x, -\Omega)\Big). \tag{3.76}$$

Since $I^{\text{even}}(x, -\Omega) = I^{\text{even}}(x, \Omega)$ and $I^{\text{odd}}(x, -\Omega) = -I^{\text{odd}}(x, \Omega)$, it suffices to consider only half of the directions, e.g. $S^2_+ := \{\Omega \in S^2 \,|\, \Omega \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} > 0\}$. Using the splitting above in the RTE, results in two coupled equations

$$\Omega \cdot \nabla_x I^{\text{odd}}(x, \Omega) + \gamma I^{\text{even}}(x, \Omega) = \frac{\sigma}{2\pi} \int_{S^2_+} I^{\text{even}}(x, \Omega') d\Omega' + \kappa B(T(x)), \tag{3.77}$$

$$\Omega \cdot \nabla_x I^{\text{even}}(x, \Omega) + \gamma I^{\text{odd}}(x, \Omega) = 0. \tag{3.78}$$

Solving the second equation for I^{odd} and plugging the result into the first one gives the even parity equation.

$$\underbrace{-\Omega \cdot \nabla_x \Big(\frac{1}{\gamma} \Omega \cdot \nabla_x I^{\text{even}}(x, \Omega)\Big)}_{-\frac{1}{\gamma}\nabla_x \cdot \left(\Omega \otimes \Omega \nabla_x I^{\text{even}}(x, \Omega)\right)} + \gamma I^{\text{even}}(x, \Omega) = \frac{\sigma}{2\pi} \int_{S^2_+} I^{\text{even}}(x, \Omega') d\Omega' + \kappa B(T(x)). \tag{3.79}$$

Since this equation is of second order, it requires boundary conditions everywhere on the boundary. These are obtained from the boundary conditions of the RTE as follows. If $n(x) \cdot \Omega < 0$, the boundary condition for Ω is applied. If $n(x) \cdot \Omega > 0$, the boundary condition

for $-\Omega$ is applied. In case of homogeneous Dirichlet boundary conditions for the RTE, this results in

$$I^{\text{even}}(x,\Omega) - \frac{1}{\gamma}\Omega \cdot I^{\text{even}}(x,\Omega) = 0, \quad \text{for } \Omega \in S^2_+ \text{ with } n(x)\cdot\Omega < 0, \tag{3.80}$$

$$I^{\text{even}}(x,\Omega) + \frac{1}{\gamma}\Omega \cdot I^{\text{even}}(x,\Omega) = 0, \quad \text{for } \Omega \in S^2_+ \text{ with } n(x)\cdot\Omega > 0. \tag{3.81}$$

Applying the DOM to the even parity equation yields a system of second order elliptic equations and, hence, standard finite elements can be used for the space discretization. This is advantageous since we want to compare the results with the Galerkin discretization of the integral equation. Since the space discretization is the same, the error due to the spherical discretization in the DOM can be measured.

In a $1D$ setting the discretization of an elliptic operator with piecewise linear elements results in a tri-diagonal matrix, whose inverse is easily computable. Hence, the matrix corresponding to the DOM discretization of the integral operator K_m can be assembled explicitly. Due to the inversion of the elliptic operator the system matrix is given by

$$\mathbf{L}_m\mathbf{H}^{-1}\Big(\mathbf{H} - \omega\mathbf{L}_m^+\mathbf{L}_m\Big)\mathbf{L}_m^+ = \underbrace{\mathbf{L}_m\mathbf{L}_m^+}_{=\mathbf{I}} - \omega\mathbf{L}_m\mathbf{H}^{-1}\mathbf{L}_m^+\underbrace{\mathbf{L}_m\mathbf{L}_m^+}_{=\mathbf{I}} = \mathbf{I} - \omega\underbrace{\mathbf{L}_m\mathbf{H}^{-1}\mathbf{L}_m^+}_{=:\mathbf{K}_m}, \tag{3.82}$$

(Refer to (2.20) for the definition of the discrete operators L_m and H_m, $\mathbf{L}_m$ and $\mathbf{H}$ denote the corresponding matrices.), whereas using the Galerkin discretization for the integral equation gives the following system matrix

$$\mathbf{M} - \omega\mathbf{K}. \tag{3.83}$$

The fact that the identity operator is mapped to the identity matrix $\mathbf{I}$ in the first case and to the mass matrix $\mathbf{M}$ in the second case has to be accounted for when comparing the two matrices $\mathbf{K}_m$ and $\mathbf{K}$. For example by multiplying the matrix $\mathbf{K}$ with $\mathbf{M}^{-1}$. Concerning the eigenvalues, this means that the generalized eigenvalues w.r.t. the mass matrix are computed instead of the standard eigenvalues.

As mentioned in Section 3.2, the assembly of the Galerkin matrix for the integral operator involves the evaluation of integrals which cannot be computed analytically and, hence, quadrature schemes have to be used. In the following a very high quadrature order is used to compute a reference matrix. This matrix is assumed to be close to the best approximation of the integral operator on the considered ansatz space. Then the matrices resulting from the DOM for different m and the integral equation using a low order quadrature formula are compared with this matrix.

There are different possibilities to measure the difference between two matrices, e.g. a matrix norm or a spectral decomposition of the matrices can be used. Due to the considerations in Remark 3.9, it is reasonable to expect that the large eigenvalues and corresponding eigenvectors are approximated more accurately by the DOM than the small ones. To investigate this numerically, a spectral decomposition of the matrices is used for the comparison. Please note that the eigenvalues of $\mathbf{M}^{-1}\mathbf{K}$ are the generalized eigenvalues of $\mathbf{K}$ (see the previous section). As expected from the considerations in the previous section the eigenvectors corresponding to the large eigenvalues are smooth, and they become more and more oscillatory for smaller eigenvalues. In fact the first eigenvector has no zero-crossing, the second-one has exactly one and so on. The large eigenvalue corresponding to the smoother eigenfunctions are simple but the smaller ones have multiplicity two.

Again, due to Remark 3.9, it is expected that the quality of the DOM for the same number of ordinates is better when the medium is optically thick. Hence, the eigenvalues and eigenfunctions of the matrices are compared for different optical thicknesses.

We use the following notation: $\mathbf{K}_{\text{ref}}$ denotes the reference matrix $\mathbf{M}^{-1}\mathbf{K}$. The entries of this matrix are determined by using a tensor-product Gauss quadrature rule with $q = 10$ points in x- and y-direction. $\mathbf{K}_{\text{integral},q}$ denotes the matrix where only q-quadrature points are used. Due to the considerations in the previous section, one should use a smaller number of quadrature points in the farfield than in the nearfield. For simplicity we use a large number in the whole domain. This is feasible since the effort for computing the integrals in the $1D$ case is moderate, even when a large number of quadrature points is used. $\mathbf{K}_{\text{DOM},m}$ denotes the matrix obtained by the DOM discretization with m ordinate directions. The computational domain is denoted by $[-1, 1]$, and γ is chosen as $\gamma = \frac{\tau}{2}$ for given τ.

Remark 3.12. The following numerical computations only consider homogeneous Dirichlet boundaries. Furthermore, the Planck function is replaced by a simple source function B. We only consider two cases: in the first case B is set equal to one in the whole domain, in the second case B is set equal to one in a small part of the domain and zero else. The plots of the numerical solutions show the mean intensity $J = \frac{G}{4\pi}$ instead of the energy G. These facts facilitate the comparison of the results with the results obtained by a group at the Interdisciplinary Center of Scientific Computing (IWR) in Heidelberg.

Optical Thickness $\tau = 100.0$

We want to measure the angular discretization error. Hence, we have to ensure that the spatial discretization error is small. For an optical thickness of $\tau = 100.0$, i.e. an optically thick medium, the energy distribution may have strong gradients near the boundary, especially if the scattering albedo ω is small. To resolve this strong gradients, we choose a graded mesh with $n = 2k + 1 = 101$ grid points

$$\xi_i = (\frac{i-1}{k})^2 - 1, \quad i = 1, \dots k+1, \qquad \xi_{k+1+i} = -\xi_{k+1-i}, \quad i = 1, \dots k.$$

It turns out that the eigenvectors of $\mathbf{K}_{\text{ref}}$, $\mathbf{K}_{\text{integral},q}$ and $\mathbf{K}_{\text{DOM},m}$ for $\tau = 100$ are very similar. Hence, solely the eigenvalues can be used for a comparison. Figure 3.5 shows a plot of the eigenvalues of $\mathbf{K}_{\text{ref}}$, $\mathbf{K}_{\text{integral},3}$, $\mathbf{K}_{\text{DOM},4}$ and $\mathbf{K}_{\text{DOM},12}$. Please remember that all eigenvalues of the matrices are real. Not the whole spectrum is shown but only two cutouts: one shows the 10 smallest eigenvalues and one the 10 largest. The top line shows the spectrum of $\mathbf{K}_{\text{ref}}$, the second that of $\mathbf{K}_{\text{integral},3}$, the third that of $\mathbf{K}_{\text{DOM},12}$ and the fourth that of $\mathbf{K}_{\text{DOM},4}$.

We make the following observations:

- The eigenvalues of $\mathbf{K}_{\text{ref}}$ lie in the interval $[\lambda_{\min}, \lambda_{\max}] = [0.0152, 0.9996]$. $\lambda_{\max}$ is in good agreement with the estimate $\|K\|_{L^2} \leq 0.9997$ in (2.68). Since K is a compact operator, its eigenvalues converge to zero, but this convergence is rather slow in the optically thick case.
- The difference between $m = 4$ and $m = 12$ is very small. That is a very small number of ordinate directions is sufficient in the optically thick case. This matches the result of the considerations in Remark 3.9.
- The error in the DOM approximation for the large eigenvalues is very small, about 10^{-4}. It becomes larger for the small eigenvalues. In fact the small eigenvalues are completely wrong. This is again in good agreement with Remark 3.9.

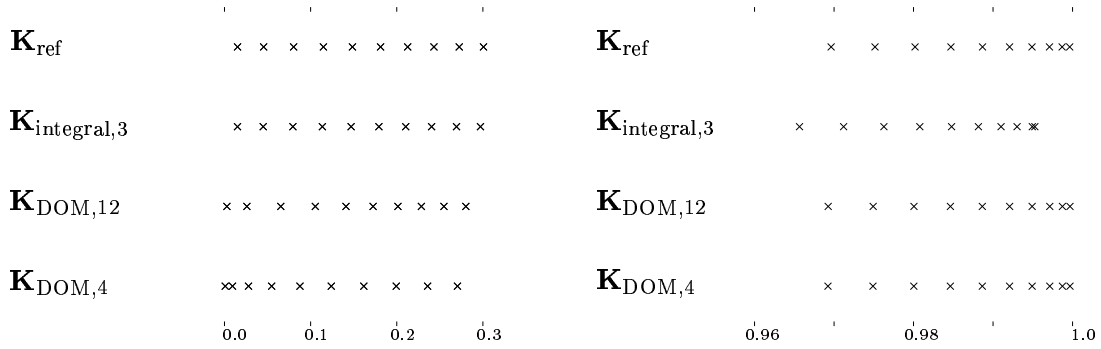

Figure 3.5: Comparison of the spectrum of the different matrices, the left figure shows the 10 smallest eigenvalues and the right one the 10 largest

- The error for $\mathbf{K}_{\text{integral},3}$ is almost equally distributed over the whole spectrum, except for the smallest eigenvalues where the error is much smaller. When comparing the error with that of the DOM, it is observed that the error of $\mathbf{K}_{\text{DOM},12}$ compared with that of $\mathbf{K}_{\text{integral},3}$ is much smaller for the large eigenvalues and becomes larger for the smaller eigenvalues. This gives rise to the fact that the error $\|\mathbf{K}_{\text{DOM},12} - \mathbf{K}_{\text{ref}}\|_p$ is larger than $\|\mathbf{K}_{\text{integral},3} - \mathbf{K}_{\text{ref}}\|_p$ for any matrix norm $\|\cdot\|_p$, $p = 1, 2, \infty$ and $\|\cdot\|_F$ (Frobenius norm).

To assess the impact of the errors in the spectrum on the solution of the linear system, two facts have to be regarded:

1. The eigenfunctions corresponding to the large eigenvalues are smoother than those corresponding to the small eigenvalues. A decomposition of the right hand side into the eigenfunctions has in general larger coefficients for the smooth eigenfunctions. Thus, errors in the eigenvalues, corresponding to these eigenfunctions, have a stronger impact on the solution than errors in the eigenvalues, corresponding to the highly oscillating eigenfunctions.

2. When the system is strongly scattering, i.e. scattering albedo $\omega \approx 1$, the system matrix is given by $\mathbf{A} \approx \mathbf{M}(\mathbf{I} - \mathbf{M}^{-1}\mathbf{K})$. Thus, the eigenvalues of $\mathbf{M}^{-1}\mathbf{K}$ which are close to one become very small eigenvalues of $\mathbf{A}$. Hence, the same absolute error becomes a large relative error. In other words, the system matrix is ill-conditioned. This is important when the eigenfunctions corresponding to the largest eigenvalues of $\mathbf{M}^{-1}\mathbf{K}$ represent a big portion of the solution of the linear system.

To investigate the considerations above numerically, the solution of the linear system for a constant source $B \equiv 1.0$ and two different values of ω is computed. Figures 3.6 and 3.7 compare the numerical solutions obtained by the different methods. Please remember: q denotes the number of quadrature points used to compute the matrix entries, when the Galerkin method is used to discretize the integral equation; m denotes the number of ordinate directions used in case of the DOM. Since the solution is symmetric, only the solution over half of the domain is plotted.

For $\omega = 0.5$ (see Figure 3.6) the error in the solution is very small, when $\mathbf{K}_{\text{DOM},m}$ ($m = 4, 12$) instead of $\mathbf{K}_{\text{ref}}$ is used. The relative error is smaller than 0.1% over a large portion of the domain. That is, the errors in the small eigenvalues do not have a big influence on the

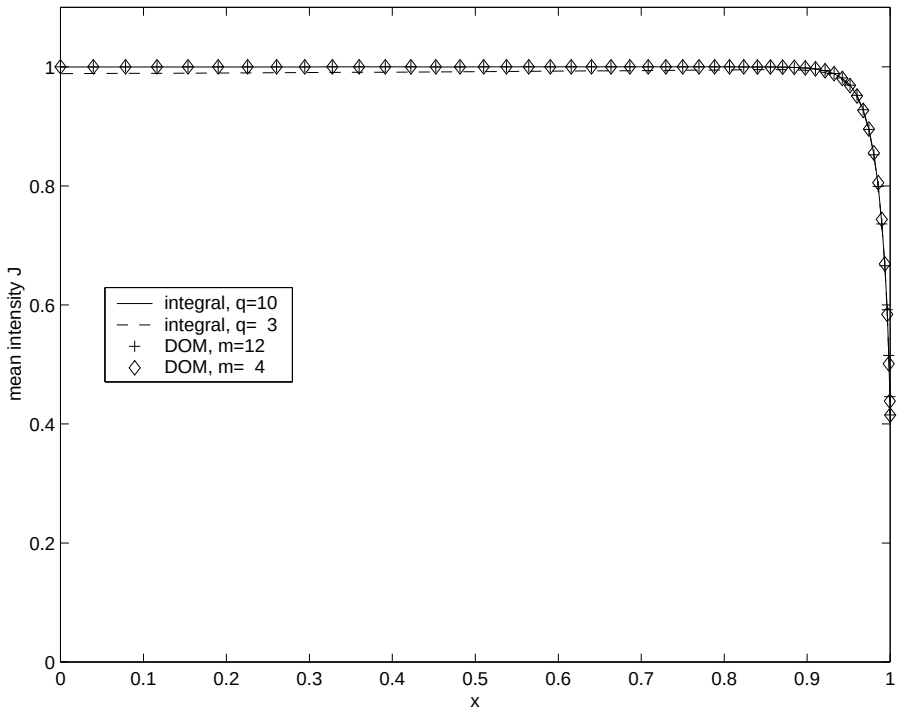

Figure 3.6: Slab Geometry with $\tau = 100.0$, $\omega = 0.5$ and constant source

solution. Only near the boundary, where the solution has a steep gradient, the error becomes a little larger. In this region the error for $m = 12$ is much smaller than that for $m = 4$.

The relative error when $\mathbf{K}_{\text{integral},3}$ is used is approximately 1%. That is, the error is much larger than that of the DOM even though the error in the matrix norm is smaller. This is due to the considerations about the spectral decomposition of the right hand side (see the first item in the enumeration above). Note that, the largest error occurs in a region where the solution is almost constant, whereas the steep gradients near the boundaries are resolved well. Furthermore, the numerical solution has a local minimum in the middle of the slab which is physically impossible. For a constant source and homogeneous Dirichlet boundaries the solution obtains its global maximum in the middle of the slab.

For $\omega = 0.99$ (see Figure 3.7) the solution obtained when $\mathbf{K}_{\text{DOM},m}$ $(m = 4, 12)$ is used is even smaller than in the case $\omega = 0.5$. This is due to the fact that there are no steep gradients at the boundary when the medium is strongly scattering.

The solution obtained when $\mathbf{K}_{\text{integral},3}$ is used is completely wrong. This is due to the considerations about the large relative error in the small eigenvalues of the system matrix (see the second item in the enumeration above).

The large condition number of the system matrix has two effects: first, the convergence of CG is slow and second, small errors in the matrix entries can cause large errors in the solution. The first problem can be solved by using DSA-preconditioning, but this does not remove the second problem.

As shown above, the error in the solution might be smaller for one approximation of the matrix even though the error in any of the widely used matrix norms is larger than for the second approximation. That means that it is not easy, to give a criterion when a matrix is a good approximation of a given matrix. Or in other words, it is hard to estimate the

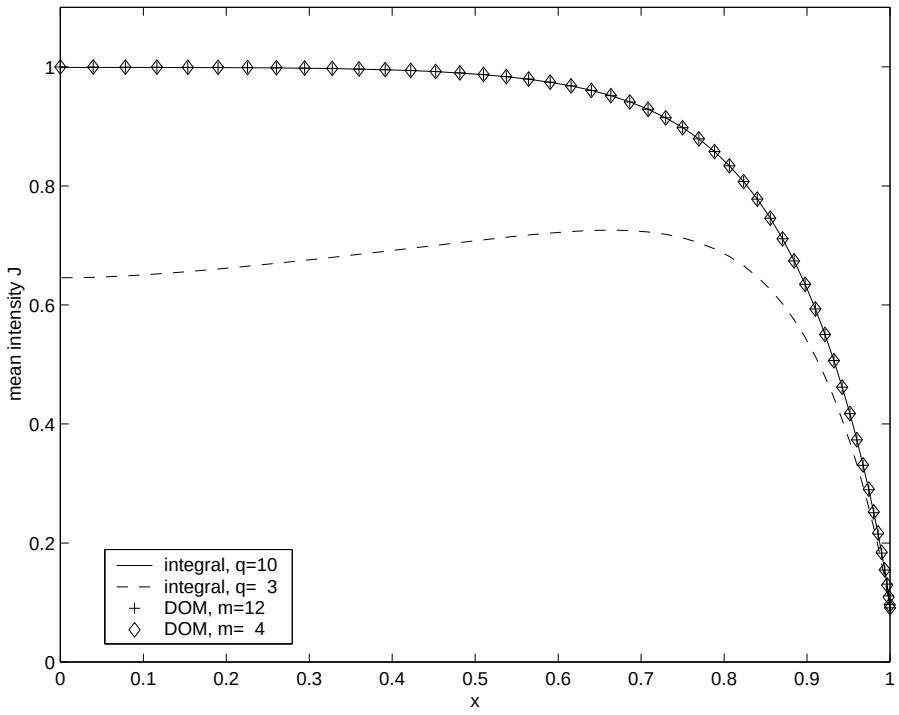

Figure 3.7: Slab Geometry with $\tau = 100.0$, $\omega = 0.99$ and constant source

influence of the error in the matrix entries on the solution of the linear system. This problem also occurs in the next chapter when methods are discussed which use approximations to the matrix which are easier to handle numerically than the matrix itself.

Optical thickness $\tau = 0.1$

For an optical thickness of $\tau = 0.1$, i.e. an optically thin medium, the eigenvectors of $\mathbf{K}_{\mathrm{DOM},12}$ do not resemble the eigenvectors of $\mathbf{K}_{\mathrm{ref}}$ as well as in the optically thick case, but the qualitative behavior is still in order.

Figure 3.8 shows a cutout of the eigenvalues of $\mathbf{K}_{\mathrm{ref}}$ (first row), $\mathbf{K}_{\mathrm{integral},3}$ (second row) and $\mathbf{K}_{\mathrm{DOM},12}$ (third row). The 10 largest eigenvalues are plotted.

The eigenvalues of $\mathbf{K}_{\mathrm{ref}}$ lie in the interval $[1.5 \cdot 10^{-5}, 0.1635]$. That is, for an optically thin medium the largest eigenvalue is much smaller than in the optically thick case and the convergence of the eigenvalues to zero is much faster.

The spectra of $\mathbf{K}_{\mathrm{ref}}$ and $\mathbf{K}_{\mathrm{integral},3}$ are almost identical, but for $\mathbf{K}_{\mathrm{DOM},12}$ already the second largest eigenvalue has an relative error of 5%. The approximation of the other eigenvalues is even less accurate. This effect can be seen when comparing the solutions obtained by the DOM for a different number of ordinate directions and by the Galerkin discretization of the integral equation (see Figure 3.9). The figure shows the numerical solutions obtained for a constant source $B \equiv 1$ and $\omega = 0.99$. But the situation is similar for other values of ω.

The error when $\mathbf{K}_{\mathrm{integral},2}$ is used is very small. This has two reasons: first, as shown in Section 3.2, the quadrature error, when the same number of quadrature points is used, is much smaller than in the optically thick case; second, the condition number of the system matrix is much smaller.

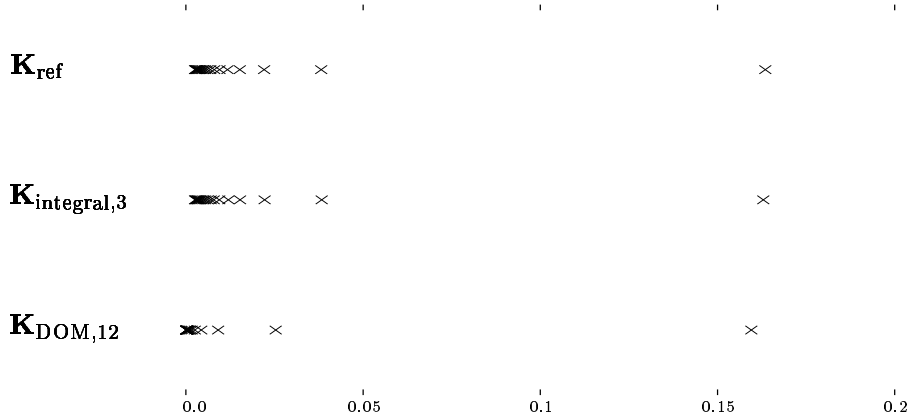

Figure 3.8: Comparison of the 10 largest eigenvalues of the different matrices

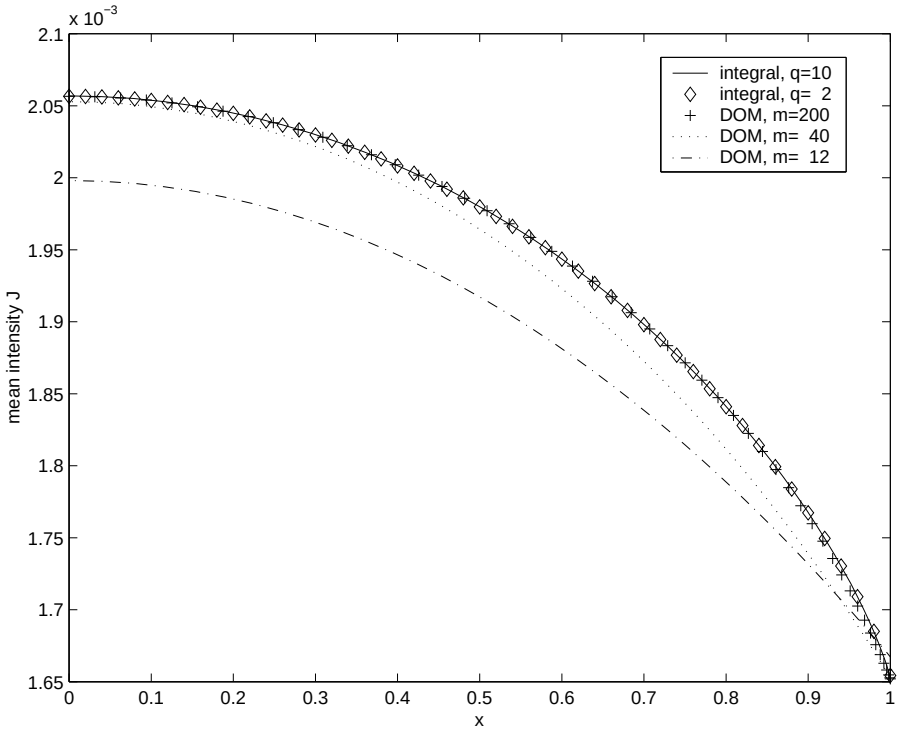

Figure 3.9: Slab Geometry with $\tau = 0.1$, $\omega = 0.99$ and constant source

The solution obtained by the DOM with $m = 12$ ordinate directions is completely wrong. For a higher number of directions the solution converges slowly to the solution obtained by the discretization of the integral equation.

That is, in the optically thin case the DOM needs a large number of ordinates to yield good results. In a $1D$ setting this can be explained by the following considerations. A ray corresponding to the smallest quadrature point μ_i leaves the plain parallel plate after a certain distance. Hence, a circular disc is cut out of the infinitely extended slab. The rest of the domain is not considered. If the medium is optically thin the radius of the disc is small and, hence, the error is high. This explanation only works in the $1D$ setting, hence, it is not clear

if the statement that in an optically thin medium more ordinates are needed, is true also in the $3D$ case. The numerical examples below show that the statement holds also in the $3D$ case, but the effect is not as strong as in the $1D$ case.

The $3D$ case

Before discussing the results obtained for the $3D$ case, we want to say the following. The problem with the ill-conditioned system matrix described above for the $1D$ case also occurs in the $3D$ case. That is, when discretizing the integral equation the matrix entries have to be computed very accurately. The only difference is, that the costs for the high order cubature formulae are much higher than in the $1D$ case. We do not want to investigate this phenomenon further, therefore, the computations in this subsection always handle situations with a scattering albedo $\omega = 0.8$. Furthermore, we again only consider the case of homogeneous Dirichlet boundary conditions and constant coefficients κ and σ.

For the $1D$ setting three different codes are at our disposal: the DOM code based on the even parity formulation, the code which discretizes the integral formulation directly and a third code which combines the DOM with a semi-analytical solution of the transport equation along the rays. This code has been made available to the author by the ITWM. As shown above, all three codes yield the same result when the grid size for the spatial mesh is sufficiently small and when the number of ordinate directions is sufficiently high in case of the DOM and when the matrix entries are evaluated sufficiently accurate in case of discretizing the integral equation.

In the $3D$ case the situation is different. Since the CPU time and storage requirements grow very fast with decreasing grid size, it is often hard to guarantee that the resolution of the spatial and angular discretization is high enough. The solution methods based on the integral formulation have to make use of the matrix compression methods discussed in the next chapter since it is impossible to store the whole system matrix for a sufficiently fine grid. These methods introduce an additional error. Furthermore, we do not have a DOM code at our disposal, but only some benchmark solutions for special reference situations which have been made available to the author by Erik Meinköhn from the IWR in Heidelberg. These results are obtained by the discrete ordinate method using the streamline diffusion Galerkin discretization described in Section 1.2.

In order to be able to calculate a reference solution, we consider a quasi-$1D$ geometry as a first test. In this case one of the $1D$ codes with a very high resolution can be used to obtain a reference solution. The test case is according to Test II in [46]. It considers a plane-parallel slab, which can be handled by the $1D$ code. For the $3D$ calculations the infinitely extended layer is approximated by a parallelepiped with a very high aspect ratio $1 : 100 : 100$. The spatial grid consists of 32^3 hexahedral elements with the same aspect ratio as the computational domain. Piecewise trilinear elements are used as ansatz functions.

When discretizing the integral equation directly, the anisotropy of the cells has a bad influence on the quadrature error for the methods used to approximate the matrix entries. Therefore, we have to choose a lower aspect ratio for the computational domain or we have to use a larger number of elements in the directions with the large extension. Using a smaller aspect ratio is in order, when the optical thickness is large, say $\tau \geq 10$. In this case the damping of the radiation very high, therefore, an aspect ratio of $1 : 10 : 10$ is sufficient. When the medium is optically thin, e.g. $\tau = 0.1$, the aspect ratio $1 : 10 : 10$ is not adequate to approximate the infinitely extended layer. Therefore, we use an aspect ratio of $1 : 100 : 100$ and a larger number of elements in the directions with the large extension.

The first test case considers an optical thickness of $\tau = 20$ and a constant source. The discretization of the integral equation uses 32^3 hexahedral elements on a computational domain with aspect ratio $1 : 10 : 10$. The ansatz functions are the same as for the DOM i.e. piecewise trilinear. The DOM results are obtained on a computational domain with aspect ratio $1 : 100 : 100$ using 32^3 and 64^3 elements, i.e. $n = 33^3$ and $n = 65^3$ vertices respectively. The elements have the same aspect ratio as the computational domain. The number of ordinates used, is $m = 320$ in both cases. To compare the $3D$ results with the $1D$ results the mean intensities $J = \frac{G}{4\pi}$ are plotted along a straight line parallel to the x_1 axes through the middle of the parallelepiped (see Figure 3.10). The solution obtained by the $1D$ code is denoted by exact.

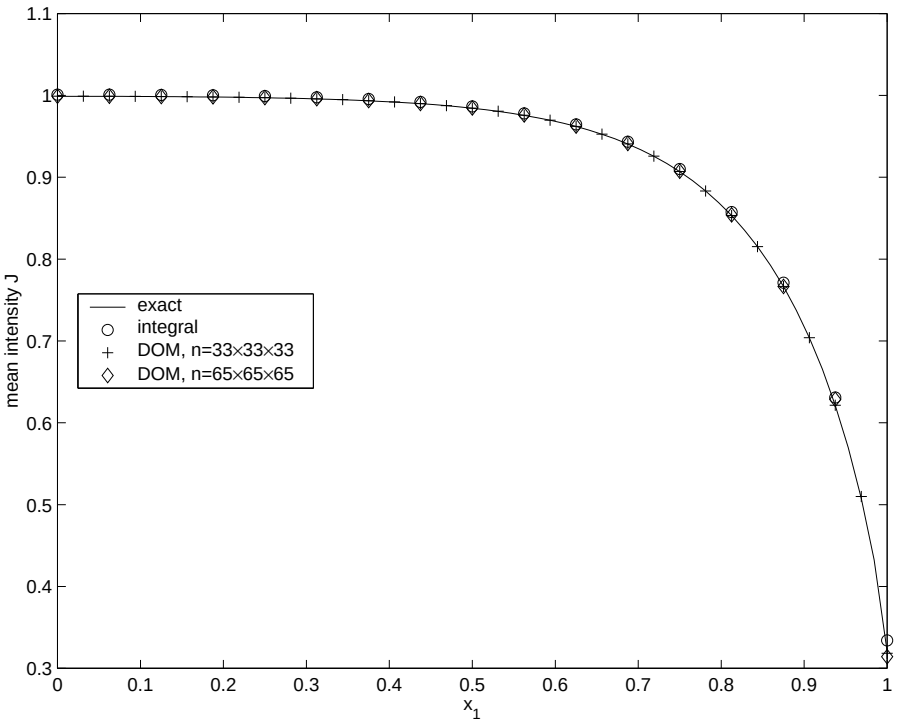

Figure 3.10: Slab geometry with $\tau = 20.0$, $\omega = 0.8$ and constant source

Both methods yield good results, but the result obtained by the DOM is slightly more accurate than that of the discretization of the integral equation. The latter is a little too large. The value in the middle of the domain is 1.001, i.e. greater than the constant source $B \equiv 1$, which is physically impossible. The true solution is 0.999, that is, the relative error is smaller than 0.3%. This error becomes bigger near the boundary: about 1.5% at the boundary point. This is due to the steep gradients at the boundary. A graded mesh should be used to resolve this gradients more accurately. The solution obtained by the DOM is about 0.5% too small at the boundary point.

Figure 3.11 shows the results obtained for the case $\tau = 0.1$. The discretization of the integral equation uses $16 \times 64 \times 64$ hexahedral elements on a computational domain with aspect ratio $1 : 100 : 100$. The DOM uses a spatial grid with 32^3 elements and $m = 320$ ordinate directions.

The integral equation gives rather good results, whereas the DOM fails completely. This is in good agreement with the $1D$ case where a very large number of ordinate directions is needed to obtain good results in the optically thin case.

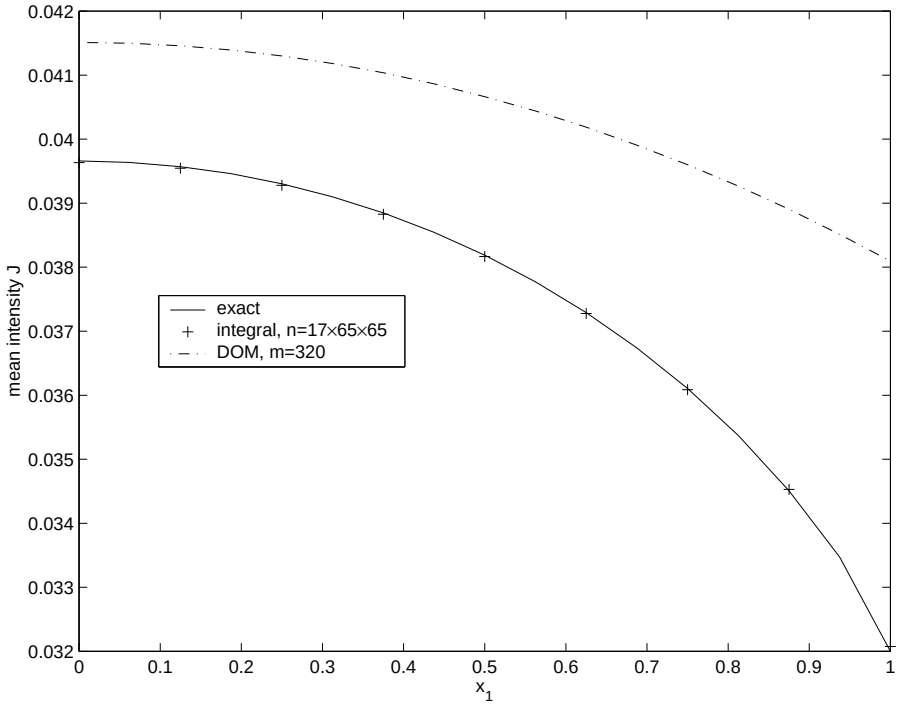

Figure 3.11: Slab geometry with $\tau = 0.1$, $\omega = 0.8$ and constant source

The next test problem should show that the discretization of the integral equation is able to handle situations where the source function is non smooth. To this end we consider the following source:

$$B(x_1, x_2, x_3) = \begin{cases} 1 & \text{if } x_1 \in [-0.25, 0.25], \\ 0 & \text{else.} \end{cases} \tag{3.84}$$

Figure 3.12 shows the results obtained for the case $\tau = 10.0$. The discretization of the integral equation uses the same grid as in the case $\tau = 20.0$. The result is very accurate. Results obtained by the DOM are not available for this case.

This section is concluded with some results for a real $3D$ problem. We consider a $3D$ computational domain $(x_1, x_2, x_3) \in [-1, 1]^3$ with constant source. The optical parameters are chosen as follows: $\omega = 0.8$ and $\gamma = 0.05$, i.e. $\tau = 0.1$ if we define the optical thickness along the coordinate directions. All numerical computations use a uniform spatial grid with 32^3 cells and piecewise trilinear elements.

To allow some quantitative evaluations, the mean intensity $J = \frac{G}{4\pi}$ is plotted along a selected path through the computational domain. The path runs through the midpoint of the domain parallel to the x_1-axis. Figure 3.13 shows the numerical results obtained by the discretization of the integral equation and by the DOM with $m = 80$ directions.

The two solutions show relatively large deviations in the middle of the domain and at the boundary (approximately 2%). In the region in between the solutions are very similar. For this $3D$ problem we do not have a reference solution. Hence, it cannot be said with certainty which solution is more accurate. On the other hand, the following argument indicates that the error in the solution computed by the discretization of the integral equation is smaller.

It is possible to derive an upper bound for the mean intensity $J = \frac{G}{4\pi}$ in terms of the source term B. The integral equation for J in case of homogeneous Dirichlet boundary conditions

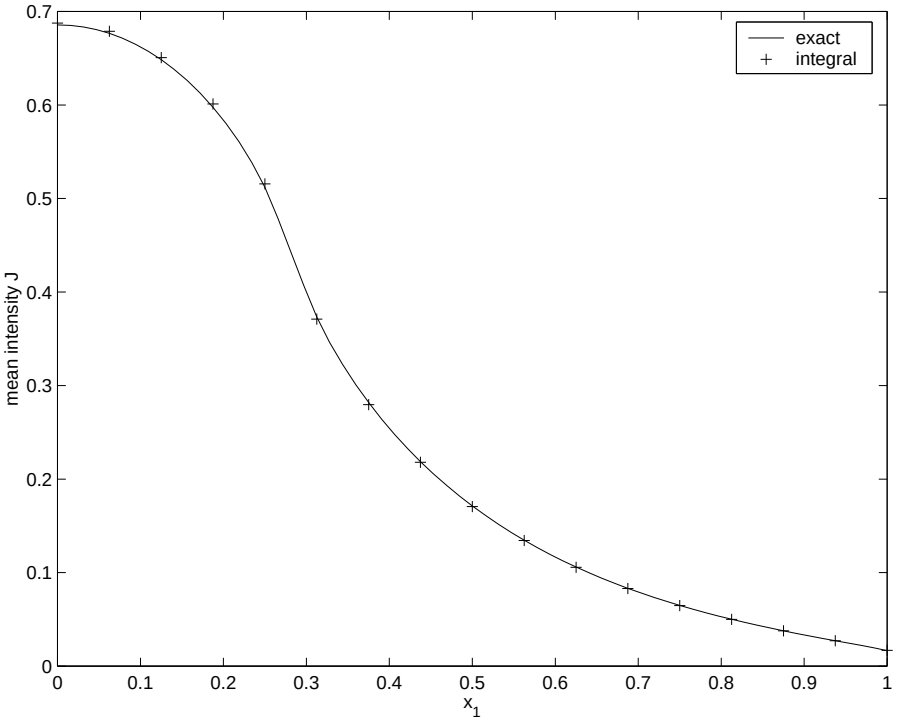

Figure 3.12: Slab geometry with $\tau = 10.0$, $\omega = 0.8$ and local source

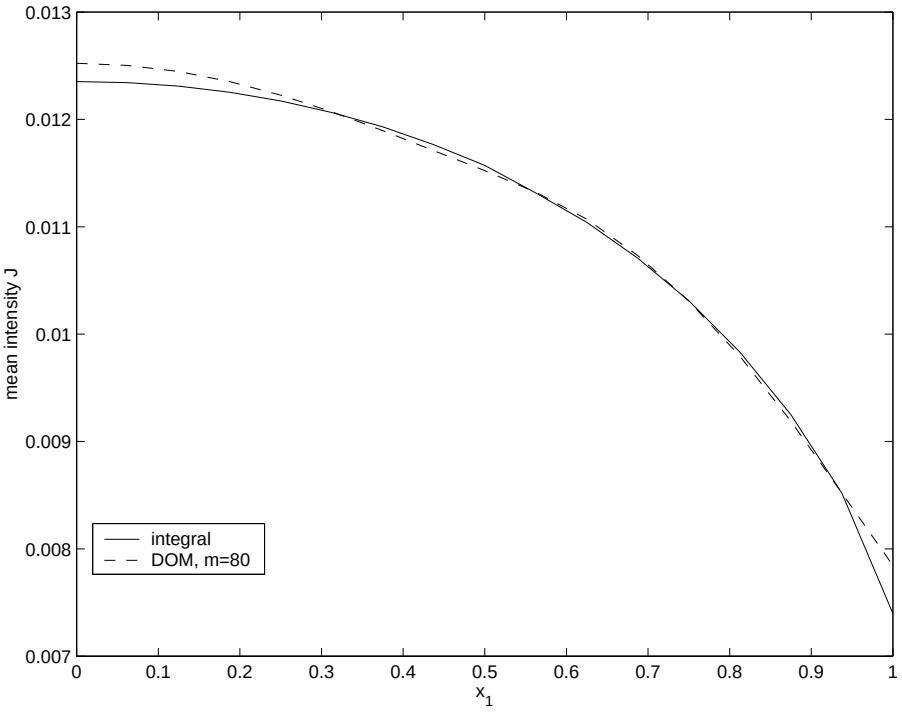

Figure 3.13: Cube with $\tau = 0.1$, $\omega = 0.8$ and constant source

is given by (see 2.7)

$$J(x) - \omega(KJ)(x) = (1-\omega)\underbrace{(KB)(x)}_{=:f(x)}, \tag{3.85}$$

where the integral operator K is defined as

$$(KJ)(x) = \frac{\gamma}{4\pi} \int\limits_D \frac{e^{-\gamma\|x-y)\|}}{\|x-y\|^2} J(y)dy. \tag{3.86}$$

Since the integral operator K is positive, the following lower bound for $\|J\|_\infty$ holds

$$\|J\|_\infty \geq (1-\omega)\|f\|_\infty. \tag{3.87}$$

The operator norm $\|K\|_\infty$ is given by

$$\|K\|_\infty = c := \max_{x\in[-1,1]^3} \int\limits_{[-1,1]^3} \frac{\gamma}{4\pi} \frac{e^{-\gamma\|x-y\|}}{\|x-y\|^2} dy \quad \in [0,1]. \tag{3.88}$$

Using a Neumann series argument, yields an upper bound for $\|J\|_\infty$

$$\|J\|_\infty \leq \frac{1-\omega}{1-\omega c}\|f\|_\infty. \tag{3.89}$$

Using $\|f\|_\infty \leq \|K\|_\infty\|B\|_\infty$ on the right hand side, gives the estimate

$$\|J\|_\infty \leq \frac{(1-\omega)c}{1-\omega c}\|B\|_\infty. \tag{3.90}$$

For a constant source this upper bound might be rather accurate. It remains to compute the value of c. The maximum in (3.88) is obtained in the middle of the domain, that is for $x = 0$. Due to symmetry considerations, the integral over the eight octants is the same. Thus

$$c = 8\frac{\gamma}{4\pi} \int\limits_{[0,1]^3} \frac{e^{-\gamma\|y\|}}{\|y\|^2} dy.$$

The domain $[0,1]^3$ can be split up into three pyramids. Again due to symmetry considerations, the value over each pyramid is the same. Hence, we only have to compute the integral over the pyramid $\{0 < y_1 < 1, 0 < y_2 < y_1, 0 < y_3 < y_1\}$. The integrand is only singular at the origin. Hence, we can apply a Duffy transformation (see Section 3.2) to render the integrand analytic

$$c = 3\frac{2\gamma}{\pi} \int\limits_{[0,1]^3} \frac{e^{-\gamma\omega_1\sqrt{1+\eta_1^2+\eta_2^2}}}{\omega_1^2(1+\eta_1^2+\eta_2^2)} \omega_1^2 d\omega_1 d\eta_1 d\eta_2.$$

This value is computed by using a tensor product Gauss quadrature formula with $q = 10$ points in every direction; this yields: $c = 0.620086$. (Note that the first 7 digits are the same when $q = 5$ is used. Therefore, the value is very accurate.) Using this value in (3.87) and (3.90) regarding the fact that $B \equiv 1$, yields the estimate:

$$0.011842 \leq \|J\|_\infty \leq 0.012432. \tag{3.91}$$

The numerical solution obtained by the discretization of the integral equation lies in this interval: $\|J\|_\infty = 0.0123$, but the DOM solution is too large: $\|J\|_\infty = 0.0125$. Hence, this solution seems to be less accurate. Unfortunately, we do not have results obtained by the

DOM with more than $m = 80$ directions. Maybe the solution lies below the upper bound when more than 80 directions are used.

The setup for the next problem is the same as for the previous one, except that the emission region is restricted to a small cube in the middle of the domain:

$$B(x_1, x_2, x_3) = \begin{cases} 1 & \text{if } (x_1, x_2, x_3) \in [-0.25, 0.25]^3, \\ 0 & \text{else.} \end{cases} \tag{3.92}$$

This source is denoted by "local source" later.

Figure 3.14 shows the numerical results obtained by the discretization of the integral equation and by the DOM with $m = 20, 80, 320$ ordinate directions. The solution for $m = 20$ shows strong deviations from that for $m = 320$ especially at the bottom of the slope (maximal relative error: 4.5%). The difference between $m = 80$ and $m = 320$ is very small except near the boundary (maximal relative error: 1.7%).

The deviation between the DOM solution and the one obtained by the integral formulation is even larger than in the previous example (approximately 7%). The solution obtained by the DOM is always smaller than that of the integral formulation.

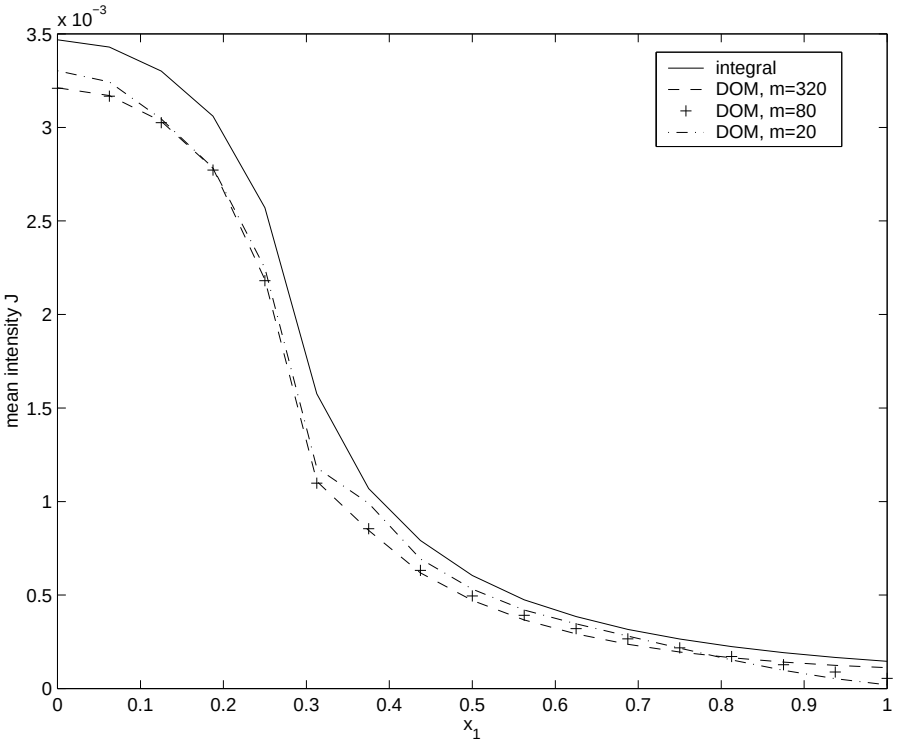

Figure 3.14: Cube with $\tau = 0.1$, $\omega = 0.8$ and local source

Again some estimates for the maximum of the solution can be given. The function $f = KB$ on the right hand side of (3.87) and (3.90) obtains its maximum in the middle of the emission region, i.e. for $x = 0$

$$\|f\|_\infty = f(0) = \frac{\gamma}{4\pi} \int\limits_{[-1,1]^3} \frac{e^{-\gamma\|y\|}}{\|y\|^2} B(y)\,dy. \tag{3.93}$$

Since the source (3.92) is symmetric, the same splittings of the domain as in the previous case and a Duffy transformation can be used to render the integrand analytic. Due to the

simple structure of the source term the integral can be evaluated very accurately by using a high order quadrature scheme. One has to regard the fact that the numerical methods use piecewise trilinear ansatz functions. Therefore, the source is approximated by a piecewise trilinear function $\tilde{B}$, i.e. the jump is approximated by a steep gradient. This modified source $\tilde{B}$ has to be used in Equation (3.93). This complicates the evaluation a little, but it is still possible to compute $f(0)$ very accurately. Using this value in (3.87) and (3.89), yields

$$0.0033958 \leq \|J\|_\infty \leq 0.0035646. \tag{3.94}$$

The numerical solution obtained by the discretization of the integral equation lies in this interval: $\|J\|_\infty = 0.00347$, but the DOM solution (for $m = 20$, 80 and 320) is too small: $\|J\|_\infty = 0.00321$ for $m = 320$. Hence, the solution obtained by the DOM seems to be less accurate again.

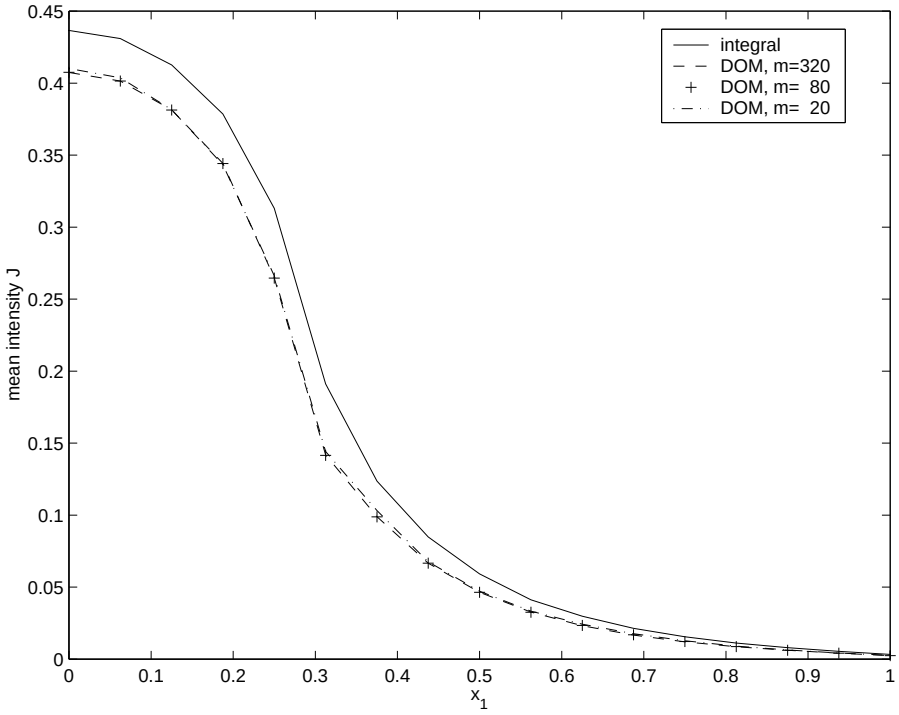

Figure 3.15: Cube with $\tau = 10.0$, $\omega = 0.8$ and local source

Figure 3.15 shows the numerical results obtained for an optical thickness $\tau = 10.0$. The deviations between the solutions obtained for a different number of ordinates is much smaller than in the optically thin case (maximal relative error: 1.0% for $m = 20$ and 0.06% for $m = 80$). This is in good agreement with the considerations of the previous section. That is $m = 80$ directions are sufficient for the given spatial resolution. On the other hand, the difference compared with the solution obtained by the integral formulation is about as big as in the optically thin case (approximately 6%). We do not have an explanation for that. The interval between the lower and upper bound for $\|J\|_\infty$ is much larger than in the optically thin case. Both solutions lie in this interval. Therefore, we cannot assess which solution is more accurate.

To indicate that the solution obtained by the integral equation is reliable, we consider the same test as above, but with $\omega = 0$, i.e. the purely absorbing case. Figure 3.16 shows the numerical result obtained by the discretization of the integral equation.

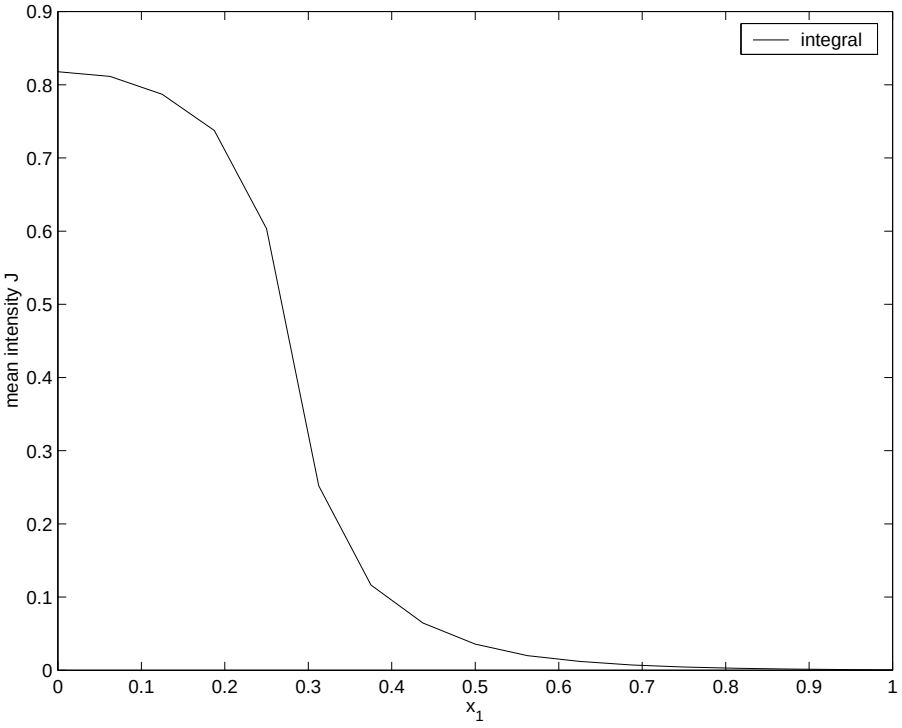

Figure 3.16: Cube with $\tau = 10.0$, $\omega = 0.0$ and local source

In this case the solution of the integral equation reduces to the evaluation of the parameter integral:

$$\frac{\kappa}{4\pi} \int\limits_{[-1,1]^3} \frac{e^{-\gamma\|x-y\|}}{\|x-y\|^2} \tilde{B}(y)dy, \quad x \in [-1,1]^3, \tag{3.95}$$

where $\tilde{B}$ is the piecewise trilinear approximation of B. This integral can be evaluated for given x by using a high order quadrature scheme. Evaluating the integral in the point $x = 0.0$, gives 0.8148, whereas the Galerkin discretization yields 0.8178. That is the relative error in the numerical solution is 0.0036.

Subsuming our observations, we can say the following. In all the considered cases we could not find any indication that the solutions obtained by our implementation of the discretization of the integral formulation are less accurate than that obtained by the implementation of the DOM. On the other hand, we do not have an explanation for the difference compared with the solution of the DOM in case of a local source. Especially in the optically thick case where there is almost no difference between $m = 80$ and $m = 320$ directions.

Remark 3.13. Unfortunately, there is only a small number of results obtained by the DOM at our disposal. It would be nice to have results for the optically thin case using more than 80 directions, or results on a finer spatial grid for the case of a local source. In the case of the integral formulation we cannot use a finer grid, say 64^3 elements, due to the enormous storage requirements. When piecewise constant elements are used instead of piecewise linear elements, the resulting matrix has a Toeplitz-type structure (see Section 4.1) which can be handled with very low storage requirements. Numerical results using 32^3 and 64^3 elements show almost no difference for the test cases with the local source. Hence, 32^3 elements seem to be sufficient in case of the discretization of the integral equation.

Based on the numerical results in this section, we draw the following conclusions.

- The discretization of the integral equation may yield completely wrong results in the case $\tau \gg 1$ and $\omega \approx 1$ when the matrix entries are evaluated not accurately enough. This is due to the high condition number of the system matrix in this case. In all other cases, the solutions obtained by the discretization of the integral formulation are at least as accurate as that obtained by the DOM.
- The DOM needs many directions to obtain accurate results in the optically thin case; $m = 80$ directions are not enough. The situation is even worse if the computational domain is highly anisotropic. In this case the solution is completely wrong even for $m = 320$ directions.
- The implementation of the DOM gives bad results when problems with local sources are considered, even in the optically thick case.

Chapter 4

Matrix Compression Methods

In the previous chapter the Galerkin method is applied to the integral equation as derived from the RTE. Furthermore, the stability and convergence properties of the method are examined. In this chapter we address ourselves to an efficient implementation of the method.

The discretization of an integral equation leads to a full matrix even if basis functions with local support are used. This is due to the integral kernel $k(\cdot,\cdot)$ not being local. Therefore, the size of memory needed to store the entries of the matrix is of order $O(n^2)$ and a multiplication of the matrix with a vector requires $O(n^2)$ operations. If n is large, this effort is not feasible.

When iterative methods are used for the solution of the discrete system, the explicit knowledge of the entries of the matrix is not needed. Instead, matrix-vector multiplications appear as elementary operations. Furthermore, the discrete solution is only an approximation to the solution of the continuous problem. Hence, an additional error introduced by approximating the discrete system need not diminish the overall accuracy if it is smaller than the discretization error.

There are different approaches for reducing the complexity of a matrix-vector multiplication: the first ones exploit the algebraic structure of the matrix, the second ones use an approximation of the operator based on geometrical and analytical considerations.

First, if the domain is very regular and the kernel is translation invariant, i.e. $k(x,y) = k(x-y)$, the resulting matrix has a Toeplitz-type structure. Therefore, a matrix-vector multiplication can be evaluated efficiently by performing fast Fourier transformations (see Section 4.1). This method does not introduce an additional error.

Second, the so-called matrix compression methods exploit another property of the kernel, namely, that it has a singularity for $x = y$ and that it is very smooth outside the diagonal. This corresponds to the physical fact that two points which are near together have a stronger influence on each other than two points further apart. On parts of the domain where the kernel is smooth, i.e. away from the diagonal, an approximation of the kernel can be used. To this end, the elements are gathered together to form clusters. If the distance between two clusters is large a suitable approximation of the kernel is used which makes the corresponding sub-matrix a low-rank matrix. This kind of matrix allows a fast matrix-vector multiplication (see Section 4.1). The exact requirements for the applicability of these methods and the involved procedures are explained in more detail in the following sections.

There are different possibilities to construct the approximation. The first one is the fast multipole method which was originally introduced to enable a fast simultaneous evaluation of the potentials of n point charges. The approximation is obtained by expanding the kernel

into spherical harmonics. The first article on this topic was published by Rokhlin in 1985 [47]. Since then the method has been improved substantially (see e.g. [13]).

The second method is the panel clustering method by Hackbusch and Nowak [28]. In this article Taylor expansion is used to approximate the kernel. The computational complexity of the method is reduced further in [34] and [50]. In more recent papers polynomial interpolation is used instead of Taylor expansion [26], which has the advantage that no explicit knowledge of an expansion of the kernel is required. Hence, the implementation does not depend on the considered kernel.

The third method is the adaptive cross approximation (ACA) [7]. It tries to find a pseudo-skeleton approximation of the matrix (see [21]), i.e. only a few of its columns and rows are used. This method also does not need any explicit knowledge of the kernel. Furthermore, it is applied to approximate the matrix entries directly instead of the kernel.

In a series of papers Hackbusch et. al. look at the problem from a more algebraic point of view. This provides a common basis for the above approaches, and thus facilitates the comparison of the methods. Furthermore, it enables the extension of the ideas to new fields of application e.g. the inversion of sparse matrices [9].

A method which implicitly exploits the smoothness of the kernel outside the diagonal is the matrix compression based on multiple scales. Instead of the usual nodal basis, a wavelet basis is used. The wavelets are orthogonal to low-order polynomials. Hence, the smoothness of the kernel with sufficient distance from the diagonal causes most entries of the matrix with respect to the wavelet basis to be very small. These entries can be neglected without introducing a significant error. Since it can be determined a priori, which entries are negligible, only the remaining entries need to be computed. For more details refer to [14], [53] or [43].

The matrix compression methods have been developed in context of the boundary element method (BEM), where a boundary value problem in a domain is transformed into an equivalent integral equation on its surface. The kernels of these integral equations are very similar to the kernels of the integral equation describing the radiative transfer. Therefore, with some modifications, the matrix compression methods are readily applicable to the latter equation.

After the introduction of the methods in Section 4.2, they are applied to the integral equation (2.26) being derived from the RTE in Section 4.3. It is demonstrated what is different from the integral equations arising from the BEM and how to cope with this. Section 4.4 concludes with a comparison of the computational complexity of the matrix compression algorithms with that of the DOM.

4.1 Preliminary Considerations

In this section it is shown that in some cases a fast matrix-vector multiplication is possible even if the matrix is fully populated. Two simple examples are the Toeplitz-type matrices and the low-rank matrices. As mentioned above low-rank matrices can only be used to approximate sub-matrices of the whole matrix arising from the discretization of integral equations. In order to keep the selection of this sub-blocks algorithmically manageable, a hierarchical structure, called cluster tree, is introduced.

Toeplitz-type matrices

A standard Toeplitz matrix is a matrix $\mathbf{T} = (t_{ij})_{i,j} \in \mathbb{C}^{n\times n}$ with $t_{ij} = t_{j-i}$, i.e. the matrix entries are constant along the diagonals. The following example shows, how such matrices

can arise in radiative transfer.

Example 4.1. Consider the integral equation resulting from the RTE for a $1D$ slab geometry with constant coefficients, isotropic scattering and homogeneous Dirichlet boundaries (see (2.13)). Let the Galerkin method with piecewise constant elements be applied to discretize the equation. Furthermore, let an equidistant grid consisting of n intervals, where $n = 2^p$ is a power of 2, be used. Then the entries of the resulting matrix are of the form

$$k_{ij} = \frac{\gamma}{2} \int_{\frac{i-1}{nl}}^{\frac{i}{n}l} \int_{\frac{j-1}{n}l}^{\frac{j}{n}l} E_1(\gamma|x-y|)dydx = \frac{\gamma}{2} \int_{0}^{\frac{1}{n}} \int_{\frac{j-i}{n}l}^{\frac{j-i+1}{n}l} E_1(\gamma|x-y|)dydx.$$

Hence, the matrix $\mathbf{K}$ is a Toeplitz matrix. Note that this property is lost if we use piecewise linear ansatz functions because the support of the basis functions corresponding to the points on the left and the right boundary is only half as big as the support of the other basis functions.

Besides the low storage requirements, $2n-1$ in the non-symmetric and n in the symmetric case, a Toeplitz matrix allows a fast matrix-vector multiplication. To show this, the definition of circulant matrices is recalled.

Definition 4.1. A circulant matrix is a Toeplitz matrix $\mathbf{C} = (c_{j-i})_{i,j} \in \mathbb{C}^{n\times n}$ with the property

$$c_k = c_{n+k}, \qquad 1-n \leq k < 0,$$

i.e.

$$\mathbf{C} = \begin{bmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & \cdots & c_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ c_1 & \dots & c_{n-1} & c_0 \end{bmatrix}.$$

Circulant matrices have the favorable feature that they are easily diagonalizable: any circulant matrix is associated with the polynomial $p_c(z) := c_0 + c_1 z + \ldots + c_{n-1}z^{n-1}$ $(z \in \mathbb{C})$ and has a diagonal representation in a Fourier basis

$$\mathbf{C} = \mathbf{F}_n^* \mathbf{\Lambda}_c \mathbf{F}_n \quad \text{with } \mathbf{\Lambda}_c = \operatorname{diag}\{p_c(1), \ldots, p_c(\omega^{n-1})\}, \quad \omega = e^{-\frac{2\pi i}{n}}.$$

The eigenvector corresponding to the eigenvalue $p_c(\omega^{j-1})$ is given by the j-th column of $\mathbf{F}_n$

$$\mathbf{v}_j = \frac{1}{\sqrt{n}}[1, \omega^{j-1} \ldots, \omega^{(n-1)(j-1)}]^T,$$

i.e. $\mathbf{v}_j$ is the j-th Fourier vector. For a proof refer to [29].

The eigenvalues can be rewritten in vector notation

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \sqrt{n}\mathbf{F}_n \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Hence, Fast-Fourier-Transform (FFT) can be used for a fast computation of the eigenvalues, $3n\log_2 n$ (real) additions and $2n\log_2 n$ (real) multiplications are needed if $n = 2^p$ is a power of 2. Since the eigenvectors are Fourier vectors, a transformation into diagonal form can also be obtained via FFT. Hence, a fast matrix-vector multiplication $\mathbf{v} = \mathbf{Cu}$ looks as follows:

1. computation of eigenvalues: $[\lambda_1, \ldots, \lambda_n] = \mathrm{FFT}([c_0, \ldots, c_{n-1}])$,
2. transformation into diagonal form: $\mathbf{x} = \mathrm{FFT}(\mathbf{u})$,
3. multiplication with eigenvalues: $y_j = \lambda_j x_j$, $j = 1, \ldots, n$ (complex multiplication),
4. backward transformation: $\mathbf{v} = \mathrm{IFFT}(\mathbf{y})$.

Hence, 3 FFTs are needed for the computation of $\mathbf{v} = \mathbf{C}\mathbf{u}$, i.e. $O(3n \log_2 n)$ operations. The first step of the algorithm has to be performed only once, when several products of the same matrix with different vectors are computed.

An arbitrary Toeplitz matrix can be embedded into a circulant matrix of twice the dimension:

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{T} & \boldsymbol{E} \\ \boldsymbol{E} & \boldsymbol{T} \end{bmatrix}, \quad \text{where } \mathbf{E} = \begin{bmatrix} 0 & t_{1-n} & \cdots & t_{-1} \\ t_{n-1} & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & c_{1-n} \\ t_1 & \ldots & t_{n-1} & 0 \end{bmatrix}.$$

By using $\mathbf{C}$, $\mathbf{Tu}$ can be computed as follows:

$$\begin{bmatrix} \boldsymbol{Tu} \\ \boldsymbol{Eu} \end{bmatrix} = \boldsymbol{C} \begin{bmatrix} \boldsymbol{u} \\ \mathbf{0} \end{bmatrix}.$$

Thus, the multiplication needs $O(4n \log_2 n)$ operations. If the Toeplitz matrix is symmetric like in Example 4.1, this property is transfered to the circulant matrix. Hence, its eigenvalues are real and the complex multiplication in the third step of the algorithm can be replaced by a real multiplication.

We are not interested in solving the RTE in $1D$ but in $3D$. Assuming homogeneous Dirichlet boundary conditions and constant coefficients, leads to a volume integral equation with a translation invariant kernel. Furthermore, assume that the computational domain is a parallelepiped and a uniform tensor product grid together with piecewise constant finite elements is used for the discretization. Let n_x, n_y and n_z denote the number of elements in x-, y- and z-direction respectively. Each element can be characterized by a $3D$ index vector (i, j, k), denoting the number of the cell in x-, y- and z-direction respectively. The $3D$ index vector is transformed into a scalar number via

$$\mathrm{index}(i, j, k) = i + jn_x + k(n_x n_y).$$

This gives a linear ordering of the indices, and leads to a recursive block structure of the system matrix. The translation invariance of the kernel implies that the small blocks of size $n_x \times n_x$ are Toeplitz matrices. A translation of j indices in y direction corresponds to a shift by jn_x in the matrix. Thus, the translation invariance in y-direction implies that the small blocks themselves are sub-blocks of a block-Toeplitz matrix.

Definition 4.2. A $n_1 n_2 \times n_1 n_2$ matrix $\mathbf{M} = (\mathbf{T}_{ij})_{i,j=1}^{n_2}$ with $n_1 \times n_1$ Toeplitz matrices $\mathbf{T}_{ij}$ has a two-level $n_2 \times n_2$ block-Toeplitz structure if $\mathbf{T}_{ij} = \mathbf{T}_{j-i}$ depends on $j - i$ only.

This formal definition is illustrated by the following example.

Example 4.2. Let $n_1 = 3$, $n_2 = 2$. Then a two-level block-Toeplitz matrix looks as follows.

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}^{(0)} & \mathbf{T}^{(1)} \\ \mathbf{T}^{(-1)} & \mathbf{T}^{(0)} \end{bmatrix}, \quad \text{where } \mathbf{T}^{(i)} = \begin{bmatrix} t_0^{(i)} & t_1^{(i)} & t_2^{(i)} \\ t_{-1}^{(i)} & t_0^{(i)} & t_1^{(i)} \\ t_{-2}^{(i)} & t_{-1}^{(i)} & t_0^{(i)} \end{bmatrix}, \quad i = -1, \ldots, 1.$$

Proceeding as in the above definition a q-level block-Toeplitz matrix can be recursively defined as a block matrix with Toeplitz structure, where the single blocks are (q-1)-level block-Toeplitz matrices. The matrix resulting from the $3D$ RTE is a three-level block-Toeplitz matrix. Like in the standard case, a q-level block-Toeplitz matrix can be embedded into a q-level block-circulant matrix, which is defined analogous to a block-Toeplitz matrix. Since the embedding has to be done at every level, the dimension is increased by a factor 2^q. The fast diagonalization of a block-circulant matrix proceeds as follows (example for $q = 2$)

1. Use FFT to transform the n_2 different circulant matrices of size n_1 into diagonal form in $n_2 O(n_1 \log_2 n_1)$ operations. This results in a matrix with circulant block structure, where the single blocks have diagonal form.
2. Use a permutation to transform the matrix into a block diagonal form with n_1 blocks, which are circulant matrices of size n_2 .
3. Use FFT to transform the n_1 circulant matrices of size n_2 into diagonal form in $n_1 O(n_2 \log_2 n_2)$ operations. This results in a diagonal matrix.

By using this procedure, a matrix-vector multiplication with the three-level block-Toeplitz matrix resulting from the $3D$ RTE can be performed in $O(16n \log_2 n)$ operations provided that the eigenvalues have been computed beforehand.

Remark 4.1. Due to the restriction on parallel piped domains and uniform grids this method is not very useful for the solution of practical problems. The matrix compression methods, described later in this chapter, are applicable to more general situations but they introduce an additional error. In order to measure this error for special test cases, the algorithms for Toeplitz-type matrices can be used.

Remark 4.2. If the assumption of homogeneous Dirichlet boundary conditions is dropped, the system matrix is given by $\mathbf{K} = \left(\begin{smallmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{smallmatrix}\right)$. The matrix $\mathbf{K}_{11}$ is a three-level block-Toeplitz matrix and $\mathbf{K}_{12}$ is a union of two-level block-Toeplitz matrices. $\mathbf{K}_{22}$ can be split into blocks corresponding to pairs of the six rectangular facets of the parallelepiped. If two of the facets are parallel, the corresponding block is a two-level block-Toeplitz matrix. In case of neighboring facets, translation invariance only holds in direction of the common edge. The corresponding matrix is a tensor product of a Toeplitz matrix with a matrix, which could be approximated by the methods described below. Algorithms for these kind of blended matrices are introduced in [27].

Low-rank matrices

A matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is called low-rank matrix if $\operatorname{rank}(\mathbf{A}) \leq k$ where $k \ll \min(n, m)$. Matrices of this kind can be represented in the following form: there are $\mathbf{x}_\nu \in \mathbb{R}^n, \mathbf{y}_\nu \in \mathbb{R}^m$ ($\nu = 1, \ldots, k$) such that

$$\mathbf{A} = \sum_{\nu=1}^{k} \mathbf{x}_\nu \mathbf{y}_\nu^T = \mathbf{X}\mathbf{Y}^T, \qquad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k] \in \mathbb{R}^{n \times k}, \ \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k] \in \mathbb{R}^{m \times k}. \tag{4.1}$$

The amount of storage for a low-rank matrix is $k(n+m)$ and the amount of work for a matrix-vector multiplication is $k(n+m)$ multiplications plus $k(m+n)-(k+n)$ additions, i.e. approximately $2k(n+m)$ operations. Hence, the computational load is reduced to linear complexity even though the matrix is fully populated.

As mentioned in the introduction, we are going to use low-rank matrices to approximate submatrices of a bigger matrix, i.e. more than one low-rank matrix is considered. In this case the following representation might be advantageous. Fix to spaces $V = \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subset \mathbb{R}^n$, $W = \operatorname{span}\{\mathbf{y}_1, \ldots, \mathbf{y}_k\} \subset \mathbb{R}^m$ with $\dim V = \dim W = k$ and define the tensor product space

$$V \otimes W = \operatorname{span}\{\mathbf{x}_\nu \mathbf{y}_\mu^T \mid 1 \leq \nu, \mu \leq k\}. \tag{4.2}$$

Then $\mathbf{A} \in V \otimes W$ has a representation of the form

$$\mathbf{A} = \sum_{\nu,\mu=1}^{k} \zeta_{\nu\mu} \mathbf{x}_\nu \mathbf{y}_\mu^T = \mathbf{X}\mathbf{Z}\mathbf{Y}^T, \quad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k],\ \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k],\ \mathbf{Z} \in \mathbb{R}^{k\times k}. \tag{4.3}$$

Note that $\mathbf{X}$, $\mathbf{Y}$ have to be stored only once if more than one matrix $\mathbf{A} \in V \otimes W$ is considered. The amount of storage needed is $k(n+m)$ for $\mathbf{X}$, $\mathbf{Y}$ and k^2 for every $\mathbf{Z}$. A matrix-vector multiplication needs $2k(n+m+k)$ operations.

The following example shows how low-rank matrices could arise when considering the discretization of integral equations.

Example 4.3. Assume that the kernel of the integral equation is separable, i.e.

$$k(x,y) = \sum_{\nu=1}^{k} \Phi^{(\nu)}(x)\Psi^{(\nu)}(y).$$

Kernels of this type are also called degenerate. Let $\{b_1, \ldots, b_n\}$ be the basis of the ansatz space. Then the entries of the Galerkin matrix $\mathbf{K}$ are given by

$$\begin{aligned} k_{i,j} = \langle b_i, Kb_j\rangle &= \int\int b_i(x) \sum_{\nu=1}^{k} \Phi^{(\nu)}(x)\Psi^{(\nu)}(y) b_j(y)\,dy\,dx \\ &= \sum_{\nu=1}^{k} \underbrace{\int \Phi_\nu(x) b_i(x)\,dx}_{=:x_{\nu,i}} \underbrace{\int \Psi_\nu(y) b_j(y)\,dy}_{=:y_{\nu,i}}. \end{aligned}$$

Hence,

$$\mathbf{K} = \sum_{\nu=1}^{k} \mathbf{x}_\nu \mathbf{y}_\nu^\top = \mathbf{X}\mathbf{Y}^T, \quad \mathbf{X},\ \mathbf{Y} \in \mathbb{R}^{n\times k}.$$

Unfortunately the kernels in the integral equation arising from the RTE are not degenerate. Furthermore, because of the singularity for $x = y$, it is impossible to find a separable kernel, which is a good global approximation. On the other hand, the kernel is very smooth if x and y are sufficiently separate. Hence, it is possible to find a local approximation in this case. This is demonstrated in the next example.

Example 4.4. Consider the integral equation (2.13) resulting from the RTE for a $1D$ slab geometry with isotropic scattering and homogeneous Dirichlet boundary conditions. Consider

two intervals $I_x = [x_0 - r_x, x_0 + r_x]$, $I_y = [y_0 - r_y, y_0 + r_y]$. The size of the intervals should be smaller than their distance:

$$\max\{r_x, r_y\} \leq \eta \operatorname{dist}(I_x, I_y), \quad \eta < \frac{1}{2}, \tag{4.4}$$

where $\operatorname{dist}(I_x, I_y)$ denotes the distance between the two intervals (see Definition 4.4 below). Define the difference variable $z := x - y$ and the center of the difference domain $z_0 = x_0 - y_0$ (we assume w.l.o.g. that $z_0 > 0$). Expanding the kernel into a Taylor series around z_0 results in

$$E_1(\gamma z) = T^{(p)}(z; z_0) + R^{(p)}(z; z_0),$$

where the approximation is given by

$$T^{(p)}(z; z_0) = \sum_{\nu+\mu \leq p} \frac{d^{\nu+\mu}}{dz^{\nu+\mu}} E_1(\gamma z_0) \frac{(x - x_0)^\nu}{\nu!} \frac{(y_0 - y)^\mu}{\mu!},$$

and the remainder satisfies the following estimate where $\zeta = z_0 + t(z - z_0)$, $t \in [0, 1]$

$$|R^{(p)}(z; z_0)| = p! \underbrace{\Big(\sum_{\nu=0}^{p} \frac{(\gamma\zeta)^\nu}{\nu!} \Big) e^{-\gamma\zeta}}_{\leq 1} \frac{1}{\zeta^{p+1}} \frac{(z - z_0)^{p+1}}{(p+1)!} \leq \frac{1}{p+1} (2\eta)^{p+1},$$

since $|z - z_0| \leq |x - x_0| + |y - y_0| < \frac{2\eta}{1+2\eta} z_0$ by (4.4).

This estimate for the remainder $R^{(p)}$ decays exponentially fast to zero, when $\eta < \frac{1}{2}$. This is due to the smoothness of the kernel away from the diagonal: the derivatives decay faster than the polynomial terms in the Taylor expansion grow. Note that the estimate is a worst case estimate, in practice exponential convergence occurs for arbitrary $\eta < 1$.

The approximation $T^{(p)}(\cdot; z_0)$ is a separable kernel, which depends on the considered intervals I_x and I_y. The condition (4.4) shows that the size of the domain, where the approximation is valid, grows with growing distance between the two domains. Hence, a non-tensor partition of the $2D$ domain in (x, y)-direction is needed, where the size of the sub-domains grows with the distance to the diagonal. An example of such a partition is given in Figure 4.1. On the small blocks near the diagonal, there is no suitable approximation. Hence, the entries of the Galerkin matrix are computed in the usual way. On each farfield block an appropriate local approximation can be used to get a low-rank matrix.

The non-tensor partition of the (x, y)-domain is the basis for the approximation of the kernel in Example 4.4. Approximating the kernel is only a means to find an approximation of the matrix. To this end, it is more efficient to look at the problem from an algebraic point of view by using a decomposition of the matrix instead of the (x, y)-domain. For piecewise constant ansatz functions the two approaches are equivalent, but for piecewise linear ansatz functions the support of the basis functions intersect. When simplex elements are used, the number of elements is much higher than the number of basis functions (= number of nodes): factor two in $2D$ and factor six in $3D$. Hence, the basis oriented approach is much more efficient. To construct a non-tensor partition, it is useful to equip the index set with a hierarchical structure. As shown later, this hierarchy can also be used to save further operations.

Cluster Tree and Block Partitioning

Let $I = \{1, \ldots, n\}$ denote the index set corresponding to the basis functions. A subset $c \in \mathcal{P}(I)$ ($\mathcal{P}(I)$ denotes the power set of I.) is called a cluster. A tree structure $T(I)$ is

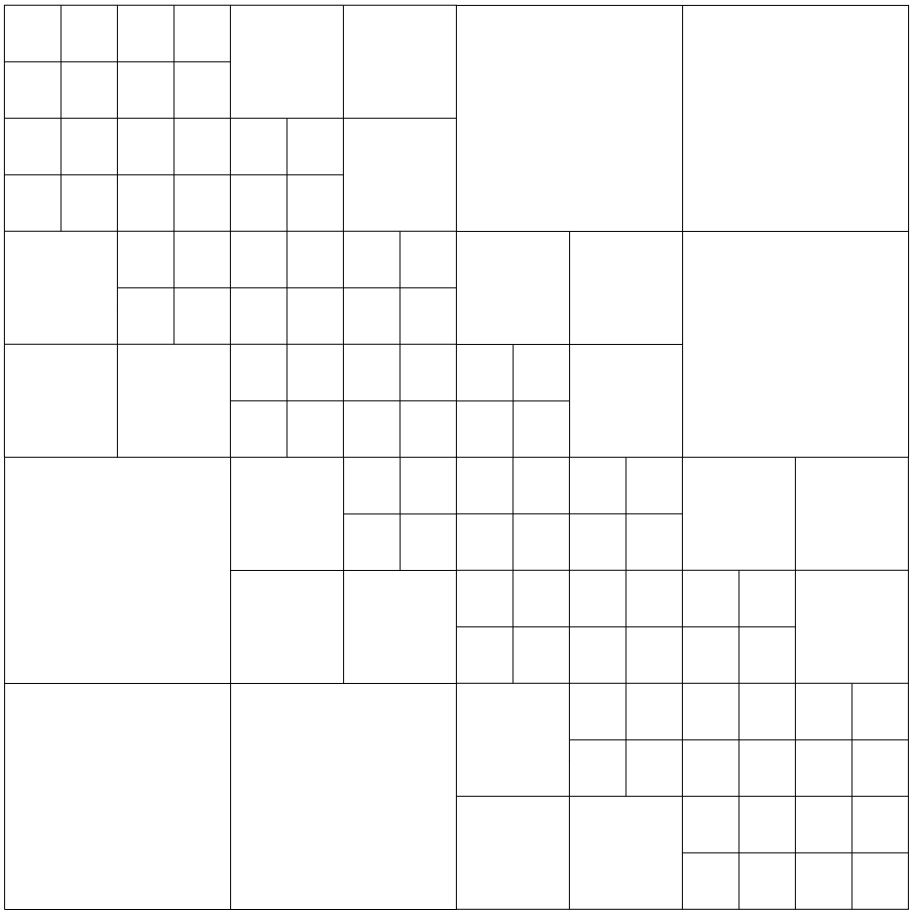

Figure 4.1: Non-Tensor Partition

obtained by associating a set of sons $\sigma(c)$ with each cluster c. c is called a leaf of the tree if $\sigma(c) = \emptyset$. The set of leaves is denoted by $\mathcal{L}(T(I))$.

Definition 4.3. A set $T(I)$ of clusters is called a cluster tree if the following conditions are satisfied:

1. $T(I) \subseteq \mathcal{P}(I)$,
2. $I \in T(I)$, (I is the root of $T(I)$),
3. for $c \in T(I)$ with $\sigma(c) \neq \emptyset$ it holds: $c = \overset{\circ}{\bigcup}_{\tilde{c} \in \sigma(c)} \tilde{c}$
   i.e. every cluster is the disjoint union of its sons.
4. There is a constant deg independent of I s. t. $\sharp\,\sigma(c) \leq deg \, \forall c \in T(I)$,
   i.e. the number of sons of each cluster is bounded.
5. There is a constant C_L independent of I s. t. $\sharp\, c \leq C_L$ for every $c \in \mathcal{L}(T(I))$, i.e. the size of each leaf is bounded.

By the above properties, $\mathcal{L}(T(I))$ is a partition of I.

The support of a cluster $c \in T(I)$ is given by the union of the supports of the basis functions corresponding to its elements, i.e.

$$\operatorname{supp} c := \bigcup_{i \in c} \operatorname{supp} b_i.$$

The level of a cluster is given by the function $\operatorname{level} : T(I) \to \mathbb{N}_0$ recursively defined via

$$\operatorname{level}(I) := 0 \quad \text{and} \quad \operatorname{level}(\tilde{c}) := \operatorname{level}(c) + 1, \text{ for } \tilde{c} \in \sigma(c).$$

The depth of $T(I)$ is given by $l_{\max} = \max\{\operatorname{level}(c) \,|\, c \in T(I)\}$. $T(I)$ can be partitioned into

$$T(I) = \bigcup_{l=0}^{l_{\max}} T_l(I), \qquad \text{where } T_l(I) = \{c \in T(I) \,|\, \operatorname{level}(c) = l\}.$$

Remark 4.3. The independent variables in the integral equation (2.26) stemming from the RTE are the energy G and the outgoing heat flux q_{out}. G is defined in the interior of the domain, whereas q_{out} is only defined on the boundary. Hence, we need to define one cluster tree for the index set I_1 corresponding to the basis functions for G and one for I_2 corresponding to q_{out}.

The definition of the cluster tree is independent of the corresponding basis functions, but having in mind Condition (4.4) for the $1D$ example, it makes sense to look for clusters, such that the support is contained in a ball with a small diameter. For the construction of the cluster tree, it is useful to associate each index $i \in I$ with a reference point P_i in $\mathbb{R}^3$. In case of piecewise linear ansatz functions, each basis function and, hence, each index corresponds to a node of the triangulation. This point can be chosen as reference point. For piecewise constant ansatz functions, each basis function and, hence, each index corresponds to a volume or surface element of the triangulation. The center of this cell can be chosen as reference point. There are basically two possibilities for the construction of the cluster tree:

- top-down: This algorithm starts with a cluster c representing the whole index set. c is split into two sons by the following procedure: the coordinate direction in which the cluster has the largest extension is determined and a hyper-plane through the middle of the cluster with normal pointing into the chosen direction is defined. The indices $i \in c$ whose reference point P_i lies on one side of the hyper plane form one son of c. The remaining indices form the other son. This algorithm is recursively applied to each cluster until the size of the cluster is smaller than a prescribed value. This gives a binary tree which has $O(\log_2 n)$ levels. Since each index is regarded exactly once at each level, the complexity of the whole algorithm is $O(n \log_2 n)$.

- bottom-up: Let Q denote the smallest parallelepiped with edges parallel to the coordinate axes, containing the computational domain D. An auxiliary Cartesian grid is defined on Q, where the number of grid cells in each coordinate direction is a power of two. A binary cluster tree for this auxiliary grid is easily obtained by in every step selecting the coordinate direction with the smallest grid size and merging each cluster with one of its neighbors along this direction. This is possible since the number of clusters in each direction at one level of the auxiliary tree is a power of two, a property which is maintained during the construction algorithm. Each index $i \in I$ is associated with a cell of the auxiliary grid, e.g. by choosing the cell which contains the reference point P_i. Indices associated with the same cell are put together to form a cluster. These

clusters form the leaves of a cluster tree of the index set I and each of these clusters is associated with exactly one cell of the auxiliary grid, i.e. with a leaf of the auxiliary tree. By these assignments, the auxiliary tree defines in a canonical way a cluster tree for the index set I. Since a binary tree has $2n-1$ clusters and every cluster is regarded only once, the complexity of the algorithm is $O(n)$.

Since the axis of splitting in the top-down algorithm is chosen individually for each cluster, whereas the axis of merging in the bottom-up algorithm is chosen for the whole level, the top-down algorithm might yield a better balanced cluster tree, i.e. one, where clusters at the same level have approximately the same size. This is especially true for surfaces.

The energy G is defined in an open set $D \subset \mathbb{R}^3$, therefore, the bottom-up algorithm yields quite good results for the construction of $T(I_1)$. The cluster tree $T(I_2)$ for the boundary indices can be constructed as follows. An index $i_2 \in I_2$ corresponds to a basis function on the boundary. In case of piecewise linear basis functions, i_2 is associated with the index $i_1 \in I_1$ corresponding to the same node of the triangulation. In case of piecewise constant basis functions, i_2 corresponds to a surface element $\tau^{(s)}$. There is exactly one volume element $\tau^{(v)}$ such that $\tau^{(s)}$ is a facet of $\tau^{(v)}$. The corresponding index $i_1 \in I_1$ is associated with i_2. By using this assignment, the cluster tree $T(I_1)$ for the volume indices implies in a canonical way a cluster tree $T(I_2)$ for the boundary indices.

Next, an admissibility condition is needed, that can be used to partition the product index sets $I_i \times I_j$, $i,j \in \{1,2\}$, into pairs $b = (c_1, c_2) \in T(I_i) \times T(I_j)$ such that the kernel $k_{ij}(\cdot,\cdot)$ is smooth enough on the associated domain $\operatorname{supp} c_1 \times \operatorname{supp} c_2$. A pair $b = (c_1, c_2)$ is called a block.

Definition 4.4. Let $\eta \in (0,1)$. A block $b \in T(I_i) \times T(I_j)$ is called η-admissible if

$$\frac{1}{2}\max\{\operatorname{diam}(c_1), \operatorname{diam}(c_2)\} \leq \eta \operatorname{dist}(c_1, c_2), \tag{4.5}$$

where the diameter and distance for two clusters are given by

$$\operatorname{diam}(c) = \max_{x,y \in \operatorname{supp} c} \|x-y\| \quad \text{and} \quad \operatorname{dist}(c_1,c_2) = \min_{\substack{x \in \operatorname{supp} c_1 \\ y \in \operatorname{supp} c_2}} \|x-y\|.$$

For unstructured grids, the computation of the diameter of a cluster and especially of the distance between two clusters is too complicated and time-consuming. Therefore, the condition (4.5) is replaced by a stronger variant which is easier to handle numerically, e.g. by using super-sets of $\operatorname{supp} c$.

By construction of the cluster tree the support of the volume clusters, at least of the big ones, has nearly the shape of a parallelepiped with axes parallel to the coordinate system. Such a parallelepiped is called a box. For a cluster c it is quite simple to determine the minimal box $Q(c)$ containing c. Such a bounding box can be characterized by only six real values, and its diameter and the distance between two boxes are easily computable. Hence, the complicated admissibility condition can be replaced by an easy-to-handle condition which yields nearly the same results.

For the surface clusters the use of bounding boxes may result in large over estimations when computing the distance between two clusters. A two dimensional example is given by a discretization of the unit circle using a uniform grid with 16 piecewise constant elements (see left side of Figure 4.2). In this figure some of the boxes representing the clusters are illustrated: the boxes on the finest level are depicted by solid lines, the ones on the next level

by dashed lines and the ones on the third level by dotted lines. The box, representing the cluster containing all elements in the right top quarter, and the one, representing the cluster containing the left bottom quarter, intersect, whereas the clusters themselves have a large distance.

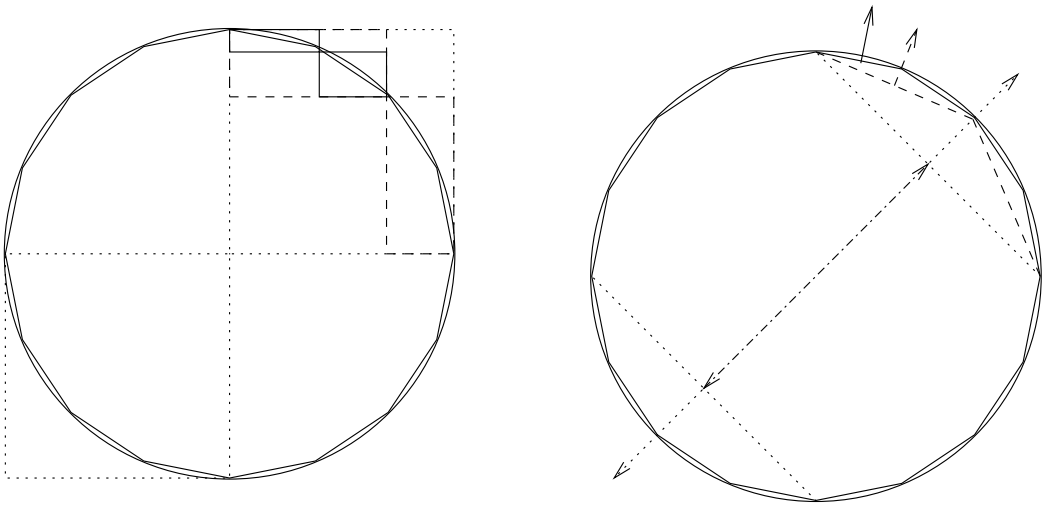

Figure 4.2: Approximation of surface clusters in $2D$ using boxes (left) and using line segments (right).

The following approach exploits the fact that we deal with convex domains only. The example above shows that it would be advantageous to represent a cluster by a line segment. In the right picture of Figure 4.2 some of the line segments representing the clusters are illustrated: the line segments on the finest level are depicted by solid lines, the ones on the next level by dashed lines and the ones on the third level by dotted lines. Since we are only considering convex domains, the area on one side of the line segment represents the cluster. The arrows in the figure illustrate this side. The distance between two such line segments can be easily computed, we only have to take care that the two clusters are on the correct side of each other. In the above example the distance between the two dotted lines, depicted by the dash-dotted arrow, is exactly two times the radius of the two line segments. Hence, when the line segments are used to represent the clusters, the two clusters are η-admissible if $\eta \geq \frac{1}{2}$. Furthermore, when the bottom-up algorithm is used to construct the cluster tree, the line segment for a cluster resulting from merging two smaller clusters can be determined by only considering the line segments of the two sons, i.e. without considering the individual elements contained in the cluster.

Unfortunately the situation gets more complicated in the $3D$ case. The line is replaced by a plane. The segment may be replaced by a rectangle, but the orientation inside the plane of this rectangle is somewhat arbitrary. There are much more possibilities for the position of two such rectangles with respect to each other in $3D$ compared with the case of line segments in $2D$. Therefore, we have to distinguish several cases when we want to find a representation of a cluster by using the representation of its sons only. The same holds for the computation of the distance between two clusters. Therefore, it seems to be advantageous to use simply the admissibility for bounding boxes.

The index set $I_i \times I_j$ is partitioned into blocks (c_1, c_2) by the following recursive algorithm

```
procedure BuildBlockPartitioning(c1,c2, N, F)
begin
    if (c1, c2) is admissible then
        F := F ∪ {(c1, c2)}
    else
        if σ(c1) ≠ ∅ and σ(c2) ≠ ∅ then
            for c̃1 ∈ σ(c1), c̃2 ∈ σ(c2) do
                BuildBlockPartitioning(c̃1,c̃2, N, F)
        else
            N := N ∪ {(c1, c2)}
end
```

Calling this procedure with $c_1 = I_i$, $c_2 = I_j$, $i, j \in \{1, 2\}$ and $N = F = \emptyset$ creates a block partitioning $P_{ij} = N \cup F$ of $I_i \times I_j$ consisting of non-admissible blocks corresponding to the leaves of $T(I)$ (nearfield N) and admissible blocks (farfield F). By construction the clusters c_1 and c_2 of a block $b = (c_1, c_2)$ belong to the same level. For $0 \leq l \leq l_{max}$ the farfield levels $F(l)$ are defined by

$$F(l) = \{(c_1, c_2) \in F \,|\, level(c_1) = level(c_2) = l\}.$$

Since the admissibility condition (4.5) is symmetric in c_1 and c_2, the cases $i = 1, j = 2$ and $i = 2, j = 1$ are equivalent. For the same reason the partitioning P_{ii}, $i = 1, 2$ is symmetric, i.e. $(c_1, c_2) \in P_{ii}$ implies $(c_2, c_1) \in P_{ii}$. Hence, it is sufficient to compute and store only half of the partitioning.

Assumption 4.1. We assume that the resulting partitionings are sparse, i.e. that there exists a constant C_{sp} satisfying

$$\max_{c_1 \in T(I_i)} \sharp\{c_2 \in T(I_j) \,|\, (c_1, c_2) \in P_{ij}\} \leq C_{\mathrm{s\,p}}, \tag{4.6}$$

$$\max_{c_2 \in T(I_j)} \sharp\{c_1 \in T(I_i) \,|\, (c_1, c_2) \in P_{ij}\} \leq C_{\mathrm{s\,p}}. \tag{4.7}$$

For standard situations with quasi-uniform meshes, this estimate has been established in [50].

4.2 Hierarchical Matrices

In this section we restrict ourselves to one integral operator i.e. one index set $I = \{1, \ldots, n\}$. The system of integral equations is considered in Section 4.3.

The following definition is due to Hackbusch [26]. It introduces different kinds of data-sparse matrices, all of which are based on a block partitioning into sub-matrices, where each sub-matrix is a low-rank matrix. For a $\mathcal{H}$-matrix the sub-matrices are completely independent of each other. Hence, Representation (4.1) is used. For a uniform $\mathcal{H}$-matrix the sub-matrices are restricted to certain subspaces of the form (4.2). Hence, Representation (4.3) is used.

Remark 4.4. The article [26] does not distinguish between nearfield and farfield blocks: a sub-matrix corresponding to a nearfield block is a low-rank matrix because of its small size. On the other hand, for these matrices the representation as a full matrix is much more efficient than the Representations (4.1) and (4.3). Hence, it is advantageous to handle the nearfield and the farfield separately as it is done e.g. in [50]. Therefore, we decided to use matrices which are equal to zero on all nearfield blocks in the following definition of hierarchical matrices.

Definition 4.5. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a matrix and $P = N \cup F$ be a block partitioning of $I \times I$ consisting of admissible blocks and leaf blocks.

1. Let $k \in \mathbb{N}$. $\mathbf{A}$ is called $\mathcal{H}$-matrix of rank k if for all blocks $b = (c_1, c_2) \in F$, there are matrices $\mathbf{X}_b, \mathbf{Y}_b \in \mathbb{R}^{n\times k_b}$, $k_b \leq k$ such that

$$\mathbf{A} = \sum_{(c_1,c_2)\in F} \mathbf{X}_b \mathbf{Y}_b^T, \tag{4.8}$$

   where only the rows i of $\mathbf{X}_b$ with $i \in c_1$ contain non-zero entries, analog for $\mathbf{Y}_b$. That is the sum in (4.8) constitutes a decomposition of the matrix into blocks $b = (c_1, c_2) \in F$.

2. A family $V = (\mathbf{V}_c)_{c\in T(I)}$ is called a cluster basis if there is a $k_c \in \mathbb{N}$ for each $c \in T(I)$ such that $\mathbf{V}_c \in \mathbb{R}^{n\times k_c}$, where again only the rows i of $\mathbf{V}_c$ with $i \in c$ contain non-zero entries.
   Let X and Y be cluster bases. $\mathbf{A}$ is called a uniform $\mathcal{H}$-matrix with respect to X and Y if for each $(c_1, c_2) \in F$, there is a matrix $\mathbf{Z}_{(c_1,c_2)} \in \mathbb{R}^{k_{c_1}\times k_{c_2}}$ satisfying

$$\mathbf{A} = \sum_{(c_1,c_2)\in F} \mathbf{X}_{c_1} \mathbf{Z}_{(c_1,c_2)} \mathbf{Y}_{c_2}^T \tag{4.9}$$

   In this context, X is called column basis and Y is called row basis.

3. A cluster basis $V = (\mathbf{V}_c)_{c\in T(I)}$ is called nested if there are transfer matrices $\mathbf{S}_{\tilde{c},c} \in \mathbb{R}^{k_{\tilde{c}}\times k_c}$ for all $c \in T(I)$ and $\tilde{c} \in \sigma(c)$ satisfying

$$\mathbf{V}_c|_{\tilde{c}\times k_c} = \left(\mathbf{V}_{\tilde{c}} \mathbf{S}_{\tilde{c},c}\right)|_{\tilde{c}\times k_c}, \tag{4.10}$$

   where $\cdot|_{\tilde{c}\times k_c}$ denotes the restriction of the matrix to the rows corresponding to the son $\tilde{c}$. Especially it holds

$$\mathbf{V}_c = \sum_{\tilde{c}\in\sigma(c)} \mathbf{V}_{\tilde{c}} \mathbf{S}_{\tilde{c},c}. \tag{4.11}$$

   $\mathbf{A}$ is called $\mathcal{H}^2$-matrix w.r.t. X and Y if it is a uniform $\mathcal{H}$-matrix w.r.t. to these cluster bases and if both bases are nested.

Remark 4.5. Most of the matrices in the above definition are never formed explicitly. Only the non zero entries have to be stored. The matrices are a means for a compact representation, which is useful for an illustrative description of the algorithm used to perform a multiplication of a hierarchical matrix by a vector (see below).

Remark 4.6. If $\mathbf{A}$ is a symmetric matrix the cluster bases X and Y are identical and $\mathbf{Z}_{(c_2,c_1)} = \mathbf{Z}_{(c_1,c_2)}^T$ for all blocks $(c_1, c_2) \in F$. (Remember: F is symmetric, i.e. $(c_1, c_2) \in F \Rightarrow (c_2, c_1) \in F$.) Hence, the amount of storage for the matrices $\mathbf{Z}_{(c_1,c_2)}$ is reduced by a factor two.

By using one of the above representations of the matrix $\mathbf{A}$, the entries of $\mathbf{A}$ are not known explicitly, i.e. direct solvers cannot be applied to solve the corresponding linear system of equations. For iterative solvers, only the computation of matrix-vector multiplications is needed. These can be performed efficiently by using one of the above representations of the matrix.

The restriction of a $\mathcal{H}$-matrix to a block $b \in F$ is a low-rank matrix. Hence, performing the multiplication of such a block with a vector restricted to c_2 is cheap as shown in Section 4.1. The result for the whole matrix is obtained by adding up the results for the individual blocks.

For a $\mathcal{H}$-matrix both matrices in the low-rank representation $\mathbf{X}_b\mathbf{Y}_b^T$ depend on the block $b = (c_1, c_2)$. Hence, the multiplication for one sub-matrix is completely independent of the other blocks. Since a certain $c \in T(I)$ usually occurs in more than one block $b \in F$, it is desirable to do as much work on the "cluster level" and as little as possible on the "block level" in order to reduce the total amount of work. This is done in case of uniform $\mathcal{H}$-matrices by choosing the cluster bases X and Y, i.e. fixing the approximation space for the block $b = (c_1, c_2)$ to range $\mathbf{X}_{c_1} \otimes$ range $\mathbf{Y}_{c_2}$ (see (4.2)). A matrix-vector multiplication $\mathbf{v} = \mathbf{A}\mathbf{u}$ in this case is performed in three steps:

1. forward transformation: for all $c_2 \in T(I)$: $\mathbf{u}_{c_2} = \mathbf{Y}_{c_2}^T \mathbf{u}$,
2. block multiplication: for all $c_1 \in T(I)$: $\mathbf{v}_{c_1} = \sum_{c_2 : (c_1,c_2) \in F} \mathbf{Z}_{(c_1,c_2)} \mathbf{u}_{c_2}$,
3. backward transformation: $\mathbf{v} = \sum_{c_1 \in T(I)} \mathbf{X}_{c_1} \mathbf{v}_{c_1}$.

Only the second step has to be performed on the block partitioning P. The first and third step can be performed on the cluster tree.

In case of $\mathcal{H}^2$-matrices the computational complexity is reduced further by exploiting the hierarchical structure of the nested bases. By using Equation (4.11) the forward transformation can be performed as follows

$$\mathbf{u}_{c_2} = \begin{cases} \mathbf{Y}_{c_2}^T \mathbf{u} & \text{if } c_2 \in \mathcal{L}(T(I)), \\ \sum_{\tilde{c}_2 \in \sigma(c_2)} \mathbf{S}_{\tilde{c}_2, c_2}^T \mathbf{u}_{\tilde{c}_2} & \text{else,} \end{cases}$$

where it is assumed that $\mathbf{u}_{\tilde{c}_2}$ is computed before it is used on the right hand side. That is, the computation starts at the leaves and proceeds bottom-up. The coefficient vectors $\mathbf{u}_{c_2}$ at higher levels can be computed by using the coefficient vectors of its immediate sons only. Hence, the matrices $\mathbf{Y}_{c_2}$ have only to be assembled for $c_2 \in \mathcal{L}(T(I))$. At higher levels the transfer matrices $\mathbf{S}_{\tilde{c}_2, c_2}$ are used.

The approach to reduce the complexity in step three is adjoint to the algorithm for step one. Consider an index $i \in I$. There is a sequence of clusters $c^{(L)}, \ldots, c^{(0)}$ at different levels of the cluster tree with

$$\{i\} \subseteq c^{(L)} \subset \ldots \subset c^{(0)} = I.$$

All these clusters contribute to the i-th entry v_i of $\mathbf{v}$ in the sum in step three. Adding up these contributions separately for all indices needs $O(n \log_2 n)$ operations because the index set has the size n and the cluster tree has $O(\log_2 n)$ levels. To reduce the computational complexity, as much operations as possible have to be performed simultaneously for more than one index, i.e. for all indices belonging to one cluster. This can again be done by exploiting the property (4.11) of nested bases. To this end, the sum in step three is split up according to the levels of the cluster tree:

$$\mathbf{v} = \sum_{l=0}^{l_{\max}} \sum_{c_1 \in T_l(I)} \mathbf{X}_{c_1} \mathbf{v}_{c_1} \overset{(4.11)}{=} \sum_{c_1 \in T_0(I)} \sum_{\tilde{c}_1 \in \sigma(c_1)} \mathbf{X}_{\tilde{c}_1} \mathbf{S}_{\tilde{c}_1, c_1} \mathbf{v}_{c_1} + \sum_{l=1}^{l_{\max}} \sum_{c_1 \in T_l(I)} \mathbf{X}_{c_1} \mathbf{v}_{c_1}.$$

For $c_1 \in T_0(I)$ the sons $\tilde{c}_1 \in \sigma(c_1)$ belong to $T_1(I)$. Hence, the terms of the first sum can be combined with the terms of the first level in the second sum:

$$\mathbf{v} = \sum_{c_1 \in T_1(I)} \mathbf{X}_{c_1} \Big(\underbrace{\mathbf{v}_{c_1} + \mathbf{S}_{c_1, \hat{c}_1} \mathbf{v}_{\hat{c}_1}}_{=: \tilde{\mathbf{v}}_{c_1}} \Big) + \sum_{l=2}^{l_{\max}} \sum_{c_1 \in T_l(I)} \mathbf{X}_{c_1} \mathbf{v}_{c_1},$$

where $\hat{c}_1$ denotes the father of c_1.

Proceeding further until reaching the leaf level leads, to the following algorithm

$$\bar{\mathbf{v}}_{c_1} = \begin{cases} \mathbf{v}_{c_1} & \text{if } c_1 = I, \\ \mathbf{v}_{c_1} + \mathbf{S}_{c_1,\hat{c}_1} \bar{\mathbf{v}}_{\hat{c}_1} & \text{where } \hat{c}_1 \text{ is the father of } c_1, \end{cases}$$

where it is assumed that $\bar{\mathbf{v}}_{\hat{c}_1}$ is computed before it is used on the right hand side. That is, the computation starts at the root of the tree and proceeds top-down by shift operations. It follows that

$$\mathbf{v} = \sum_{c_1 \in T(I)} \mathbf{X}_{c_1} \mathbf{v}_{c_1} = \sum_{c_1 \text{ a leaf}} \mathbf{X}_{c_1} \bar{\mathbf{v}}_{c_1}.$$

Note that the matrices $\mathbf{X}_{c_1}$ again have only to be assembled if $c_1 \in \mathcal{L}(I)$.

Complexity of the Algorithm

Let $k_c = k_{\text{const}}$ $\forall c \in T(I)$. Furthermore, we assume that $k_{\text{const}} \leq \sharp c$ $\forall c \in T(I)$, i.e. the leaf clusters have a certain size. Beneath this size it is more efficient to use the representation as a full matrix instead of a low-rank representation. Due to this assumption, there are no more than $\frac{n}{k_{\text{const}}}$ leaves in the binary tree $T(I)$ and, hence, no more than $\frac{2n}{k_{\text{const}}}$ clusters. Each matrix $S_{\hat{c},c}$ requires k_{const}^2 storage, so that all transformation matrices together need $O(nk_{\text{const}})$ storage. The storage requirements for $\mathbf{X}_c$ for c a leaf is given by $\sum_{c \in \mathcal{L}(T(I))} k_{\text{const}} \sharp c = nk_{\text{const}}$. Analog for $\mathbf{Y}_c$. Since the block partitioning P is sparse (see Assumption 4.1), the farfield F contains less than $\frac{n}{k_{\text{const}}} C_{\text{sp}}$ elements. Hence, the matrices $\mathbf{Z}_b$ for all blocks in F need $O(nk_{\text{const}})$ storage.

In the algorithm for the multiplication of a vector by a $\mathcal{H}^2$-matrix every entry of the above matrices is used exactly once, therefore, the complexity of the algorithm is also $O(nk_{\text{const}})$.

Approximation of the Kernel

After the considerations about the fast multiplication of a hierarchical matrix by a vector, we want to show how to construct a $\mathcal{H}^2$-matrix to approximate a matrix stemming from the Galerkin discretization of an integral equation. The matrix is decomposed as follows

$$\mathbf{K} = \mathbf{N} + \mathbf{F} \approx \tilde{\mathbf{K}} = \mathbf{N} + \tilde{\mathbf{F}}. \tag{4.12}$$

The nearfield matrix $\mathbf{N}$ corresponds to the nearfield N. Due to the small size of the blocks $b \in N$, it is not efficient to use a low-rank representation on these blocks. Hence, the entries are computed in the usual way and no approximation is used. Due to the sparseness of the partitioning P (see Assumption 4.1), $\mathbf{N}$ contains only $O(1)$ entries per row, i.e. $\mathbf{N}$ is a sparse matrix. The part of the matrix $\mathbf{K}$ corresponding to the farfield F is approximated by the hierarchical matrix $\tilde{\mathbf{F}}$.

Motivated by Example 4.3, we make the following assumption about the integral kernel.

Assumption 4.2. Let $1 < \eta < \bar{\eta} < 1$, and let P be a η-admissible block partitioning.There exist positive constants $C_1 < \infty$ and $C_2 < \frac{1}{\eta}$, only depending on η, such that for every

$p \in \mathbb{N}_0$ and every $b = (c_1, c_2) \in F$, there exists an approximation $k_b^{(p)}$ of the kernel function k satisfying

$$|k(x,y) - k_b^{(p)}(x,y)| \leq C_1(C_2\eta)^{p+1} \max_{\substack{x \in \operatorname{supp} c_1, \\ y \in \operatorname{supp} c_2}} \|x-y\|^{-\alpha} \forall x \in \operatorname{supp} c_1, y \in \operatorname{supp} c_2, \quad (4.13)$$

where α denotes the order of the singularity of the kernel. The approximation k_b is of the form

$$k_b^{(p)}(x,y) = \sum_{(\nu,\mu) \in J_p \times J_p} \kappa_{\nu,\mu}^{(p)}(c_1,c_2) \Phi_\nu^{c_1,p}(x) \Psi_\mu^{c_2,p}(y), \quad (4.14)$$

with index sets J_p satisfying $k(p) := \sharp J_p \leq (p+1)^3$.

The exponential convergence of the approximation error in (4.13) is crucial for the efficiency of the approximation method.

The coefficients $\kappa_{\nu,\mu}^{(p)}(c_1,c_2)$ depend on the block (c_1,c_2), but the expansion functions $\Phi_\nu^{c_1,p}$ in x-direction only depend on the cluster c_1 in x-direction but not on c_2. An analogous statement holds for the y-direction. This is crucial for obtaining a uniform $\mathcal{H}$-matrix.

The approximation of the kernel leads to the following approximation of the integral operator

$$\begin{aligned} \langle Ku, v \rangle \approx \langle \tilde{K}u, v \rangle = & \sum_{(c_1,c_2) \in N} \int_{\operatorname{supp} c_1} \int_{\operatorname{supp} c_2} v(x) k(x,y) u(y) dx dy \\ & + \sum_{(c_1,c_2) \in F} \sum_{(\nu,\mu) \in J_p \times J_p} \kappa_{\nu,\mu}^{(p)}(c_1,c_2) \int_{\operatorname{supp} c_1} \Phi_\nu^{c_1,p}(x) v(x) dx \int_{\operatorname{supp} c_2} \Psi_\mu^{c_2,p}(y) u(y) dy. \quad (4.15) \end{aligned}$$

The expansion order p might be different on different clusters, i.e. $p = p_c$, but it should be the same on clusters belonging to the same block. The dependency of p on the cluster c is omitted for notational convenience.

The Galerkin discretization of the operator $\tilde{K}$ leads to a uniform $\mathcal{H}$-matrix in the farfield: let (c_1,c_2) be admissible, then the matrix entry $\tilde{f}_{ij}$ for $(i,j) \in (c_1,c_2)$ is given by

$$\tilde{f}_{ij} = \sum_{(\nu,\mu) \in J_p \times J_p} \kappa_{\nu,\mu}^{(p)}(c_1,c_2) \int_{\operatorname{supp} b_i} \Phi_\nu^{c_1,p}(x) b_i(x) dx \int_{\operatorname{supp} b_j} \Psi_\mu^{c_2,p}(y) b_j(y) dy. \quad (4.16)$$

Introducing matrices $\mathbf{X}_{c_1} \in \mathbb{R}^{n \times k(p)}$, $\mathbf{Y}_{c_2} \in \mathbb{R}^{n \times k(p)}$ and $\mathbf{Z}_{(c_1,c_2)} \in \mathbb{R}^{k(p) \times k(p)}$ (i.e. $k_{c_1} = k_{c_2} = k(p)$) with

$$x_{i,\nu}^{(c_1)} := \int_{\operatorname{supp} b_i} \Phi_\nu^{c_1,p}(x) b_i(x) dx, \text{ if } i \in c_1 \text{ and zero else}, \quad (4.17)$$

$$y_{j,\mu}^{(c_2)} := \int_{\operatorname{supp} b_j} \Psi_\mu^{c_2,p}(y) b_j(y) dy, \text{ if } j \in c_2 \text{ and zero else}, \quad (4.18)$$

$$z_{\nu,\mu}^{(c_1,c_2)} := \kappa_{\nu,\mu}^{(p)}(c_1,c_2), \quad (4.19)$$

yields

$$\tilde{f}_{ij} = \left(\mathbf{X}_{c_1} \mathbf{Z}_{(c_1,c_2)} \mathbf{Y}_{c_2}^T \right)_{ij \in (c_1,c_2)}, \quad (4.20)$$

i.e. a representation of a uniform $\mathcal{H}$-matrix.

In order to get a $\mathcal{H}^2$-matrix, the expansion functions $\Phi_\nu^{c,p}$ and $\Psi_\mu^{c,p}$ for clusters c at different levels in the cluster tree must be organized hierarchically. This is discussed below in more detail when some examples of expansion systems are introduced.

In the literature two different families of functions are used to approximate the kernel: polynomials and spherical harmonics. Expansions into spherical harmonics are used in the fast multipole method and also in the panel clustering method. An expansion of the kind (4.14) for the kernel $k(x,y) = \|x-y\|^{-1}$ corresponding to the single layer potential of the Laplace equation in $3D$ can be found in [34]. Using spherical harmonics has the advantage that only $O(p^2)$ terms in the expansion are needed to obtain the requirement (4.13) instead of $O(p^3)$ in case of polynomials. An additional advantage is that an expansion into spherical harmonics is well suited for the approximation of oscillating kernels which appear for example when the BEM is used for the Helmholtz equation. In this case expansion functions are used which depend on the wave number [20]. According to our knowledge, an expansion of the kernels in Chapter 2 into spherical harmonics is not known, therefore, we use polynomials.

The polynomial approximations used in the literature are obtained by interpolation or Taylor expansion. Regarding interpolation, a family $\{x_\nu^c\}_{\nu\in J_p}$ of interpolation points in $Q(c)$ (minimal box containing c) is fixed for each cluster $c \in T(I)$ (e.g. tensor products of the zeros of Chebychev polynomials). $\{p_\nu^c\}_{\nu\in J_p}$ denotes the family of corresponding Lagrange polynomials. (Each polynomial p_ν^c, $\nu \in J_p$, depends on the polynomial degree p: $p_\nu^c = p_\nu^{c,p}$, but this dependency is omitted for notational convenience.) $\{p_\nu^c\}_{\nu\in J_p}$ is a basis of the space $\mathcal{Q}_p$ of tensor product polynomials of a degree up to p

$$\mathcal{Q}_p := \{p(x) = \sum_{0\le\alpha_1,\alpha_2,\alpha_3\le p} \lambda_\alpha x^\alpha,\ \lambda_\alpha \in \mathbb{R}\}, \quad \alpha \text{ a multi-index.} \tag{4.21}$$

When the family $\{p_\nu^c\}_{\nu\in J_p}$ is used as expansion functions, the kernel $k(\cdot,\cdot)$ on a given admissible block $(c_1,c_2) \in F$ is approximated by its interpolant

$$k_b^{(p)}(x,y) = \sum_{\nu\in J_p^{\mathcal{Q}}} \sum_{\mu\in J_p^{\mathcal{Q}}} k(x_\nu^{c_1}, y_\mu^{c_2}) p_\nu^{c_1}(x) p_\mu^{c_2}(y). \tag{4.22}$$

where $J_p^{\mathcal{Q}} = \{\nu \in \mathbb{N}_0^3 \mid \nu_i \le p, i = 1,\dots,3\}$, i.e. $k^{\mathcal{Q}}(p) = (p+1)^3$.

The above expansion is of the required form (4.14). Furthermore, if the kernel $k(\cdot,\cdot)$ is asymptotically smooth, i.e. if there are constants $C_{\text{as}}(n,m)$ such that

$$|\partial_x^\alpha \partial_y^\beta k(x,y)| \le C_{\text{as}}(|\alpha|,|\beta|)\|x-y\|^{-(|\alpha|+|\beta|)}|k(x,y)| \tag{4.23}$$

for all multi-indices α, $\beta \in \mathbb{N}_0^d$ and all x, $y \in \mathbb{R}^d$ with $x \ne y$, then Estimate (4.13) is satisfied for the Expansion (4.22) (for a proof refer [26]).

Since the same expansion functions are used in x and y-direction, the column basis X and the row basis Y are identical, which reduces the amount of storage. It remains to compute the transformation matrices $\mathbf{S}_{\tilde{c},c}$ in (4.10) to obtain a $\mathcal{H}^2$-matrix. Let interpolation polynomials up to the same degree p be used on all clusters. Then the polynomials corresponding to father clusters can be expressed in terms of polynomials corresponding to son clusters without loss. Due to the properties of Lagrange polynomials (i.e. $p_\mu(x_\nu) = \delta_{\mu,\nu}$) and the uniqueness of the interpolation polynomial, the following equality holds for a cluster $c \in T(I)$ with son $\tilde{c} \in T(I)$

$$p_\mu^c(x) = \sum_{\tilde{\mu}\in J_{\tilde{c}}} p_\mu^c(x_{\tilde{\mu}}^{\tilde{c}}) p_{\tilde{\mu}}^{\tilde{c}}(x). \tag{4.24}$$

Hence, the matrix $\mathbf{S}_{\tilde{c},c}$ is given by

$$s^{(\tilde{c},c)}_{\tilde{\mu},\mu} := p^c_\mu(x^{\tilde{c}}_{\tilde{\mu}}). \tag{4.25}$$

As shown below, it might be useful to use a higher polynomial degree on larger clusters. If the polynomial order corresponding to the cluster $\tilde{c}$ is lower than that of its father c, Equation (4.24) yields only an approximation of the higher-order polynomial p^c_μ by lower-order polynomials $p^{\tilde{c}}_{\tilde{\mu}}$. The right hand side of Equation (4.24) can be used to define a modified expansion system $\tilde{p}^c_\mu$. Under the condition that the growth of the polynomial order is slow enough, Estimate (4.13) is satisfied also for the modified expansion system.

For the derivation of the Taylor expansion, we assume that the kernel is translation invariant, i.e. $k(x,y) = k(x-y)$. Not all kernels in Chapter 2 are of this form, but in Section 4.3 it is shown, how to cope with this situation. Define the difference variable $z := x - y$ and the center of the difference domain $z_b = M_{c_1} - M_{c_2}$ where M_{c_i} is the center of cluster c_i. Expanding the kernel into a Taylor series around z_b results in

$$k^{(p)}_b(x,y) = \sum_{|\nu+\mu|\le p} \frac{d^{|\nu+\mu|}}{dz^{\nu+\mu}} k(z_b) \underbrace{\frac{(x-M_{c_1})^\nu}{\nu!}}_{=:\Phi^{c_1}_\nu(x)} \underbrace{\frac{(M_{c_2}-y)^\mu}{\mu!}}_{=:\Psi^{c_2}_\mu(y)}. \tag{4.26}$$

Under the assumption that the kernel $k(\cdot,\cdot)$ is asymptotically smooth, it can again be shown that the Taylor expansion (4.26) satisfies the estimate (4.13) (see [50]).

The Taylor expansion is of the required form (4.14), but it has some additional properties. First, the expansion functions Φ^c_ν and Ψ^c_μ do not depend on the used approximation order p. This is advantageous when a different polynomial degree p is used on different blocks. In this case, the degree p_c can be defined as

$$p_c = \max\{p(b) \,|\, b = (c_1,c_2) \in F \ \wedge\ c \in \{c_1,c_2\}\}.$$

That is, the polynomial degree p on a block is determined to meet a certain accuracy requirement and this defines the degree on the clusters. This is opposed to the definition of $\mathcal{H}^2$-matrices where the rank on the clusters defines the rank on the block.

Second, the index set $J^{\mathcal{P}}_p$ is given by $J^{\mathcal{P}}_p = \{\mu \in \mathbb{N}^3_0 \,|\, |\mu| \le p\}$, i.e. $k^{\mathcal{P}}(p) = \sharp J^{\mathcal{P}}_p = \binom{p+3}{3}$, but the coefficients $\kappa^{(p)}_{\nu,\mu}(c_1,c_2)$ are non zero only if $(\nu,\mu) \in \mathcal{J}_p := \{|\mu+\nu| \le p\} \subsetneq J^{\mathcal{P}}_p \times J^{\mathcal{P}}_p$,

$$\kappa^{(p)}_{\nu,\mu}(c_1,c_2) = \begin{cases} \frac{d^{|\nu+\mu|}}{dz^{\nu+\mu}} k(z_b) & \text{if } (\nu,\mu) \in \mathcal{J}_p, \\ 0 & \text{else,} \end{cases} \tag{4.27}$$

i.e. the matrix $\mathbf{Z}_{(c_1,c_2)}$ is not fully populated and the coefficients $\kappa^{(p)}_{\nu,\mu}(c_1,c_2)$ only depend on the sum $\nu+\mu$ and not on the pair (ν,μ).

As in case of Lagrange polynomials, the expansion system is organized hierarchical

$$\frac{(M_c-y)^\mu}{\mu!} = \sum_{\tilde{\mu}\le\mu} \frac{(M_c-M_{\tilde{c}})^{\mu-\tilde{\mu}}}{(\mu-\tilde{\mu})!} \frac{(M_{\tilde{c}}-y)^{\tilde{\mu}}}{\tilde{\mu}!}, \tag{4.28}$$

i.e. the shift matrix $\mathbf{S}_{\tilde{c},c}$ is given by

$$s^{(\tilde{c},c)}_{\tilde{\mu},\mu} := \begin{cases} \frac{(M_c-M_{\tilde{c}})^{\mu-\tilde{\mu}}}{(\mu-\tilde{\mu})!} & \text{if } \tilde{\mu} \le \mu, \\ 0 & \text{else.} \end{cases} \tag{4.29}$$

That is, the shift matrix $\mathbf{S}_{\tilde{c},c}$ is not fully populated and the entries only depend on $\mu - \tilde{\mu}$.

If the polynomial order $p_{\tilde{c}}$ corresponding to the cluster $\tilde{c}$ is lower than that of its father c, the indices $\tilde{\mu}$ on the right hand side of Equation (4.28) must fulfill the additional condition $|\tilde{\mu}| \leq p_{\tilde{c}}$. That is the higher-order polynomials $\frac{(M_c - y)^{\mu}}{\mu!}$, $p_{\tilde{c}} < |\mu| \leq p_c$, are approximated by lower-order polynomials $\frac{(M_{\tilde{c}} - y)^{\tilde{\mu}}}{\tilde{\mu}!}$, $|\tilde{\mu}| \leq p_{\tilde{c}}$. Restricting the sum on the right hand side of Equation (4.28) to the index set $\{\tilde{\mu} \,|\, \tilde{\mu} \leq \mu \wedge |\tilde{\mu}| \leq p_{\tilde{c}}\}$ defines a modified expansion system $\tilde{\Psi}^c_{\mu}$. Under the condition that the growth of the polynomial order is slow enough, Estimate (4.13) is satisfied also for the modified expansion system (for a proof refer to [50]).

The main advantage of using polynomial interpolation, compared with using Taylor expansion, is its simplicity: only pointwise evaluations of the kernel and of simple polynomials have to be implemented. On the other hand, if we only consider one integral equation the computation of the Taylor expansion has to be done only once, e.g. by using a computer algebra system like Maple. Furthermore, Taylor expansion has some advantages over polynomial interpolation:

First, as mentioned above the asymptotical approximation properties of Taylor expansion and tensor product Lagrangian interpolation of the same polynomial degree are the same. The Taylor polynomial lies in the space $\mathcal{P}_p$, i.e. the number k of terms in the expansion is given by $k^{\mathcal{P}}(p) = \binom{p+3}{3}$ whereas the tensor product polynomials used for interpolation lie in the space $\mathcal{Q}_p$, $k^{\mathcal{Q}}(p) = (p+1)^3$. Table 4.1 shows that Taylor expansion needs much less terms to obtain the same asymptotic convergence rate.

p	0	1	2	3	4	5
$k^{\mathcal{Q}}(p)$	1	8	27	64	125	216
$k^{\mathcal{P}}(p)$	1	4	10	20	35	56

Table 4.1: Size of index sets for $\mathcal{Q}_p$ and $\mathcal{P}_p$

Second, when Taylor expansion is used, many of the coefficients in the matrix $\mathbf{Z}_{(c_1,c_2)}$ are equal to zero (see (4.27)). The number of non-zero entries is given by $\sharp\,\mathcal{J}_p = \binom{p+6}{6}$ whereas in case of polynomial interpolation, $\mathbf{Z}_{(c_1,c_2)}$ is a full matrix of dimension $k^{\mathcal{Q}}(p)$, i.e. it has $k^{\mathcal{Q}}(p)^2$ non-zero entries. This leads to a further reduction in the number of operations for the block multiplication (see first two lines in Table 4.2).

Third, assuming that the kernel is of the form $k(x,y) = k(x-y)$ the coefficients $\kappa^{(k)}_{\nu,\mu}(c_1,c_2)$ only depend on the sum $\nu+\mu$ and not on the pair (ν,μ), which reduces the amount of storage to $k^{\mathcal{P}}(p) = \binom{p+3}{3}$ (see third line in Table 4.2).

p	0	1	2	3	4
$k^{\mathcal{Q}}(p)^2$	1	64	729	4096	15625
$\sharp\,\mathcal{J}_p$	1	7	28	84	210
$k^{\mathcal{P}}(p)$	1	4	10	20	35

Table 4.2: Number of non-zero entries in $\mathbf{Z}$

Fourth, the shift matrices $\mathbf{S}_{\tilde{c},c}$ are not fully populated and the entries only depend on the difference $\mu - \tilde{\mu}$ (see (4.28).

The above reasons lead to large savings in memory and CPU requirements when Taylor expansion is used in comparison with tensor product polynomial interpolation. Therefore,

we have decided to use Taylor expansion. For an efficient implementation which exploits that the matrices $\mathbf{Z}_b$ and $\mathbf{S}_{\tilde{c}_2,c_2}$ are not fully populated and that many of the entries are identical, it is useful to introduce an one-dimensional ordering for the multi-indices (see [33])

$$\mu \prec \nu \quad \Leftrightarrow \quad \Big(|\mu| < |\nu| \vee \Big[|\mu| = |\nu| \wedge (\exists i \; s.t. \, \mu_j = \nu_j \forall j < i \wedge \mu_i < \nu_i)\Big]\Big). \tag{4.30}$$

When this ordering is used, only the non-zero coefficients of the matrices have to be stored in one-dimensional arrays. During the algorithms one has to iterate through the indices, e.g. for the block multiplication in step two of the matrix-vector multiplication, the position of $\mu + \nu$ in the linear ordering (4.30) has to be computed for given μ and ν. This makes the implementation of the algorithms rather complicated in comparison with the standard matrix form.

On the other hand, the testing of the implementation is simple. Only the entries $z_{\nu,\mu}^{(c_1,c_2)}$, $(\nu,\mu) \in \mathcal{J}_p$ of the matrix $\mathbf{Z}_{(c_1,c_2)}$ depend on the kernel. The entries of the matrices $\mathbf{X}_{c_1}$, $\mathbf{Y}_{c_1}$ and $\mathbf{S}_{\tilde{c},c}$ are independent of the kernel. When polynomials of degree p are used as test kernels, Taylor expansion of order p is exact. Therefore, the $\mathcal{H}^2$-approximation is an exact representation of the matrix. Hence, the result for a matrix-vector multiplication obtained by using the algorithm for $\mathcal{H}^2$-matrices for the fast computation of a matrix-vector multiplication has to be the same as when the full representation of the matrix is used.

Error Analysis

Using the $\mathcal{H}^2$-approximation in the farfield in (4.15) introduces an additional error. This error should not increase the discretization error of the Galerkin method. As in case of the cubature error in Section 3.2, the impact of this error can be estimated by the first Strang lemma (Lemma 3.2). This leads to the requirement

$$|\langle (K - \tilde{K})u, v\rangle| \leq TOL \|v\|_{L^2(D)} \|u\|_{L^2(D)} \quad \forall u, v \in V_h, \tag{4.31}$$

with $TOL = C_S O(h^{k+1})$ (k = polynomial degree of the ansatz functions used for the Galerkin discretization). This global error can be split up into local errors on individual blocks. To this end, it is useful to define the restriction of the operator K onto a cluster $c \in T(I)$

$$(K_c u)(x) = \int\limits_{\operatorname{supp} c} k(x,y)u(y)dy. \tag{4.32}$$

For $(c_1, c_2) \in F$, let $\varepsilon_{(c_1,c_2)}$ denote the smallest value such that

$$|\langle (K_{c_2} - \tilde{K}_{c_2})u, v\rangle_{L^2(\operatorname{supp} c_1)}| \leq \varepsilon_{(c_1,c_2)} \|v\|_{L^2(\operatorname{supp} c_1)} \|u\|_{L^2(\operatorname{supp} c_2)} \quad \forall v \in L^2(\operatorname{supp} c_1),\, u \in L^2(\operatorname{supp} c_2). \tag{4.33}$$

This local error can be estimated by using the Estimate (4.13) for the error of the kernel approximation

$$\begin{aligned} |\langle (K_{c_2} - \tilde{K}_{c_2})u, v\rangle_{L^2(\operatorname{supp} c_1)}| &\leq \int\limits_{\operatorname{supp} c_1} \int\limits_{\operatorname{supp} c_2} |v(x)| \underbrace{|k(x,y) - k_b^{(p)}(x,y)|}_{\leq C_1 (C_2\eta)^{p+1} \operatorname{dist}(c_1,c_2)^{-\alpha}} |u(y)| dx\, dy \\ &\leq C_1 (C_2\eta)^{p+1} \sqrt{|\operatorname{supp} c_1||\operatorname{supp} c_2|}\, \operatorname{dist}(c_1,c_2)^{-\alpha} \|v\|_{L^2(\operatorname{supp} c_1)} \|u\|_{L^2(\operatorname{supp} c_2)}. \end{aligned} \tag{4.34}$$

That is

$$\varepsilon_{(c_1,c_2)} \leq C_1(C_2\eta)^{p+1}\sqrt{|\operatorname{supp} c_1||\operatorname{supp} c_2|}\,\operatorname{dist}(c_1,c_2)^{-\alpha}. \tag{4.35}$$

Assuming w.l.o.g. that $\operatorname{diam}(c_1) = \max\{\operatorname{diam}(c_1), \operatorname{diam}(c_2)\}$, gives $\sqrt{|\operatorname{supp} c_1||\operatorname{supp} c_2|} \leq C\operatorname{diam}(c_1)^d$, where d denotes the dimension of the domain. Exploiting the admissibility condition (4.5) yields:

$$\varepsilon_{(c_1,c_2)} \leq C_1(C_2\eta)^{p+1}C(2\eta)^{\alpha}\operatorname{diam}(c_1)^{d-\alpha}. \tag{4.36}$$

The kernels considered in this thesis have the property that $\alpha = d-1$. Therefore, the estimate (4.36) of the error on a single block grows linearly with $\operatorname{diam}(c_1)$ when the same p is used on all blocks.

Remark 4.7. The estimate in (4.36) is only a worst case estimate based on the assumption about the approximation of the kernel (see (4.13)). It does not regard the actual form of the kernel and the separable approximation, e.g. the exponential damping term in the kernels of the integral equations in Chapter 2 does not influence the estimate (4.13). A sharper upper bound can be computed by using Cauchy-Schwarz inequality

$$\varepsilon_{(c_1,c_2)} \leq \Big(\int_{\operatorname{supp} c_1}\int_{\operatorname{supp} c_2} |k(x,y) - k_b^{(p)}(x,y)|^2 dx\,dy\Big)^{\frac{1}{2}}. \tag{4.37}$$

Applying Cauchy-Schwarz inequality might still result in large over estimation. Assuming that v and u are constant an estimate for the error can be computed which might yield better approximations in praxis:

$$\varepsilon_{(c_1,c_2)} \approx \frac{1}{\sqrt{|\operatorname{supp} c_1||\operatorname{supp} c_2|}}\int_{\operatorname{supp} c_1}\int_{\operatorname{supp} c_2} |k(x,y) - k_b^{(p)}(x,y)|dx\,dy. \tag{4.38}$$

The values for the estimates above can be approximated by a quadrature rule. By this, the polynomial degree on every block can be chosen adaptively to meet a certain requirement.

There are different strategies of how to prescribe tolerances for the local errors such that the global error fulfills the condition (4.31). In the following we require that the error on blocks belonging to the same level satisfies the same estimate:

$$\varepsilon_{(c_1,c_2)} \leq \varepsilon_l \quad \forall (c_1,c_2) \in F(l). \tag{4.39}$$

This local estimate leads to the following global estimate

$$\langle (K-\tilde{K})u, v\rangle \leq \underbrace{CC_{\mathrm{sp}}\Big(\sum_{l=0}^{L}\varepsilon_l\Big)}_{=:err}\|v\|_{L^2(D)}\|u\|_{L^2(D)}, \tag{4.40}$$

where C_{sp} is defined in (4.6). (The proof is deferred to Appendix A.7.)

In [50] the polynomial degree on a block is chosen as

$$p_{(c_1,c_2)} = a(l_{\max} - l) + p_0, \qquad \text{for } (c_1,c_2) \in F(l), \tag{4.41}$$

with suitable constants a, $p_0 \in \mathbb{N}$. That is, the polynomial degree only depends on the level of the block and is larger on larger blocks. In [50] it is shown that using a polynomial degree according to (4.41) in Estimate (4.36) for the block error, gives

$$\varepsilon_{(c_1,c_2)} \leq \varepsilon_l = h\hat{C}_1\hat{C}_2^{l_{\max}-l}, \qquad \text{for } (c_1,c_2) \in F(l). \tag{4.42}$$

Using this estimate in (4.40), yields

$$err \leq C_{\text{sp}} \frac{\hat{C}_1}{1-\hat{C}_2} h \overset{!}{\leq} TOL = C_S O(h^{k+1}). \tag{4.43}$$

That is, the constants a and p_0 can be chosen independently of the number of unknowns to yield an asymptotic error of $O(h)$. Furthermore, it is shown in [50] that this choice for the polynomial degree leads to a complexity of $O(n)$ for the storage requirement and for the number of operations needed for a matrix-vector multiplication, i.e. to optimal complexity.

In case of piecewise linear ansatz functions an asymptotic error rate of $O(h^2)$ has to be required. Therefore, the value of p_0 in (4.41) has to be chosen in dependency on n, due to the exponential convergence in Estimate (4.13), it holds $p_0 \sim \log(n)$. In this case the overall complexity is $O(n(\log n)^d)$.

Adaptive cross approximation (ACA)

In the previous subsection a function $g(\cdot,\cdot)$ of two variables is approximated by a sum of products of functions out of a given family of functions of one variable. Instead of this approximation, Bebendorf and Rjasanow [6] suggest to use the values of the function itself. A separable approximation is generated by fixing one of the two variables, i.e. by using functional skeletons.

Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ denote two sets of pairwise distinct points in $\mathbb{R}^d$, D_X and D_Y their convex hulls and g a function $g : D_X \times D_Y \to \mathbb{R}$. For brevity of notation let

$$g(x,[y]_k) = \begin{bmatrix} g(x,y_{\mu_1}) \\ \vdots \\ g(x,y_{\mu_k}) \end{bmatrix} \in \mathbb{R}^k, \quad g([x]_k,y) = \begin{bmatrix} g(x_{\nu_1},y) \\ \vdots \\ g(x_{\nu_k},y) \end{bmatrix} \in \mathbb{R}^k.$$

Then, we are looking for a separable approximation of the form

$$g(x,y) = \underbrace{g(x,[y]_k)^T M_k g([x]_k,y)}_{=:h_k(x,y)} + r_k(x,y). \tag{4.44}$$

A sequence of such approximations can be constructed by the following rule:

$$h_0(x,y) := 0, \qquad r_0(x,y) := g(x,y), \tag{4.45}$$

and for $k = 0, 1, \ldots$

$$h_{k+1}(x,y) := h_k(x,y) + \gamma_{k+1} r_k(x,y_{\mu_{k+1}}) r_k(x_{\nu_{k+1}},y), \tag{4.46}$$

$$r_{k+1}(x,y) := r_k(x,y) - \gamma_{k+1} r_k(x,y_{\mu_{k+1}}) r_k(x_{\nu_{k+1}},y). \tag{4.47}$$

where ν_{k+1}, μ_{k+1} are chosen such that $r_k(x_{\nu_{k+1}},y_{\mu_{k+1}}) \neq 0$ and γ_{k+1} is defined by $\gamma_{k+1} := r_k(x_{\nu_{k+1}},y_{\mu_{k+1}})^{-1}$. The indices ν_{k+1} and μ_{k+1} are called pivot indices.

By construction, h_k is of the required form (4.44). Furthermore, it is easy to see that

$$\begin{aligned} h_k(x,y) + r_k(x,y) &= g(x,y) \qquad \forall k, \\ r_k(x,y_{\mu_l}) = r_k(x_{\nu_l},y) &= 0 \qquad \forall l \leq k, \end{aligned} \tag{4.48}$$

that is, h_k gradually interpolates g. If g is asymptotically smooth and x_{ν_k} is chosen such that

$$|r_{k-1}(x_{\nu_k}, y_{\mu_k})| \geq |r_{k-1}(x, y_{\mu_k})| \quad \forall x \in D_X, \tag{4.49}$$

it is proven in [6] that h_k fulfills an error estimate similar to (4.13), where the constant C_1 grows with growing p. Therefore, the estimate for the asymptotical convergence rate is worse compared with that of polynomial approximations, but in practical applications the ACA yields much better results than polynomial approximations of the same rank. This is illustrated by the following $1D$ Example.

Example 4.5. Let $\gamma = 0.001$ and let the intervals I_x and I_y be given by $I_x = [2, 3]$ and $I_y = [0, 1]$. Figure 4.3 shows the error $k(x, y) - \tilde{k}(x, y)$ when approximating the kernel on the $2D$ domain $I_x \times I_y$, using Taylor expansion and using the ACA to construct the separable kernel $\tilde{k}(x, y)$. The approximation rank is $k = 2$. In every direction 21 equidistant points are used. For the Taylor expansion this is only of interest for plotting the graph, but for the ACA, the chosen points are used in the algorithm, since the ACA can only be used to obtain an approximation in a finite number of points (see considerations below).

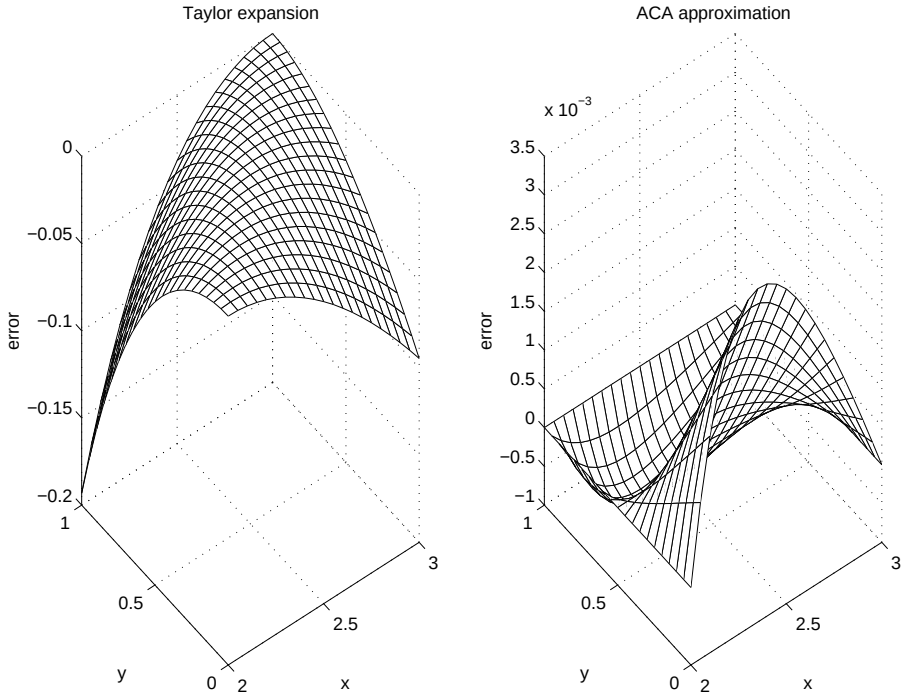

Figure 4.3: Error when approximating the $1D$ RTE-kernel using Taylor expansion (left) and the ACA (right)

It is easily observable which indices have been used as pivot indices in the ACA construction algorithm: $\nu_1 = 1$, $\nu_2 = 21$, $\mu_1 = 11$ and $\mu_2 = 21$. The error in the rows and columns corresponding to these indices is zero. The error for the Taylor expansion is zero for all (x, y) with $x - y = z_0$ and gets larger for $x - y \neq z_0$.

Looking at the scale of the error axis, it stands out that the maximal error when the ACA is used is about two orders of magnitude smaller than when Taylor expansion is used. The L^2-norm of the relative error on the domain $I_x \times I_y$ is given by 0.0070 in case of Taylor expansion and 0.00018 in case of the ACA (about factor 40).

Further numerical results show that Taylor expansion with approximation rank $k = 4$ yields approximately the same error in the L^2-norm as the ACA for $k = 2$. On the other hand, using Taylor expansion of the same rank gives faster algorithms than the ACA because the ACA only leads to a $\mathcal{H}$-matrix instead of a $\mathcal{H}^2$-matrix (see considerations below).

The results strongly depend on the extinction coefficient γ. In case of Taylor expansion the relative error becomes larger for larger γ. The exponential damping leads to strong gradients, which are difficult to approximate by polynomials. Therefore, the ACA has to be used in case of large γ when a high accuracy is required. The dependency of the error on γ, will be examined in more detail in the next section.

In $3D$ the difference between the two methods is not as big as in $1D$. For a similar situation as above the L^2-norm of the relative error for the ACA is about factor 3 smaller than for Taylor expansion. We have chosen to display the error in the $1D$ case, since then the (x, y)-domain is two dimensional and, therefore, well suited for a graphical illustration.

An expansion into polynomials or spherical harmonics leads to a parameterized representation, i.e. once the coefficients in the expansion have been determined, it can be evaluated for arbitrary x and $y \in \mathbb{R}^3$. This is not true for the approximation $h_k(x, y)$. The construction rule (4.46) has to be used for a pointwise evaluation of $h_k(x, y)$. Hence, $h_k(x, y)$ cannot be used to approximate the integral kernel $k(x, y)$. Instead, the algorithm above is applied to construct an approximation $\mathbf{H}_b^{(k)}$ of the matrix $\mathbf{A}_b$ directly. ($\mathbf{A}_b$ denotes the restriction of the matrix $\mathbf{A}$ to the block $b \in F$.) Assuming that the matrix entries are generated by a function g: $a_{ij} = g(x_i, y_j)$, we are looking for an approximation of g in the points (x_i, y_j) only. Hence, the matrix $\mathbf{H}_b^{(k)} = (h_{ij}^{(k)})_{ij}$, given by $h_{ij}^{(k)} = h_k(x_i, y_j)$, can be computed by using only a small number of rows and columns of $\mathbf{A}_b$, namely, the rows corresponding to the indices ν_l and the columns corresponding to the indices μ_l, $l = 1, \ldots, k$. The algorithm can be reformulated such that it returns a low-rank representation (4.1) of the matrix $\mathbf{H}_b^{(k)}$. The fact that the ACA is used to approximate the matrix itself instead of approximating the kernel can be advantageous in some applications (see next section).

Remark 4.8. In [7] a new algorithm for the partitioning of the matrix is introduced. It is based on a modified admissibility condition:

$$\frac{1}{2}\operatorname{diam}(c_2) \leq \eta \operatorname{dist}(c_1, c_2). \tag{4.50}$$

This admissibility condition can be used instead of (4.5) since the ACA error can be estimated by that of polynomial interpolation in y direction only (see [7]). The modified admissibility condition reduces the number of blocks in the farfield dramatically. The disadvantage of this new method is, that it destroys the symmetry of the matrix. Therefore, we use the admissibility condition (4.5) when handling matrices resulting from a Galerkin discretization of an integral equation with a symmetric kernel.

The functional skeleton $g(x, [y]_k)$ in x-direction in the separable approximation (4.44) depends on the points $y_{\mu_1}, \ldots, y_{\mu_k}$ chosen in y-direction. Hence, this kind of approximation does not lead to a uniform $\mathcal{H}$-matrix but to a $\mathcal{H}$-matrix only. That is, on the one hand, the ACA yields a better approximation compared with polynomial expansion when the same approximation order is used. On the other hand, it only leads to a $\mathcal{H}$-matrix, which does not allow as much savings of operations as a $\mathcal{H}^2$-matrix obtained when Taylor expansion is used. Therefore, it is a priori not clear which method yields the better compression factors.

Remark 4.9. In a recent paper [25] a method for computing an optimal $\mathcal{H}^2$-approximation of a given $\mathcal{H}$-matrix in $O(n)$ operations is introduced. Maybe using the ACA to compute a

$\mathcal{H}$-matrix and transforming this matrix into a $\mathcal{H}^2$-matrix yields the best compression rates. We are not aware if numerical tests have been performed in this direction.

This subsection is concluded with some remarks about error estimation and error balancing strategies when the ACA is used. As described above, the method is applied to yield an approximation of the matrix directly instead of the kernel. Hence, the methods for deriving a local error estimate, described in Remark 4.7, cannot be used. The ACA yields a $\mathcal{H}$-matrix, i.e. every block is a low-rank matrix of the form

$$\mathbf{H}_b^{(k)} = \sum_{\nu=1}^{k} \mathbf{x}_\nu \mathbf{y}_\nu^T, \quad \mathbf{x} \in \mathbb{R}^n, \, \mathbf{y} \in \mathbb{R}^m.$$

This representation allows to compute the Frobenius norm of the matrix by using only $(m+n)k^2$ operations:

$$\|\mathbf{H}_b^{(k)}\|_F^2 = \sum_{\nu,\mu=1}^{k} \Big(\sum_{i=1}^{n} (\mathbf{x}_\nu)_i (\mathbf{x}_\mu)_i \Big) \Big(\sum_{j=1}^{m} (\mathbf{y}_\nu)_j (\mathbf{y}_\mu)_j \Big).$$

Therefore, it is rather natural to use the Frobenius norm to measure the error: $\mathbf{R}_b^{(k)} = \mathbf{A}_b - \mathbf{H}_b^{(k)}$. The following criterion may be used to determine the required rank k on each block:

$$\|\mathbf{R}_b^{(k)}\|_F \leq \varepsilon \|\mathbf{A}_b\|_F \quad \forall b \in F. \tag{4.51}$$

The Frobenius norm has the nice property that if the local estimate (4.51) is fulfilled, the same estimate holds for the total matrix without using a further approximation:

$$\|\mathbf{R}\|_F^2 = \sum_{b \in F} \|\mathbf{R}_b\|_F^2 \leq \varepsilon^2 \sum_{b \in F} \|\mathbf{A}_b\|_F^2 = \varepsilon^2 \|\mathbf{A}_F\|_F^2 \leq \varepsilon^2 \|\mathbf{A}\|_F^2. \tag{4.52}$$

Please note that due to the fact that no approximation is used in the nearfield (except for using cubature formulae), the relative error in the total matrix is in general smaller than the local errors.

The stopping criterion (4.51) for determining the approximation rank cannot be used in practical applications since the matrix A_b and, hence, also $\mathbf{R}_b$ is not available. The error can be estimated by using two approximations of different rank: $\mathbf{H}_b^{(k)}$ and $\mathbf{H}_b^{(\hat{k})}$ with $\hat{k} > k$. If the error $\|\mathbf{A}_b - \mathbf{H}_b^{(\hat{k})}\|_F$ is much smaller than $\|\mathbf{A}_b - \mathbf{H}_b^{(k)}\|_F$, then $\|\mathbf{H}_b^{(\hat{k})} - \mathbf{H}_b^{(k)}\|_F$ is a good estimate for the error $\|\mathbf{A}_b - \mathbf{H}_b^{(k)}\|_F$. Since the difference $\mathbf{H}_b^{(\hat{k})} - \mathbf{H}_b^{(k)}$ is a low rank matrix, its Frobenius norm is easily computable. Furthermore, the norm $\|\mathbf{A}_b\|_F$ can be estimated by $\|\mathbf{H}_b^{(\hat{k})}\|_F$. Hence, the non practicable stopping criterion (4.51) might be replaced by the following

$$\|\mathbf{H}_b^{(\hat{k})} - \mathbf{H}_b^{(k)}\|_F \leq \varepsilon \|\mathbf{H}_b^{(\hat{k})}\|_F. \tag{4.53}$$

In practical applications the error estimator described above works much better than the ones described in Remark 4.7. For uniform $\mathcal{H}$-matrices the Representation (4.3) is used instead of (4.1). Therefore, the Frobenius norm of a sub-matrix cannot be computed as efficiently as in case of a $\mathcal{H}$-matrix. For $\mathcal{H}^2$-matrices even the Representation (4.3) is not known explicitly. Therefore, the above approach to estimate the block error in a matrix norm can only be applied if a matrix norm based on matrix-vector multiplications is used. Since computing a matrix norm in this way is costly, the above approach seems to be applicable

only to $\mathcal{H}$-matrices. That is, the fact that the individual sub-matrices of a $\mathcal{H}^2$-matrices share information, complicates the error estimation.

Using the Frobenius norm of the error has two advantages: first, when the low-rank Representation (4.1) is used, the relative error is easy to estimate; second, when computing the impact of the local error on the global error, no additional estimates are needed as opposed to the case when the L^2-norm of the error in the continuous operator is used. (See the procedure used to proof (4.40) in Appendix A.7) . That is, the estimate is more accurate.

On the other hand, one is not really interested in the error measured in the Frobenius norm but in the Euclid norm. These two norms fulfill the estimate:

$$\|\mathbf{M}\|_2 \leq \|\mathbf{M}\|_F \leq \sqrt{n}\|\mathbf{M}\|_2 \qquad \forall \mathbf{M} \in \mathbb{R}^{n\times n} \tag{4.54}$$

Hence, using the Frobenius norm may lead to a large over estimation of the error.

In this section two different methods to prescribe a tolerance for the local error have been introduced. In the first case, the absolute error on the block is related to the norm of the global operator. (Please note that the constant C_s occurring in the first Strang lemma corresponds to the L^2-norm of the operator.) In the second case, the requirement (4.51) is used. This error tolerance is relative to the norm of the matrix restricted to the considered block. If this norm is small, the required error tolerance is unnecessarily restrictive. These considerations will be resumed in the next section when the application to radiative transfer is discussed.

4.3 Application to Radiative Transfer

First of all, we note that all kernels in Chapter 2 are asymptotically smooth, i.e. they satisfy the condition (4.23). Hence, one of the methods discussed in the previous section can be used to obtain a data sparse approximation of the system matrix.

In the following, we assume that the coefficients κ, σ and ρ are constant in space. This is essential, because we want to use Taylor expansion to approximate the kernel. In case of variable coefficients the kernels involve the integrals

$$\tau(x,y) = \|x-y\| \int_0^1 \gamma(tx + (1-t)y)dt. \tag{4.55}$$

Taylor expansion of the kernels (assuming that γ is smooth enough) would depend on γ. Hence, the implementation would have to be changed when another γ is considered. This is not acceptable. Hence, polynomial interpolation or the ACA should be used in case of variable coefficients.

Since σ is assumed to be constant, the kernel k_{11} is translation invariant, i.e. $k_{11}(x,y) = k_{11}(x-y)$. Hence, when Taylor expansion is used, the methods discussed in the previous section can be applied directly to this kernel. The situation for the kernel

$$k_{12}(x,y) = \frac{\rho}{4\pi}\frac{n(y)\cdot(y-x)}{\|x-y\|^3}e^{-\gamma\|x-y\|}$$

is slightly more complicated. A Taylor expansion for a given function has to be computed explicitly. This is acceptable when the function depends on the underlying equation. On the other hand, the function must not depend on the computational domain, because then the

algorithm would have to be changed when another domain is considered. Since the outer normal $n(y)$ depends on the surface, it cannot be included into a Taylor expansion. The situation is very similar when tensor product polynomial interpolation is used. In this case an approximation of a function on a $3D$ parallelepiped is constructed. Since $n(y)$ is only defined on the surface, it cannot be included into the approximation. On the other hand, $n(y)$ only depends on y. Hence, this part of the kernel is already separable

$$k_{12}(x,y) = \frac{\rho}{4\pi} \sum_{j=1}^{3} n_j(y) \frac{(y-x)_j e^{-\gamma\|x-y\|}}{\|x-y\|^3}.$$

Using a separable approximation for each of the three terms in this sum, gives a separable approximation of the whole kernel, but the number of terms in the expansion is increased by a factor three. This expansion is called direct expansion, as opposed to the indirect expansion derived below.

According to Remark 2.3, the integral operator K_{12} can be written as a double layer potential

$$k_{12}(x,y) = -\frac{\rho}{4\pi} \frac{\partial}{\partial n(y)} k_{\text{pot}}(x,y).$$

Hence, a separable approximation of $k_{12}(x,y)$ can be obtained by applying the differential operator $\frac{\partial}{\partial n(y)}$ to an approximation of k_{pot}:

$$\tilde{f}_{ij} = \sum_{(\nu,\mu)\in J_p\times J_p} \kappa^{(p)}_{\nu,\mu}(c_1,c_2) \int_{\operatorname{supp} b_i} \Phi^{c_1}_{\nu}(x) b_i(x)dx \int_{\operatorname{supp} b_j} \underbrace{\frac{\partial}{\partial n(y)} \Psi^{c_2}_{\mu}(y)}_{=:\hat{\Psi}^{c_2}_{\mu}(y)} b_j(y)dy, \tag{4.56}$$

where the coefficients $\kappa^{(p)}_{\nu,\mu}(c_1,c_2)$ are given according to k_{pot}.

Due to the differentiation of the expansion, one order of accuracy is lost in the asymptotically convergence rate: exponent p instead of $p+1$ in (4.13). In [33] the error for the indirect expansion and the direct expansion for the double layer potential of the Laplace operator are computed numerically. These computations show that the indirect expansion yields better results for $\eta \geq 0.3$ and $p \geq 1$. These constraints on the parameters η and p are usually fulfilled when hierarchical approximations are used. Hence, the indirect expansion seems to be preferable. A further advantage of the indirect expansion is that, when Taylor expansion is used, the coefficients $\kappa^{(p)}_{\nu,\mu}(c_1,c_2)$ in (4.56) only depend on the sum $\mu+\nu$ since k_{pot} is translation invariant.

It remains to show that the new expansion functions $\hat{\Psi}^{c_2}_{\mu}(y)$ allow a shifting over the cluster tree. The entries of the row basis $\hat{Y}$ corresponding to this expansion function are given by

$$\hat{y}^{(c)}_{j,\mu} = \int_{\operatorname{supp} b_j} \frac{\partial}{\partial n(y)} \Psi^{c}_{\mu}(y) b_j(y)dy, \text{ for } j \in c.$$

In case of Taylor expansion, it holds

$$\frac{d}{dy_i}\Psi^{c}_{\mu}(y) = \frac{d}{dy_i}\frac{(M_c-y)^{\mu}}{\mu!} = -\frac{(M_c-y)^{\mu-e_i}}{(\mu-e_i)!} = -\Psi^{c}_{\mu-e_i}(y), \text{ for } \mu \geq e_i,$$

where e_i denotes the multi-index in $\mathbb{N}^3$ which has a one at the i-th position and zeros else. Let $\tilde{c}$ denote the son of c that contains the index j. Then it holds

$$\hat{y}_{j,\mu}^{(c)} = \sum_{\substack{1\le i\le 3\\ e_i\le\mu_i}} \int_{\operatorname{supp} b_j} n_i(y)[-\Psi_{\mu-e_i}^{c}(y)]b_j(y)dy$$

$$= -\sum_{\substack{1\le i\le 3\\ e_i\le\mu_i}} \sum_{e_i\le\tilde{\mu}\le\mu} s_{\tilde{\mu}-e_i,\mu-e_i}^{\tilde{c},c} \int_{\operatorname{supp} b_j} n_i(y)\Psi_{\tilde{\mu}-e_i}^{\tilde{c}}(y)b_j(y)dy.$$

According to (4.29), $s_{\tilde{\mu},\mu}^{\tilde{c},c} = s_{\mu-\tilde{\mu}}^{\tilde{c},c}$ and, hence, $s_{\tilde{\mu}-e_i,\mu-e_i}^{\tilde{c},c} = s_{\tilde{\mu},\mu}^{\tilde{c},c}$ is independent of i. Therefore

$$\hat{y}_{j,\mu}^{(c)} = -\sum_{\tilde{\mu}\le\mu} s_{\tilde{\mu},\mu}^{\tilde{c},c} \sum_{\substack{1\le i\le 3\\ e_i\le\tilde{\mu}}} \int_{\operatorname{supp} b_j} n_i(y)\Psi_{\tilde{\mu}-e_i}^{\tilde{c}}(y)b_j(y)dy = -\sum_{\tilde{\mu}\le\mu} s_{\tilde{\mu},\mu}^{\tilde{c},c}\hat{y}_{j,\tilde{\mu}}^{(\tilde{c})}.$$

That is, $\hat{Y}$ is a nested basis and the transfer matrices are except for the sign the same as for Y.

Since the operator K_{21} is, except for a constant factor, the adjoint of K_{12}, the corresponding matrices are transposed. The approximation of the matrix $\mathbf{K}_{12}$ is given by

$$\tilde{\mathbf{K}}_{12} = \mathbf{N} + \sum_{(c_1,c_2)\in F_{12}} \mathbf{X}_{c_1}\mathbf{Z}_{(c_1,c_2)}\hat{\mathbf{Y}}_{c_2}^T.$$

Hence,

$$\tilde{\mathbf{K}}_{12}^T = \mathbf{N}^T + \sum_{(c_1,c_2)\in F_{12}} \hat{\mathbf{Y}}_{c_2}\mathbf{Z}_{(c_1,c_2)}^T\mathbf{X}_{c_1}^T$$

is an $\mathcal{H}^2$-approximation of $\mathbf{K}_{21}$. That is, no additional storage for the operator K_{21} is needed.

By using a direct expansion of the kernel

$$k_{22}(x,y) = \frac{\rho}{\pi}n(x)\cdot(x-y)n(y)\cdot(y-x)\underbrace{\frac{e^{-\gamma\|x-y\|}}{\|x-y\|^4}}_{=:\hat{k}(x,y)},$$

i.e. expanding $\hat{k}$ and multiplying with $n_i(x)(x-y)_i n_j(y)(y-x)_j$, $(i,j)\in\{1,2,3\}^2$, the number of terms in the expansion is increased by a factor nine. To apply a similar trick as for the kernel k_{12}, we need a scalar function k_{pot} such that

$$\frac{\partial^2}{\partial_{x_i}\partial_{y_j}}k_{\mathrm{pot}}(x,y) = \underbrace{(x-y)_i(y-x)_j\hat{k}(x,y)}_{=:h_{ij}(x,y)}, \tag{4.57}$$

which would imply that

$$k_{22}(x,y) = \frac{\rho}{\pi}\frac{\partial^2}{\partial n(x)\partial n(y)}k_{\mathrm{pot}}(x,y).$$

Then the differential operators in x and y direction could be applied to an expansion of k_{pot}. The Equation (4.57) states that $H = (h_{ij})_{i,j}$ is a Hesse matrix. Due to the Theorem of Schwarz, it has to fulfill the condition

$$\partial_j h_{ii} = \partial_i h_{ij}.$$

Since this equation does not hold, a function k_{pot} with the required property cannot exist. Therefore, it is only possible to use a primitive function of $(y-x)\hat{k}(x,y)$ w.r.t. y to get rid of one of the scalar products, i.e. a function k_{pot} such that

$$k_{22}(x,y) = \frac{\rho}{\pi} n(x) \cdot (x-y) \frac{\partial}{\partial n(y)} k_{\text{pot}}(x,y). \tag{4.58}$$

This reduces the number of terms from nine to three, but the symmetry of the kernel $k_{22}(x,y) = k_{22}(y,x)$ is destroyed.

The ACA introduced in Section 4.2 is applied directly to the matrix entries instead of the kernel. Hence, the number of expansion terms does not depend on the structure of the kernel. Therefore, it seems to be preferable to use this method for the approximation of the matrix $\mathbf{K}_{22}$. Since both methods, the $\mathcal{H}^2$-approximation based on Taylor expansion and the ACA, use a partitioning of the index set, it is no problem to use one method on some part of the matrix and the other on some other part. This is not the case, when piecewise linear ansatz functions and "classical" panel clustering is used. This method is based on clusters of elements instead of clusters of basis indices. Therefore, two kinds of clusters would have to be used, which would increase the computational complexity.

The main difference between the integral equations stemming from the RTE and from the BEM is the fact that the first one is a volume integral equation whereas the second one is posed on a $2D$ manifold. In case of a $3D$ computational domain a vertex of the grid has much more neighboring vertices than in the $2D$ case. The same holds for the elements of the grid when piecewise constant ansatz functions are used. Therefore, the number of entries per row in the nearfield matrix is much higher than in the $2D$ case. This gives rise to smaller compression factors. Indeed, the storage needed for the nearfield matrix is often larger than that for the hierarchical approximation of the farfield part.

Error Analysis

When the error analysis is considered, the main difference between the kernels in the integral equation stemming from the RTE and from the BEM is the exponential damping term $e^{-\gamma\|x-y\|}$.

To get a first impression of the dependence of the approximation error on γ, Taylor expansion in the $1D$ case is considered. The formula for the error $k(x,y) - \tilde{k}(x,y)$ derived in Example 4.4 contains the term

$$\Big(\sum_{\nu=0}^{p} \frac{(\gamma\zeta)^\nu}{\nu!}\Big) e^{-\gamma\zeta}.$$

This term is bounded from above by one. This estimate is used in Example 4.4 to obtain the exponential convergence of the separable expansion. For large γ, the term is much smaller than one, resulting in a large over-estimation of the error. Assuming that $\zeta = z_0$, an upper bound $\tilde{\varepsilon}_{(c_1,c_2)}$ for the block error $\varepsilon_{(c_1,c_2)}$ (see (4.33)) can be computed. Assuming further that $\operatorname{diam}(c_1) = \operatorname{diam}(c_2) = \operatorname{dist}(c_1,c_2)$, $\tilde{\varepsilon}_{(c_1,c_2)}$ can be plotted as a function of $\operatorname{diam}(c_1)$. Figure (4.4) shows this functions for different values of γ. The approximation rank $k = 2$ has been used to obtain the plots but the qualitative behavior is similar for other values of k.

For comparison the figure contains also the error for the log-kernel: $k(x,y) = \log(|x-y|)$. This kernel occurs in the BEM for surfaces of $2D$ domains. The error for this kernel grows linearly with the diameter of the cluster. This matches the considerations above and leads to the choice of the expansion order p in [50], i.e. p is larger on larger blocks.

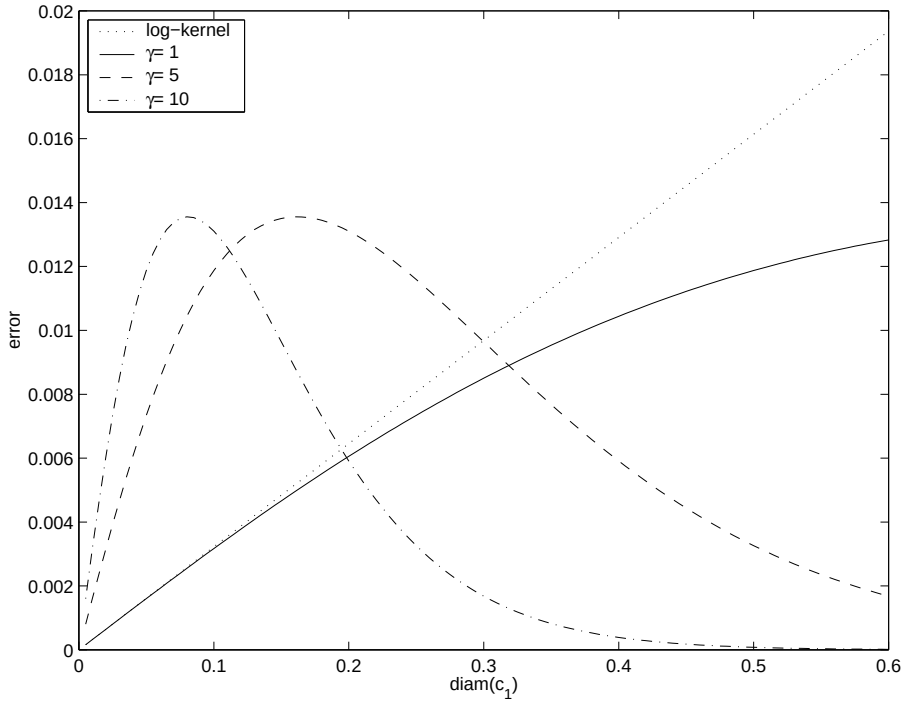

Figure 4.4: Block error as a function of the block size for different values of γ

All curves for $\gamma > 0$ look very similar. For a small size of the cluster, the error grows with growing diameter of the cluster until a maximum is reached. After that the error decays very fast. The size of the maximum is independent of the value of γ, but its location is shifted to smaller values of $\mathrm{diam}(c_1)$ for larger γ. That is, for large γ, there is no need to choose a larger p on larger blocks.

Figure 4.4 shows the absolute error of the kernel approximation. This absolute value has to be somehow related to the norm of the operator to get a meaningful relative quantity. As shown in the last section, there are two different possibilities to do that. First, the absolute error can be related to the operator norm $\|K\|_{L^2}$ of the global operator. This is done when the first Strang lemma is used. $\|K\|_{L^2}$ depends on γ, it grows with growing γ, but it lies in the interval $[0, 1]$ for all γ and the dependence on γ is not too strong, e.g. for $\gamma \in [1, 10]$ as in Figure 4.4, $\|K\|_{L^2}$ is in $(0.84, 1.0)$. Therefore, the situation does not change much when the absolute value is related to the global operator norm instead of considering the absolute error itself. That is, the statement that, for large γ, there is no need to choose a larger p on larger blocks, remains valid.

The second possibility to obtain a relative error, is to relate the error to the norm of the operator restricted to the considered block (see error estimation in case of the ACA). Due to the exponential damping for large γ, the value $k(x, y)$ decays very fast when the distance $|x - y|$ becomes large. It can be shown that this decay is faster than the decay of the error when Taylor expansion is used. Therefore, the relative error grows for larger γ. This is due to the fact that the strong gradients occurring for large γ are difficult to approximate by polynomials. For quantitative results, see the numerical results obtained in the $3D$ case below. Hence, using the local error balancing strategy (4.51) would require that a very large approximation rank is used on large blocks which is unnecessary since the farfield has a very small influence on the behavior of the total operator.

Concluding, it can be said that the exponential damping in case of large γ has two contrary effects on the approximation error when Taylor expansion is used. On the one hand, the strong gradients are difficult to approximate by polynomials. Therefore, the relative error is large. On the other hand, the kernel decays very fast such that the farfield plays only a minor role in comparison with the whole matrix. Therefore, we expect the following. For small γ, the approximation error is large if a small approximation rank k is used, but it decays very fast when k is increased. For large γ, the approximation error is small already for small k, but it cannot be improved significantly by enlarging k. The numerical results below confirm this statement.

The situation is a little different if the ACA is used to approximate the kernel. The ACA yields small relative errors even for large γ (see numerical results below). That is, the fact that the local error balancing strategy (4.51) is too restrictive on large blocks, does not have such a bad influence on the efficiency in case of the ACA.

As mentioned in the previous section, for the ACA as well as for $\mathcal{H}^2$-matrices using Taylor expansion, it is possible to use different approximation ranks on different farfield blocks. Due to the considerations above and in the previous section, the following holds. In the optically thin case, a small approximation rank can be used on small farfield blocks, while larger blocks require a higher order. On the other hand, for an optically thick medium, the approximation rank on the large blocks can be very small on the large farfield blocks whereas for the small blocks a larger approximation rank is required. Due to the small size of these blocks, it is sometimes more efficient to put them into the nearfield, i.e. to compute the corresponding matrix entries exactly. By the different choice of the expansion order in dependence on the optical thickness, one can make use of the different regimes occurring in the radiative heat transfer to construct an effective method.

After these qualitative considerations, some quantitative results obtained for the $3D$ case, with homogeneous Dirichlet boundaries and constant coefficients, are shown. The setting is as follows: the computational domain is the unit cube in $\mathbb{R}^3$. The discretization uses an equidistant Cartesian grid consisting of 32^3 hexahedral elements with piecewise constant ansatz functions. This implies that the system matrix is a block-Toeplitz matrix (see Section 4.1), and the algorithms based of fast Fourier transformation can be used for the exact evaluation of a matrix-vector multiplication.

Taylor expansion and the ACA are used to obtain a hierarchical approximation of the system matrix for different values of γ. The value of the admissibility constant η is chosen to be $\eta = \frac{\sqrt{3}}{2}$ and the minimal size of a farfield block is 64. Different values for the approximation rank k are used. Using Taylor expansion results in a rank of the form $k = \binom{p+3}{3}$ for some $p \in N_0$. For the ACA the rank can be an arbitrary natural number, but since we want to compare the two methods, only k of the above form is used here.

The quality of the approximation is measured by two quantities

$$\text{global error} = \frac{\|(\tilde{\mathbf{F}} - \mathbf{F})\mathbf{u}\|_2}{\|\mathbf{K}\mathbf{u}\|_2}, \qquad \text{local error} = \frac{\|(\tilde{\mathbf{F}} - \mathbf{F})\mathbf{u}\|_2}{\|\mathbf{F}\mathbf{u}\|_2}, \tag{4.59}$$

where $\mathbf{F}$ denotes the part of the matrix $\mathbf{K}$ corresponding to the farfield F. The first quantity is denoted by global error because it relates the approximation error to the entire matrix. This quantity gives the impact of the approximation on the solution of the discrete system by using the first Strang lemma. The second quantity is denoted by local error because it relates the approximation error to the part of the matrix corresponding to the farfield, i.e. to the part where the approximation is used. Due to the considerations above, we expect that this quantity becomes large for large γ in case of Taylor expansion.

The Euclid norm is chosen since this norm is equivalent to the L^2-norm of the error of the continuous operator. This is the norm in which the discretization error of the Galerkin method is measured. The numerical results listed in the Tables 4.3 and 4.4 below always use $\mathbf{u} = (1, \ldots, 1)^T$, but the results are similar for other values of $\mathbf{u}$.

Taylor expansion								
	$\gamma = 0.1$		$\gamma = 1.0$		$\gamma = 10.0$		$\gamma = 100.0$	
k	local	global	local	global	local	global	local	global
1	0.085	0.067	0.095	0.070	0.17	0.051	0.95	8.6e-05
4	0.023	0.017	0.033	0.024	0.14	0.043	0.95	8.6e-05
10	3.2e-03	2.5e-03	4.8e-03	3.5e-03	0.024	7.0e-03	0.76	6.8e-05
20	1.9e-03	1.5e-03	2.6e-03	1.9e-03	0.018	5.1e-03	0.75	6.7e-05
35	4.3e-04	3.3e-04	5.7e-04	4.1e-04	3.6e-03	1.0e-03	0.46	4.2e-05

Table 4.3: Approximation error when using Taylor expansion

ACA								
	$\gamma = 0.1$		$\gamma = 1.0$		$\gamma = 10.0$		$\gamma = 100.0$	
k	local	global	local	global	local	global	local	global
1	0.040	0.031	0.042	0.031	0.067	0.020	0.30	2.7e-05
4	4.4e-03	3.4e-03	3.7e-03	2.6e-03	9.4e-03	2.7e-03	3.5e-04	3.2e-08
10	3.3e-04	2.5e-04	6.0e-04	4.2e-04	2.9e-04	8.5e-05	9.8e-06	8.8e-10

Table 4.4: Approximation error when using ACA

In case of Taylor expansion the convergence is partly not as smooth as expected (e.g. for $\gamma = 10.0$), but the convergence is nevertheless exponential in the polynomial degree p. As expected from the theoretical considerations, the local error for fixed approximation rank is monotonously increasing in γ. Especially for large γ, i.e. $\gamma = 100.0$, the local error is very large, and the convergence rate when increasing the polynomial order is also much smaller than in the optically thin case. That is, Taylor expansion cannot be used when a high accuracy is required in the optically thick case.

On the other hand, due to the exponential damping, the farfield part of the matrix becomes less important in the optically thick case. For $\gamma = 100.0$ the global error for $k = 1$ is smaller than 10^{-4} even though the approximation in the farfield is completely wrong (local error $= 0.95$). Even if the farfield is approximated by zero, the error will be smaller than 10^{-4}. An accuracy of 10^{-4} seems to be good enough, but we will see below that this is not the case in certain situations.

For small γ the difference between the local and the global error becomes smaller. This is due to the fact that the farfield part of the matrix becomes more important. This can be measured e.g. by the Frobenius norm. For this norm, the norm of the global matrix can be obtained by summing up the norms of the different parts of the matrix. For smaller γ the Frobenius norm of the farfield part becomes larger compared with the that of the whole matrix.

When the ACA is used, the error is much smaller than when Taylor expansion of the same rank is used, and the convergence rates are also much smaller. This is especially true for large γ. Even for $\gamma = 100.0$ good approximations measured by the local error are obtained. The error for $k = 1$ is much larger than in the optically thin cases, but the convergence rate

is even smaller. Concluding it can be said that an approximation rank of $k = 10$ is sufficient for arbitrary γ when the ACA is used. Therefore, the Table 4.4 does not contain any results for larger k.

Remark 4.10. For the construction of the hierarchical approximation in the farfield in case of the ACA, certain entries of the Galerkin matrix have to be computed. To this end a quadrature rule is used. Due to the exponential damping, the number of quadrature points has to be sufficiently large in the optically thick case. Therefore, the time for the construction of the approximation is high in this case. If the number of quadrature points is to small, the approximation error becomes larger, but it is still smaller than in case of Taylor expansion.

Before we consider the storage and CPU time requirements of the two methods, we want to examine the impact of the approximation error on the solution of the discrete system more deeply. Remember the integral equation in case of homogeneous Dirichlet boundary conditions and constant coefficients (see (2.7)):

$$(Id - \omega K)G = 4\pi(1-\omega)KB. \tag{4.60}$$

The operator K occurs on the left and and on the right hand side. The situation for the right hand side is covered by the above consideration but for the left hand side, the following relative error should have been used

$$\frac{\|\omega(\tilde{\mathbf{F}} - \mathbf{F})\mathbf{u}\|_2}{\|(\mathbf{M} - \omega\mathbf{K})\mathbf{u}\|_2}. \tag{4.61}$$

For small optical thickness or small ω this error is smaller than the error for evaluating the operator. On the other hand, as shown in Section 3.3, for large optical thickness and $\omega \approx 1$ the system matrix $\mathbf{A} = \mathbf{M} - \omega\mathbf{K}$ is ill-conditioned and therefore small errors in the matrix have a big influence on the result of the matrix-vector multiplication. This effect has already occurred when the impact of errors due to the used cubature formulae has been discussed (see Section 3.4).

Table 4.5 shows the error for a matrix-vector multiplication in case of $\gamma = 100.0$ and $\omega = 0.99$. It follows that Taylor expansion should not be used in this case when an accuracy of less than 10^{-3} is required.

k	Taylor	ACA
1	$2.1 \cdot 10^{-3}$	$6.4 \cdot 10^{-4}$
4	$2.1 \cdot 10^{-3}$	$7.6 \cdot 10^{-7}$
10	$1.6 \cdot 10^{-3}$	$2.1 \cdot 10^{-8}$
20	$1.6 \cdot 10^{-3}$	-
35	$1.0 \cdot 10^{-3}$	-

Table 4.5: Relative error for a matrix-vector multiplication in case of $\gamma = 100.0$ and $\omega = 0.99$

As mentioned in the last section, the ACA only leads to a $\mathcal{H}$-matrix whereas Taylor expansion leads to a $\mathcal{H}^2$-matrix. Therefore, it is not correct to compare the accuracy obtained for the same approximation rank. We have to compare the results obtained when the same amount of resources is used. Table 4.6 shows the storage and CPU time requirements for the ACA and Taylor expansion for different approximation ranks.

Concerning the storage, the amount of memory needed to store all the data for the nearfield matrix and the hierarchical approximation of the farfield matrix is listed in row one, the

second row contains the compression factor achieved in comparison with storing the full matrix.

Concerning the CPU time, the following can be said. The time for constructing the hierarchical approximation of the farfield part of the matrix is much larger in case of the ACA since for the computation of the matrix entries which are used for the approximation $6D$ integrals have to be approximated when the Galerkin method is used for the discretization. As described in the next section, the time for the construction of the approximation is not of big importance if the RTE is coupled with a time dependent process. In this case, the time for the repeated solution of the linear system is much larger. Since iterative solvers are used, performing matrix-vector multiplications is the most time consuming part. To have a meaningful quantity, the time needed for one matrix-vector multiplication by using one of the hierarchical approximations should be compared with the one when the full matrix is used. Since the full matrix needs too much memory to be stored on a computer being available to us, we are not able to perform a matrix-vector multiplication with the full matrix. Therefore, we decided to simply count the number of multiplications necessary to perform one matrix-vector multiplication when one of the approximations is applied and compare this with the one when the full matrix is used.

Taylor			
k	Storage		CPU time
1	54 MB	1.3%	0.7%
4	63 MB	1.5%	1.0%
10	79 MB	1.9%	1.9%
20	103 MB	2.5%	4.1%
35	140 MB	3.4%	9.1%

ACA			
k	Storage		CPU time
1	105 MB	2.5%	1.8%
4	262 MB	6.4%	5.7%
10	507 MB	12.4%	10.7%

Table 4.6: Storage and CPU time requirements when Taylor expansion and the ACA are used

Since using Taylor expansion results in a $\mathcal{H}^2$-matrix, the storage and CPU requirements are much smaller compared with that of the ACA when the same approximation rank is used. In case of Taylor expansion with large k, the compression factor for the storage is much smaller than that for the number of multiplications needed for a matrix-vector multiplication. This is due to the fact that the expansion coefficients that have to be stored only depend on $\nu + \mu$ instead on the pair (ν, μ) (see Section 4.2).

Using the Tables 4.3, 4.4 and 4.6 shows that Taylor expansion seems to be preferable in the optically thin case, say $\gamma \leq 5.0$, especially if the storage requirement is of more importance than the CPU time and if the accuracy requirement is small. In this case, the savings, due to the fact that a $\mathcal{H}^2$-matrix is obtained, compensate the worse accuracy when Taylor expansion is used. For an optically thick medium, say $\gamma \geq 5.0$, the ACA seems to be preferable, especially if a high accuracy is required.

The comparison above uses a constant approximation rank on all farfield blocks. The reason for this is, that only then, the accuracy of the two methods using the same approximation rank can be compared. Furthermore, the convergence rate of the two methods when different ranks are used can be shown. On the other hand, the considerations above imply that it is advantageous to use variable approximation orders. The following numerical results obtained for the case $\gamma = 10.0$ confirm this.

Using the ACA with the stopping criterion (4.53) and a local error tolerance $\varepsilon = 0.02$ yields a global error $6.6 \cdot 10^{-4}$ with a storage requirement of 301 MB. That is, the achieved accuracy is nearly as good as in the case $k = 10 = const$, but the storage requirement is reduced by 40%.

By using Taylor expansion with variable approximation order , a global error $1.2 \cdot 10^{-3}$ is achieved with a storage requirement of 96 MB memory. That is, the achieved accuracy is smaller than for $k = 20 = const$, using a little less storage. The improvement is not as good as in case of the ACA. This is due to the fact that the storage requirements for a $\mathcal{H}^2$-matrix is very low, such that a reduction of the rank does not lead to such big savings of memory as in case of a $\mathcal{H}$-matrix.

4.4 Comparison With DOM: Computational Complexity

As mentioned above, memory and CPU requirements are very large when modeling $3D$ radiative transfer problems. This holds for the DOM which discretizes the RTE directly and also for the discretization of the integral formulation, even when one of the matrix compression methods described in this chapter is used. The effort for the two methods should be compared in this section. Regarding the DOM, the streamline diffusion Galerkin discretization described in Section 1.2 is used. Furthermore, only the case of homogeneous Dirichlet boundary conditions and constant coefficients is considered.

Concerning the CPU time, there are two different tasks which have to be considered:

- the time for assembling the system matrix,
- the time for the iterative solution of the linear system. This time is given by the time needed for one iteration step times the number of iterations. The costs for one iteration step are mainly determined by the costs for a matrix-vector multiplication.

As mentioned in Section 1.2, the system matrix is not stored explicitly in case of the DOM. Therefore, the costs for the "assembly" of the system matrix are small. This is not true for the discretization of the integral equation even when matrix compression methods are used. The computation and storage of the entries in the nearfield matrix is rather time consuming when the Galerkin method is used (evaluation of $6D$ integrals). This is especially true, when the medium is optically thick since then a high quadrature order has to be used (see Section 3.2).

When the radiative heat transfer is coupled with an underlying time dependent process, the same radiative transfer equation has to be solved several times with different sources. In this case the assembly of the system matrix has to be done only once, whereas the iterative solution of the linear system has to be performed several times. Hence, the time for assembling the system matrix becomes less important.

This means that discretizing the integral equation is more suitable when a time dependent process is considered. This is the case, when a heat transfer problem with coupled conduction and radiation is considered. Since radiation changes instantaneously with the underlying temperature distribution, the two processes can be decoupled. In every time step for the solution of the heat equation, the RTE is solved for the given temperature distribution and the computed radiative heat flux is used in the modified heat equation.

For the comparison the same problem as in Section 3.4 is considered. That is, the computational domain is the cube $[-1, 1]^3$, and a uniform grid consisting of 32^3 hexahedral elements

with piecewise trilinear ansatz functions is used for the discretization. In case of the discretization of the integral equation, Taylor expansion is used to obtain a hierarchical approximation of the system matrix since the ACA has not been implemented for piecewise trilinear ansatz functions.

Before the numerical results are presented, we want to make the following remarks:

- The integral formulation is valid only in case of isotropic or linear anisotropic scattering. As mentioned in Section 1.2, an efficient implementation of the DOM can exploit the fact that a simple scattering phase function is used to reduce the number of operations needed for a matrix-vector multiplication with the scattering matrix tremendously. The implementation used for the comparison can handle arbitrary scattering and therefore, does not use the tricks for the reduction of CPU time.
- The implementation of the DOM avoids the explicit computation and storage of the entire system matrix. Instead, the matrix entries are computed only for some large reference elements (see Section 1.2). This leads to a large reduction of the storage requirements. On the other hand, the result of a matrix-vector multiplication has to be computed by summing up the results for all element matrices. Therefore, the time for performing a matrix-vector multiplication is approximately increased by a factor two (see Section 1.2).
- The implementation of the discretization of the integral equation makes use of the fact that a Cartesian grid is used, e.g. there is no need to store the elements of the grid and the corresponding grid points explicitly. The implementation of the DOM works for arbitrary grids consisting of hexahedral elements. That is, hanging nodes are allowed. This is necessary to handle locally refined grids. The implementation of such a grid needs more storage than a simple Cartesian grid.
- Due to Section 3.4 the numerical results obtained by the two methods on the same grid are rather different. The results obtained by the integral formulation seem to be more accurate but this cannot be said with certainty.
- The numerical experiments were performed on different computers. The computations using the integral formulation were performed on a Dual Pentium III 850 MHz. The computations using the DOM were performed on the Pentium III 650 MHz cluster of the Interdisciplinary Center of Scientific Computing (IWR) in Heidelberg. The cluster consists of 16 single nodes with 1 GB memory for each processor. The results were obtained using 10 of these processors. Due to Erik Meinköhn the efficiency of the parallelization is almost 100% for the considered problem, i.e. the time needed when 10 processors are used can be multiplied by 10 to get a feeling what time would have been needed on a single processor.

For all that reasons, the following comparison only gives a rough indicator which method might be preferable in which situation.

Before we compare the resources needed by the DOM and by the discretization of the integral equation using matrix compression, we have to determine the parameters which are needed to obtain a sufficient accuracy. That is, the number of ordinates in case of the DOM and the approximation rank in case of matrix compression. As shown in Section 3.4 and the previous section the choice of the parameters depends on the optical thickness.

We start with the DOM. Table 4.7 depicts the difference between the solutions obtained by using a different number of ordinates. The relative difference in the maximum norm between

the solutions obtained by using $m = 20$ and $m = 80$ ordinates compared with that for $m = 320$ ordinates is displayed.

	$m = 20$	$m = 80$
$\tau = 0.1$	0.045	0.014
$\tau = 10.0$	0.010	6.4e-04

Table 4.7: Difference between the solutions obtained by using $m = 20$ and $m = 80$ ordinates compared with that for $m = 320$ ordinates.

This table shows that it is sufficient to use $m = 80$ directions in the optically thick case, but in the optically thin case at least $m = 320$ directions are needed. (Remember that the choice of the angular discretization described in Section 1.2 implies that m is equal to 20, 80, 320 or 1280.) By using the data for 10 processors, the time for one iteration step on a single processor can be estimated. This results in approximately 40 seconds for $m = 80$ and 200 seconds for $m = 320$.

Concerning the storage, the following can be said. The number of unknowns is very large, namely $n \cdot m$, where n is the number of vertices of the spatial grid and m is the number of ordinate directions. The BiCGSTAB method, which is used to solve the linear system, needs to store six auxiliary vectors of this size, leading to a storage requirement of 132 MB and 526 MB in case of $m = 80$ and $m = 320$ directions respectively. That is, even though the system matrix is not stored explicitly, the needed amount of storage is rather high. Unfortunately, no data concerning the needed overall storage is available to us.

In case of the discretization of the integral equation, the assembly of the system matrix involves two different parts: the computation of the nearfield matrix and the construction of the hierarchical approximation in the farfield. For the first task a quadrature formula has to be used. This involves much more operations than the construction of the approximation in the farfield. Due to the considerations in Section 3.2 the required number of quadrature points depends on the optical thickness. The numerical experiments in Section 3.4 show that the discretization error when 32^3 grid cells are used is approximately $5 \cdot 10^{-3}$. That is, it is sufficient to compute the nearfield entries such that the error when solving the modified system is less than 10^{-3}. Numerical experiments show that to this end $q = 2$ quadrature points in every coordinate direction are sufficient in the optically thin case, $\tau = 0.1$, resulting in a relative error $9.3 \cdot 10^{-4}$, whereas $q = 3$ points are necessary for $\tau = 10.0$ (relative error: $2.8 \cdot 10^{-4}$). The used quadrature order strongly influences the time for the assembly of the matrix: 2950 seconds in case of $q = 2$ and 7320 seconds in case of $q = 3$.

The results in Section 3.4 have been obtained by using the following parameters: admissibility constant $\eta = 1.0$, minimal size of farfield blocks 64 and approximation rank $k = 35$. It cannot be said with certainty how large the approximation error due to this choice of parameters is, but the numerical experiments for the case of piecewise constant elements in the last section and a comparison with the full matrix in case of 16^3 elements indicate that the approximation error is less than 10^{-3} which is smaller than the discretization error.

Using the above parameters leads to a storage requirement of 380 MB, and performing a CG-iteration step needs approximately 3.8 seconds. That is, the storage requirements are much larger compared with the results for piecewise constant elements in the last section (even if we regard the fact that the number of unknowns is increased by 10%). This can be explained as follows. In case of piecewise linear elements the number of entries per row in the nearfield matrix is approximately 870, which is much larger than the approximately 190

entries in case of piecewise constant elements. This is due to the fact that the supports of the basis functions are larger and that they intersect. Therefore, the amount of storage needed for the nearfield matrix is larger than that for the $\mathcal{H}^2$-approximation in the farfield. The ratio between the storage needed for the nearfield and the farfield is approximately $60 : 40$. Thus, the needed storage is much larger than in case of piecewise constant elements. The same holds for the time needed for a matrix-vector multiplication.

Remark 4.11. The nearfield matrix is also the reason for the large memory requirements compared with hierarchical approximations in context of the BEM. Since the computational domain is three dimensional, a vertex of the grid has much more neighboring vertices than in case of a two dimensional manifold, resulting in a large number of entries in the nearfield.

The number of entries in the nearfield matrix can be reduced by enlarging the admissibility constant η. Furthermore, the effort for the farfield approximation can be reduced by using a variable approximation rank. Using $\eta = 2.0$ and a variable approximation rank reduces the storage requirements to 250 MB and the time for one iteration step to 2.1 seconds. The approximation error compared with the solution, obtained for $\eta = 1.0$ and $k = 35$, is $3.6 \cdot 10^{-3}$ in case of $\tau = 0.1$ and $8.5 \cdot 10^{-3}$ in case of $\tau = 10.0$. That is, if the accuracy requirements are not too restrictive, say relative error $\leq 5 \cdot 10^{-3}$, the storage requirements and CPU time can be reduced by a factor one and a half, at least in the optically thin case.

To get the overall time for the solution of the linear system. The time for one iteration step has to be multiplied with the number of iterations. Table 4.8 displays the number of iterations needed to obtain a relative residual of 10^{-4}. (Please remember that the unknowns are the energies $G(x_i)$ in case of the integral equation and the intensities $I(x_i, \Omega_j)$ in case of the DOM. Therefore, the same relative residual does not mean that the same accuracy is achieved, but it should be almost the same.) When the integral equation is used, much less iterations are needed, especially when the medium is optically thin. In this case, preconditioning the system by solving the transport equation without scattering, should improve the speed of convergence in case of the DOM.

	integral	DOM
$\tau = 0.1$	9	59
$\tau = 10.0$	7	26

Table 4.8: Number of iterations needed to obtain a relative residual of 10^{-4} for different optical thicknesses

Concluding, the following can be said about the storage and CPU time requirements of the two methods. Even though the system matrix is not stored explicitly in case of the DOM (see Section 1.2), the storage requirements are comparable with that of the matrix compression methods.

Having in mind the remarks at the beginning of this section about the difficulties of the comparison of the computing time, especially, regarding the fact that a faster computer is used in case of the discretization of the integral equation, we can nevertheless say the following. The assembly of the system matrix needs more time in case of the integral equation, especially in the optically thick case where a high order quadrature rule has to be used for the computation of the nearfield entries. On the other hand, the time for performing one iteration step of a Krylov method is about one order of magnitude smaller in case of the integral equation, at least if $m = 320$ directions are used for the DOM. Furthermore, the number of iteration steps needed to obtain the same accuracy is considerably smaller.

Conclusions

In this work radiative transfer in scattering media is described by a system of weakly singular integral equations of the second kind. The unknowns are the energy G, which is defined on the whole domain, and the outgoing radiative heat flux q_{out}, which is only defined on the boundary. It is shown that the integral formulation constitutes a well posed problem. That is, it has a unique solution which continuously depends on the input data. Furthermore, the Galerkin as well as the collocation method lead to a stable discretization of the equation.

Since the assembly of the corresponding system matrix involves the evaluation of integrals that cannot be computed analytically, cubature formulae have to be used to approximate the matrix entries. By this, an additional error besides the discretization error of the method is introduced. Furthermore, to yield reasonable requirements of storage and computing time, it is necessary to use matrix compression methods which perform a matrix-vector multiplication only approximately. This also leads to additional errors. Although these errors are small, they can cause large errors in the solution of the linear system when the condition number of the matrix is large. This is the case when the medium is optically thick and strongly scattering. Therefore, it is not recommendable to use the discretization of the integral equation in this case.

The DOM which discretizes the RTE directly, does not have this problem. If the DOM is interpreted as a discretization of the integral equation, it can be seen that the critical eigenvalues of the matrix are approximated much better than in case of discretizing the integral equation using low order quadrature formulae or matrix compression methods. Furthermore, only a small number of ordinate directions is needed to yield accurate results in the optically thick case.

On the other hand, in the optically thin case, small errors in the entries of the matrix do not have a big influence on the solution. Therefore, discretizing the integral equation and using a matrix compression method yields good results. On the other hand, when the DOM is used, a large number of ordinates is needed to obtain accurate results in the optically thin case, especially if the geometry is anisotropic. Furthermore, the numerical results in Section 3.4 indicate that the DOM does not yield accurate results if there are local heat sources, even if the medium is optically thick. Subsuming our observations, it can be said that using the DOM is preferable in the optically thick case, especially if the medium is strongly scattering and if there are no local heat sources, whereas using the discretization of the integral equation is preferable in the optically thin case or if there are local sources.

The discretization of an integral equation leads to a full matrix even if basis functions with local support are used. Since the dimension of the matrix is large, it is not recommendable to assemble and store the entire matrix. Instead, matrix compression methods have to be used. In case of the integral equation being derived from the RTE, the efficiency of these methods strongly depends on the optical thickness. The exponential damping term in the integral kernels has two contrary effects on the approximation error when Taylor expansion

is used to obtain a separable approximation. On the one hand, the strong gradients in the optically thick case are difficult to approximate by polynomials. Therefore, the relative error is large. On the other hand, the kernels decay very fast such that the farfield plays only a minor role in comparison with the whole matrix, i.e. in extreme cases the farfield can be neglected completely.

In the optically thin case, the exponential damping has only a minor effect. Therefore, the expansion order should be chosen as follows: for small farfield blocks, a lower order is sufficient, while larger blocks require a higher order. By the different choice of the expansion order in dependence on the optical thickness, one can make use of the different regimes occurring in the radiative heat transfer to construct an effective method.

When the CPU time requirements are compared with that of the DOM, the following can be said. The assembly of the system matrix for the integral equation needs more time than for the DOM even if matrix compression methods are used. This is worthwhile if the radiative transfer for the same problem has to be calculated several times, like it is the case if the sources change due to an underlying time dependent process.

When iterative methods are used for the solution of the linear system, the number of steps needed to obtain a certain accuracy is considerably smaller in case of the integral equation than in case of the DOM. This is already true for the optically thick case but the difference is even more important for the optically thin one.

The time needed for one iteration step is mainly determined by the time needed to perform a matrix-vector multiplication. This time depends on the parameters used. That is, on the number of ordinates in case of the DOM and on the approximation rank in case of the matrix compression methods. In the optically thin case the DOM needs many directions to yield accurate results leading to large requirements in computing time. Hence, matrix compression methods are much faster in this case. Furthermore, even in the optically thick case, where less directions are needed for the DOM, a matrix-vector multiplication still needs considerably less time when matrix compression methods are used. On the other hand, the time for a matrix-vector multiplication in case of the implementation of the DOM could be reduced by optimizing the code for isotropic scattering.

Concluding, it can be said that the discretization of the integral equation using matrix compression methods may lead to improvements in accuracy and computing time as compared with the DOM, especially in the optically thin case and if there are local heat sources.

Appendix A

Mathematical Means

For the sake of completeness, this chapter contains some important definitions and assertions which have been used in the previous chapters.

A.1 Characterization of Surfaces

Definition A.1. Let $D \subset \mathbb{R}^d$ be a bounded domain and let $\Gamma = \partial D$ denote its boundary. Γ is said to be in $C^{k,\alpha}$ if there exists a neighborhood $U_x \subset \Gamma$ of x such that U_x is the image of an injective mapping $\Psi : \mathbb{R}^{d-1} \supset B_1(0) \to U_x$ with $\Psi \in C^{k,\alpha}(B_1(0))$.

An important property of smooth surfaces is the following.

Lemma A.1. *Let $\Gamma \in C^{1,\alpha}$. Then there exits a constant C_Γ such that*

$$|(y-x)\cdot n(y)| \le C_\Gamma \|x-y\|^{1+\alpha} \quad \forall x, y \in \Gamma. \tag{A.1}$$

Proof. refer to [24] □

A.2 Spherical Harmonics

Let $n \in \mathbb{N}_0$ and $|m| \le n$. The Legendre polynomials are defined by

$$P_n(t) = \frac{1}{2^n n!}\Big(\frac{d}{dt}\Big)^n (t^2-1)^n, \quad t \in \mathbb{R}, \tag{A.2}$$

the associated Legendre functions by

$$P_n^{|m|}(t) = (1-t^2)^{\frac{|m|}{2}}\Big(\frac{d}{dt}\Big)^{|m|} P_n(t), \quad t \in [-1,1], \tag{A.3}$$

and the spherical harmonics by

$$Y_n^m(\Omega) := \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos\Theta) e^{im\varphi}, \quad \Omega = \begin{pmatrix} \sin\Theta\cos\varphi \\ \sin\Theta\sin\varphi \\ \cos\Theta \end{pmatrix} \in S^2. \tag{A.4}$$

The following addition theorem holds:

$$\frac{2n+1}{4\pi} P_n(\Omega\cdot\Omega') = \sum_{m=-n}^{n} Y_n^m(\Omega) Y_n^{-m}(\Omega'), \quad \Omega, \Omega' \in S^2 \tag{A.5}$$

Proof. refer to [19] □

A.3 Differentiation of a Parameter Integral

Lemma A.2. *Let $f(x,y) \in C(D \times D)\backslash\{(x,y) \mid \|x-y\| \leq r_0\}$, then it holds for $r > r_0$*

$$\frac{\partial}{\partial x_i} \int\limits_{D\backslash B_r(x)} f(x,y)dy = \int\limits_{D\backslash B_r(x)} \frac{\partial f(x,y)}{\partial x_i} dy + \int\limits_{\partial B_r(x)} f(x,y)\frac{x_i - y_i}{r} do(y) \tag{A.6}$$

Proof. refer to [56] □

A.4 The Mapping $F_{\mathrm{col_d}}$

Let $d \in N$. Let $\mathrm{col_d} : \{2,\ldots,d\} \to \{1,\ldots,d-1\}$ be a mapping with $\mathrm{col_d}(i) \leq i-1$. Let the numbers $c_{i,j}$, $i = 2,\ldots,d$, $j = 1,\ldots,i-1$ be defined via

$$c_{i,j} := \begin{cases} 1 & \text{if } j = \mathrm{col_d}(i), \\ 0 & \text{else.} \end{cases} \tag{A.7}$$

and $e_{i,j}$, $i = 3,\ldots,d$, $j = 1,\ldots,i-2$ be recursively defined via

$$e_{i,i-2} := c_{i,i-1}, \tag{A.8}$$

$$e_{i,j} := c_{i,j+1} + \sum_{k=j+1}^{i-2} c_{k+1,j+1} e_{i,k}, \quad j = i-3,\ldots,1, \tag{A.9}$$

We are going to proof that the mapping $F_{\mathrm{col_d}}$ defined in Section 3.2 fulfills the prerequisite of the substitution rule. To this end, some preliminary results are needed.

From (A.7) follows

$$\sum_{j=1}^{i-1} c_{i,j} = 1 \tag{A.10}$$

The definition of the e_{ij} implies the following relationship for $c_{i,1}$:

$$\sum_{k=1}^{i-2} c_{k+1,1} e_{i,k} = 1 - c_{i,1}, \quad i \in \{2,\ldots,d\}. \tag{A.11}$$

Proof. Using the equality (see (A.10))

$$c_{k+1,1} = 1 - \sum_{j=2}^{k} c_{k+1,j}$$

yields

$$\begin{aligned} \sum_{k=1}^{i-2} c_{k+1,1} e_{i,k} &= \sum_{k=1}^{i-2} e_{i,k} - \sum_{k=1}^{i-2}\sum_{j=2}^{k} c_{k+1,j} e_{i,k} \\ &= \sum_{k=1}^{i-2} e_{i,k} - \sum_{j=2}^{i-2} \underbrace{\sum_{k=j}^{i-2} c_{k+1,j} e_{i,k}}_{\overset{(A.9)}{=} e_{i,j-1} - c_{i,j}} \overset{(A.10)}{=} e_{i,i-2} + 1 - c_{i,1} - c_{i,i-1} \overset{(A.8)}{=} 1 - c_{i,1} \end{aligned}$$

□

The assertion above is needed to prove the following lemma.

Lemma A.3. *Let*

$$F_{\mathrm{col_d}} : \left\{ \begin{array}{ccc} \mathbb{R}^d & \to & \mathbb{R}^d \\ (\omega_1, \eta_1, \ldots, \eta_{d-1}) & \to & (z_1, \ldots, z_d) \end{array} \right. ,$$

be defined by

$$\begin{aligned} z_1 &:= \omega_1, \\ z_i &:= \omega_1 \prod_{j=1}^{i-2} \eta_j^{e_{ij}} \eta_{i-1}, \quad i = 2, \ldots, d. \end{aligned}$$

Then the following equality holds for $z = F_{\mathrm{col_d}}(\omega_1, \eta)$:

$$z_i = \eta_{i-1} \prod_{j=1}^{i-1} z_j^{c_{i,j}} = \eta_{i-1} z_{\mathrm{col_d}(i)}, \quad i = 2, \ldots, d, \tag{A.12}$$

or equivalently, if $z_i \neq 0 \, \forall i$

$$\eta_j = z_{j+1} \prod_{k=1}^{j} z_k^{-c_{j+1,k}} = \frac{z_{j+1}}{z_{\mathrm{col_d}(j+1)}}, \quad j = 1, \ldots, d-1. \tag{A.13}$$

Proof by induction. $i = 2$:

$$z_2 = \omega_1 \eta_1 = z_1 \eta_1 \quad (\text{note: } c_{2,1} = 1).$$

$i - 1 \to i$:

$$\begin{aligned} z_i \overset{\text{def.}}{=} \omega_1 \eta_{i-1} \prod_{j=1}^{i-2} \eta_j^{e_{i,j}} &\overset{(A.13)}{=} z_1 \eta_{i-1} \prod_{j=1}^{i-2} \Big(z_{j+1} \prod_{k=1}^{j} z_k^{-c_{j+1,k}} \Big)^{e_{i,j}} \\ &= z_1 \eta_{i-1} \Big(\prod_{j=1}^{i-2} z_{j+1}^{e_{i,j}} \Big) \quad \Big(\prod_{k=1}^{i-2} \prod_{j=k}^{i-2} z_k^{-c_{j+1,k} e_{i,j}} \Big) \\ &= \eta_{i-1} z_1^{1 - \sum_{j=1}^{i-2} c_{j+1,1} e_{i,j}} \prod_{k=2}^{i-2} z_k^{e_{i,k-1} - \sum_{j=k}^{i-2} c_{j+1,k} e_{i,j}} z_{i-1}^{e_{i,i-2}} = \eta_{i-1} \prod_{k=1}^{i-1} z_k^{c_{i,k}} \end{aligned}$$

The last equality holds due to (A.11), (A.8) and (A.9). □

Lemma A.4. *$F_{\mathrm{col_d}} : S^{(1)} \times C^{(d-1)} \to P_{\mathrm{col_d}}$ is injective, $\det(F'_{\mathrm{col_d}}) \neq 0$, and $F_{\mathrm{col_d}}(S^{(1)} \times C^{(d-1)}) = P_{\mathrm{col_d}}$.*

Proof. $F_{\mathrm{col_d}}$ is injective due to its triangular structure: the first component depends only on ω_1, the second on ω_1 and η_1 and so on. Concerning the determinant of the Jacobian, it holds

$$\det(F'_{\mathrm{col_d}}) = \prod_{j=1}^{d-2} \eta_j^{\sum_{i=j+2}^{d} e_{i,j}} > 0, \quad \text{since } \eta_j > 0.$$

It holds $F_{\mathrm{col_d}}(S^{(1)} \times C^{(d-1)}) \subset P_{\mathrm{col_d}}$, since:
for $(\Omega_1, \eta) \in F_{\mathrm{col_d}}(S^{(1)} \times C^{(d-1)})$ and $z := F_{\mathrm{col_d}}(\omega_1, \eta)$ the Equality (A.12) implies

$$\begin{aligned} &0 < z_1 = \omega_1 < 1, \\ &0 < z_i = z_{\mathrm{col_d}(i)}\eta_{i-1} < z_{\mathrm{col_d}(i)}, \quad i = 2, \ldots, d, \end{aligned}$$

i.e. $z \in P_{\mathrm{col_d}}$.

It holds $F_{\mathrm{col_d}}(S^{(1)} \times C^{(d-1)}) \supset P_{\mathrm{col_d}}$, since:
for $z \in P_{\mathrm{col_d}}$, define $0 < \omega_1 := z_1 < 1$ and $0 < \eta_j := \frac{z_{j+1}}{z_{\mathrm{col_d}(j+1)}} < 1$, $j = 1, \ldots, d-1$. Then $z = F_{\mathrm{col_d}}(\omega_1, \eta)$ due to (A.12). □

A.5 Regularizing Coordinate Transformations

In this section the regularizing coordinate transformations for the remaining cases are specified.

1. (c) The Case of a Common Edge

After introducing the relative coordinate $\hat{z}_1 = \hat{x}_1 - \hat{y}_1$ and interchanging the order of integration, the resulting domain can be split into two parts to which a $5D$ Duffy transformation can be applied. The transformations result in the following integral with analytic integrand:

$$I^{(1)}_{\tau,t}(u,v) = \sum_{j=1}^{2} \int_{C^{(4)}} \int_{S^{(2)}} \omega_1^4 f^{(j)}(\eta)\hat{v}(\hat{x}^{(j)})\hat{k}(\hat{x}^{(j)}, \hat{y}^{(j)})\hat{u}(\hat{y}^{(j)}) d\omega d\eta, \tag{A.14}$$

where the transformations $(\omega, \eta) \to (\hat{x}^{(j)}, \hat{y}^{(j)})$ are given by:

$$\begin{pmatrix} \hat{x}_1^{(1)} \\ \hat{x}_2^{(1)} \\ \hat{x}_3^{(1)} \\ \hat{y}_1^{(1)} \\ \hat{y}_2^{(1)} \\ \hat{y}_3^{(1)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 \\ \omega_1 \\ \eta_1\omega_1 \\ (1-\eta_2)\omega_1 + \omega_2 \\ \eta_3(1-\eta_2)\omega_1 \\ \eta_4\eta_3(1-\eta_2)\omega_1 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(2)} \\ \hat{x}_2^{(2)} \\ \hat{x}_3^{(2)} \\ \hat{y}_1^{(2)} \\ \hat{y}_2^{(2)} \\ \hat{y}_3^{(2)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 \\ \eta_1\omega_1 \\ \eta_1\eta_2\omega_1 \\ (1-\eta_3)\omega_1 + \omega_2 \\ (1-\eta_3)\omega_1 \\ (1-\eta_3)\eta_4\omega_1 \end{pmatrix},$$

and $f^{(j)}(\eta)$ is given by

$$f^{(1)}(\eta) = (1-\eta_2)^2\eta_3, \quad f^{(2)}(\eta) = (1-\eta_3)\eta_1$$

1. (d) The Case of a Common Vertex

In this case only the splitting in $\hat{y}_1 < \hat{x}_1$ and $\hat{y}_1 > \hat{x}_1$ is needed. A $6D$ Duffy transformation yields:

$$I^{(1)}_{\tau,t}(u,v) = \int_{C^{(2)}} \int_{S^{(4)}} \omega_1^5\eta_1^2\eta_2\eta_3\hat{v}(\hat{x}^{(1)})\hat{k}(\hat{x}^{(1)}, \hat{y}^{(1)})\hat{u}(\hat{y}^{(1)}) d\omega_1 d\eta, \tag{A.15}$$

where the transformation $(\omega_1, \eta) \to (\hat{x}^{(1)}, \hat{y}^{(1)})$ is given by:

$$\begin{pmatrix} \hat{x}_1^{(1)} \\ \hat{x}_2^{(1)} \\ \hat{x}_3^{(1)} \\ \hat{y}_1^{(1)} \\ \hat{y}_2^{(1)} \\ \hat{y}_3^{(1)} \end{pmatrix} = \begin{pmatrix} \omega_1 \\ \eta_2\omega_1 \\ \eta_2\eta_4\omega_1 \\ \eta_1\omega_1 \\ \eta_1\eta_3\omega_1 \\ \eta_1\eta_3\eta_5\omega_1 \end{pmatrix}.$$

2. The case of one volume and one surface element

In the following we consider the case that τ is a volume element and t is a surface element. In this case the situation is not symmetric in x and y. Hence, the splitting of the integral $I_{\tau,t}$ into the two parts $\hat{y}_1 < \hat{x}_1$ and $\hat{y}_1 > \hat{x}_1$ (see Remark 3.2) does not yield two cases which can be handled by simply interchanging the roles of x and y. Instead, both cases have to be handled separately.

2. (a) The Case of a Common Face

After introducing the relative coordinates $\hat{z}_i = \hat{x}_i - \hat{y}_i$, $i = 1, 2$, and interchanging the order of integration, the resulting domain can be split into seven parts to which a $3D$ Duffy transformation can be applied. The transformations result in the following integral with analytic integrand:

$$I_{\tau,t}(u,v) = \sum_{j=1}^{7} \int_{C^{(2)}} \int_{S^{(3)}} \omega_1^2 f^{(j)}(\eta)\hat{v}(\hat{x}^{(j)})\hat{k}(\hat{x}^{(j)}, \hat{y}^{(j)})\hat{u}(\hat{y}^{(j)}) d\omega d\eta, \tag{A.16}$$

where the transformations $(\omega, \eta) \to (\hat{x}^{(j)}, \hat{y}^{(j)})$ are given by:

$$\begin{pmatrix} \hat{x}_1^{(1)} \\ \hat{x}_2^{(1)} \\ \hat{x}_3^{(1)} \\ \hat{y}_1^{(1)} \\ \hat{y}_2^{(1)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 \\ \eta_1\omega_1 + \omega_2 \\ \eta_1\omega_1 \\ (1-\eta_2)\omega_1 + \omega_2 + \omega_3 \\ (1-\eta_2)\omega_1 + \omega_2 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(2)} \\ \hat{x}_2^{(2)} \\ \hat{x}_3^{(2)} \\ \hat{y}_1^{(2)} \\ \hat{y}_2^{(2)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 \\ \omega_1 + \omega_2 \\ \eta_1\omega_1 \\ (1-\eta_2)\omega_1 + \omega_2 + \omega_3 \\ \omega_2 \end{pmatrix},$$

$$\begin{pmatrix} \hat{x}_1^{(3)} \\ \hat{x}_2^{(3)} \\ \hat{x}_3^{(3)} \\ \hat{y}_1^{(3)} \\ \hat{y}_2^{(3)} \end{pmatrix} = \begin{pmatrix} (1-\eta_1)\omega_1 + \omega_2 + \omega_3 \\ (1-\eta_1)\omega_1 + \omega_2 \\ (1-\eta_1)\omega_1 \\ \omega_1 + \omega_2 + \omega_3 \\ \eta_2\omega_1 + \omega_2 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(4)} \\ \hat{x}_2^{(4)} \\ \hat{x}_3^{(4)} \\ \hat{y}_1^{(4)} \\ \hat{y}_2^{(4)} \end{pmatrix} = \begin{pmatrix} \eta_1\omega_1 + \omega_2 + \omega_3 \\ \eta_1\omega_1 + \omega_2 \\ \eta_1\eta_2\omega_1 \\ \omega_1 + \omega_2 + \omega_3 \\ \omega_2 \end{pmatrix},$$

$$\begin{pmatrix} \hat{x}_1^{(5)} \\ \hat{x}_2^{(5)} \\ \hat{x}_3^{(5)} \\ \hat{y}_1^{(5)} \\ \hat{y}_2^{(5)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 \\ \omega_1 + \omega_2 \\ \omega_1 \\ (1-\eta_1\eta_2)\omega_1 + \omega_2 + \omega_3 \\ (1-\eta_1)\omega_1 + \omega_2 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(6)} \\ \hat{x}_2^{(6)} \\ \hat{x}_3^{(6)} \\ \hat{y}_1^{(6)} \\ \hat{y}_2^{(6)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 \\ \eta_1\omega_1 + \omega_2 + \\ \eta_1\eta_2\omega_1 \\ \omega_2 + \omega_3 \\ \omega_2 \end{pmatrix},$$

$$\begin{pmatrix} \hat{x}_1^{(7)} \\ \hat{x}_2^{(7)} \\ \hat{x}_3^{(7)} \\ \hat{y}_1^{(7)} \\ \hat{y}_2^{(7)} \end{pmatrix} = \begin{pmatrix} (1-\eta_1\eta_2)\omega_1 + \omega_2 + \omega_3 \\ (1-\eta_1)\omega_1 + \omega_2 \\ (1-\eta_1)\omega_1 \\ \omega_1 + \omega_2 + \omega_3 \\ \omega_1 + \omega_2 \end{pmatrix},$$

and $f^{(j)}(\eta)$ is given by

$$f^{(1)}(\eta) = f^{(2)}(\eta) = f^{(3)}(\eta) = 1, \quad f^{(4)}(\eta) = f^{(5)}(\eta) = f^{(6)}(\eta) = f^{(7)}(\eta) = \eta_1.$$

2. (b) The Case of a Common Edge

After introducing the relative coordinate $\hat{z}_1 = \hat{x}_1 - \hat{y}_1$ and interchanging the order of integration, the resulting domain can be split into four parts to which a $4D$ Duffy transformation can be applied. The transformations result in the following integral with analytic integrand:

$$I_{\tau,t}(u,v) = \sum_{j=1}^{4} \int_{C^{(3)}} \int_{S^{(2)}} \omega_1^3 f^{(j)}(\eta) \hat{v}(\hat{x}^{(j)}) \hat{k}(\hat{x}^{(j)}, \hat{y}^{(j)}) \hat{u}(\hat{y}^{(j)}) d\omega d\eta, \tag{A.17}$$

where the transformations $(\omega, \eta) \to (\hat{x}^{(j)}, \hat{y}^{(j)})$ are given by:

$$\begin{pmatrix} \hat{x}_1^{(1)} \\ \hat{x}_2^{(1)} \\ \hat{x}_3^{(1)} \\ \hat{y}_1^{(1)} \\ \hat{y}_2^{(1)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 \\ \omega_1 \\ \eta_1\omega_1 \\ (1-\eta_2)\omega_1 + \omega_2 \\ (1-\eta_2)\eta_3\omega_1 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(2)} \\ \hat{x}_2^{(2)} \\ \hat{x}_3^{(2)} \\ \hat{y}_1^{(2)} \\ \hat{y}_2^{(2)} \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 \\ \eta_1\omega_1 \\ \eta_1\eta_2\omega_1 \\ (1-\eta_3)\omega_1 + \omega_2 \\ (1-\eta_3)\omega_1 \end{pmatrix},$$

$$\begin{pmatrix} \hat{x}_1^{(3)} \\ \hat{x}_2^{(3)} \\ \hat{x}_3^{(3)} \\ \hat{y}_1^{(3)} \\ \hat{y}_2^{(3)} \end{pmatrix} = \begin{pmatrix} (1-\eta_1)\omega_1 + \omega_2 \\ (1-\eta_1)\eta_2\omega_1 \\ (1-\eta_1)\eta_2\eta_3\omega_1 \\ \omega_1 + \omega_2 \\ \omega_1 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(4)} \\ \hat{x}_2^{(4)} \\ \hat{x}_3^{(4)} \\ \hat{y}_1^{(4)} \\ \hat{y}_2^{(4)} \end{pmatrix} = \begin{pmatrix} (1-\eta_2)\omega_1 + \omega_2 \\ (1-\eta_2)\omega_1 \\ (1-\eta_2)\eta_3\omega_1 \\ \omega_1 + \omega_2 + \omega_3 \\ \eta_1\omega_1 \end{pmatrix},$$

and $f^{(j)}(\eta)$ is given by

$$f^{(1)}(\eta) = f^{(4)}(\eta) = 1 - \eta_2 \quad f^{(2)}(\eta) = \eta_1 \quad f^{(3)}(\eta) = (1-\eta_1)^2\eta_2.$$

2. (c) The Case of a Common Vertex

In this case again only the splitting in $\hat{y}_1 < \hat{x}_1$ and $\hat{y}_1 > \hat{x}_1$ is needed. A $5D$ Duffy transformation yields:

$$I_{\tau,t}(u,v) = \sum_{j=1}^{2} \int_{C^{(4)}} \int_{S^{(1)}} \omega_1^4 f^{(j)}(\eta) \hat{v}(\hat{x}^{(j)}) \hat{k}(\hat{x}^{(j)}, \hat{y}^{(j)}) \hat{u}(\hat{y}^{(j)}) d\omega d\eta, \tag{A.18}$$

where the transformations $(\omega_1, \eta) \to (\hat{x}^{(j)}, \hat{y}^{(j)})$ are given by:

$$\begin{pmatrix} \hat{x}_1^{(1)} \\ \hat{x}_2^{(1)} \\ \hat{x}_3^{(1)} \\ \hat{y}_1^{(1)} \\ \hat{y}_2^{(1)} \end{pmatrix} = \begin{pmatrix} \omega_1 \\ \eta_2\omega_1 \\ \eta_2\eta_4\omega_1 \\ \eta_1\omega_1 \\ \eta_1\eta_3\omega_1 \end{pmatrix}, \quad \begin{pmatrix} \hat{x}_1^{(2)} \\ \hat{x}_2^{(2)} \\ \hat{x}_3^{(2)} \\ \hat{y}_1^{(2)} \\ \hat{y}_2^{(2)} \end{pmatrix} = \begin{pmatrix} \eta_2\omega_1 \\ \eta_2\eta_3\omega_1 \\ \eta_2\eta_3\eta_4\omega_1 \\ \omega_1 \\ \eta_1\omega_1 \end{pmatrix},$$

and $f^{(j)}(\eta)$ is given by

$$f^{(1)}(\eta) = \eta_1\eta_2 \quad f^{(2)}(\eta) = \eta_2^2\eta_3.$$

A.6 Singular Value Decomposition

Let $(\sigma_k, x_k, y_k)_{k=1}^{\infty}$, denote the singular value decomposition of a bijective compact operator H, i.e.

$$Hx_k = \sigma_k y_k, \quad H^* y_k = \sigma_k x_k, \qquad \langle y_k, y_j \rangle = \delta_{ij},$$

then

$$H^{-1} f = \sum \langle f, y_k \rangle H^{-1} y_k = \sum \frac{1}{\sigma_k} \langle f, y_k \rangle x_k. \tag{A.19}$$

A.7 Impact of Local Errors

This section is devoted to the proof of estimate (4.40). The impact of the local error on the global error is also estimated in [50]. In this article a partition of the integration domain, instead of a partition of the index set is used. Therefore, the proof requires a slight modification.

Let $v = \sum_{i \in I} v_i b_i \in V_h$ and let $\mathbf{v} = (v_1, \ldots, v_n)^T$ denote the vector of coefficients. If the grid is quasi-uniform and shape regular, there exist constants C_1 and C_2 independent of h such that

$$C_1 \|v\|_{L^2}^2 \leq h^d \|\mathbf{v}\|_2^2 \leq C_2 \|v\|_{L^2}^2 \qquad \forall v \in V_h. \tag{A.20}$$

(For a proof refer to [10].) In the following, we use the abbreviation $v_c := \sum_{i \in c} v_i b_i$ for a cluster $c \in T(I)$. The corresponding index vector is given by $\mathbf{v}_c = (v_i)_{i \in c}$. By construction the clusters at the same level of the cluster tree are disjoint. Therefore, it holds:

$$\sum_{c \in T(l)} \|\mathbf{v}_c\|_2^2 \leq \|\mathbf{v}\|_2^2 \qquad \forall l = 0, \ldots, l_{\max}. \tag{A.21}$$

Combining the two estimates above yields

$$\sum_{c \in T(l)} \|v_c\|_{L^2}^2 \leq \frac{C_2}{C_1} \|v\|_{L^2}^2 \qquad \forall l = 0, \ldots, l_{\max}. \tag{A.22}$$

After these preliminary considerations, we are able to proof the estimate (4.40).

$$\begin{aligned}
\langle (K-\tilde{K})u, v\rangle &= \sum_{i,j\in I} v_i u_j \langle (K-\tilde{K})b_j, b_i\rangle \\
&= \sum_{l=0}^{L} \sum_{(c_1,c_2)\in F(l)} \sum_{i\in c_1} \sum_{j\in c_2} v_i u_j \langle (K-\tilde{K})b_j, b_i\rangle = \sum_{l=0}^{L} \sum_{(c_1,c_2)\in F(l)} \langle (K-\tilde{K})u_{c_2}, v_{c_1}\rangle \\
&\overset{(4.33),(4.39)}{\leq} \sum_{l=0}^{L} \varepsilon_l \sum_{(c_1,c_2)\in F(l)} \|v_{c_1}\|_{L^2} \|u_{c_2}\|_{L^2} \leq \sum_{l=0}^{L} \varepsilon_l \Big(\sum_{(c_1,c_2)\in F(l)} \|v_{c_1}\|_{L^2}^2 \Big)^{\frac{1}{2}} \Big(\sum_{(c_1,c_2)\in F(l)} \|u_{c_2}\|_{L^2}^2 \Big)^{\frac{1}{2}} \\
&= \sum_{l=0}^{L} \varepsilon_l \Big(\sum_{c_1\in T(l)} \|v_{c_1}\|_{L^2}^2 \underbrace{\sum_{c_2:(c_1,c_2)\in F(l)} 1}_{\overset{(4.6)}{\leq} C_{\mathrm{sp}}} \Big)^{\frac{1}{2}} \Big(\sum_{c_2\in T(l)} \|u_{c_2}\|_{L^2}^2 \underbrace{\sum_{c_1:(c_1,c_2)\in F(l)} 1}_{\overset{(4.7)}{\leq} C_{\mathrm{sp}}} \Big)^{\frac{1}{2}} \\
&\overset{(A.22)}{\leq} \frac{C_2}{C_1} C_{\mathrm{sp}} \Big(\sum_{l=0}^{L} \varepsilon_l \Big) \|v\|_{L^2(D)} \|u\|_{L^2(D)}. \qquad \text{(A.23)}
\end{aligned}$$

Notations

Sets

D, $\bar{D}$	open bounded domain in $\mathbb{R}^3$ and topological closure
∂D	surface of D
S^2	$= \{\Omega \in \mathbb{R}^3 \mid \|\Omega\|_2 = 1\}$, unit sphere in $\mathbb{R}^3$
$B_r(x)$	$= \{y \in \mathbb{R}^d \mid \|y - x\| < r\}$
$\mathrm{conv}\{x_1, \ldots, x_l\}$	convex hull of $x_1, \ldots, x_l \in R^d$
$S^{(d)}$	d-dimensional simplex (see (3.17))
$T^{(d)}$	d-dimensional simplex (see (3.18))
$C^{(d)}$	d-dimensional unit-cube (see (3.19))
$\mathcal{G}$	triangulation of $D \cup \partial D$
N	nearfield
F	farfield

Optical Parameters

κ	absorption coefficient
σ	scattering coefficient
γ	$= \kappa + \sigma$, total extinction coefficient
ω	$= \frac{\sigma}{\gamma}$, scattering albedo
ρ	diffuse reflection coefficient

Variables and Functions

$B(T, \lambda)$	Planck's function (see (1.2))
$\Phi(\Omega, \Omega')$	scattering phase function
P_n	Legendre polynomials (see Appendix A.2)
Y_n^m	spherical harmonics (see Appendix A.2)
χ_τ	local parameterization of $\tau \in \mathcal{G}$
k_{ij}, $\bar{k}_{ij}$	integral kernels (see page 29)

Operators

Id	identity operator

Δ, ∇, $\nabla\cdot$	Laplace operator, gradient, divergence
H_Ω	transport operator in direction Ω (see (1.24))
Σ_Ω, Σ	scattering operators (see (1.24))
T	radiative transfer operator (see (1.24))
T_1, $\hat{T}_1$	diffusion operators (see (1.63) and (1.65))
K	integral operator
$L(U,V)$	space of linear operators from U to V
$L(V)$	$=L(V,V)$
$K(U,V)$	space of compact operators from U to V
$K(V)$	$=K(V,V)$
T^*	adjoint operator of $T \in L(U,V)$, U, V Hilbert spaces
	Matrices corresponding to the operators are denoted by bold face symbols.

Function Spaces

$\mathcal{P}_k$	polynomials up to degree k (see (3.5))
$\mathcal{Q}_k$	tensor product polynomials up to degree k (see (4.21))
$C^{k,\alpha}(\mu)$	space of k times differentiable functions whose k-th derivative is Hölder continuous with coefficient α
$L^p(\mu)$	$=L^p(D,\mu_1)\oplus_p L^p(\partial D,\mu_2)$ (see Definition 2.1)
H^s	Sobolev space for $s \in \mathbb{R}$

Miscellaneous

$\lvert\tau\rvert$	= volume of τ
$\sharp\, c$	= cardinality of the set c
$\overset{\circ}{\bigcup_i} c_i$	= disjoint union, i.e. $c_i \cap c_j = \emptyset$, if $i \neq j$
$Q_{\tau,t}$	quadrature formula
diam	diameter of a set in $\mathbb{R}^3$ (see Definition 4.4)
dist	distance between to sets in $\mathbb{R}^3$ (see Definition 4.4)
δ_{ij}	Kronecker symbol
O	Landau symbol

Bibliography

[1] H.W. Alt. *Lineare Funktionalanalysis*. Springer, Heidelberg, 3rd edition, 1999.

[2] S.F. Ashby, P.N. Brown, M.R. Dorr, and A.C. Hindmarsh. A linear algebra analysis of DSA for the Boltzmann transport equation. *SIAM J. Numer. Anal.*, 32:128–178, 1995.

[3] K. Atkinson and G. Chandler. The collocation method for solving the radiosity equation for unoccluded surfaces. *J. Integral Equations Appl.*, 10(3):253–290, 1998.

[4] C. Bardos, F. Golse, and B. Perthame. The Rosseland approximation for the radiative transfer equation. *Comm. Pure. Appl. Math*, 40:691–720, 1987.

[5] C. Bardos, R. Santos, and R. Sentis. Diffusion approximation and computation of the critical size. *Trans. Amer. Math. Soc.*, 284:617–649, 1984.

[6] M. Bebendorf. Low-rank approximation of boundary element matrices. Preprint 1, Universität des Saarlandes, Fachrichtung Mathematik, 1999.

[7] M. Bebendorf and S. Rjasanow. Adaptive low-rank approximation of collocation matrices. Preprint 39, Universität des Saarlandes, Fachrichtung Mathematik, 2001.

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

[9] S. Börm, L. Grasedyk, and W. Hackbusch. Introduction to hierarchical matrices with applications. Preprint 18, MPI for Mathematics in the Sciences, Leipzig, 2002.

[10] D. Braess. *Finite Elemente*. Springer, Heidelberg, 2nd edition, 1997.

[11] P.N. Brown. A linear algebra development of DSA for $3d$ transport equations. *SIAM J. Numer. Anal.*, 32:179–214, 1995.

[12] B.G. Carlson and K.D. Lathrop. Transport theory - the method of discrete ordinates. In H. Greenspan, C.N. Kelber, and D. Okrent, editors, *Computing Methods in Reactor Physics*, pages 37–62, New York, 1968. Gordon & Breach.

[13] H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *Journal of Computational Physics*, 155:468–498, 1999.

[14] W. Dahmen, S. Prössdorf, and R. Schneider. Wavelet approximation methods for pseudodifferential equations II: Matrix compression and fast solution. *Adv. Comp. Math*, 1:259–335, 1993.

[15] P.J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Academic Press, New York, 1975.

[16] B. Davison. *Neutron Transport Theory.* The International Series of Monographs on Physics. Clarendon Press, Oxford, 1957.

[17] M.G. Duffy. Quadrature over a pyramid or cube of integrands with singularity at a vertex. *SIAM J. Numer. Anal.*, 19(6):1260–1262, 1982.

[18] W.A. Fiveland. The selection of discrete ordinate quadrature sets for anisotropic scattering. In *Fundamentals of Radiation Heat Transfer*, volume 160 of *ASME HTD*, pages 89–96. 1991.

[19] W. Freeden. *Multiscale modeling of spaceborne geodata.* European Consortium for Mathematics in Industry. Stuttgart, B. G. Teubner, 1999.

[20] K. Giebermann. *Schnelle Summationsverfahren zur numerischen Lösung von Integralgleichungen für Streuprobleme im* $\mathbb{R}^3$. PhD thesis, University of Karlsruhe, 1997.

[21] S.A. Goreinov, E.E. Tyrtyshnikov, and N.L. Zamarashkin. A theory of pseudo-skeleton approximations. *Linear Algebra Appl*, 261:1–21, 1997.

[22] T. Götz. Untersuchungen zur Kopplung von Wärmeleitung und Wärmestrahlung. Diploma thesis, University of Kaiserslautern, 1997.

[23] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen.* B.G.Teubner, Stuttgart, 2nd edition, 1996.

[24] W. Hackbusch. *Integralgleichungen: Theorie und Numerik.* B.G. Teubner, Stuttgart, 2nd edition, 1997.

[25] W. Hackbusch and S. Börm. Data sparse approximation by adaptive $\mathcal{H}^2$-matrices. Preprint 86, MPI for Mathematics in the Sciences, Leipzig, 2001.

[26] W. Hackbusch and S. Börm. $\mathcal{H}^2$- matrix approximation of integral operators by interpolation. Preprint 104, MPI for Mathematics in the Sciences, Leipzig, 2001.

[27] W. Hackbusch and B. Khoromskij. Blended kernel approximation in the $\mathcal{H}$-matrix techniques. Preprint 66, MPI for Mathematics in the Sciences, Leipzig, 2000.

[28] W. Hackbusch and Z.P. Nowak. On the fast matrix multiplication in the boundary element method by panel clustering. *Numer. Math*, 54:463–491, 1989.

[29] M. Hanke. *Numerische Mathematik II.* Scriptum, University of Karlsruhe, 1997.

[30] F. Harris. *The Theory of Branching Process.* Springer, Berlin, 1963.

[31] G. Kanschat. *Parallel and Adaptive Galerkin Methods for Radiative Transfer Problems.* PhD thesis, University of Heidelberg, 1996.

[32] M. Kerker. *The Scattering of Light and Other Electromagnetic Radiation.* Academic Press, New York, 1969.

[33] C. Lage. *Software Entwicklung zur Randelementemethode: Analyse und Entwurf effizienter Techniken.* PhD thesis, University of Kiel, 1995.

[34] C. Lage and C. Schwab. Advanced boundary element algorithms. Technical Report 99-11, ETH, Seminar for Applied Mathematics, Zürich, 1999.

[35] M. Laitinen and T. Tiihonen. Integro-differential equation modelling heat transfer in conducting, radiating and semitransparent materials. *Math. Meth. Appl. Sci.*, 21(5):375–392, 1998.

[36] E.W. Larsen. Diffusion-synthetic acceleration methods for discrete-ordinates problems. *Trans. Th. Stat. Phys.*, 13:107–126, 1984.

[37] F. Lentes and N. Siedow. Three-dimensional radiative heat transfer in glass cooling processes. *Glass Sci. Technol.*, 72(6):188–196, 1999.

[38] Th. Manteuffel and V. Faber. A look at transport theory from the viewpoint of linear algebra. In P. Nelson, editor, *Transport Theory, Invariant Embedding and Integral Equations*, pages 37–62, New York, 1989. Marcel Dekker.

[39] W. McLean. *Strongly Elliptic Systems and Boundary Integral Equations.* Univ. Press, Cambridge, 2000.

[40] M.F. Modest. *Radiative Heat Transfer.* McGraw-Hill Series in Mechanical Engineering. McGraw-Hill, New York, 1993.

[41] M.F. Modest and F.H. Azad. The influence and treatment of Mie-anisotropic scattering in radiative heat transfer. *ASME Journal of Heat Transfer*, 102:92–98, 1980.

[42] M.N. Özisik. *Radiative Transfer and Interactions with Conduction and Convection.* John Wiley & Sons, 1973.

[43] T.v. Petersdorff and C. Schwab. Fully discrete multiscale Galerkin BEM. Technical Report 95-08, ETH, Seminar for Applied Mathematics, Zürich, 1995.

[44] J. Pitkäranta. Estimates for the derivatives of the solutions to weakly singular fredholm integral equations. *SIAM J. Numer. Anal.*, 11(6):952–968, 1980.

[45] J. Pitkäranta. Error estimates for the combined spatial and angular approximations of the transport equation for slab geometry. *SIAM J. Numer. Anal.*, 20(5):922–950, 1983.

[46] S. Richling, E. Meinköhn, N. Kryzhevoi, and G. Kanschat. Radiative transfer with finite elements: I. basic method and tests. *Astronomy Astrophysics*, 380:776–788, 20001.

[47] V. Rokhlin. Rapid solution of integral equations of classical potential theory. *Journal of Computational Physics*, 60(2):187–207, 1985.

[48] S. Rosseland. *Radiative Transfer and Interactions with Conduction and Convection.* John Wiley & Sons, New York, 1936.

[49] S.A. Sauter. *Über die effiziente Verwendung des Galerkinverfahrens zur Lösung Fredholmscher Integralgleichungen.* PhD thesis, University of Kiel, 1992.

[50] S.A. Sauter. Variable order panel clustering. *Computing*, 64(3):223–261, 2000.

[51] S.A. Sauter and S. Erichsen. Efficient automatic quadrature in 3-d Galerkin BEM. *Comput. Methods. Appl. Mech. Engrg.*, 157:215–224, 1998.

[52] S.A. Sauter and A. Krapp. On the efficient computation of singular and nearly singular surface integrals arising in 3d-Galerkin BEM. *Numer. Math.*, 74(3):337–359, 1996.

[53] R. Schneider. *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur Lösung großer vollbesetzter Gleichungssysteme.* Habilitationsschrift, Technische Hochschule Darmstadt, 1995.

[54] R. Siegel and J.R. Howell. *Thermal Radiation Heat Transfer.* Hemisphere Publishing Corporation, Washington, 1992.

[55] O. Steinbach and W. Wendland. The construction of some efficient preconditioners in the boundary element method. *Adv. Comput. Math.*, 9(1-2):191–216, 1998.

[56] G. Vainikko. *Multidimensional Weakly Singular Integral Equations.* Springer, Heidelberg, 1993.

[57] D. Werner. *Funktionalanalysis.* Springer, Heidelberg, 3rd edition, 2000.

[58] F. Zingsheim. *Numerical Solution Methods for Radiative Transfer in Semitransparent Media.* PhD thesis, University of Kaiserslautern, 1999.

Wissenschaftlicher Werdegang

19. Juni 1974	Geboren in Karlsruhe
1980 - 1984	Grundschule in Maximiliansau
1984 - 1993	Europagymnasium Wörth
1993	Abitur
1994 - 1999	Studium der Technomathematik an der Universität Karlsruhe
1997 - 1998	6 monatige Werkstudenten-Tätigkeit bei der Siemens AG in Erlangen
1999	Diplom in Technomathematik
1999 -	Promotionsstudium in Mathematik an der Universität Kaiserslautern als Stipendiat des Fraunhofer Instituts für Techno- und Wirschaftsmathematik