UNIVERSIDADE D
COIMBRA

Ana Catarina Quitério Lourenço

RUNGE-KUTTA DISCONTINUOUS GALERKIN METHODS FOR MAXWELL'S EQUATIONS

VOLUME 1

Dissertação no âmbito do Mestrado em Matemática, Ramo Análise Aplicada e Computação, orientada pelo Professor Doutor Adérito Luís Martins Araújo e apresentada ao Departamento de Matemática da Faculdade de Ciências e Tecnologia.

Julho de 2019

Runge-Kutta Discontinuous Garlerkin methods for Maxwell's equations

Ana Catarina Quitério Lourenço

UNIVERSIDADE D
COIMBRA

Master in Mathematics

Mestrado em Matemática

MSc Dissertation | Dissertação de Mestrado

Julho 2019

Abstract

The main motivation behind this work is to study a method that simulates the propagation of an electromagnetic wave through the layers of the retina during an Optical Coherence Tomography (OCT). Simulating the full complexity of the retina, in particular the variation of the size and shape of each structure, distance between them and the respective refractive indexes, requires a rigorous approach that can be achieved by solving Maxwell's equations. These are the set of partial differential equations that govern the behaviour of an electromagnetic field in free space and in media. To fully model the problem, we introduce the constitutive relations, boundary conditions and initial conditions.

In this work, we establish a fully discrete method that allows us to obtain a numerical solution for the problem in a homogeneous isotropic medium with perfect electric conductor (PEC) boundary conditions. First the spatial discretization is achieved by resorting to the strong formulation of the discontinuous Galerkin (dG) method with central fluxes and we prove the stability and convergence of this method. Then the explicit Euler method is used in order to achieve the temporal discretization of the semi-discrete scheme obtained with the dG method and we prove the stability and the convergence of the temporal discretization.

The spatial convergence rate of the dG method is numerically corroborated for the bidimensional case by implementing the dG method in Matlab. Here the fully discrete method is implemented using a Runge-Kutta method for the temporal discretization. Finally, we consider a numerical example for a two dimensional model which tries to represent a single nucleus of the outer nuclear layer (ONL), that comprises the cells bodies of light sensitive photoreceptors cells.

Resumo

A principal motivação por detrás deste trabalho é estudar um método que simule a propagação de uma onda electromagnética através das camadas da retina durante uma Tomografia de Coerência Óptica (OCT). A simulação da complexidade total da retina, em particular da variação do tamanho e da forma de cada estrutura, da distância entre estruturas e dos respectivos índices refractivos, requer uma abordagem rigorosa que pode ser alcançada resolvendo equações de Maxwell. Estas são o conjunto de equações com derivadas parciais que ditam o comportamento de um campo electromagnético quer no vácuo quer num meio não vazio. Para modelar completamente o problema, introduzimos as relações constitutivas, condições de fronteira e condições iniciais.

Neste trabalho, estabelecemos um método totalmente discreto que nos permite obter uma solução numérica para o problema num meio isotrópico homogéneo com condições de fronteira para um condutor eléctrico perfeito (PEC). Primeiro, a discretização espacial é obtida recorrendo à formulação forte do método de Galerkin descontínuo (dG) com fluxos centrais e provamos a estabilidade e convergência deste método. De seguida, o método de Euler explícito é usado para proceder à discretização temporal do esquema semi-discreto que resulta do método dG e provamos a estabilidade e convergência da discretização temporal.

A ordem de convergência espacial do método dG é confirmada numericamente para o caso bidimensional implementando o método dG em Matlab. Aqui o método totalmente discreto é implementado usando um método de Runge-Kutta para a discretização temporal. Por fim, consideramos um exemplo numérico para um modelo bidimensional que tenta representar um único núcleo da camada nuclear externa (ONL), que é constituída pelos corpos celulares de células de fotorreceptores sensíveis à luz.

Table of contents

List of figures ix

List of tables xi

1 Introduction 1

2 Maxwell's equations 5

2.1 The partial differential equations . . . 5
2.2 The constitutive relations . . . 6
2.3 Boundary conditions . . . 7
2.3.1 Tangential continuity condition . . . 7
2.3.2 Perfect electric conductor boundary condition . . . 8
2.4 The three dimensional problem . . . 8
2.5 Homogeneous isotropic media . . . 9
2.5.1 Reduction to two dimensions . . . 9
2.6 Mathematical aspects of Maxwell's equations . . . 10
2.6.1 The state and graph spaces . . . 11
2.6.2 Boundary conditions in the graph space . . . 12
2.6.3 Well-Posedness . . . 13

3 Discontinuous Garlerkin methods 17

3.1 Concepts about meshes . . . 17
3.2 Homogeneous medium and normalized form . . . 19
3.3 Concepts about discrete bilinear forms . . . 20
3.4 Boundness of discrete bilinear forms . . . 24
3.5 Discrete operators . . . 29
3.6 Stability . . . 30
3.7 Convergence . . . 31
3.8 Full discretization . . . 34
3.8.1 The explicit Euler method . . . 34
3.8.2 Stability . . . 35
3.8.3 Convergence . . . 37

4 Computational results **41**
4.1 The problem . . . 41
4.1.1 The numerical flux . . . 42
4.1.2 The boundary conditions . . . 42
4.2 Implementation of dG method . . . 42
4.3 Spatial order of convergence . . . 44
4.4 Scattered field formulation . . . 45

5 Conclusion **49**

References **51**

Appendix A Auxiliary results **53**
A.1 Broken polynomial spaces . . . 53
A.1.1 The broken Sobolev space $H^m(T_h)$ . . . 54
A.1.2 The broken graph space $H(curl, T_h)$ . . . 54
A.1.3 Broken polynomial space . . . 55
A.1.4 The broken polynomial space $\mathbb{P}_d^k(T_h)$ . . . 56
A.2 Admissible Mesh Sequences . . . 56
A.2.1 Geometric properties . . . 56
A.2.2 Inverse and trace inequality . . . 57
A.2.3 Polynomial approximation . . . 57

List of figures

1.1 A scheme of the time-domain OCT [23]. 3

4.1 Example of mesh for K = 50. 44
4.2 Numerical solution computed for $N = 6$ and $K = 800$ 47

List of tables

4.1 The L^2-error and spatial order of convergence for the central flux. 46

Chapter 1

Introduction

Maxwell's equations are the fundamental set of partial differential equations that govern the behaviour of an electromagnetic field in free space and in media. When combined with constitutive relations, the equations fully describe the effect of media on the propagation of electromagnetic waves. Together with boundary conditions and the initial conditions, they complete the model of the propagation of electromagnetic radiation [25].

The importance and diversity of application of the Maxwell's equations has lead to a great interest in solving these equations. The first numerical method for solving time-dependent Maxwell's equations was the Finite-Difference Time Domain (FDTD) scheme proposed by Yee in 1966 in [5], which uses a staggered grid both in space and time and is a fully discrete method explicit in time. Similarly to all finite difference methods, FDTD is difficult to generalize to unstructured grids and can handle only regular domains, while also having other disadvantages: no adaptivity, the numerical analysis requires high regularity and it is only conditionally stable (CFL condition) [12]. In 2000 it was proposed a very efficient, unconditionally stable method based on a finite-difference scheme [6]. Finite element based methods can handle irregular domains, achieve higher order and allow adaptivity and error control, while using a variational approach which inherits many properties of the continuous problem, which makes a rigorous error analysis possible [12].

In the last years, there has been a focus on solving Maxwell's equations numerically by using discontinuous Galerkin (dG) finite element methods for the spatial discretization [4]. Some of the main advantages of dG methods are: non-conforming meshes are handled much more easily, they are highly parallelizable and the mass matrix is block diagonal, which is particularly appealing if one is interested in the simulation of wave propagation in composite materials, where the electric permittivity and the magnetic permeability are discontinuous [12]. For a fully discrete method it is necessary to also use a suitable time integration method. We can choose an explicit time integrator, which can exploit the block diagonal structure of the mass matrix of discontinuous Galerkin schemes and thus lead to fully explicit schemes. For example, the Runge-Kutta dG-methods achieve high-order convergence both in space and time by using strong stability preserving Runge-Kutta schemes in time. Yet, explicit methods have step size restrictions due to stability requirements (CFL condition). Thus, we can choose implicit methods which can be used with larger time steps at the cost of solving linear or even nonlinear systems [12].

The main goal of this thesis is to study a method that combines dG methods for spatial discretization with an explicit time integrator, thus providing a fully discrete scheme for time-dependent Maxwell's equations. Most of the work focuses on the proof of stability and convergence of this method. The method is applied to a model with heterogeneous isotropic permittivity, which is the first step to model the behaviour of electromagnetic waves during an OCT in the layers of the human retina.

Motivation behind this work

The human retina is a complex structure in the eye that is responsible for the vision. It is a part of the central nervous system and it is composed by several layers, namely the outer nuclear layer that includes the cells bodies of light sensitive photoreceptors cells, rods and cones [11].

Optical Coherence Tomography (OCT) is a recent imaging technique that has become increasingly popular as a ophthalmic diagnostic tool because of its high resolution. OCT allows us to obtain a highly detailed tridimensional map of the eye's fundus and is a non-invasive technique, thus being more comfortable for the patients and of easier access [1]. Therefore, OCT is considered a very useful and important technique for the early diagnosis of ophthamologic pathologies.

OCT is analogous to ultrasound imaging but uses light instead of sound [9]. Light and sound travel at different speeds in different materials and, as it travels from one type of material to another, part is reflected back and part continues to travel forward. The portion that is reflected back is detected. In an ultrasound the echo time delay and intensity of this signal are used in order to characterise the material that caused the rejection. As the speed of light is at least 10^6 times that of sound, direct measure of the echo time delays isn't possible for electromagnetic waves because they are in a very small scale. Therefore, OCT uses a method called interferometry [1].

An electromagnetic beam is emitted by a source and travels until it reaches a beam splitter, where it is split in two identical parts. One of the resulting beams travels to a reference mirror, where it is reflected back; the other goes through the eye and, there, it is reflected back by the eye's different structures [1].

The result of this constructive interference is detected by the photodetector. If we change the position of the mirror, we change the optical path the reference beam travels [1]. Thus it will constructively interfere with a portion of the other beam that was reflected by the sample at a different depth, which is the principle behind time-domain OCT [23], as described by Figure 1.1.

The intensity of the detected beam depends on the intensity of the reference beam and the structure's reflectivity. Different structures reflect the beam back in different proportions, and can therefore be identified by analysis of this signal [23].

A light scattering simulation allows us to model the layers of the retina and the dG method for Maxwell's equations allows us to obtain the scattered field simulating the beam reflected through the eye in the OCT.

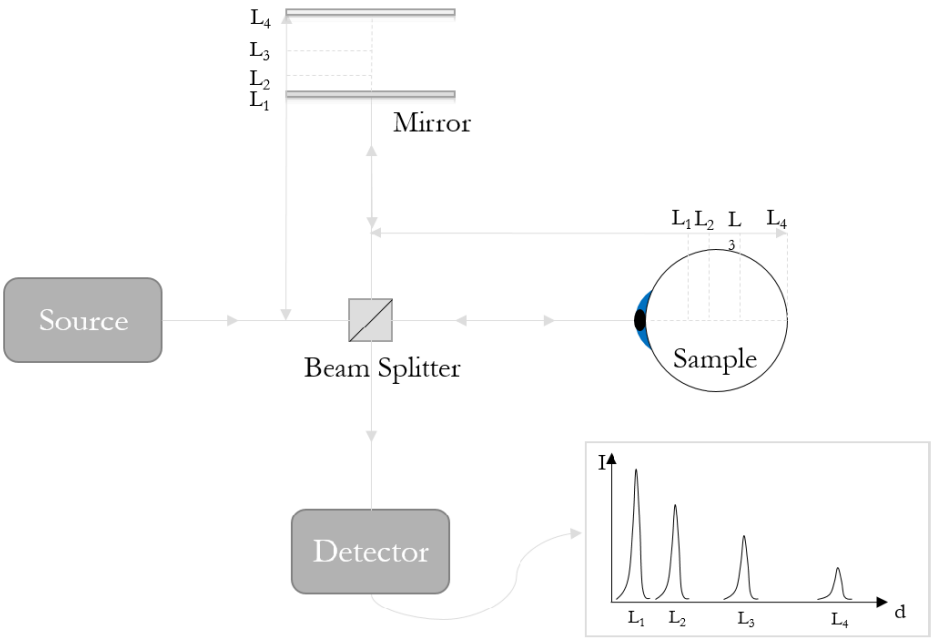

Fig. 1.1 A scheme of the time-domain OCT [23].

Chapter 2

Maxwell's equations

2.1 The partial differential equations

Maxwell's equations are the fundamental set of equations that describe the behaviour of an electromagnetic field in free space and in media. These fundamental laws were first formulated by James Clerk Maxwell in 1873 [17].

The electromagnetic field in space and time is described by four vector fields: E, H, D, B : $\mathbb{R}^+ \times \Omega \to \mathbb{R}^3$ on a set $\Omega \in \mathbb{R}^3$, where E represents the electric field, H the magnetic field, D the electric flux density and B the magnetic flux density. The SI units of these fields are volts per meter (V/m), amperes per meter (A/m), coulombs per square meter (C/m^2) and webers per square meter(Wb/m^2), respectively [25]. We consider that Ω is a polyhedron in $\mathbb{R}^3$ because this will allows us to cover Ω with meshes built of polyhedral elements and we can define the outward unit normal a.e.

Thus, the time-dependent Maxwell's equations, relating these electromagnetic fields, are stated in differential form as:

$$\frac{\partial B}{\partial t} = -\nabla \times E \tag{2.1a}$$

$$\frac{\partial D}{\partial t} = \nabla \times H - J \tag{2.1b}$$

$$\nabla \cdot D = \rho \tag{2.1c}$$

$$\nabla \cdot B = 0, \tag{2.1d}$$

where the electric current density $J : \mathbb{R}^+ \times \Omega \to \mathbb{R}^3$ and the electric charge density $\rho : \mathbb{R}^+ \times \Omega \to \mathbb{R}$ is the source generating the electromagnetic fields, with SI units amperes per square meter and coulombs per cubic meter, respectively. The notation $\nabla\times$ and $\nabla\cdot$ refers to the vector operator curl and divergence, respectively. Thus, equations (2.1a) and (2.1b) are called the curl equations and (2.1c) and (2.1d) are called the divergence equations.

The differential form of the electric charge conservation law, called continuity equation, follows from (2.1b) and (2.1c). By taking the divergence of (2.1b) and using (2.1c), we get

$$\nabla \cdot \nabla \times H = \nabla \cdot J + \nabla \cdot \frac{\partial D}{\partial t} = \nabla \cdot J + \frac{\partial}{\partial t} \nabla \cdot D = \nabla \cdot J + \frac{\partial \rho}{\partial t}. \tag{2.2}$$

Then the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot J = 0 \tag{2.3}$$

results from the identity $\nabla \cdot \nabla \times H = 0$ [15].

Differentiating the divergence equations with respect to time and using the curl equations, the continuity equation and the identity $\nabla \cdot \nabla \times H = 0$ gives

$$\frac{\partial}{\partial t}(\nabla \cdot D - \rho) = \nabla \cdot \frac{\partial D}{\partial t} - \frac{\partial \rho}{\partial t} = \nabla \cdot (\nabla \times H - J) + \nabla \cdot J = 0 \tag{2.4}$$

$$\frac{\partial}{\partial t}\nabla \cdot B = \nabla \cdot \frac{\partial B}{\partial t} = -\nabla \cdot (\nabla \times E) = 0 \tag{2.5}$$

that is, if the continuity equation holds, the curl equations allow us to imply that that the divergence equations are constant in time. Thus, if the divergence equations are satisfied for an initial time, they are satisfied for all times t, which means that, regarding the time evolution of Maxwell's equation, we only have to consider the curl equations.

In the Maxwell's equations, electric current density J and the electric charge density ρ are the sources of electromagnetic radiation. However, when analysing the propagation of electromagnetic radiation in regions far from the source, J and ρ can be considered zero. Such assumption will be used from now on, that is, we will only consider the homogeneous problem where $J = \rho = 0$.

2.2 The constitutive relations

As seen above, the Maxwell's equations are reduced to a system of six independent scalar equations (formed by the vectorial curl equations) for twelve scalar unknowns (the components of vectors E, H, D, B). Therefore, although the behaviour of a electromagnetic field in free space and in media is entirely described by the Maxwell's equations, the system of six equations is underdetermined. This is overcome by adding six more equations, defining a connection between two couple of fields: D and E, B and H.

These equations, called constitutive relations and represented by

$$D = D(E), \qquad B = B(H), \tag{2.6}$$

model the effect of the electromagnetic field on material and describe the functional dependence between vectors, considering the properties of media.

In free space, we have

$$D = \varepsilon_0 E, \qquad B = \mu_0 H, \tag{2.7}$$

where the constants ε_0 and μ_0 are the electric permittivity and magnetic permeability of free space, respectively. In SI units we have

$$\varepsilon_0 = 8.854 \times 10^{-12} F/m \quad \text{(farads per meter)} \tag{2.8}$$

$$\mu_0 = 4\pi \times 10^{-7} H/m \quad \text{(henrys per meter)}. \tag{2.9}$$

In vacuum, the speed of light c_0 is given by $\frac{1}{\sqrt{\varepsilon_0 \mu_0}}$.

In homogeneous and isotropic media, where the electric and magnetic properties are uniform in all directions, the constitutive relations are given by

$$D = \varepsilon E, \qquad \varepsilon = \varepsilon_r \varepsilon_0, \tag{2.10}$$

$$B = \mu H, \qquad \mu = \mu_r \mu_0, \tag{2.11}$$

where the dimensionless scalar ε_r and μ_r are the relative permittivity and relative permeability of the medium and ε and μ are refered to as the permittivity and the permeability of the medium.

In inhomogeneous isotropic media, the relative permittivity and relative permeability of the medium are scalar functions of the position, $\varepsilon_r,\ \mu_r : \mathbb{R}^3 \to \mathbb{R}$.

Now, as stated in section 2.1, if the continuity equation (2.3) holds, the divergence equations (2.1c) and (2.1d) are satisfied and, using the constitutive relations, in three-dimensional spaces for heterogeneous isotropic linear media with no source, the Maxwell's system is

$$\mu \frac{\partial H}{\partial t} = -\nabla \times E \tag{2.12a}$$

$$\varepsilon \frac{\partial E}{\partial t} = \nabla \times H \tag{2.12b}$$

2.3 Boundary conditions

In order to properly simulate electromagnetic wave propagation, it is necessary to take into consideration the boundary conditions. In reality, an electromagnetic wave propagates until it eventually dies out or is totally absorbed by an obstacle. As it would be too computationally expensive to define a domain large enough to simulate the dying out of an electromagnetic wave, we have to truncate the space, by selecting a suitable artificial boundary and regions that define a finite domain.

Therefore the boundary conditions introduced on the computational domain in an electromagnetic wave propagation simulation are of three types: reflecting, absorbing and periodic boundary conditions.

Reflecting boundary conditions, such as Perfect Electric Conductor (PEC) and Perfect Magnetic Conductor (PMC), are boundary conditions that reflect all incident radiation and are used to model cavities and to introduce symmetry planes into the system [3]. The absorbing boundary conditions, such as the Silver-Müller Absorbing Boundary Condition (SM-ABC), mimic an infinite computational domain by partially absorbing outgoing radiation. In this thesis, we will use only PEC boundary conditions.

2.3.1 Tangential continuity condition

Before we can solve the Maxwell's equations in proximity of the boundaries, we must establish conditions relating the field components at the interface between two media of different properties [13].

According to [2], [13] and [14], the tangential continuity conditions

$$n \times (E_1 - E_2) = 0, \tag{2.13}$$
$$n \times (H_1 - H_2) = 0, \tag{2.14}$$

where n is the normal unit vector and indexes 1 and 2 represent the field component inside and outside the domain, respectively, guarantee that the tangential component of the field vectors E and H is continuous on either side of the boundary.

Regarding the continuity in the normal components of the field vectors D and B, it is ensured by the divergence equations

$$n \cdot (D_1 - D_2) = 0, \tag{2.15}$$
$$n \cdot (B_1 - B_2) = 0. \tag{2.16}$$

Owing to the constitutive relations (2.10) and (2.11), equations (2.15) and (2.16) are equivalent to

$$n \cdot (\varepsilon_1 E_1 - \varepsilon_2 E_2) = 0, \tag{2.17}$$
$$n \cdot (\mu_1 H_1 - \mu_2 H_2) = 0, \tag{2.18}$$

this is, the continuity in normal direction of B and D is equivalent to the continuity of the normal direction of E and H.

Equations (2.13)-(2.16) are known as the interface conditions.

2.3.2 Perfect electric conductor boundary condition

The perfect electric conductor boundary condition, usually used to model metallic cavities, is a reflective boundary condition where the tangential component of E outside of the domain is zero and there is no field propagation into the outside.

Thus, in a domain surrounded by a PEC medium, the tangential component of E and the normal component of B disappear at the boundary, which is translated into the fact that the interface conditions yield the following boundary conditions:

$$n \times E = 0 \quad \text{on } \partial\Omega, \tag{2.19}$$
$$n \cdot B = 0 \quad \text{on } \partial\Omega, \tag{2.20}$$

where $\partial\Omega$ is the boundary of Ω and n is the outward unit normal.

2.4 The three dimensional problem

Combining the Maxwell's equations with the constitutive relations and the boundary conditions (dropping the divergence conditions) and adding the necessary initial values, we obtain the following

reduced problem: Solve for $E,\ H:\ \mathbb{R}^+ \times \Omega \to \mathbb{R}^3$ such that

$$\mu \frac{\partial H}{\partial t} = -\nabla \times E \qquad \text{in } \mathbb{R}^+ \times \Omega \tag{2.21a}$$

$$\varepsilon \frac{\partial E}{\partial t} = \nabla \times H \qquad \text{in } \mathbb{R}^+ \times \Omega \tag{2.21b}$$

$$E(0,x,y,z) = E_0 \qquad \text{in } \Omega \tag{2.21c}$$

$$H(0,x,y,z) = H_0 \qquad \text{in } \Omega \tag{2.21d}$$

$$n \times E = 0 \qquad \text{on } \mathbb{R}^+ \times \partial\Omega \quad \text{(for PEC)} \tag{2.21e}$$

where $E = (E_x, E_y, E_z)$, $H = (H_x, H_y, H_z)$ and $\Omega \in \mathbb{R}^3$, the permittivity ε and the permeability μ of the medium are space-dependent and the E_0 and H_0 are the initial values for the electric field and the magnetic field, respectively.

For each equation (2.21a)–(2.21b), we write the vector components of the curl operators in Cartesian coordinates. This yields the following three scalar equations for (2.12a) and three scalar equations for (2.12b).

2.5 Homogeneous isotropic media

In order to simplify the first approach to the Maxwell's equations, we will study the equations in homogeneous isotropic media. Therefore, we suppose that $\varepsilon,\ \mu$ are positive constants.

2.5.1 Reduction to two dimensions

The underlying physical system modeled by the Maxwell's equations may have some symmetries which allow to reduce the dimensions of the system. Oftentimes, the system is homogeneous in one direction, for instance, if the structure extends to infinity in the z-direction with no change in the shape or position of its transverse cross section. If the incident wave is also uniform in the z-direction, then all partial derivatives of the fields with respect to z are equal to zero [25]. The scalar equations are reduced to

$$\mu \frac{\partial H_x}{\partial t} = -\frac{\partial E_z}{\partial y} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.22a}$$

$$\mu \frac{\partial H_y}{\partial t} = \frac{\partial E_z}{\partial x} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.22b}$$

$$\mu \frac{\partial H_z}{\partial t} = \frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.22c}$$

for (2.12a), and

$$\varepsilon \frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.23a}$$

$$\varepsilon \frac{\partial E_y}{\partial t} = -\frac{\partial H_z}{\partial x} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.23b}$$

$$\varepsilon \frac{\partial E_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.23c}$$

for (2.12b). These equations can be grouped according to field vector components. This allows us to create two sets of three equations:

- **TE polarization:** The first set, which involves only E_x, E_y and H_z, is

$$\varepsilon \frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.24a}$$

$$\varepsilon \frac{\partial E_y}{\partial t} = -\frac{\partial H_z}{\partial x} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.24b}$$

$$\mu \frac{\partial H_z}{\partial t} = \frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.24c}$$

  and is designated the transverse electric (TE) mode. TE mode describes the propagation when the electric field lies in the plane of propagation.

- **TM polarization:** The second set, which involves only E_z, H_x and H_y, is

$$\varepsilon \frac{\partial E_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.25a}$$

$$\mu \frac{\partial H_x}{\partial t} = -\frac{\partial E_z}{\partial y} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.25b}$$

$$\mu \frac{\partial H_y}{\partial t} = \frac{\partial E_z}{\partial x} \qquad \text{in } \mathbb{R}^+ \times \Omega, \tag{2.25c}$$

$$\tag{2.25d}$$

  and is designated the transverse magnetic (TM) mode. TM mode describes the propagation when the electric field is perpendicular to the plane of propagation.

Note that the TE and TM modes contain no common field vector components.

2.6 Mathematical aspects of Maxwell's equations

We will only consider bounded domains Ω with a Lipschitz-continuous boundary $\partial\Omega$. Thus, the outward unit normal n is defined almost everywhere (a.e.) on Ω.

2.6.1 The state and graph spaces

Definition 2.1. *(The state space V) We define the state space V as*

$$V = \mathrm{L}^2(\Omega)^3 \times \mathrm{L}^2(\Omega)^3. \tag{2.26}$$

We assign it the inner product: for $[H_1, E_1]^T, [H_2, E_2]^T \in V$,

$$\left(\begin{bmatrix} H_1 \\ E_1 \end{bmatrix}, \begin{bmatrix} H_2 \\ E_2 \end{bmatrix} \right)_V = \int_\Omega (\mu H_1 \cdot H_2 + \varepsilon E_1 \cdot E_2)\, dx \tag{2.27}$$

and the associated norm: for $[H, E]^T \in V$,

$$\left\| \begin{bmatrix} H \\ E \end{bmatrix} \right\|_V = \left(\begin{bmatrix} H \\ E \end{bmatrix}, \begin{bmatrix} H \\ E \end{bmatrix} \right)_V^{1/2}. \tag{2.28}$$

Assuming that μ, ε are positive constants allows us to assert that the V-inner product is equivalent to the standard L^2-inner product. Therefore, $(V, (,\cdot,\cdot))$ is a Hilbert space [8]. We choose to use the V-inner product instead of the standard L^2-inner product because its induced norm represents electromagnetic energy.

We have to define a new meaning to the curl operator in (2.21a), (2.21b) since the standard curl operator is only defined for continuous differentiable functions and functions in V do not have to satisfy this condition. In a L^2 space, for $[H, E]^T \in V$, $\nabla \times E, \nabla \times H \in L^2(\Omega)^3$ is a sufficient condition to ensure that (2.21a), (2.21b) are well-defined. Let $C_0^\infty(\Omega)^3$ denote the space of infinitely differentiable functions with compact support in Ω.

Definition 2.2. *(The variational curl) A function $F \in L^2(\Omega)^3$ has a variational curl in $L^2(\Omega)^3$ if there exists a function $G \in L^2(\Omega)^3$ such that*

$$\int_\Omega G \cdot \varphi = \int_\Omega F \cdot (\nabla \varphi), \qquad \forall \varphi \in C_0^\infty(\Omega)^3. \tag{2.29}$$

We write $\nabla \times F = G$. If the variational curl exists, it is unique since the space $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$.

Definition 2.3. *(The graph space $H(curl, \Omega)$) The graph space of the curl-operator is defined as*

$$H(curl, \Omega) := \{F \in L^2(\Omega)^3 | \nabla \times F \in L^2(\Omega)^3\}, \tag{2.30}$$

with the inner product: For $F, G \in H(curl, \Omega)$,

$$(F, G)_{H(curl,\Omega)} := (F, G)_{L^2(\Omega)^3} + (\nabla \times F, \nabla \times G)_{L^2(\Omega)^3}, \tag{2.31}$$

and the associated graph norm

$$\|F\|_{H(curl,\Omega)} := (F, F)^{1/2}_{H(curl,\Omega)}$$

.

Theorem 2.4. *The graph space $H(curl,\Omega)$ is a Hilbert space.*

Proof. Let (F_n) be a Cauchy sequence in $H(curl,\Omega)$. Then, (F_n) and $(\nabla\times F_n)$ are Cauchy sequences in $L^2(\Omega)^3$ and therefore convergent. By the definition of $H(curl,\Omega)$, for all $\varphi\in C_0^\infty(\Omega)$ and all $n\in\mathbb{N}$, we have

$$\int_\Omega(\nabla\times F_n)\cdot\varphi=\int_\Omega F_n\cdot(\nabla\varphi). \tag{2.32}$$

Let F and G be limits of F_n and $\nabla\times F_n$ in $L^2(\Omega)^3$, respectively. Taking the limit $n\to\infty$, (2.32) yields

$$\int_\Omega G\cdot\varphi=\int_\Omega F\cdot(\nabla\varphi). \tag{2.33}$$

Thus we conclude from (2.29) that $F\in H(curl,\Omega)$ and $G=\nabla\times F$. □

Theorem 2.5. *[18, Theorem 3.26] If Ω is a bounded Lipschitz domain in $\mathbb{R}^3$, then the closure of $C_0^\infty(\overline{\Omega})^3$ in the norm $\|\cdot\|_{H(curl,\Omega)}$ is $H(curl,\Omega)$.*

2.6.2 Boundary conditions in the graph space

Now we add the boundary condition (2.21e) in the graph space $H(curl,\Omega)$.

Definition 2.6. *(The space $H_0(curl,\Omega)$) The space $H_0(curl,\Omega)$ is defined as the closure of $C_0^\infty(\Omega)^3$ in the norm $\|\cdot\|_{H(curl,\Omega)}$.*

$H_0(curl,\Omega)$ is a closed subspace of the Hilbert space $H(curl,\Omega)$. Thus, $H_0(curl,\Omega)$ is a Hilbert space.

Lemma 2.7. *[18, Lemma 3.27] Let $F\in H(curl,\Omega)$ be such that for every $\varepsilon\in C^\infty(\overline{\Omega})^3$ it holds*

$$(\nabla\times F,\varepsilon)_{L^2(\Omega)^3}=(F,\nabla\times\varepsilon)_{L^2(\Omega)^3}. \tag{2.34}$$

Then, $F\in H_0(curl,\Omega)$.

We can rewrite Lemma 2.7 for functions with more regularity, such as $\tilde{F}\in H^1(\Omega)^3$. Using integration by parts, we have

$$\int_\Omega(\nabla\times\tilde{F})\cdot\varepsilon=\int_\Omega\tilde{F}\cdot(\nabla\times\varepsilon)+\int_{\partial\Omega}(n\times\tilde{F})\cdot\varepsilon,\qquad\forall\varepsilon\in C^\infty(\overline{\Omega})^3. \tag{2.35}$$

Due to (2.34), we have

$$\int_{\partial\Omega}(n\times\tilde{F})\cdot\varepsilon=0,\qquad\forall\varepsilon\in C^\infty(\overline{\Omega})^3. \tag{2.36}$$

This is equivalent to

$$n\times\tilde{F}=0\text{ a.e. on }\partial\Omega \tag{2.37}$$

because $C^\infty(\overline{\Omega})^3$ is dense in $L^2(\Omega)^3$.

Then we may conclude the following generalization of Green's theorem holds for functions on $H(curl,\Omega)$ (see [18, Remark 3.28]).

Lemma 2.8. *(Green's theorem) Let $H \in H(curl,\Omega)$ and $E \in H_0(curl,\Omega)$. Then we have*

$$(H,\nabla\times E)_{L^2(\Omega)^3} = (\nabla\times H,E)_{L^2(\Omega)^3}. \tag{2.38}$$

2.6.3 Well-Posedness

Definition 2.9. *(Maxwell operator) We define the Maxwell operator A as*

$$A : D(A) \to V, \quad \begin{bmatrix} H \\ E \end{bmatrix} \mapsto \begin{bmatrix} \mu^{-1}\nabla\times E \\ -\varepsilon^{-1}\nabla\times H \end{bmatrix}, \tag{2.39}$$

where the domain of A is given by

$$D(A) := H(curl,\Omega)\times H_0(curl,\Omega). \tag{2.40}$$

We define the following graph norm for $D(A)$: For $[H,E]^T \in D(A)$,

$$\left\| \begin{bmatrix} H \\ E \end{bmatrix} \right\|_A^2 := \left\| \begin{bmatrix} H \\ E \end{bmatrix} \right\|_V^2 + \left\| A \begin{bmatrix} H \\ E \end{bmatrix} \right\|_V^2. \tag{2.41}$$

Then, we have

$$\left\| \begin{bmatrix} H \\ E \end{bmatrix} \right\|_A^2 = \left\| \begin{bmatrix} \mu^{1/2}H \\ \varepsilon^{1/2}E \end{bmatrix} \right\|_{L^2(\Omega)^6}^2 + \left\| \begin{bmatrix} \mu^{-1/2}\nabla\times E \\ \varepsilon^{-1/2}\nabla\times H \end{bmatrix} \right\|_{L^2(\Omega)^6}^2 \tag{2.42}$$

Due to the fact that the coefficients μ and ε are assumed to be a positive constant, we can conclude that the norms $\|\cdot\|_A$ and $\|\cdot\|_{H(curl,\Omega)\times H(curl,\Omega)}$ are equivalent. We also conclude that $(D(A),\|\cdot\|_A)$ is a Hilbert space, thus A is a closed operator.

Homogeneous evolution equation

We write (2.12) in a more compact form, the abstract evolution equation: For a given initial value $u_0 = [H_0,E_0]^T \in D(A)$ we search for $u = [H,E]^T \in C^1(\mathbb{R}^+,V)\cap C(\mathbb{R}^+,D(A))$ such that

$$\frac{\partial u}{\partial t} + Au = 0, \quad t \geq 0, \tag{2.43a}$$

$$u(0) = 0. \tag{2.43b}$$

In order to prove the well-posedness of (2.43), we resort to Stone's theorem A.1. As stated previously, V is a Hilbert space, thereby we only need to show that the domain $D(A)$ is dense in V and the operator A is skew-adjoint in order to conclude that A generates a C_0-group of unitary operators.

From (2.5) and (2.6) we know that $C^\infty(\overline{\Omega})^3 \times C_0^\infty(\overline{\Omega})^3$ is a subset of $D(A)$. As both $C^\infty(\Omega)^3$ and $C_0^\infty(\Omega)^3$ are dense in $L^2(\Omega)^3$ with respect to the L^2-norm and the L^2-norm and V-norm are equivalent. Then the density of the domain $D(A)$ in V follows.

Theorem 2.10. *(Skew-adjointness of A) The Maxwell operator A is skew-adjoint with respect to the V-inner product.*

Proof. We follow the proof in [19, Proposition 3.1]. In order to prove that A is skew-adjoint with respect to the V-inner product, we have to show that the domain of A and the domain of its adjoint A^* coincide, $D(A) = D(A^*)$, and that A is skew-symmetric, that is, for all $v_1, v_2 \in D(A)$ we have

$$(Av_1, v_2)_V = -(v_1, Av_2)_V. \tag{2.44}$$

We will prove the skew-symmetry of A first. For $v_1 = [H_1, E_1]^T$, $v_2 = [H_2, E_2]^T \in D(A)$ we have

$$\begin{aligned}(Av_1, v_2)_V &= \left(\begin{bmatrix} \mu^{-1}\nabla \times E_1 \\ -\varepsilon^{-1}\nabla \times H_1 \end{bmatrix}, \begin{bmatrix} H_2 \\ E_2 \end{bmatrix} \right)_V && (2.45)\\ &= \left(\begin{bmatrix} \nabla \times E_1 \\ -\nabla \times H_1 \end{bmatrix}, \begin{bmatrix} H_2 \\ E_2 \end{bmatrix} \right)_{L^2(\Omega)^6} && (2.46)\\ &= (\nabla \times E_1, H_2)_{L^2(\Omega)^3} - (\nabla \times H_1, E_2)_{L^2(\Omega)^3}. && (2.47)\end{aligned}$$

Using the Green's Theorem 2.8, we have

$$\begin{aligned}(Av_1, v_2)_V &= (E_1, \nabla \times H_2)_{L^2(\Omega)^3} - (H_1, \nabla \times E_2)_{L^2(\Omega)^3} && (2.48)\\ &= -\left(\begin{bmatrix} H_1 \\ E_1 \end{bmatrix}, \begin{bmatrix} \mu^{-1}\nabla \times E_2 \\ -\varepsilon^{-1}\nabla \times H_2 \end{bmatrix} \right)_V && (2.49)\\ &= -(v_1, Av_2)_V. && (2.50)\end{aligned}$$

Thus, the skew-symmetry of A is shown. Next we prove the coincidence of the domains of A and A^*, that is, we show that both domains contain each other, $D(A) \subset D(A^*)$ and $D(A^*) \subset D(A)$. The domain of the adjoint A^* is

$$D(A^*) = \{v_2 \in V \,|\, \exists v_3 \in V \ \forall v_1 \in D(A) : (Av_1, v_2)_V = (v_1, v_3)_V\}. \tag{2.51}$$

Let $v_2 \in D(A)$ and set $v_3 = -Av_2$. Then, for all $v_1 \in D(A)$ we have

$$(Av_1, v_2)_V = (v_1, v_3)_V \tag{2.52}$$

and $D(A) \subset D(A^*)$. Now let $v_2 = [H_2, E_2]^T \in D(A^*)$. By the definition of $D(A^*)$, there is $v_3 = [H_3, E_3]^T \in V$ such that for all $v_1 = [H_1, E_1]^T \in D(A)$ we have

$$(Av_1, v_2)_V = (v_1, v_3)_V, \tag{2.53}$$

or equivalently

$$(\nabla \times E_1, H_2)_{L^2(\Omega)^3} - (\nabla \times H_1, E_2)_{L^2(\Omega)^3} = (\mu H_1, H_3)_{L^2(\Omega)^3} + (\varepsilon E_1, E_3)_{L^2(\Omega)^3}. \tag{2.54}$$

Choosing $H_1 = 0$, we have

$$(\nabla \times E_1, H_2)_{L^2(\Omega)^3} = (\varepsilon E_1, E_3)_{L^2(\Omega)^3}. \tag{2.55}$$

Then (2.55) holds for all $E_1 \in H_0(curl,\Omega)$, so it also holds for all $E_1 \in C_0^\infty(\Omega)^3$. Then we have

$$\int_\Omega (\nabla \times E_1)\cdot H_2 = \int_\Omega \varepsilon E_1 \cdot E_3, \qquad \forall E_1 \in C_0^\infty(\Omega)^3. \tag{2.56}$$

The Definition 2.2 of variational curl allows us to we conclude that $\nabla \times H_2 = \varepsilon E_3 \in L^2(\Omega)^3$ and thus $H_2 \in H(curl,\Omega)$.

If we choose $E_1 = 0$ in (2.54), we get

$$-(\nabla \times H_1, E_2)_{L^2(\Omega)^3} = (\mu H_1, H_3)_{L^2(\Omega)^3}, \tag{2.57}$$

and, using the same argument as above, $E_2 \in H(curl,\Omega)$ with $\nabla \times E_2 = -\mu H_3 \in L^2(\Omega)^3$ and

$$\int_\Omega (\nabla \times H_1)\cdot E_2 = \int_\Omega H_1 \cdot (\nabla \times E_2), \qquad \forall H_1 \in H(curl,\Omega). \tag{2.58}$$

Theorem 2.5 allows us to conclude that this equation also holds for all function $H_1 \in C^\infty(\overline{\Omega})^3$, that is,

$$\int_\Omega (\nabla \times H_1)\cdot E_2 = \int_\Omega H_1 \cdot (\nabla \times E_2), \qquad \forall H_1 \in C^\infty(\overline{\Omega})^3. \tag{2.59}$$

But according to Lemma 2.7 we have $E_2 \in H_0(curl,\Omega)$. Thus we have proven that $v_2 = [H_2, E_2]^T \in D(A)$ and consequently the inclusion $D(A^*) \subset D(A)$ is shown. □

From this theorem we reach the following result:

Corollary 2.11. *We have $(Av, v)_V = 0$ for all $v \in D(A)$.*

Now we have all the necessary results to prove the well-posedness of the evolution equation (2.43).

Theorem 2.12. *(Well-posedness) The operator $-A$ generates a C_0-group of unitary operator*

$$T : \mathbb{R} \to L(V,V), \qquad t \mapsto e^{-tA}. \tag{2.60}$$

Thereafter, for every initial value $u_0 \in D(A)$ the homogeneous evolution equation (2.43) has a unique solution $u \in C^1(\mathbb{R}^+, V) \cap C(\mathbb{R}^+, D(A))$ given by

$$u(t) = T(t)u_0. \tag{2.61}$$

Moreover, the electromagnetic energy is conserved,

$$\|u(t)\|_V = \|u_0\|_0, \qquad \forall t \geq 0. \tag{2.62}$$

Proof. In [21, Theorem 2.2], it is shown that the homogeneous evolution equation (2.43) is well-posed if and only if the operator $-A$ generates a C_0-semigroup, $T(\cdot)$. Thus, the solution of (2.43) is given by $u = T(\cdot)u_0$, for every initial value $u_0 \in D(A)$. Theorem 2.10 and Stone's Theorem A.1 allow us to conclude that $-A$ even generates a C_0-group of unitary operators.

Conservation of the electromagnetic energy is an immediate consequence of the unitary property of the C_0-group. □

Chapter 3

Discontinuous Garlerkin methods

In order to construct a spatial discretization of Maxwell's equations by dG methods, we need to construct finite dimensional function spaces in which we search for an approximate solution. First we discretize the domain Ω using a mesh and then we construct the approximation space as the space of all functions which are polynomials on each mesh element. This leads to the necessity of introducing some concepts about meshes, broken polynomials spaces (see Appendix A.1) and admissible mesh sequences (see Appendix A.2).

Then we construct the discrete bilinear forms of the dG method and prove the stability and convergence of the method based on [24] and [19].

3.1 Concepts about meshes

Let us begin by introducing some concepts about meshes from [20].

Definition 3.1. *(Simplex) Let $\{x_0,\dots,x_d\}$ be a set of $d+1$ points in $\mathbb{R}^d$ such that the vectors $\{x_1-x_0,\dots,x_d-x_0\}$ are linearly independent. Then, the interior of the convex hull of $\{x_0,\dots,x_d\}$ is called a non-degenerate simplex of $\mathbb{R}^d$ and the points $\{x_0,\dots,x_d\}$ its vertices.*

In dimension 1, a non-degenerate simplex is an open interval, in dimension 2 a triangle and in dimension 3 a tetrahedron.

Definition 3.2. *(Simplicial mesh) A finite set $T=\{K\}$ is called a simplicial mesh of the domain Ω if*

i) *every $K\in T$ is a non-degenerate simplex;*

ii) *the set T forms a partition of Ω, that is*

$$\overline{\Omega}=\bigcup_{K\in T}\overline{K}, \tag{3.1}$$

and for every $K_1,K_2\in T,\ K_1\neq K_2$, it holds

$$K_1\cap K_2=\emptyset. \tag{3.2}$$

Each $K\in T$ is called a mesh element.

Definition 3.3. *(General mesh) A general mesh T of the domain Ω is a finite collection of polyhedra $T = \{K\}$ satisfying condition **ii)** of Definition 3.2. Each element $K \in T$ is called a mesh element.*

Definition 3.4. *(Element diameter and meshsize) Let T be a general mesh of the domain Ω. We denote with h_K the diameter of a mesh element $K \in T$. Moreover, we define the meshsize h as the largest diameter in the mesh*

$$h := \max_{K \in T} h_k. \tag{3.3}$$

From now on we will denote a mesh T with meshsize h as T_h.

Definition 3.5. *(Element outward normal) Let T_h be a mesh of the domain Ω and $K \in T_h$. We define n_K a.e. on ∂K as the unit outward normal to K.*

Now we introduce some further concepts related to meshes frequently used in the dG method which are also from [20].

Definition 3.6. *(Mesh faces) Let T_h be a mesh of the domain Ω. We say that a closed subset F of $\overline{\Omega}$ is a mesh face if F has positive $(d-1)$-measure and either one of the following conditions is satisfied:*

- ***i)*** *There are distinct mesh elements $K_1, K_2 \in T_h$ such that $F = \partial K_1 \cap \partial K_2$. In this case, we call F an interface.*
- ***ii)*** *There is a mesh element $K \in T_h$ such that $F = \partial K \cap \partial \Omega$. In this case, we call F a boundary face.*

We group interfaces in the set F_h^i and boundary faces in the set F_h^b. Hereinafter, we set

$$F_h := F_h^i \cup F_h^b. \tag{3.4}$$

Moreover, for any mesh element $K \in T_h$ we group the mesh faces composing the boundary of K in the set

$$F_K := \{F \in F_h | F \subset \partial K\}. \tag{3.5}$$

At last, we denote the maximum number of mesh faces composing the boundary of mesh elements by

$$N_\partial := \max_{K \in T_h} \mathrm{card}(F_K). \tag{3.6}$$

Let us introduce the following notation: For every mesh element $K \in T_h$ and every corresponding interface $F \in F_K \cap F_h^i$ we denote the neighbouring mesh element with respect to F with K_F.

Definition 3.7. *(Face normals) For all $F \in F_h$ we define the unit normal n_F to F as follows:*

- ***i)*** *The unit normal n_K to F pointing from K to K_F if $F \in F_h^i$. The orientation of n_F is arbitrary depending on the choice of K, but kept fixed in what follows.*
- ***ii)*** *The outward unit normal n to Ω if $F \in F_h^b$.*

Next, we introduce averages and jumps across interfaces of piecewise smooth functions. Let us begin by introducing the following notation

$$v_K := v|_K. \qquad v_{K_F} := v|_{K_F}. \tag{3.7}$$

Definition 3.8. *(Interface averages and jumps) Let v be a scalar-valued function and assume that for every mesh element $K \in T_h$ its restriction $v|_K$ is smooth enough to admit a trace a.e. on the boundary ∂K. Then, for all $F \in F_h^i$ the function v admits a possible two-valued trace and we define*

i) *the average of v on F as*

$$\{\{u\}\}_F := \frac{1}{2}\left((v_k)|_F + (v_{K_F})|_F\right), \tag{3.8}$$

ii) *the jump of v on F as*

$$[[u]]_F := (v_{K_F})|_F - (v_K)|_F. \tag{3.9}$$

When v is vector-valued, the above average and jump operators act componentwise on v.

3.2 Homogeneous medium and normalized form

As from now on we are considering a homogeneous medium, we assume that the coefficients ε and μ are positive constants. As stated previously, we have the homogeneous evolution problem: Given $u_0 = [H_0, E_0]^T \in D(A)$ we search for $u = [H, E]^T \in C^1(0,T;V) \cap C(0,T;D(A))$ with $u(0) = u_0$ and such that

$$\frac{\partial H}{\partial t} + \mu^{-1}\nabla \times E = 0, \qquad \text{in } (0,T)\times\Omega, \tag{3.10a}$$

$$\frac{\partial E}{\partial t} - \varepsilon^{-1}\nabla \times H = 0, \qquad \text{in } (0,T)\times\Omega. \tag{3.10b}$$

From now on we restrict our considerations to bounded time intervals. The fact that the coefficients ε and μ are constant allows us to rewrite (3.10) as

$$\frac{\partial \widetilde{H}}{\partial t} + c_0\nabla \times \widetilde{E} = 0, \qquad \text{in } (0,T)\times\Omega, \tag{3.11a}$$

$$\frac{\partial \widetilde{E}}{\partial t} - c_0\nabla \times \widetilde{H} = 0, \qquad \text{in } (0,T)\times\Omega, \tag{3.11b}$$

where $c_0 := (\varepsilon\mu)^{-1/2}$ is the speed of light in the medium and we set

$$\widetilde{H} := \mu^{1/2}H, \quad \widetilde{E} := \varepsilon^{1/2}E. \tag{3.12}$$

In (3.11), all quantities are normalized to the same physical unit.

The space discretization will be based on the following formulation of (3.11): Given $\widetilde{u}_0 = [\widetilde{H}_0, \widetilde{E}_0]^T \in D(A)$ we search for $\widetilde{u} = [\widetilde{H}, \widetilde{E}]^T \in C^1(0,T;V) \cap C(0,T;D(A))$ such that for all test functions in the state space $\varphi = [\phi, \psi]^T \in V$ we have

$$\left(\frac{\partial \widetilde{H}}{\partial t}, \phi\right)_{L^2(\Omega)^3} + c_0\left(\nabla \times \widetilde{E}, \phi\right)_{L^2(\Omega)^3} + \left(\frac{\partial \widetilde{E}}{\partial t}, \psi\right)_{L^2(\Omega)^3} - c_0\left(\nabla \times \widetilde{H}, \psi\right)_{L^2(\Omega)^3} = 0. \tag{3.13}$$

In the following formulation, we group the inner products in two bilinear forms according to the kind of derivative they involve and we write (3.13) in a compact form: We search for $\widetilde{u} \in C^1(0,T;V) \cap$

$C(0,T;D(A))$ such that

$$m\left(\frac{\partial \widetilde{u}}{\partial t},\varphi\right)+a(\widetilde{u},\varphi)=0,\qquad \forall\varphi\in V, \tag{3.14}$$

where we define the bilinear forms $m,\ a:\ D(A)\times V\to\mathbb{R}$ as follows: For $v=[H,E]^T$ and $\varphi=[\phi,\psi]^T$,

$$m(v,\varphi):=(H,\phi)_{L^2(\Omega)^3}+(E,\psi)_{L^2(\Omega)^3}, \tag{3.15a}$$
$$a(v,\varphi):=c_0(\nabla\times E,\phi)_{L^2(\Omega)^3}-c_0(\nabla\times H,\psi)_{L^2(\Omega)^3}. \tag{3.15b}$$

3.3 Concepts about discrete bilinear forms

In the dG discretization we replace the continuous bilinear forms by discretized ones, thus allowing for the approximation to the continuous problem (3.14) in a finite dimensional space, which can be solved computationally.

The dG discretization works with discontinuous elementwise smooth functions and we will frequently have to consider averages and jumps over interfaces of such functions. However, functions in the graph space $D(A)$ do not necessarily admit an L^2-trace. Therefore, we will require more regularity from the exact solution. Thus, we assume that the exact solution $\widetilde{u}=[H,E]^T$ of (3.14) satisfies

$$\widetilde{u}\in V_*:=D(A)\cap\left(H^1(T_h)^3\times H^1(T_h)^3\right). \tag{3.16}$$

From Remark **??**, we have that the $\widetilde{E}$-field vanishes on boundary faces, that is,

$$n\times\widetilde{E}=0,\qquad \forall F\in F_h^b. \tag{3.17}$$

Moreover, using Lemma A.9 we conclude that the exact solution only admits zero tangential jumps on interfaces,

$$n_F\times[[\widetilde{H}]]_F=n_F\times[[\widetilde{E}]]_F=0,\qquad \forall F\in F_h^i. \tag{3.18}$$

There are two more spaces needed to construct the discrete bilinear forms. First, we want to contruct the discrete solution in the broken polynomial space $\mathbb{P}_3^N(T_h)^3\times\mathbb{P}_3^N(T_h)^3$ defined in (A.18), assuming that T_h belongs to an admissible mesh sequence (see Appendix A.2). Thus, we define the discrete solution space as

$$V_h:=\mathbb{P}_3^N(T_h)\times\mathbb{P}_3^N(T_h)^3. \tag{3.19}$$

This discrete solution space is not contained in the continuous space, that is, $V_h\not\subset V_*$ (see Lemma A.9), which characterizes dG methods as non-conforming methods, hence the necessity of introducing the additional space

$$V_{*h}:=V_*+V_h, \tag{3.20}$$

which contains both the exact and the discrete solutions. Thus, V_{*h} also contains the error function of the discretization, that is, it contains the difference between the exact and the discrete solution. This guarantees that the error function can be plugged into the first argument of the discrete bilinear forms, which is essential for the later convergence analysis.

Now, in order to find the discrete bilinear forms, we begin by assuming that the global solution is approximated by the piecewise N-order polynomial approximation $\widetilde{u}_h = [\widetilde{H}_h, \widetilde{E}_h]^T$

$$\widetilde{u}(x,y,z,t) \simeq \widetilde{u}_h(x,y,z,t) = \bigoplus_{k=1}^{K} \widetilde{u}_{h,k}(x,y,z,t) \tag{3.21}$$

defined as the direct sum of the local polynomial solutions $\widetilde{u}_{h,k} = [\widetilde{H}_{h,k}, \widetilde{E}_{h,k}]^T$.

In order to achieve the semi-discrete scheme, we begin by substituting $\widetilde{u}_h$ by $\widetilde{u}_{h,k}$ in (3.11), multiplying the equations by a test function $\varphi_h = [\phi_h, \psi_h]^T \in V_h$ and integrating over each element K. We get

$$\int_K \left(\frac{\partial \widetilde{H}_K}{\partial t} \cdot \phi_h + c_0 (\nabla \times \widetilde{E}_K) \cdot \phi_h \right) = 0, \tag{3.22a}$$

$$\int_K \left(\frac{\partial \widetilde{E}_K}{\partial t} \cdot \psi_h - c_0 (\nabla \times \widetilde{H}_K) \cdot \psi_h \right) = 0, \tag{3.22b}$$

where we have dropped the index h in writing H_K and E_K instead of $H_{h,K}$ and $E_{h,K}$, respectively. Integrating by parts, we obtain

$$\int_K \left(\frac{\partial \widetilde{H}_K}{\partial t} \cdot \phi_h + c_0 \widetilde{E}_K \cdot (\nabla \times \phi_h) \right) + c_0 \int_{\partial K} (n_K \times \widetilde{E}_K) \cdot \phi_h = 0, \tag{3.23a}$$

$$\int_K \left(\frac{\partial \widetilde{E}_K}{\partial t} \cdot \psi_h - c_0 \widetilde{H}_K \cdot (\nabla \times \psi_h) \right) - c_0 \int_{\partial K} (n_K \times \widetilde{H}_K) \cdot \psi_h = 0. \tag{3.23b}$$

Since $\widetilde{H}_h, \widetilde{E}_h$ are not continuous in tangential directions on the boundary of elements, boundary integrals would not be well defined. Therefore we replace $n_K \times \widetilde{E}_K$ and $n_K \times \widetilde{H}_K$ by numerical fluxes $(n_K \times \widetilde{E}_K)^*$ and $(n_K \times \widetilde{H}_K)^*$, respectively. Let $F = \partial K \cap \partial K_F$. The simplest choice for the numerical fluxes is the central flux, that is,

$$(n_K \times \widetilde{E}_K)^*\big|_F = n_K \times \frac{\widetilde{E}_K + \widetilde{E}_{K_F}}{2}, \qquad (n_K \times \widetilde{H}_K)^*\big|_F = n_K \times \frac{\widetilde{H}_K + \widetilde{H}_{K_F}}{2}. \tag{3.24}$$

Inserting this into (3.23), we get the semi-discrete scheme

$$\int_K \left(\frac{\partial \widetilde{H}_K}{\partial t} \cdot \phi_h + c_0 \widetilde{E}_K \cdot (\nabla \times \phi_h) \right) + c_0 \int_{\partial K} \left(n_K \times \frac{\widetilde{E}_K + \widetilde{E}_{K_F}}{2} \right) \cdot \phi_h = 0, \tag{3.25a}$$

$$\int_K \left(\frac{\partial \widetilde{E}_K}{\partial t} \cdot \psi_h - c_0 \widetilde{H}_K \cdot (\nabla \times \psi_h) \right) - c_0 \int_{\partial K} \left(n_K \times \frac{\widetilde{H}_K + \widetilde{H}_{K_F}}{2} \right) \cdot \psi_h = 0. \tag{3.25b}$$

This is called the local weak form of a general dG method for Maxwell's equations. Integrating by parts once more and taking function values from the elements K, we derive the local strong form of

the dG method:

$$\int_K \left(\frac{\partial \widetilde{H}_K}{\partial t} \cdot \phi_h + c_0 (\nabla \times \widetilde{E}_K) \cdot \phi_h \right) + c_0 \int_{\partial K} \frac{1}{2} \phi_h \cdot \left((n_K \times [[\widetilde{E}_K)]]_F \right) = 0, \tag{3.26a}$$

$$\int_K \left(\frac{\partial \widetilde{E}_K}{\partial t} \cdot \psi_h - c_0 (\nabla \times \widetilde{H}_K) \cdot \psi_h \right) - c_0 \int_{\partial K} \frac{1}{2} \psi_h \cdot \left((n_K \times [[\widetilde{H}_K)]]_F \right) = 0. \tag{3.26b}$$

To obtain a global formulation we have to sum over all elements. On each inner face $F \in F_h^i$ we get

$$\frac{1}{2} \int_F \left(n_F \times [[\widetilde{E}_h]]_F \right) \cdot (\phi_K + \phi_{K_F}), \quad \frac{1}{2} \int_F \left(n_F \times [[\widetilde{H}_h]]_F \right) \cdot (\psi_K + \psi_{K_F}), \tag{3.27}$$

respectively. If $F = \partial K \cap \partial K_F$ is a boundary face, we model the boundary conditions in the following way:

$$\begin{aligned} &(n_F \times \widetilde{E}_h)^* \big|_F = 0 \ \text{ since we have } n_F \times \widetilde{E} = 0 \text{ for the exact solution,} \\ &(n_F \times \widetilde{H}_h)^* \big|_F = (n_F \times \widetilde{H}) \big|_F \ \text{ since we have no boundary for } \widetilde{H}. \end{aligned}$$

Therefore we get the following global formulation:

$$\begin{aligned} \int_\Omega \left(\frac{\partial \widetilde{H}_h}{\partial t} \cdot \phi_h + c_0 (\nabla_h \times \widetilde{E}_h) \cdot \phi_h \right) + c_0 \sum_{F \in F_h^i} \int_F \left(n_F \times [[\widetilde{E}_h)]]_F \right) \cdot \{\{\phi_h\}\}_F \\ + c_0 \sum_{F \in F_h^b} \int_F -(n_F \times \widetilde{E}_h) \cdot \psi_h = 0, \end{aligned} \tag{3.28a}$$

$$\int_K \left(\frac{\partial \widetilde{E}_K}{\partial t} \cdot \psi_h - c_0 (\nabla \times \widetilde{H}_K) \cdot \psi_h \right) - c_0 \sum_{F \in F_h^i} \int_F \left(n_F \times [[\widetilde{H}_h)]]_F \right) \cdot \{\{\psi_h\}\}_F = 0. \tag{3.28b}$$

Now we can write the discretization of (3.14): We search for $\widetilde{u}_h = [\widetilde{H}_h, \widetilde{E}_h]^T \in C^1(0,T;V_h)$ such that

$$m_h \left(\frac{\partial \widetilde{u}_h}{\partial t}, \varphi_h \right) + a_h(\widetilde{u}_h, \varphi_h) = 0, \quad \forall \varphi_h \in V_h, \tag{3.29}$$

where we define the bilinear forms $m_h,\ a_h :\ V_{*h} \times V_h \to \mathbb{R}$ as follows: For $v = [H,E]^T$ and $\varphi_h = [\phi_h, \psi_h]^T$,

$$m_h(v, \varphi_h) := (H, \phi_h)_{L^2(\Omega)^3} + (E, \psi_h)_{L^2(\Omega)^3}, \tag{3.30a}$$

$$\begin{aligned} a_h(v, \varphi_h) := {} & c_0 (\nabla_h \times E, \phi_h)_{L^2(\Omega)^3} - c_0 (\nabla_h \times H, \psi_h)_{L^2(\Omega)^3} \\ & + c_0 \sum_{F \in F_h^i} \left[(n_F \times [[E]]_F, \{\{\phi_h\}\}_F)_{L^2(F)^3} - (n_F \times [[H]]_F, \{\{\psi_h\}\}_F)_{L\ 2(F)^3} \right] \\ & + c_0 \sum_{F \in F_h^b} \left[-(n \times E, \phi_h)_{L^2(F)^3} \right]. \end{aligned} \tag{3.30b}$$

This bilinear form is clearly consistent, that is, for the exact solution $\widetilde{u} \in V_*$ we have

$$a_h(\widetilde{u}, \varphi_h) = a(\widetilde{u}, \varphi_h), \qquad \forall \varphi_h \in V_h, \tag{3.31}$$

because of (3.17) and (3.18). This is true for every $v \in V_*$.

This fact and the following lemma ensures we have constructed a meaningful discrete bilinear form a_h.

Lemma 3.9. *(Skew-adjointness of a_h) The discrete bilinear form a_h is skew-adjoint on V_h, that is,*

$$a_h(v_h, \widehat{v}_h) = -a_h(\widehat{v}_h, v_h), \qquad \forall v_h, \widehat{v}_h \in V_h. \tag{3.32}$$

Proof. Let $v_h = [H_h, E_h]^T$, $\widehat{v}_h = [\widehat{H}_h, \widehat{E}_h]^T \in V_h$. We integrate by parts the curl terms in (3.30b)

$$\begin{aligned}(\nabla_h \times E_h, \widehat{H}_h)_{L^2(\Omega)^3} - &(\nabla_h \times H_h, \widehat{E}_h)_{L^2(\Omega)^3} \\ &= (E_h, \nabla_h \times \widehat{H}_h)_{L^2(\Omega)^3} - (H_h, \nabla_h \times \widehat{E}_h)_{L^2(\Omega)^3} \\ &= \sum_{K \in T_h} \sum_{F \in F_K} \left[(n_K \times E_K, \widehat{H}_K)_{L^2(F)^3} - (n_K \times H_K, \widehat{E}_K)_{L^2(F)^3} \right]. \end{aligned} \tag{3.33}$$

We can write the last sum as

$$\begin{aligned} &\sum_{F \in F_h^i} \left[-(n_F \times [[E_h]]_F, \{\{\widehat{H}_h\}\}_F)_{L^2(F)^3} + (n_F \times [[\widehat{H}_h]], \{\{E_h\}\}_F)_{L^2(F)^3} \right] \\ &+ \sum_{F \in F_h^i} \left[(n_F \times [[H_h]]_F, \{\{\widehat{E}_h\}\}_F)_{L^2(F)^3} - (n_F \times [[\widehat{E}_h]], \{\{H_h\}\}_F)_{L^2(F)^3} \right] \\ &+ \sum_{K \in F_h^b} \left[(n \times E_h, \widehat{H}_h)_{L^2(F)^3} - (n \times H_h, \widehat{E}_h)_{L^2(F)^3} \right]. \end{aligned}$$

Using this in (3.30b), we have

$$\begin{aligned} a_h(v_h, \widehat{v}_h) =& c_0 (E_h, \nabla : h \times \widehat{H}_h)_{L^2(\Omega)^3} - c_0 (H_h, \nabla_h \times \widehat{E}_h)_{L^2(\Omega)} \\ &+ c_0 \sum_{F \in F_h^i} \left[(n_F \times [[H_h]]_F, \{\{\widehat{E}_h\}\}_F)_{L^2(F)^3} - (n_F \times [[\widehat{E}_h]], \{\{H_h\}\}_F)^{L^2(F)^3} \right] \\ &+ c_0 \sum_{K \in F_h^b} \left[-(n \times H_h, \widehat{E}_h)_{L^2(F)^3} \right] \\ =& - a_h(\widehat{v}_h, v_h). \end{aligned} \tag{3.34}$$

□

From now on we have the following equivalent representation of a_h where we use the convention $n_F \times [[\psi_h]]_F = -n \times \psi_h$ for boundary faces $F \in F_h^b$.

The discrete bilinear form a_h can be equivalently written as: For $v = [H,E]^T \in V_{*h}$ and $\varphi_h = [\phi_h, \psi_h]^T \in V_h$,

$$\begin{aligned} a_h(v,\varphi_h) =& c_0(E, \nabla_h \times \phi_h)_{L^2(\Omega)^3} - c_0(H, \nabla_h \times \psi_h)_{L^2(\Omega)^3} \\ &+ c_0 \sum_{F \in F_h^i} \left[(\{\{E\}\}_F, n_F \times [[\phi_h]]_F)_{L^2(\Omega)^3} - (\{\{H\}\}_F, n_F \times [[\psi_h]]_F)_{L^2(\Omega)^3} \right] \quad (3.35) \\ &+ c_0 \sum_{F \in F_h^b} \left[-(H, n_F \times [[\psi_h]]_F)_{L^2(\Omega)^3} \right]. \quad (3.36) \end{aligned}$$

Now we construct a seminorm using the bilinear form $s_h : V_{*h} \times V_h \to \mathbb{R}$ given as: For $v = [H,E]^T \in V_{*h}$ and $\varphi_h = [\phi_h, \psi_h]^T \in V_h$,

$$\begin{aligned} s_h(v,\varphi_h) :=& c_0 \sum_{F \in F_h^i} \left[\frac{1}{2}(n_F \times [[H]]_F, n_F \times [[\phi_h]]_F)_{L^2(F)^3} + \frac{1}{2}(n_F \times [[E]]_F, n_F \times [[\psi_h]]_F)_{L^2(F)^3} \right] \\ &+ c_0 \sum_{F \in F_h^b} (n \times E, n \times \varphi_h)_{L^2(F)^3}. \quad (3.37) \end{aligned}$$

Definition 3.10. *(S-seminorm) We define the S-seminorm on V_{*h} as*

$$|v|_S := \left(s_h(v,v)\right)^{1/2}, \qquad \forall v \in V_{*h}. \quad (3.38)$$

Obviously, this is not a norm. We see from (3.17) and (3.18) that for all functions $v \in V_*$ there holds

$$|v|_S = 0. \quad (3.39)$$

3.4 Boundness of discrete bilinear forms

It is now necessary to add a new assumption to the concepts and results about the mesh sequence in Appendix A.2.

We assume that the mesh sequence T_H is quasi-uniform, meaning that there is a constant C_{qu} such that for all $h \in H$ there holds

$$\max_{K \in T_h} h_K \leq C_{qu} \min_{K \in T_h} h_K. \quad (3.40)$$

For a function $v = [H,E]^T$ we use the convention

$$\nabla \times v = \begin{bmatrix} \nabla \times H \\ \nabla \times E \end{bmatrix}, \quad (3.41)$$

and analogously for $\nabla_h \times v$, $\{\{v\}\}_F$ and $[[v]]_F$. Let us begin by presenting a result for the boundness of the central fluxes bilinear form.

Theorem 3.11. *(Boundness of a_h) For the central fluxes bilinear form we have for all $v \in V_{*h}$ and for all $\varphi_h \in V_h$,*

$$|a_h(v,\varphi_h)| \leq \left(c_0^2 \|\nabla_h \times v\|_V + C_{bnd} c_0^{1/2} h^{-1/2} |v|_S \right) \|\varphi_h\|_V, \quad (3.42)$$

with $C_{bnd} = (\max\{\varepsilon^{-1/2}, \mu^{-1/2}\} N_\partial^{1/2} + \mu^{-1/2}) C_{tr} C_{qu}^{-1/2}$ *independent of* h.

Proof. Let $v = [H,E]^T \in V_{*h}$ and $\varphi_h = [\phi_h, \psi_h]^T \in V_h$. First, we bound the two curl terms in $a_h(v, \varphi_h)$. We have

$$\begin{aligned}
&c_0(\nabla_h \times E, \phi_h)_{L^2(\Omega)^3} - c_0(\nabla_h \times H, \psi_h)_{L^2(\Omega)^3} \\
&\qquad = c_0 \sum_{K \in T_h} \left(\begin{bmatrix} \nabla \times E \\ \nabla \times H \end{bmatrix}, \begin{bmatrix} \phi_h \\ \psi_h \end{bmatrix} \right)_{L^2(K)^6} \\
&\qquad = c_0 \sum_{K \in T_h} \varepsilon^{-1/2} \mu^{-1/2} \left(\begin{bmatrix} \varepsilon^{1/2} \nabla \times E \\ \mu^{1/2} \nabla \times H \end{bmatrix}, \begin{bmatrix} \mu^{1/2} \phi_h \\ -\varepsilon^{1/2} \psi_h \end{bmatrix} \right)_{L^2(K)^6} \\
&\qquad \leq c_0 \sum_{K \in T_h} c_0 \left\| \begin{bmatrix} \varepsilon^{1/2} \nabla \times E \\ \mu^{1/2} \nabla \times H \end{bmatrix} \right\|_{L^2(K)^6} \left\| \begin{bmatrix} \mu^{1/2} \phi_h \\ -\varepsilon^{1/2} \psi_h \end{bmatrix} \right\|_{L^2(K)^6} \\
&\qquad = c_0^2 \sum_{K \in T_h} \|\nabla \times v\|_{V(K)} \, \|\varphi_h\|_{L(K)} \,,
\end{aligned}$$

where the inequality is obtained by the Cauchy-Schwarz inequality. Applying the Cauchy-Schwarz inequality again, it follows

$$\begin{aligned}
c_0^2 \sum_{K \in T_h} \|\nabla \times v\|_{V(K)} \, \|\varphi_h\|_{L(K)} &\leq c_0^2 \left(\sum_{K \in T_h} \|\nabla \times v\|_{V(K)}^2 \right)^{1/2} \left(\sum_{K \in T_h} \|\varphi_h\|_{V(K)}^2 \right)^{1/2} \\
&= c_0^2 \, \|\nabla \times v\|_V \, \|\varphi_h\|_V \,.
\end{aligned}$$

Next we bound the sum over the interfaces in $a_h(v, \varphi_h)$, that is, the terms

$$c_0 \sum_{F \in F_h^i} \left[(n_F \times [[E]]_F, \{\{\phi_h\}\}_F)_{L^2(F)^3} - (n_F \times [[H]]_F, \{\{\psi_h\}\}_F)_{L^2(F)^3} \right]. \tag{3.43}$$

Recalling the definition of average (3.8), we have

$$\{\{\phi_h\}\}_F = \frac{1}{2}(\phi_K - \phi_{K_F}), \qquad \{\{\psi_h\}\}_F = \frac{1}{2}(\psi_K - \psi_{K_F}). \tag{3.44}$$

Thus, a single summand can be written as

$$\begin{aligned}
&(n_F \times [[E]]_F, \{\{\phi_h\}\}_F)_{L^2(F)^3} - (n_F \times [[H]]_F, \{\{\psi_h\}\}_F)_{L^2(F)^3} \\
&= \left(\begin{bmatrix} \frac{1}{\sqrt{2}} n_F \times [[E]]_F \\ \frac{1}{\sqrt{2}} n_F \times [[H]]_F \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}} (\phi_K - \phi_{K_F}) \\ \frac{1}{\sqrt{2}} (\psi_K - \psi_{K_F}) \end{bmatrix} \right)_{L^2(F)^6}.
\end{aligned}$$

Let us denote this summand with S_F. The Cauchy-Schwarz inequality reveals

$$c_0 \sum_{F\in F_h^i} S_F \leq \left(c_0 \sum_{F\in F_h^i} \left\| \begin{bmatrix} \frac{1}{\sqrt{2}} n_F \times [[E]]_F \\ \frac{1}{\sqrt{2}} n_F \times [[H]]_F \end{bmatrix} \right\|^2_{L^2(F)^6} \right)^{1/2} \times \left(c_0 \sum_{F\in F_h^i} \left\| \begin{bmatrix} \frac{1}{\sqrt{2}} (\phi_K - \phi_{K_F}) \\ \frac{1}{\sqrt{2}} (\psi_K - \psi_{K_F}) \end{bmatrix} \right\|^2_{L^2(F)^6} \right)^{1/2}.$$

The first factor on the right-hand side (RHS) is equal to

$$\left(c_0 \sum_{F\in F_h^i} \left[\frac{1}{2} \|n_F \times [[E]]_F\|_{L^2(F)^3} + \frac{1}{2} \|n_F \times [[H]]_F\|_{L^2(F)^3} \right] \right)^{1/2}, \tag{3.45}$$

which can be clearly bounded by $|v|_S$. Then, using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we infer that the second factor can be estimated by

$$\left(\frac{c_0}{2} \sum_{F\in F_h^i} \left\| \begin{bmatrix} \phi_K - \phi_{K_F} \\ \psi_K - \psi_{K_F} \end{bmatrix} \right\|^2_{L^2(F)^6} \right)^{1/2} \leq \left(\frac{c_0}{2} \sum_{F\in F_h^i} \left[2 \left\| \begin{bmatrix} \phi_K \\ \psi_K \end{bmatrix} \right\|^2_{L^2(F)^6} + 2 \left\| \begin{bmatrix} \phi_{K_F} \\ \psi_{K_F} \end{bmatrix} \right\|^2_{L^2(F)^6} \right] \right)^{1/2}. \tag{3.46}$$

Due to the fact that $1 \leq \max\left\{\varepsilon^{-1}, \mu^{-1}\right\} \varepsilon$ and $1 \leq \max\left\{\varepsilon^{-1}, \mu^{-1}\right\} \mu$, we have

$$\begin{aligned} &\left(\frac{c_0}{2} \sum_{F\in F_h^i} \left\| \begin{bmatrix} \phi_K - \phi_{K_F} \\ \psi_K - \psi_{K_F} \end{bmatrix} \right\|^2_{L^2(F)^6} \right)^{1/2} \\ &\leq c_0^{1/2} \left(\sum_{F\in F_h^i} \left[\max\left\{\varepsilon^{-1}, \mu^{-1}\right\} \left\| \begin{bmatrix} \mu^{1/2} \phi_K \\ \varepsilon^{1/2} \psi_K \end{bmatrix} \right\|^2_{L^2(F)^6} + \max\left\{\varepsilon^{-1}, \mu^{-1}\right\} \left\| \begin{bmatrix} \mu^{1/2} \phi_{K_F} \\ \varepsilon^{1/2} \psi_{K_F} \end{bmatrix} \right\|^2_{L^2(F)^6} \right] \right)^{1/2} \\ &\leq c_0^{1/2} \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} \left(\sum_{F\in F_h^i} \left[\left\| \begin{bmatrix} \mu^{1/2} \phi_K \\ \varepsilon^{1/2} \psi_K \end{bmatrix} \right\|^2_{L^2(F)^6} + \left\| \begin{bmatrix} \mu^{1/2} \phi_{K_F} \\ \varepsilon^{1/2} \psi_{K_F} \end{bmatrix} \right\|^2_{L^2(F)^6} \right] \right)^{1/2}. \end{aligned} \tag{3.47}$$

Since μ and ε are constant, the discrete trace inequality Lemma A.12 remains the same for the V-norm. So we continue by applying it to (3.47), which yields the following upper bound

$$\begin{aligned} &c_0^{1/2} \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} C_{tr} \left(\sum_{F\in F_h^i} \left[h_K^{-1} \left\| \begin{bmatrix} \mu^{1/2} \phi_K \\ \varepsilon^{1/2} \psi_K \end{bmatrix} \right\|^2_{L^2(K)^6} + h_{K_F}^{-1} \left\| \begin{bmatrix} \mu^{1/2} \phi_{K_F} \\ \varepsilon^{1/2} \psi_{K_F} \end{bmatrix} \right\|^2_{L^2(K)^6} \right] \right)^{1/2} \\ &= c_0^{1/2} \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} C_{tr} \left(\sum_{F\in F_h^i} \left[h_K^{-1} \|\varphi_K\|^2_{V(K)} + h_{K_F}^{-1} \|\varphi_{K_F}\|^2_{V(K_F)} \right] \right)^{1/2}. \end{aligned} \tag{3.48}$$

By (3.40) we infer that there holds for all $K \in T_h$,

$$h_K^{-1} \leq C_{qu}^{-1} h^{-1}. \tag{3.49}$$

Thus, we can further estimate (3.48) by

$$\begin{aligned} & c_0^{1/2} \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} C_{tr} C_{qu}^{-1/2} h^{-1/2} \left(\sum_{F \in F_h^i} \left[\|\varphi_K\|_{V(K)}^2 + \|\varphi_{K_F}\|_{V(K_F)}^2 \right] \right)^{1/2} \\ & \leq c_0^{1/2} \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} C_{tr} C_{qu}^{-1/2} h^{-1/2} \left(\sum_{K \in T_h} \operatorname{card}(F_K) \|\varphi_K\|_{V(K)}^2 \right)^{1/2}. \end{aligned} \tag{3.50}$$

As stated previously, N_∂ is the maximum number of faces composing a mesh element and N_∂ is uniformly bounded with respect to the meshsize h (Lemma A.10). Thus, we have the following upper bound for (3.50)

$$\begin{aligned} & c_0^{1/2} \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} C_{tr} C_{qu}^{-1/2} N_\partial^{1/2} h^{-1/2} \left(\sum_{K \in T_h} \|\varphi_K\|_{V(K)}^2 \right)^{1/2} \\ & = c_0^{1/2} \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} C_{tr} C_{qu}^{-1/2} N_\partial^{1/2} h^{-1/2} \|\varphi_h\|_V. \end{aligned} \tag{3.51}$$

Altogether, we have shown the following bound for the sum over the interfaces

$$c_0 \sum_{F \in F_h^i} \left[(n_F \times [[E]]_F, \{\{\phi_h\}\}_F)_{L^2(F)^3} - (n_F \times [[H]]_F, \{\{\psi_h\}\}_F)_{L^2(F)^3} \right] \leq \widetilde{C} h^{-1/2} |v|_S \|\varphi_h\|_V, \tag{3.52}$$

with $\widetilde{C} = c_0^{1/2} \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} C_{tr} C_{qu}^{-1/2} N_\partial^{1/2}$.

At last, we bound the last term in $a_h(v, \varphi_h)$, that is, the sum over the boundary faces. Using the same arguments as for the interfaces, we deduce

$$\begin{aligned} c_0 \sum_{F \in F_h^b} (n \times E, \phi_h)_{L^2(F)^3} & \leq c_0 \sum_{F \in F_h^B} \left[\left(\frac{1}{\mu^{1/2}} \|n \times E\|_{L^2(F)^3} \right) \left\| \mu^{1/2} \phi_h \right\|_{L^2(F)^3} \right] \\ & \leq c_0^{1/2} \mu^{-1/2} \left(c_0 \sum_{F \in F_h^b} \|n \times E\|_{L^2(F)^3}^2 \right)^{1/2} \left(\sum_{F \in F_h^b} \left\| \mu^{1/2} \phi_h \right\|_{L^2(F)^3}^2 \right)^{1/2} \\ & \leq c_0^{1/2} \mu^{-1/2} |v|_S \left(\sum_{F \in F_h^b} \left\| \mu^{1/2} \phi_h \right\|_{L^2(F)^3}^2 \right)^{1/2} \\ & \leq c_0^{1/2} \mu^{-1/2} C_{tr} C_{qu}^{-1/2} h^{-1/2} |v|_S \left(\sum_{F \in F_h^b} \left\| \mu^{1/2} \phi_h \right\|_{L^2(K)^3}^2 \right)^{1/2} \\ & \leq c_0^{1/2} \mu^{-1/2} C_{tr} C_{qu}^{-1/2} h^{-1/2} |v|_S \|\varphi_h\|_V. \end{aligned} \tag{3.53}$$

and the proof is complete. □

Now we show a similar result for the bilinear form s_h.

Theorem 3.12. *(Boundness of s_h) Let $v \in V_{*h}$ and $\varphi_h \in V_h$. Then, for the bilinear form s_h there holds*

$$|s_h(v,\varphi_h)| \le C'_{bnd} c_0^{1/2} h^{-1/2} |v|_S \|\varphi_h\|_V , \tag{3.54}$$

where the constant $C'_{bnd} = (2^{-1/2} N_\partial^{-1/2} \max\{\varepsilon^{-1/2}, \mu^{-1/2}\} + \varepsilon^{-1/2}) C_{tr} C_{qu}^{-1/2}$.

Proof. We have

$$\begin{aligned} s_h(v,\varphi_h) =& c_0 \sum_{F\in F_h^i} \left[\frac{1}{2} (n_F \times [[H]]_F, n_F \times [[\phi_h]]_F)_{L^2(F)^3} + \frac{1}{2} (n_F \times [[E]]_F, n_F \times [[\psi_h]]_F)_{L^2(F)^3} \right] \\ &+ c_0 \sum_{F\in F_h^b} (n\times E, n\times \psi_h)_{L^2(F)^3} \\ =& \frac{c_0}{2} \sum_{F\in F_h^i} \left(\begin{bmatrix} n_F \times [[E]]_F \\ n_F \times [[H]]_F \end{bmatrix}, \begin{bmatrix} n_F \times [[\psi_h]]_F \\ n_F \times [[\phi_h]]_F \end{bmatrix} \right)_{L^2(F)^6} + c_0 \sum_{F\in F_h^b} (n\times E, n\times \psi_h)_{L^2(F)^3} . \end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned} s_h(v,\varphi_h) \le& \left(\frac{c_0}{2}\right)^{1/2} \left(\frac{c_0}{2} \sum_{F\in F_h^i} \left\| \begin{bmatrix} n_F \times [[E]]_F \\ n_F \times [[H]]_F \end{bmatrix} \right\|^2_{L^2(F)^6} \right)^{1/2} \left(\sum_{F\in F_h^i} \left\| \begin{bmatrix} n_F \times [[\psi_h]]_F \\ n_F \times [[\phi_h]]_F \end{bmatrix} \right\|^2_{L^2(F)^6} \right)^{1/2} \\ &+ c_0^{1/2} \left(c_0 \sum_{F\in F_h^b} \|n\times E\|^2_{L^2(F)^3} \right)^{1/2} \left(\sum_{F\in F_h^b} \|n_F \times \psi_h\|^2_{L^2(F)^3} \right)^{1/2} , \end{aligned}$$

The first factors can be estimated by $|v|_S$ and using $|n_F| = 1$ we get

$$s_h(v,\varphi_h) \le \left(\frac{c_0}{2}\right)^{1/2} |v|_S \left(\sum_{F\in F_h^i} \left\| \begin{bmatrix} [[\psi_h]]_F \\ [[\phi_h]]_F \end{bmatrix} \right\|^2_{L^2(F)^6} \right)^{1/2} + c_0^{1/2} |v|_S \left(\sum_{F\in F_h^b} \|\psi_h\|^2_{L^2(F)^3} \right)^{1/2} . \tag{3.55}$$

Using $1 \le \max\{\varepsilon^{-1}, \mu^{-1}\}\varepsilon$ and $1 \le \max\{\varepsilon^{-1}, \mu^{-1}\}\mu$, we have

$$\begin{aligned} s_h(v,\varphi_h) \le& \left(\frac{c_0}{2}\right)^{1/2} |v|_S \max\left\{\varepsilon^{-1/2}, \mu^{-1/2}\right\} \left(\sum_{F\in F_h^i} \left\| \begin{bmatrix} \varepsilon^{1/2}[[\psi_h]]_F \\ \mu^{1/2}[[\phi_h]]_F \end{bmatrix} \right\|^2_{L^2(F)^6} \right)^{1/2} \\ &+ c_0^{1/2} \varepsilon^{-1/2} |v|_S \left(\sum_{F\in F_h^b} \left\| \varepsilon^{1/2} \psi_h \right\|^2_{L^2(F)^3} \right)^{1/2} . \end{aligned} \tag{3.56}$$

According to Definition 3.8, we have

$$[[\psi_h]]_F := \psi_{K_F} - \psi_K, \qquad [[\phi_h]]_F := \phi_{K_F} - \phi_K. \tag{3.57}$$

Using this property in (3.56) together with inequality $(a+b)^2 \leq 2a^2+2b^2$ yields

$$s_h(v,\varphi_h) \leq \left(\frac{c_0}{2}\right)^{1/2} |v|_S \max\left\{\varepsilon^{-1/2},\mu^{-1/2}\right\} \left(\sum_{F\in F_h^i} \left[2 \left\| \begin{bmatrix} \varepsilon^{1/2}\psi_K \\ \mu^{1/2}\phi_K \end{bmatrix} \right\|^2_{L^2(F)^6} + 2 \left\| \begin{bmatrix} \varepsilon^{1/2}\psi_{K_F} \\ \mu^{1/2}\phi_{K_F} \end{bmatrix} \right\|^2_{L^2(F)^6} \right] \right)^{1/2}$$
$$+ c_0^{1/2}\varepsilon^{-1/2}|v|_S \left(\sum_{F\in F_h^b} c \left\| \varepsilon^{1/2}\psi_h \right\|^2_{L^2(F)^3} \right)^{1/2}.$$

Using the same arguments as in the deduction from (3.47) to (3.53) in the previous proof we infer

$$s_h(v,\varphi_h) \leq \left(2^{-1/2}N_\partial^{1/2}+\varepsilon^{-1/2}\right) C_{tr}C_{qu}^{-1/2}c_0^{1/2}h^{-1/2}|v|_S \,\|\varphi_h\|_V\,.$$

□

Now we can easily get the following corollary.

Corollary 3.13. *(Bound for S-seminorm) For all $v \in V_{*h}$ we have*

$$|v|_S \leq C_S' h^{-1/2}\|v\|_V, \tag{3.58}$$

with $C_S' = C_{bnd}' c_0^{1/2}$.

Proof. Let $v \in V_{*h}$ and let $v_* \in V_*$, $v_h \in V_h$ such that $v = v_* + v_h$. It was shown in (3.39) that it holds $|v_*|_S = 0$. Moreover, using Theorem 3.12, we infer

$$|v_h|_S^2 = s_h(v_h,v_h) \leq C_{bnd}' c_0^{1/2} h^{-1/2}|v_h|_S\|v_h\|_V, \tag{3.59}$$

whence by dividing through $|v_h|_S$ we see

$$|v_h|_S \leq C_{bnd}' c_0^{1/2} h^{-1/2}\|v_h\|_V. \tag{3.60}$$

Then, the corollary follows from the triangle inequality. □

3.5 Discrete operators

We have formulated Maxwell's equations as the abstract evolution problem (2.43): Search for $u \in C^1(0,T;V_*)\cap C(0,T;D(A))$ such that $u(0)=u_0$ and

$$\frac{\partial u}{\partial t} + Au = 0. \tag{3.61}$$

For the convergence analysis it is useful to write the central fluxes discretization (3.29) in an operator based notation.

Thus, we define the operator $A_h : V_{*h} \to V_h$ as

$$(A_h v, \varphi_h)_V := a_h(v,\varphi_h), \qquad \forall \varphi_h \in V_h.$$

Moreover, we define the V-projection onto V_h as $\pi_h^V : V \to V_h$ such that

$$(\pi_h^V v, \varphi_h)_V = (v, \varphi_h)_V, \qquad \forall \varphi_h \in V_h. \tag{3.62}$$

In the following sections we use the V-inner product and thus omit the index V and always assume that π_h denotes the V-projection π_h^V. Note that we have for all $v \in V$

$$\|\pi_h v\|_V = \sup_{\substack{\varphi_h \in V_h \\ \|\varphi_h\|_V = 1}} (\pi_h v, \varphi_h)_V = \sup_{\substack{\varphi_h \in V_h \\ \|\varphi_h\|_V = 1}} (v, \varphi_h)_V \leq \sup_{\substack{\varphi_h \in V_h \\ \|\varphi_h\|_V = 1}} \|v\|_V \|\varphi_h\|_V = \|v\|_V. \tag{3.63}$$

The consistency, boundness and skew-adjointness results shown in the previous sections transfer from the discrete bilinear forms to the discrete operators.

We have proven previously in Lemma 3.9 that, for the exact solution $u \in V_*$ of (2.43)

$$(A_h u, \varphi_h)_V = a_h(u, \varphi_h) = a(u, \varphi_h) = (Au, \varphi_h)_V, \qquad \forall \varphi_h \in V_h. \tag{3.64}$$

Thus, from (3.62) follows

$$A_h u = \pi_h A u. \tag{3.65}$$

The same arguments apply for all $v \in V_*$. Thus, the consistency of A_h is proven.

For $v \in V_{*h}$ we have

$$\begin{aligned} \|A_h v\|_V &= \sup_{\substack{\varphi_h \in V_h \\ \|\varphi_h\|_V = 1}} |(A_h v, \varphi_h)_V| = \sup_{\substack{\varphi_h \in V_h \\ \|\varphi_h\|_V = 1}} |a_h(v, \varphi_h)| \\ &\leq \sup_{\substack{\varphi_h \in V_h \\ \|\varphi_h\|_V = 1}} \left[\left(c_0^2 \|\nabla_h \times v\|_V + C_{bnd} c_0^{1/2} h^{-1/2} |v|_S \right) \|\varphi_H\|_V \right] \\ &= c_0^2 \|\nabla_h \times v\|_V + C_S h^{-1/2} |v|_S, \end{aligned} \tag{3.66}$$

with $C_S = C_{bnd} c_0^{1/2}$, where we used the boundness of a_h (3.42). Thus the boundness of A_h is proven.

The skew-adjointness of A_h on V_h follows directly from Lemma 3.9. Thus, for all $v_h, \widehat{v}_h \in V_h$ it holds

$$(A_h v_h, \widetilde{v_h})_V = -(A_h \widehat{v_h}, v_h)_V. \tag{3.67}$$

The discretization in operator form is: We search for $u_h \in C^1(0, T; V_h)$ such that there holds

$$\frac{\partial u_h}{\partial t} + A_h u_h = 0, . \tag{3.68}$$

We use the projection of u_0 as initial value, that is, we require $u_h(0) = \pi_h u_0$.

3.6 Stability

The following theorem states that the discrete scheme (3.68) is stable in the same sense as the continuous problem (see (2.62)).

Theorem 3.14. *(Stability of discrete scheme) Let $u_h \in V_h$ be the solution of (3.68). Then, for all $t \in [0,T]$ the following result holds:*

$$\|u_h(t)\|_V = \|\pi_h u_0\|_V, \tag{3.69}$$

In the inhomogeneous case, that is, if $u_h \in V_h$ is the solution of: We search for $u_h \in C^1(0,T;V_h)$ such that there holds

$$\frac{\partial u_h}{\partial t} + A_h u_h = \pi_h g, \tag{3.70}$$

we have

$$\|u_h(t)\|_V \leq C_0 \left(\|u_0\|_V^2 + T \int_0^t \|g(s)\|_V^2 \, ds \right), \tag{3.71}$$

with $C_0 := e$.

Proof. Multiplying (3.70) by u_h, we get

$$\left(\frac{\partial u_h}{\partial t}, u_h \right)_V + (A_h u_h, u_h)_V = (\pi_h g, u_h)_V . \tag{3.72}$$

Using the identity $\left(\frac{\partial v}{\partial t}, v \right)_V = \frac{1}{2} \frac{d}{dt} \|v\|_V^2$ and the skew-adjointness of A_h yields

$$\frac{d}{dt} \|u_h\|_V^2 = 2 (\pi_h g, u_h)_V . \tag{3.73}$$

Clearly, for $g \equiv 0$, we get assertion (3.69) by integrating (3.73) from 0 to t.

In the inhomogeneous case, we apply the weighed Young's inequality Theorem A.2 with $\gamma = T$ to (3.73) yielding

$$\frac{d}{dt} \|u_h(t)\|_V^2 \leq T \|g(t)\|_V^2 + \frac{1}{T} \|u_h(t)\|_V^2 , \tag{3.74}$$

where we also used (3.63). Integrating from 0 to t gives the inequality

$$\|u_h(t)\|_V^2 \leq \|u_h(0)\|_V^2 + T \int_0^t \|g(s)\|_V^2 \, ds + \frac{1}{T} \int_0^t \|u_h(s)\|_V^2 \, ds, \tag{3.75}$$

to which the continuous Gronwall Lemma A.3 Lemma apllies. Thus, we have

$$\|u_h(t)\|_V^2 \leq e^{t/T} \left(\|u_h(0)\|_V^2 + T \int_0^t \|g(s)\|_V^2 \, ds \right). \tag{3.76}$$

Obviously, there holds $e^{t/T} \leq e$ for $t \in [0,T]$ and by (3.63) we see $\|u_h(0)\|_V \leq \|u_0\|_V$. Hence (3.71) is proven. □

3.7 Convergence

In the following section we use the notation

$$V_{*,N+1} := D(A) \cap H^{N+1}(T_h)^6. \tag{3.77}$$

We begin by analysing the types of errors in the discretization (3.68).

Definition 3.15. *(Error types) Let $u \in V_*$ denote the exact solution of (2.43) and $u_h \in V_h$ denote the discrete solution of (3.68). We define the spatial discretization error*

$$e(t) := u(t) - u_h(t). \tag{3.78}$$

Moreover, we split the error into two parts

$$e(t) = e_\pi(t) - e_h(t), \tag{3.79}$$

where $e_\pi(t)$ is the projection error

$$e_\pi(t) := u(t) - \pi_h u(t), \tag{3.80}$$

and $e_h(t)$ is defined as

$$e_h(t) := u_h(t) - \pi_h u(t). \tag{3.81}$$

By (3.62) there holds

$$(e_\pi(t), \varphi_h)_V = 0, \qquad \forall \varphi_h \in V_h. \tag{3.82}$$

The projection error e_π arises from replacing the continuous space V_* with the finite space V_h. The projection error e_π is the minimal error we can obtain because π_h is the best approximation to u in V_h.

Lemma 3.16. *(Bounds for the projection error) Let $v \in H^{N+1}(T_h)^6$. Then, the projection error is bounded by*

$$\|v - \pi_h v\|_V \leq C_\pi h^{N+1} |v|_{H^{N+1}(T_h)^6}, \tag{3.83}$$

and its broken curl by

$$\|\nabla_h \times (v - \pi_h v)\|_V \leq C_\pi h^N |v|_{H^{N+1}(T_h)^6}. \tag{3.84}$$

The constant is given by $C_\pi := C'_{app} \max\{\mu^{1/2}, \varepsilon^{1/2}\}$ and is independent of the meshsize h.

Proof. Let $v = [H, E]^T \in H^{N+1}(T_h)^6$ and set $\xi = v - \pi_h v$, that is,

$$\xi_\pi = \begin{bmatrix} \xi_\pi^H \\ \xi_\pi^E \end{bmatrix} = \begin{bmatrix} H - \pi_h H \\ E - \pi_h E \end{bmatrix}. \tag{3.85}$$

Then, we have

$$\|\xi_\pi\|_V = \left\| \begin{bmatrix} \mu^{1/2} \xi_\pi^H \\ \varepsilon^{1/2} \xi_\pi^E \end{bmatrix} \right\|_{L^2(\Omega)^6} \leq \max\{\mu_\infty^{1/2}, \varepsilon_\infty^{1/2}\} \, \|\xi_\pi\|_{L^2(\Omega)^6}. \tag{3.86}$$

By using Lemma A.15 on each mesh element K we get

$$\|\xi_\pi\|_V \leq C'_{app} \max\{\mu_\infty^{1/2}, \varepsilon_\infty^{1/2}\} |v|_{H^{k+1}(T_h)^6}. \tag{3.87}$$

Thus, the first assertion follows. For the second assertion note that $\|\nabla_h \times \xi_\pi\|_{L^2(\Omega)^6} \leq |\xi_\pi|_{H^1(T_h)^6}$ and thus

$$\|\nabla_h \times \xi_\pi\|_V \leq \max\{\mu_\infty^{1/2}, \varepsilon_\infty^{1/2}\} \, |\xi_\pi|_{H^1(T_h)^6}. \tag{3.88}$$

The result follows from Lemma A.15. □

If the exact solution satisfies $u \in V_{*,N+1}$, Lemma 3.16 provides the bounds

$$\|e_\pi\|_V \leq C_\pi h^{N+1} |u|_{H^{N+1}(T_h)^6}, \tag{3.89}$$

and

$$\|\nabla_h \times e_\pi\|_V \leq C_\pi h^N |u|_{H^{N+1}(T_h)^6}. \tag{3.90}$$

Lemma 3.17. *(Error equations) For the error e_h we have the following discrete evolution equation*

$$\frac{\partial e_h}{\partial t} + A_h e_h = A_h e_\pi, \qquad e_h(0) = 0. \tag{3.91}$$

Proof. Projecting the continuous problem (2.43) onto V_h gives

$$\frac{\partial \pi_h u}{\partial t} + \pi_h A u = 0, \qquad \pi_h u(0) = \pi_h u_0, \tag{3.92}$$

where we used the fact that the projection operator is independent of the time t and therefore it holds $\frac{\partial \pi_h}{\partial t} = \pi_h$. Due to the consistency of the operator A_h (see 3.65)), this is equivalent to

$$\frac{\partial \pi_h u}{\partial t} + A_h u = 0. \tag{3.93}$$

For the discrete solution u_h we have $u_h(0) = \pi_h u_0$ as well as

$$\frac{\partial u_h}{\partial t} + A_h u_h = 0. \tag{3.94}$$

Obviously there holds $e_h(0) = 0$ and by substracting (3.93) from (3.94) we get

$$\frac{\partial e_h}{\partial t} + A_h e = 0. \tag{3.95}$$

Thus, (3.91) follows by the splitting of the error, that is, by $e = e_\pi - e_h$. □

The convergence of the central fluxes discretization is proven by combining the error equation (3.91) with the stability result.

Theorem 3.18. *(Convergence for central fluxes) Let $u \in C^1(0,T;V) \cap C(0,T;V_{*,N+1})$ be the exact solution of (2.43) and $u_h \in C^1(0,t;V_h)$ be the discrete solution of (3.68). Then, for the error there holds*

$$\|e(t)\|_V^2 \leq CTh^{2N} \int_0^t |u(s)|^2_{H^{N+1}(T_h)^6} \, ds + C' h^{2N+2} |u(t)|^2_{H^{N+1}(T_h)^6}, \tag{3.96}$$

with $C = 2C_0 C_\pi^2 (c_0^2 + C_S C_S'')^2$ and $C' = 2C_\pi^2$ both independent of h.

Proof. We apply the stability result for the central flux scheme (3.71) to the error equation (3.91) and obtain

$$\|e_h(t)\|_V^2 \leq C_0 T \int_0^t \|A_h e_\pi(s)\|_V^2 \, ds. \tag{3.97}$$

Using the boundness of the operator A_h (3.66) and the bound of the S-seminorm (3.58), we get

$$\begin{aligned}\|e_h(t)\|_V^2 &\leq C_0 T \int_0^t \left(c_0^2 \|\nabla_h \times e_\pi(s)\|_V + C_S h^{-1/2} |e_\pi(s)|_S\right)^2 ds \\ &\leq C_0 T \int_0^t \left(c_0^2 \|\nabla_h \times e_\pi(s)\|_V + C_S C_S'' h^{-1} \|e_\pi(s)\|_V\right)^2 ds.\end{aligned}$$

Next, we use the bounds on the projection errors (3.89) and (3.90) to infer

$$\|e_h(t)\|_V^2 \leq C_0 C_\pi^2 \left(c_0^2 + C_S C_S''\right)^2 H^{2N} T \int_0^t |u(s)|^2_{H^{N+1}(T_h)^6} ds. \tag{3.98}$$

By using Young's inequality Theorem A.2, we get the following result for the full error

$$\|e_h(t)\|_V^2 \leq 2\|e_h(t)\|_V^2 + 2\|e_\pi(t)\|_V^2 \leq 2\|e_h(t)\|_V^2 + 2C_\pi^2 h^{2N+2} |u(t)|^2_{H^{N+1}(T_h)^6}, \tag{3.99}$$

where we used (3.89) in the second inequality. Combining (3.98) and (3.99) yields the assertion. □

3.8 Full discretization

The use of the dG method for the spatial discretization of the Maxwell's equations homogeneous problem (2.43) led us to the following discrete evolution problem: Search for $u_h \in V_h$ such that

$$\frac{\partial u_h}{\partial t} = -A_h u_h(t), \qquad t \in (0,T). \tag{3.100a}$$

$$u_h(0) = \pi_h u_0. \tag{3.100b}$$

Now, in order to obtain a fully discrete problem that can be solved computationally, we discretize the semi-discrete problem (3.100) in time with an explicit Euler method. We will present the method and prove its stability and convergence based on [19].

3.8.1 The explicit Euler method

We begin by dividing the time interval $[0,T]$ into M subintervals by the equidistants points

$$0 = t_0 < t_1 < \cdots < t_M = T, \tag{3.101}$$

where $t_m = m\Delta t$ for $m = t_0, \ldots, t_M$.

We want to approximate the semi-discrete solution $u_h^m \approx u_h(t_m)$ on the discrete set $\{t_m\}_{m=1}^M$. The explicit Euler (EE) method is a single step method that uses the initial value $u_h^0 = u_h(0)$ to compute consecutively a sequence of approximations $\{u_h^m\}_m$ resorting only to the approximation from the previous step to compute the current approximation.

The EE method for (3.100) to advance from $\mathbf{u}_h^m$ to $\mathbf{u}_h^{m+1}$, separated by a time step Δt can be written as follows:

$$u_h^{m+1} = u_h^m - \Delta t A_h u_h^m. \tag{3.102}$$

In order to prove the stability and convergence of the method we need two results:

Lemma 3.19. *(Energy identity) Let $\{u_h^m\}_m$ be the EE approximation of (3.100). Then, there holds*

$$\left\|u_h^{m+1}\right\|_V^2 + 2\Delta t\,(u_h^m, A_h u_h^n)_V = \|u_h^m\|_V^2 + \|\Delta t A_h u_h^m\|_V^2\,. \tag{3.103}$$

Proof. We calculate the norm of u_h^{m+1} using the recursion (3.102) and get

$$\left\|u_h^{m+1}\right\|_V^2 = \|u_h^m - \Delta t A_h u_h^m\|_V^2 = \|u_h^m\|_V^2 - 2\Delta t\,(u_h^m, A_h u_h^n)_V + \|\Delta t A_h u_h^m\|_V^2\,. \tag{3.104}$$

□

Theorem 3.20. *(Boundness of A_h on V_h) For all $v_h \in V_h$ there holds*

$$\|A_h v_h\|_V \leq C_h c_0 h^{-1} \|v_h\|_V, \tag{3.105}$$

with the associated constant $C_h = c_0 C_{inv} + C_{bnd} C'_{bnd}$.

Proof. We show with (3.66) and Corollary 3.13 that for all $v_h \in V_h$ there holds

$$\|A_h v_h\|_V \leq c_0^2 \|\nabla_h \times v_h\|_V + C_S C'_S h^{-1} \|v_h\|_V. \tag{3.106}$$

As we defined $C_S = C_{bnd} c_0^{1/2}$ and $C'_S = C'_{bnd} c_0^{1/2}$, the second term in (3.106) meets the bound (3.105). Now, according to the definition of the broken curl and the V-norm we have for all $v_h = [H_h, E_h]^T \in V_h$,

$$\begin{aligned}\|\nabla_h \times v_h\|_V^2 &= \sum_{K\in T_h} \|\nabla \times v_h\|_{V(K)}^2 = \sum_{K\in T_h} \left\| \begin{bmatrix} \mu^{1/2}\nabla \times H_h \\ \varepsilon^{1/2}\nabla \times E_H \end{bmatrix} \right\|_{L^2(K)^6}^2 \\ &= \sum_{K\in T_h} \left(\mu \|\nabla \times H_h\|_{L^2(K)^3}^2 + \varepsilon \|\nabla \times E_h\|_{L^2(K)^3}^2 \right). \end{aligned} \tag{3.107}$$

We have

$$\|\nabla \times H_h\|_{L^2(K)^3} \leq |H_h|_{H^1(K)^3} = \|\nabla H_h\|_{L^2(K)^{3\times 3}} \leq C_{inv}^2 h_K^{-2} \|H_h\|_{L^2(K)^3}, \tag{3.108}$$

where we used the inverse inequality (A.21) componentwise in the last estimate. Analogously, we have

$$\|\nabla \times E_h\|_{L^2(K)^3} \leq C_{inv}^2 h_K^{-2} \|E_h\|_{L^2(K)^3}, \tag{3.109}$$

Inserting this into (3.107), we get

$$\begin{aligned}\|\nabla_h \times v_h\|_V^2 &= C_{inv}^2 \sum_{K\in T_h} h_K^{-2} \left(\mu \|\nabla H_h\|_{L^2(K)^3}^2 + \varepsilon \|E_h\|_{L^2(K)^3}^2 \right). \\ &= C_{inv}^2 \sum_{K\in T_h} h_K^{-2} \|v_h\|_{V(K)}^2 \leq C_{inv}^2 h^{-2} \sum_{K\in T_h} \|v_h\|_V(K)^2 = C_{inv}^2 h^{-2} \|v_h\|_V^2, \end{aligned} \tag{3.110}$$

where we used (3.40) in the last inequality. □

3.8.2 Stability

We will see that each method is stable only if the time step size is bounded with respect to the meshsize h. This condition is called the CFL condition.

Definition 3.21. *(CFL-conditions). Let ρ be a positive number. We say that the step size Δt satisfies the usual CFL condition if it holds*

$$\Delta t \leq \rho \left(\frac{h}{c_0}\right)^2. \tag{3.111}$$

Lemma 3.22. *(Stability for EE) Let $\{u_h^m\}_m$ be the EE approximation of (3.100). Then, under the CFL condition (3.111), there holds*

$$\|u_h^m\|_V^2 \leq C_1 \|u_h^0\|_V^2, \tag{3.112}$$

with the associated constant $C_1 = \exp\left(C_h^2 \rho t_m\right)$.

Proof. From the energy identity (3.103) we have

$$\left\|u_h^{m+1}\right\|_V^2 + 2\Delta t \left(u_h^m, A_h u_h^n\right)_V = \|u_h^m\|_V^2 + \|\Delta t A_h u_h^m\|_V^2 . \tag{3.113}$$

The term $\|\Delta t A_h u_h^m\|_V^2$ can be estimated by the bound (3.105) of the discrete operator A_h,

$$\left\|u_h^{m+1}\right\|_V^2 + 2\Delta t \left(u_h^m, A_h u_h^n\right)_V \leq \|u_h^m\|_V^2 + C_h^2 c_0^2 \Delta t^2 h^{-2} \|u_h^m\|_V^2 . \tag{3.114}$$

Using the CFL-condition yields

$$\left\|u_h^{m+1}\right\|_V^2 - \|u_h^m\|_V^2 + 2\Delta t \left(u_h^m, A_h u_h^n\right)_V \leq C_h^2 \rho \Delta t \|u_h^m\|_V^2 . \tag{3.115}$$

Let us denote $C_0 := C_h^2 \rho$. Summing (3.115 from 0 to $m-1$ gives

$$\|u_h^m\|_V^2 + 2\Delta t \sum_{l=0}^{m-1} \left(u_h^l, A_h u_h^l\right)_V \leq \|u_h^0\|_V^2 + C_0 \Delta t \sum_{l=0}^{m-1} \|u_h^l\|_V^2, \tag{3.116}$$

which implies

$$\|u_h^m\|_V^2 \leq \|u_h^0\|_V^2 + C_0 \Delta t \sum_{l=0}^{m-1} \|u_h^l\|_V^2. \tag{3.117}$$

This inequality meets the assumptions of the discrete Gronwall Lemma A.4 and its use provides the estimate

$$\|u_h^m\|_V^2 \leq e^{C_0 m \Delta t} \|u_h^0\|_V^2. \tag{3.118}$$

Inserting this bound in the RHS of (3.116) gives

$$\begin{aligned} \|u_h^m\|_V^2 + 2\Delta t \sum_{l=0}^{m-1} \left(u_h^l, A_h u_h^l\right)_V &\leq \|u_h^0\|_V^2 + C_0 \Delta t \sum_{l=0}^{m-1} e^{C_0 l \Delta t} \|u_h^0\|_V^2 \\ &\leq \left(1 + C_0 \Delta t \sum_{l=0}^{m-1} e^{C_0 l \Delta t}\right) \|u_h^0\|_V^2. \end{aligned} \tag{3.119}$$

Note that the sum $\Delta t \sum_{l=0}^{m-1} e^{C_0 l \Delta t}$ is a lower sum of the monotonically increasing function $e^{C_0 t}$ and thus we can deduce

$$\Delta t \sum_{l=0}^{m-1} e^{C_0 l \Delta t} \leq \int_0^{t_m} e^{C_0 s}\, ds = C_0^{-1} (e^{C_0 t_m} - 1). \tag{3.120}$$

Combining the last two inequalities yields

$$\|u_h^m\|_V^2 + 2\Delta t \sum_{l=0}^{m-1} \left(u_h^l, A_h u_h^l\right)_V \leq e^{C_0 t_m} \|u_h^0\|_V^2. \tag{3.121}$$

(3.112) follows from the skew-adjointness of A_h (3.67). □

3.8.3 Convergence

In order to prove the convergence of the EE method, we are going to need some preliminary results. The bounds for the projection error are known from Lemmas A.15 and A.16.

Lemma 3.23. *(Bounds for projection errors) Let $u \in H^{N'+1}(K)^6$ for some $N' \leq N$. The following error bounds hold*

$$\|e_\pi\|_{L^2(\Omega)^6} \leq C h^{N'+1} |u|_{H^{N'+1}(K)^6} \tag{3.122}$$

and

$$\|e_{\pi,K}\|_{L^2(F)^6} \leq C h^{N'+1/2} |u|_{H^{N'+1}(K)^6} \tag{3.123}$$

where C is independent of both h and K.

Lemma 3.24. *Let $u \in H^{N'+1}(T_h)^6$ for some $N' \leq N$. Then, for all $\varphi_h = [\phi_h, \psi_h]^T \in V_h$ and for all $\gamma > 0$ there holds*

$$(A_h e_\pi, \varphi_h)_V \leq C \left(\gamma h^{2N'} |u|_{H^{N'+1}(T_h)^6} + \frac{1}{\gamma} \|\varphi_h\|_V^2 \right), \tag{3.124}$$

where C is independent of h.

Proof. We have

$$\begin{aligned}(A_h e_\pi, \varphi_h)_V =& c_0 (e_\pi^E, \nabla_h \times \phi_h)_{L^2(\Omega)^3} - c_0 (e_\pi^H, \nabla_h \times \psi_h)_{L^2(\Omega)^3} \\ &+ c_0 \sum_{F \in F_h^i} \left[\left(\{\{e_\pi^E\}\}_F, n_F \times [[\phi_h]]_F\right)_{L^2(F)^3} + \left(\{\{e_\pi^H\}\}_F, n_F \times [[\psi_h]]_F\right)_{L^2(F)^3} \right] \\ &+ c_0 \sum_{F \in F_h^b} (e_\pi^H, n_F \times \psi_h)_{L^2(F)^3}. \end{aligned} \tag{3.125}$$

Since $e_\pi = [e_\pi^H, e_\pi^E]^T$ is the projection error and $\nabla_h \times \phi_h,\ \nabla_h \times \psi_h \in V_h$, we have

$$(e_\pi^E, \nabla \times \phi_h)_{L^2(\Omega)^3} = 0, \qquad (e_\pi^H, \nabla \times \psi_h)_{L^2(\Omega)^3} = 0 \tag{3.126}$$

and the first two terms vanish. The remaining terms can be bounded by using the Cauchy-Schwarz inequality, the triangle inequality, Lemma 3.23 and the discrete trace inequality A.12 as follows:

$$\begin{aligned}\left(\{\{e_\pi^E\}\}_F, n_F \times [[\phi_h]]_F\right)_{L^2(F)^3} &\leq \frac{1}{2}\|e_{\pi,K}^E + e_{\pi,K_F}^E\|_{L^2(F)^3}\|n_F \times (\phi_{K_F} - \phi_K)\|_{L^2(F)^3} \\ &\leq \frac{1}{2}\left(\|e_{\pi,K}^E\|_{L^2(F)^3} + \|e_{\pi,K_F}^E\|_{L^2(F)^3}\right)\left(\|\phi_{K_F}\|_{L^2(F)^3} + \|\phi_K)\|_{L^2(F)^3}\right) \\ &\leq C\left(h_K^{N'+1/2}|E|_{H^{N'+1}(K)^3} + h_{K_F}^{N'+1/2}|E|_{H^{N'+1}(K_F)^3}\right) \\ &\quad \cdot \left(h_K^{-1/2}\|\phi_K\|_{L^2(\Omega)^3} + h_{K_F}^{-1/2}\|\phi_{K_F}\|_{L^2(\Omega)^3}\right).\end{aligned}$$

By using $h = \max_K h_K$ and applying Young's inequality A.2 to each of the products we get

$$\left(\{\{e_\pi^E\}\}_F, n_F \times [[\phi_h]]_F\right)_{L^2(F)^3} \leq C\gamma h^{2N'}\left(|E|_{H^{N'+1}(K)^3} + |E|_{H^{N'+1}(K_F)^3}\right) + C\frac{1}{\gamma}\left(\|\phi_K\|_{L^2(\Omega)^3} + \|\phi_{K_F}\|_{L^2(\Omega)^3}\right). \tag{3.127}$$

Bounding the other terms analogously and summing over all faces, the lemma is proven. □

Theorem 3.25. *(Convergence for EE) Let u be the solution of (3.13) and u_h^m the discrete solution of the problem (3.100). Assume that $u \in C(0,T;H^{N+1}(T_h)^6)$ and $u'' \in C(0,T;V)$. Then, under the CFL condition (3.111), the error $e_h^m := u_h^m - \pi_h u(t_m)$ is bounded by*

$$\|e_h^M\|_V^2 \leq C_{EE}\left(\Delta t^2 \int_0^T \|u''(t)\|_V^2 \, dt + h^{2N} \max_{t\in[0,T]} |u(t)|_{H_{N+1}(T_h)^6}\right), \tag{3.128}$$

where the constant C_{EE} is independent of h and u.

Proof. For the exact solution, resorting to a Taylor expansion, we have

$$u(t_{m+1}) = u(t_m) + \Delta t \frac{\partial u}{\partial t}(t) + \int_{t_m}^{t_{m+1}} (t_{m+1} - t_m)\frac{\partial^2 u}{\partial t^2}(t)\, dt. \tag{3.129}$$

For $\eta^{m+1} := \int_{t_m}^{t_{m+1}} (t_{m+1} - t)\frac{\partial^2 u}{\partial t^2}(t)\, dt$ it holds

$$\|\eta^{m+1}\|_V \leq \Delta t \int_{t_m}^{t_{m+1}} \|u''(t)\|_V \, dt = \mathscr{O}(\Delta t^2). \tag{3.130}$$

As $\frac{\partial u}{\partial t} = -Au$, taking the L^2-projection of (3.129) yields

$$\pi_h u(t_{m+1}) = \pi_h u(t_m) - \Delta t A_h u(t_m) + \pi_h \eta^{m+1}. \tag{3.131}$$

Subtracting (3.131) from (3.100) gives the error recursion

$$e_h^{m+1} = e_h^m - \Delta t A_h e_h^m + \Delta t A_h e_\pi^m - \pi_h \eta^{m+1}. \tag{3.132}$$

Taking the V-inner product with e_h^m and using

$$(e_h^{m+1}, e_h^m)_V = \frac{1}{2}\left(\|e_h^{m+1}\|_V - \|e_h^{m+1} - e_h^m\|_V^2 + \|e_h^m\|_V^2\right), \tag{3.133}$$

we get

$$\|e_h^{m+1}\|_V^2 - \|e_h^m\|_V^2 + 2\Delta t(A_h e_h^m, e_h^m)_V = \|e_h^{m+1} - e_h^m\|_V^2 + 2\Delta t(A_h e_\pi^m, e_h^m)_V - 2(\pi_h \eta^{m+1}, e_h^m)_V. \quad (3.134)$$

For the second term of the RHS we use Lemma 3.24 with $\gamma = 1$ and we get

$$\|e_h^{m+1}\|_V^2 - \|e_h^m\|_V^2 \leq \|e_h^{m+1} - e_h^m\|_V^2 + C\Delta t h^{2N} |u(t_m)|^2_{H^{N+1}(T_h)^6} + C\Delta t \|e_H^m\|_V^2 - 2(\pi_h \eta^{m+1}, e_h^m)_V. \quad (3.135)$$

For the last term we use Young's inequality A.2 and the stability of the L^2-projection:

$$(\pi_h \eta^{m+1}, e_h^m)_V \leq \Delta t \left(\left\| \frac{\eta^{m+1}}{\Delta t} \right\|_V^2 + \|e_h^m\|_V^2 \right). \quad (3.136)$$

Now we bound the first term of the RHS of (3.135). From (3.132) we conclude

$$\begin{aligned} \|e_h^{m+1} - e_h^m\|_V^2 &= \| - \Delta t A_h e_h^m + \Delta t A_h e_\pi^m - \pi_h \eta^{m+1}\|_V^2 \\ &\leq 3 \left(\|\Delta t A_h e_h^m\|_V^2 + \|\Delta t A_h e_\pi^m\|_V^2 + \|\eta^{m+1}\|_V^2 \right). \end{aligned}$$

By using Theorem 3.20, we bound the first term as

$$\|\Delta t A_h e_h^m\|_V^2 \leq C\Delta t^2 h^{-2} \|e_h^m\|_V^2. \quad (3.137)$$

As $A_h e_\pi = A_h(u - \pi_h u) = \pi_h A u - A_h \pi u$, Lemma 3.16 implies

$$\|\Delta t A_h e_\pi^m\|_V^2 \leq C\Delta t^2 h^{2N+2} \|u(t_m)\|^2_{H^{N+1}(T_h)^6}. \quad (3.138)$$

Gathering all inequalities we have

$$\|e_h^{m+1}\|_V^2 - \|e_h^m\|_V^2 \leq C\Delta t(1 + \Delta t h^{-2}) \|e_h^m\|_V^2 + C\Delta t(h^{2N} + \Delta t h^{2N+2}) |u(t_m)|^2_{H^{N+1}(T_h)^6} + C\Delta t \left\| \frac{\eta^{m+1}}{\Delta t} \right\|_V^2. \quad (3.139)$$

After taking the CFL condition (3.111) into account and using (3.130), we get

$$\begin{aligned} &\|e_h^{m+1}\|_V^2 - \|e_h^m\|_V^2 \\ &\quad \leq C\Delta t \left(1 + \frac{\rho}{c_0^2}\right) \|e_h^m\|_V^2 + C\Delta t \left(h^{2N} + \frac{\rho}{c_0^2} h^{2N+4}\right) |u(t_m)|^2_{H^{N+1}(T_h)^6} + C\Delta t \int_{t_m}^{t^{m+1}} \left\|u''(t)\right\|_V^2 dt \\ &\quad \leq C\Delta t \|e_h^m\|_V^2 + C\Delta t h^{2N} |u(t_m)|^2_{H^{N+1}(T_h)^6} + C\Delta t \int_{t_m}^{t^{m+1}} \left\|u''(t)\right\|_V^2 dt. \end{aligned} \quad (3.140)$$

Summing over m, we get

$$\begin{aligned}
&\|e_h^M\|_V^2 - \|e_h^0\|_V^2 \\
&\quad \leq C\Delta t \sum_{m=0}^{M-1} \|e_h^m\|_V^2 + C\Delta t h^{2N} \sum_{m=0}^{M-1} |u(t_m)|^2_{H^{N+1}(T_h)^6} + C\Delta t \sum_{m=0}^{M-1} \int_{t_m}^{t^{m+1}} \left\|u''(t)\right\|_V^2 dt \\
&\quad \leq C\Delta t \sum_{m=0}^{M-1} \|e_h^m\|_V^2 + C\Delta t h^{2N} \max_{t\in[0,T]} |u(t)|^2_{H^{N+1}(T_h)^6} + C\Delta t \int_0^T \left\|u''(t)\right\|_V^2 dt. \qquad (3.141)
\end{aligned}$$

Applying the discrete Gronwall theorem A.4, we get

$$\begin{aligned}
\|e_h^M\|_V^2 &\leq (1+C\Delta t)^M \left[C\Delta t h^{2N} \max_{t\in[0,T]} |u(t)|^2_{H^{N+1}(T_h)^6} + C\Delta t \int_0^T \left\|u''(t)\right\|_V^2 dt\right] \\
&\leq C\left[h^{2N} \max_{t\in[0,T]} |u(t)|^2_{H^{N+1}(T_h)^6} + \Delta t^2 \max_{t\in[0,T]} \left\|u''(t)\right\|_V^2\right] \qquad (3.142)
\end{aligned}$$

and the theorem is proven. □

Chapter 4

Computational results

Having proven the stability and convergence of the dG method and using a temporal discretization method, we must be able to obtain the numerical results that corroborate the theoretical results. We are also interested in simulating the scattering through the human's retina, which can be done resorting to the scattered field formulation.

Thus, we intend to adapt the Matlab code from [12] for the two-dimensional vacuum Maxwell's equations in TM mode to the TE mode for inhomogeneous isotropic media with PEC conditions and proceed with the necessary simulations. This Matlab code used a fourth order four stage Runge-Kutta method as the temporal discretization method. This method has an order of convergence of Δt^s, where s is the number of stages of the Runge-Kutta method. We chose to use this method for the temporal discretization because in a previous work (see [16]) we used the Runge-Kutta method for the temporal discretization for the unidimensional case. We have already seen that usually the error in time is dominated by the error in space, and it is not easy to observe the order of convergence in time. Thus we will only analyse the spatial convergence.

4.1 The problem

We intend to solve the two-dimensional Maxwell's equations in the TE mode for homogeneous isotropic media with PEC boundary conditions. Thus, we have the following problem: Solve for $u = [H,E]^T : (0,T)\times\Omega \to \mathbb{R}^2$ such that

$$\varepsilon \frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y} \qquad \text{in } (0,T)\times\Omega, \tag{4.1a}$$

$$\varepsilon \frac{\partial E_y}{\partial t} = -\frac{\partial H_z}{\partial x} \qquad \text{in } (0,T)\times\Omega, \tag{4.1b}$$

$$\mu \frac{\partial H_z}{\partial t} = \frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} \qquad \text{in } (0,T)\times\Omega. \tag{4.1c}$$

$$E(0,x,y) = E_0 \qquad \text{in } \Omega \tag{4.1d}$$

$$H(0,x,y) = H_0 \qquad \text{in } \Omega \tag{4.1e}$$

$$n\times E = 0 \qquad \text{on } (0,T)\times\partial\Omega \quad \text{(for PEC)} \tag{4.1f}$$

where $E = (E_x, E_y)$, $H = (H_z)$ and $\Omega \in \mathbb{R}^2$, the permittivity ε and the permeability μ of the medium are space-dependent and the E_0 and H_0 are the initial values for the electric field and the magnetic field, respectively.

4.1.1 The numerical flux

In the implementation of the dG method we use the numerical fluxes computed using Riemann conditions according to [12]. The numerical fluxes for the three dimensionsal case are

$$-[n \times H - (n \times H)^*] = -\frac{1}{2\{\{Z\}\}} n \times (Z_{K_F}[[H]] - \alpha n \times [[E]]), \tag{4.2a}$$

$$[n \times E - (n \times E)^*] = \frac{1}{2\{\{Y\}\}} n \times (Y_{K_F}[[E]] + \alpha n \times [[H]]) \tag{4.2b}$$

for the equations for the electric and magnetic fields, respectively. This allows us the possibility of piecewise constant material coefficients, represented by

$$Z_K = \frac{1}{Y_K} = \sqrt{\frac{\mu_K}{\varepsilon_K}}, \tag{4.3a}$$

$$Z_{K_F} = \frac{1}{Y_{K_F}} = \sqrt{\frac{\mu_{K_F}}{\varepsilon_{K_F}}} \tag{4.3b}$$

as the local impedance and conductance, respectively. The parameter α is used to control dissipation, for example: $\alpha = 0$ for the central fluxes and $\alpha = 1$ for the upwind fluxes.

The numerical flux for the TE form is

$$n \cdot (F - F^*) = \frac{1}{2} \begin{cases} -\frac{1}{\{\{Z\}\}} Z_{K_F} n_y [[H_z]] - \alpha (n_x n \cdot [[E]] - [[E_x]]) \\ \frac{1}{\{\{Z\}\}} Z_{K_F} n_x [[H_z]] + \alpha (n_y n \cdot [[E]] - [[E_y]]) \\ \frac{1}{\{\{Y\}\}} Y_{K_F} (n_x [[E_y]] - n_y [[E^x]]) - \alpha [[H_z]] \end{cases} \tag{4.4}$$

4.1.2 The boundary conditions

The PEC boundary conditions are implemented by applying the mirror principle as $n \times E_{K_F} = -n \times E_K$ and $n \times H_{K_F} = n \times H_K$ [10]. Therefore the jumps at the outer boundary are set as

$$[[E_X]] = 2(E_x)_K, \quad [[E_y]] = 2(E_y)_K, \quad [[H_z]] = 0. \tag{4.5}$$

4.2 Implementation of dG method

According to the notation of the previous chapter, we have the following semi-discrete problem: We search for $\widetilde{u}_h = [\widetilde{H}_h, \widetilde{E}_h]^T \in C^1(0,T;V_h)$ such that

$$m_h \left(\frac{\partial \widetilde{u}_h}{\partial t}(t), \varphi_h \right) + a_h(\widetilde{u}_h(t), \varphi_h) = 0, \quad \forall \varphi_h \in V_h. \tag{4.6}$$

The bilinear form a_h is given in (3.30b) and m_h is given in (3.30a) and accords with

$$m_h(\frac{\partial \widetilde{u}_h}{\partial t}, \varphi_h) = \left(\frac{\partial \widetilde{u}_h}{\partial t}(t), \varphi_h\right)_V. \tag{4.7}$$

Let us choose a basis of V_h consisting of $N_p = \frac{(N+1)(N+2)}{2}$ vectors, where N is the polynomial degree we use in the dG discretization. We denote the basis with $\mathbf{V}_h = \{\varphi_1, \ldots, \varphi_{N_p}\}$. The dG approximation $\widetilde{u}_h(t)$ is an element of the space V_h. Thus there is a unique coefficient vector $\mathbf{u}_h(t) = [u_{h,1}(t), \ldots, u_{h,N_p}(t)]^T \in \mathbb{R}^{N_p}$ such that

$$\widetilde{u}_h(t) = \sum_{m=1}^{N_p} u_{h,m}(t)\varphi_m. \tag{4.8}$$

For (4.6) it is equivalent to hold for all $\varphi_h \in V_h$ or to hold for all basis functions $\varphi_m \in \mathbf{V}_h$. Using this concept and using (4.7) and (4.8), we have that the problem

$$\sum_{m=1}^{N_p} (\varphi_m, \varphi_l)_V \, u'_{h,m}(t) + \sum_{m=1}^{N_p} a_h(\varphi_m, \varphi_l)\, u_{h,m}(t) = 0, \qquad \forall l = 1, \ldots, N \tag{4.9}$$

is equivalent to (4.6). Thus we define the mass and stiffness matrices.

Definition 4.1. *(Mass and stiffness matrix) We define the mass matrix* $\widetilde{M} \in \mathbb{R}^{N_p \times N_p}$ *as*

$$\widetilde{M} := [(\varphi_m, \varphi_l)_V]_{l,m=1}^{N_p}, \tag{4.10}$$

and the stiffness matrix $\widetilde{A} \in \mathbb{R}^{N_p \times N_p}$ *as*

$$\widetilde{A} := [a_h(\varphi_m, \varphi_l)]_{l,m=1}^{N_p}, \tag{4.11}$$

Due to the definition of the space V_h, both matrices are sparse and the mass matrix is block diagonal and symmetric positive, thus invertible.

We write the following equivalent formulation of (4.9):

$$\widetilde{M}\mathbf{u}'_h(t) + \widetilde{A}\mathbf{u}_h(t) = 0. \tag{4.12}$$

As $\widetilde{M}$ is invertible, we get

$$\mathbf{u}'_h(t) = -\widetilde{M}^{-1}\widetilde{A}\mathbf{u}_h(t). \tag{4.13}$$

This is the semi-discrete scheme to which we apply a temporal discretization method, such as the Runge-Kutta method.

In order to proceed with the temporal discretiztion, we divide the time interval $[0, T]$ into M subintervals by the equidistants points

$$0 = t_0 < t_1 < \cdots < t_M = T, \tag{4.14}$$

where $t_m = m\Delta t$ for $m = t_0, \ldots, t_M$. A temporal discretization method (in this case the Runge-Kutta method) allows us obtain the discrete solution $u_h^m \approx u_h(t_m)$.

4.3 Spatial order of convergence

The computational domain is $\Omega = (-1,1)^2$, which is triangulated with K non-overlapping straight-sided triangles. An example of the triangular mesh used to space discretize the Maxwell's equations in TE mode is presented in Figure 4.1.

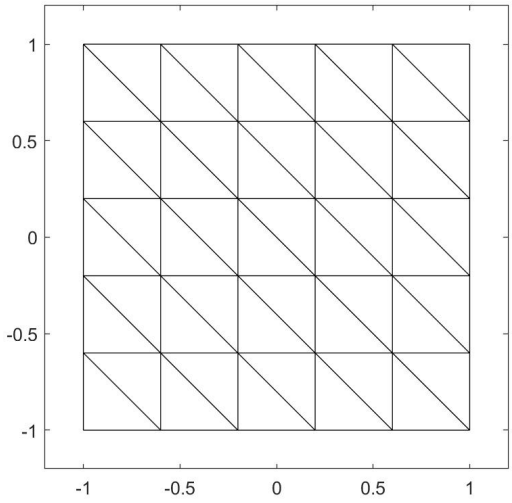

Fig. 4.1 Example of mesh for K = 50.

On each triangle we define

$$N_p = \frac{(N+1)N+2)}{2} \tag{4.15}$$

nodal points, where N is the order of the polynomial approximation.

We begin by considering the case where $\varepsilon = \mu = 1$ and $T = 1$. The initial conditions are

$$E_x(x,y,0) = 0, \tag{4.16a}$$
$$E_y(x,y,0) = 0, \tag{4.16b}$$
$$H_z(x,y,0) = \cos(\pi x)\cos(\pi y). \tag{4.16c}$$

We introduce the source terms $P(x,y,t)$ $Q(x,y,t)$ and $R(x,y,t)$ in order to more easily find examples of problems with known exact solutions, thus allowing us to compute the error of the mathematical solution. Introducing the source terms, we have the following Maxwell's equations in the TE mode:

$$\varepsilon \frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y} + P(x,y,t), \tag{4.17a}$$
$$\varepsilon \frac{\partial E_y}{\partial t} = -\frac{\partial H_z}{\partial x} + Q(x,y,t), \tag{4.17b}$$
$$\mu \frac{\partial H_z}{\partial t} = \frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} + R(x,y,t). \tag{4.17c}$$
$$\tag{4.17d}$$

The source terms are defined such that the exact solution allows for the initial conditions (4.16). If the exact solution is

$$E_x(x,y,t) = -\pi\cos(\pi x)\sin(\pi y)\sin(t), \tag{4.18a}$$
$$E_y(x,y,t) = \pi\sin(\pi x)\cos(\pi y)\sin(t), \tag{4.18b}$$
$$H_z(x,y,t) = \cos(\pi x)\cos(\pi y)\cos(t), \tag{4.18c}$$

the source terms are given by

$$P(x,y,t) = (1-\varepsilon)\pi\cos(\pi x)\sin(\pi y)\cos(t), \tag{4.19a}$$
$$Q(x,y,t) = (\varepsilon-1)\pi\sin(\pi x)\cos(\pi y)\cos(t), \tag{4.19b}$$
$$H_z(x,y,t) = (2\pi^2\mu)\cos(\pi x)\cos(\pi y)\cos(t). \tag{4.19c}$$

In order to analyse the spatial order of convergence we use a fixed time-step ($\Delta t = 10^{-4}$). We compute the error

$$\text{Error} = \|e_h^M\|_V \tag{4.20}$$

as defined in Theorem 3.25.

Now we refine the used mesh for various degrees of polynomial approximation for the central flux. The results are represented in Table 4.1. The spatial order of convergence is given by

$$\text{Order} = \frac{\log\left(\|e_h^M\|_V / \|e_{h^*}^M\|_V\right)}{\log(h/h^*)}, \tag{4.21}$$

where u_h and u_{h^*} denote the numerical solutions computed for two consecutive meshes of diameter h and h^*. The results obtained for the spatial order of convergence are approximately the expected, as we have shown in Chapter 3 that the numerical solution converges to the exact solution as the meshsize tends to zero with convergence rate h^N for the central fluxes case.

4.4 Scattered field formulation

In order to detect subcellular changes in the retina, such as the variation of size and shape of each structure, we can solve Maxwell's equations and then compute the backscattered light intensity.

We begin by separating the electromagnetic fields $u = [H,E]^T$ into two fields: the incident fields $u^i = [H^i,E^i]^T$ and the scattered components $u^s = [H^s,E^s]^T$, thus we have

$$H = H^s + H^i \quad \text{and} \quad E = E^s + E^i. \tag{4.22}$$

Assuming that the incident field u^i is also a solution of the Maxwell's equations (4.1), with coefficients ε^i and μ^i being the relative permittivity and permeability of the medium in which the incident field propagates in the absence of scatterers, we insert (4.22) into (3.10) and, using the linearity of these

Table 4.1 The L^2-error and spatial order of convergence for the central flux.

N	K	h	Error	Order
1	32	0.5	2.3202	
	50	0.4	1.8622	0.99
	200	0.2	1.1442	0.70
	800	0.1	0.6323	0.86
	3200	0.05	0.3268	0.95
2	32	0.5	1.1164	
	50	0.4	0.8348	1.30
	200	0.2	0.1191	2.31
	800	0.1	0.1684	2.08
	3200	0.05	0.0098	2.03
3	32	0.5	0.2753	
	50	0.4	0.1459	2.85
	200	0.2	0.0199	2.87
	800	0.1	0.0025	2.97
	3200	0.05	3.3291E-04	2.93
4	32	0.5	0.0602	
	50	0.4	0.0250	3.94
	200	0.2	0.0016	4.01
	800	0.1	1.1121E-04	3.81
	3200	0.05	6.1035E-05	0.87

equations, we obtain the scattered field formulation

$$\varepsilon \frac{\partial E_x^s}{\partial t} = \frac{\partial H_z^s}{\partial y} + P \tag{4.23a}$$

$$\varepsilon \frac{\partial E_y^s}{\partial t} = -\frac{\partial H_z^s}{\partial x} + Q \tag{4.23b}$$

$$\mu \frac{\partial H_z^s}{\partial t} = \frac{\partial E_x^s}{\partial y} - \frac{\partial E_y^s}{\partial x} + R, \tag{4.23c}$$

with the source terms

$$P(x,y,t) = (\varepsilon^i - \varepsilon)\frac{\partial E_x^i}{\partial t}, \tag{4.24a}$$

$$R(x,y,t) = (\varepsilon^i - \varepsilon)\frac{\partial E_y^i}{\partial t}, \tag{4.24b}$$

$$Q(x,y,t) = (\mu^i - \mu)\frac{\partial H_z^i}{\partial t}. \tag{4.24c}$$

We intend to apply the method to a real problem modelling the retina's layers. The outer nuclear layer (ONL) is mostly populated by the cells bodies of light sensitive photoreceptor cells (rods and cons) [11]. The nucleus is the biggest organelle in the photoreceptor cell's soma and presents a high refractive index difference to the surrounding medium. Therefore the light scattering is mainly caused

by the nucleus [22]. Thus the outer nuclear layer is modeled as a group of spherical nuclei (the scatterers) in a homogeneous medium.

The first approach to model this problem is to consider the case of a single nucleus in the ONL. Thus we have a two dimensional square domain which contains a circle to represent the single nucleus, that is, the circle is the part of the domain where the permittivity is different.

Let us consider the previous problem (4.1), in $\Omega = (-2,2)^2$ and $T = 1$. In order to avoid reflections we chose a domain large enough to prevent the influence of the PEC boundary conditions.

We consider the magnetic permeability and the electric permittivity to be constants $\varepsilon^i = 1$ and $\mu = 1$. We consider the permittivity $\varepsilon = 1.2$ for $\left\{(x,y) \in \Omega : x^2 + y^2 \leq 0.09\right\}$ and $\varepsilon = 1$ otherwise. We define the incident wave as the planar wave $E_y^i(x,y,t) = \cos(10(x-t))$.

The results obtained for central fluxes ($\alpha = 0$), time step $\Delta t = 0.002$, final time $T = 1$ and approximation polynomial degree $N = 6$ are illustrated in Figure 4.2, which shows the evolution of H_z with time.

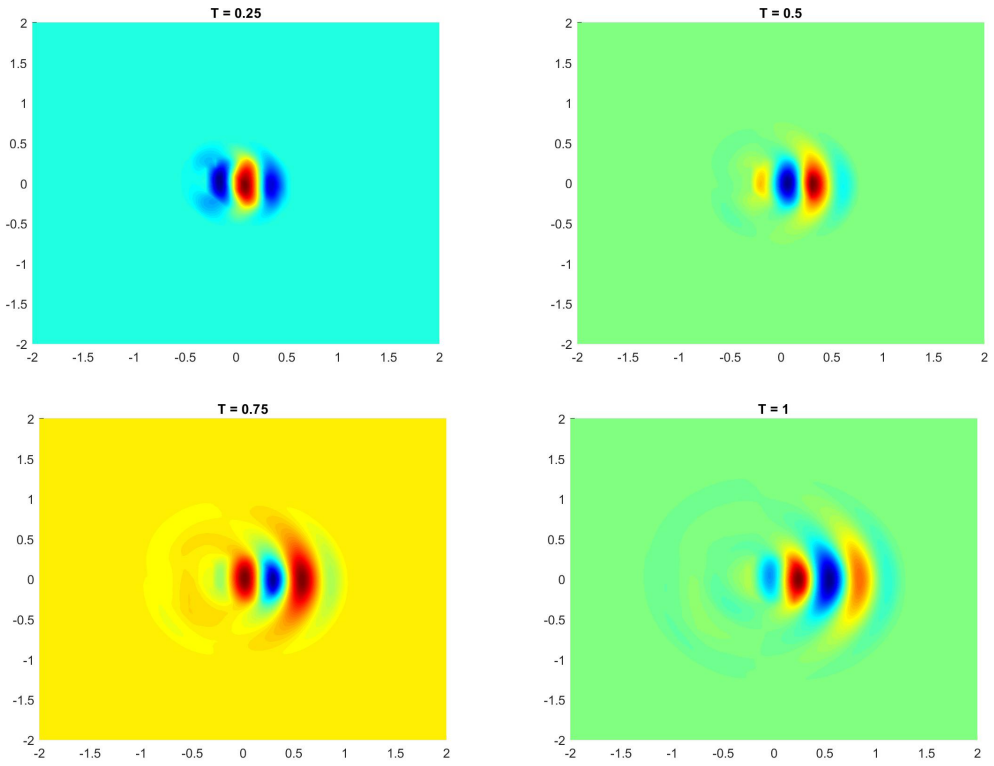

Fig. 4.2 Numerical solution computed for $N = 6$ and $K = 800$

Chapter 5

Conclusion

Throughout this work, we presented the Maxwell's equations and some properties like the constitutive relations and the boundary conditions, which allowed us to write the reduced problem for isotropic media. We intended to formulate and analyse a fully discrete method for this problem.

We established the discontinuous Galerkin method with central fluxes to deal with the spatial discretization of Maxwell's equations in a homogeneous medium with PEC boundary conditions and proved its stability and a convergence rate of $\mathcal{O}(h^N)$, where N is degree of polynomial approximation used in the discretization. For the temporal discretization we used the explicit Euler method, proving its stability and convergence rate of $\mathcal{O}(\Delta t)$ when applied to the semi-discrete scheme that results from the dG method for Maxwell's equations.

Assuming that the physical system is isotropic and homogeneous in the z-direction, we formulated the problem for two dimensions and used this formulation to obtain numerical results, that corroborate the theoretical convergence rates. We used the scattered field formulation in order to simulate the propagation of an electromagnetic wave through a domain that models the case of a single nucleus. In order to obtain the computational results, we used the fourth-order four stage Runge-Kutta method for the temporal discretization.

As the main goals of this work were to study a fully discrete method for Maxwell's equations and implement it, we can say that these goals were achieved.

The next steps in this work would be to adapt the dG method to deal with PMC and SM-ABC boundary conditions and to study the dG method with upwind fluxes, which would allow us to achieve a higher convergence rate. Then, it would be interesting to analyse the stability and convergence of the Runge-Kutta method. We could also try to adapt the dG method to the anisotropic case, in order to better model the retina's layers.

As far as the numerical results are concerned, it would be interesting to implemenent the anisotropic case and to progressively increase the level of complexity of the domain in order to better model the behaviour of a electromagnetic wave through the retina's layers during an OCT.

References

[1] Bernardes, R., Cunha-Vaz, J., , and Serranho, P. (2012). *Optical Coherence Tomography: a Concept Review*. Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg.

[2] Born, M. and Wolf, E. (1999). *Principles of Optics: Electromagnetic theory of propagation, interference and diffraction of light*. Cambridge University Press, 7th edition.

[3] Busch, K., König, M., and Niegemann, J. (2011). Discontinuous Galerkin methods in nanophotonics. *Laser & Photonics Reviews*, 5(6):773–809.

[4] Chen, Q. and Babuška, I. (1995). Approximate optimal points for polynomial interpolation of real functions in an interval and a triangle. *Computer Methods in Applied Mechanics and Engineering*, 128:405–417.

[5] Cockburn, B., Kanschat, G., Perugia, I., and Schötzau, D. (2002). Superconvergence of the local discontinuous galerkin method for elliptic problems on cartesian grids. *SIAM Journal on Numerical Analysis*, pages 264–285.

[6] Cockburn, B., Kanschat, G., and Schötzau, D. (2005). A locally conservative ldg method for the incompressible navier-stokes equations. *Math Comp.*, 74:1067–1095.

[7] Emmrich, E. (1999). *Discrete Versions of Gronwall's Lemma and Their Application to the Numerical Analysis of Parabolic Problems*, volume Preprint No. 637. Fachbereich Mathematik, TU Berlin.

[8] Evans, L. C. (1998). *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, R.I.

[9] Fujimoto, J. G. (2001). Optical coherence tomography. *Comptes Rendus de l'Académie des Sciences - Series IV - Physics*, 2(8):1099 – 1111.

[10] Ghalati, M. K. (2016). *Numerical Analysis and Simulation of Discontinuous Galerkin Methods for Time-Domain Maxwell's Equations*. PhD thesis, Universidade de Coimbra and Universidade do Porto.

[11] Gray, H. and Lewis, W. H. (1918). *Anatomy of the human body*. Lea & Febiger.

[12] Hesthaven, J. S. and Warburton, T. (2010). *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer Publishing Company, Incorporated, 1st edition.

[13] Jackson, J. D. (1962). *Classical Thermodynamics*. John Wiley & Sons, Inc., 3rd edition.

[14] Jin, J. (2002). *The Finite Element Method in Electromagnetics*. John Wiley & Sons, Inc., 2nd edition.

[15] Kovach, L. D. (1982). *Advanced Engineering Mathematics*. Addison-Wesley Publishing Company, New York.

[16] Lourenço, A. C. Q. (2019). Maxwell's equations and the discontinuous galerkin method.

[17] Maxwell, C. J. (1954). *A treatise on Electricity and Magnetism*. Dover Publications, New York.

[18] Monk, P. (2003). *Finite Element Methods for Maxwell's Equations*. Clarendon Press.

[19] Pazur, T. (2013). *Error analysis of implicit and exponential time integration of linear Maxwell's equations*. PhD thesis, Karlsruher Instituts für Technologie.

[20] Pietro, D. D. and Ern, A. (2012). *Mathematical Aspects of Discontinuous Galerkin Methods*, volume 69.

[21] Schnaubelt, R. (2018). *Lecture notes - Evolution equations*.

[22] Seet, K. Y., Nieminen, T. A., and Zvyagin, A. V. (2009). Refractometry of melanocyte cell nuclei using optical scatter images recorded by digital fourier microscopy. *Journal of Biomedical Optics*, 14(4):044031.

[23] Silva, A. (2013). Modelling light scattering in the human retina.

[24] Sturm, A. (2014). Error analysis for full discretizations of maxwell's equations with explicit runge-kutta methods.

[25] Taflove, A. and C. Hagness, S. (2005). *Computational electrodynamics: the finite-difference time-domain method*. Artech House, Inc., 3rd edition.

Appendix A

Auxiliary results

Theorem A.1. *(Stone's theorem, [21, Theorem 1.36]) Let H be a Hilbert space and $A : D(A) \to H$ be a linear operator with dense domain, that is, $D(A) = H$. Then, A generates a C_0-group of unitary operators if and only if A is skew-adjoint.*

Theorem A.2. *(Young's inequality) Let $x,\ y \geq 0$ be real numbers. Then, there holds for every $\gamma > 0$*

$$xy \leq \frac{1}{2}\gamma x^2 + \frac{1}{2}\gamma^{-1} y^2.$$

We call this inequality the weighted Young's inequality. The usual Young's inequality is obtained by choosing $\gamma = 1$.

Theorem A.3. *(Continuous Gronwall lemma, [7, Proposition 2.1]) Let $T \in \mathbb{R}_+ \cup \{\infty\}$, $f, g \in L^\infty(0,T)$ and $c \geq 0$. Furthermore, let g be a monotonically increasing, continuous function and let f satisfy*

$$f(t) \leq g(t) + c \int_0^t f(s), \qquad a.\ e.\ in\ [0,T].$$

Then, there holds

$$f(t) \leq e^{ct} g(t).$$

Theorem A.4. *(Discrete Gronwall lemma, [7, Proposition 4.1]) Let $\{a_n\}_n,\ \{b_n\}_n \subset \mathbb{R}$ be two sequences, $c \geq 0$ and $\tau > 0$ be two constants. Let $\{b_b\}_n$ be a monotonically increasing and let $\{a_n\}_n$ satisfy*

$$a_n \leq b_n + c\tau \sum_{m=0}^{n-1} a_m, \qquad n = 1, 2, \ldots,$$

with initial value $a_0 \leq b_0$. Then, there holds

$$a_n \leq (1 + c\tau)^n b_n \leq e^{cn\tau} b_n.$$

A.1 Broken polynomial spaces

We now look into broken versions of Sobolev spaces $H^m(\Omega)$ and of the graph space $H(curl, \Omega)$ as well as broken versions of the gradient and the curl operator [20].

A.1.1 The broken Sobolev space $H^m(T_h)$

Let $m \geq 0$ be an integer. We define the broken Sobolev space as

$$H^m(T_h) := \left\{ v \in L^2(\Omega) \middle| \forall K \in T_h : v|_K \in H^m(K) \right\}, \tag{A.1}$$

and endue it with the norm: For $v \in H^m(T_h)$,

$$\|v\|^2_{H^m(T_h)} := \sum_{n=0}^{m} |v|^2_{H^n(T_h)}, \qquad |v|_{H^n(T_h)} := \sum_{K \in T_h} |v|^2_{H^n(K)}. \tag{A.2}$$

Using the continuous trace inequality [20, Chapter 1], we see that for all function $v \in H^1(T_h)$ and for all mesh elements $K \in T_h$ the trace $v|_{\partial K}$ on the boundary of the elements is well defined and it holds

$$\|v\|_{L^2(\partial K)} \leq C \|v\|^{1/2}_{L^2(K)} \|v\|^{1/2}_{H^1(K)}. \tag{A.3}$$

We define a broken gradient operator acting on the broken Sobolev space $H^1(T_h)$. Obviously, this operator also acts on the broken polynomial spaces.

Definition A.5. *(Broken gradient) The broken gradient $\nabla_h : H^1(T_h) \to L^2(\Omega)^d$ is defined such that, for all $v \in H^1(T_h)$,*

$$(\nabla_h v)|_K := \nabla(V|_K), \qquad \forall K \in T_h. \tag{A.4}$$

Now we will distinguish the usual Sobolev spaces from their broken versions in more detail. The usual Sobolev spaces are subspaces of their broken versions, that is, for every integer $m \geq 0$, we have

$$H^m(\Omega) \subset H^m(T_h). \tag{A.5}$$

Moreover, it is proven in [20, Lemma 1.22] that for functions in $H^1(\Omega)$ the variational gradient coincides with the broken gradient, that is, for all $v \in H^1(\Omega)$,

$$\nabla v = \nabla_h v. \tag{A.6}$$

But, the reverse does not generally hold true. The main difference is that while the broken Sobolev spaces contain functions having nonzero jumps at interfaces, functions in the usual Sobolev spaces must have zero jumps across any interface. The exact statement translates into the following lemma:

Lemma A.6. *(Characterization of $H^1(\Omega)$, [20, Lemma 1.23]) A function $v \in H^1(\Omega)$ belongs to $H^1(\Omega)$ if and only if*

$$[[v]]_F = 0, \qquad \forall F \in F_h^i. \tag{A.7}$$

A.1.2 The broken graph space $H(curl, T_h)$

Analogously to the definition of the broken Sobolev spaces $H^m(T_h)$, we define the broken version of the graph space $H(curl, \Omega)$ as

$$H(curl, T_h) := \left\{ v \in L^2(\Omega)^3 | \forall K \in T_h : v \in H(curl, K) \right\}. \tag{A.8}$$

We also define a broken version of the curl operator.

Definition A.7. *(Broken curl) The broken curl $\nabla_h\times : H(curl,T_h) \to L^2(\Omega)^3$ is defined such that, for all $v \in H(curl,T_h)$,*

$$(\nabla_h \times v)|_K := \nabla \times (v|_K), \quad \forall K \in T_h. \tag{A.9}$$

Next we establish the relation between $H(curl,\Omega)$ and its broken version $H(curl,T_h)$ in a result equivalent to the relation between Sobolev spaces and their broken counterparts.

Lemma A.8. *(Broken curl on $H(curl,\Omega)$) We have $H(curl,\Omega) \subset H(curl,T_h)$. Furthermore, for all $v \in H(curl,\Omega)$,*

$$\nabla_h \times v = \nabla \times v. \tag{A.10}$$

Lemma A.9. *(Characterization of $H(curl\Omega)$) A function $v \in H^1(T_h)^3$ belongs to $H(curl,\Omega)$ if and only if*

$$n_F \times [[v]]_F = 0, \quad \forall F \in F_h^i. \tag{A.11}$$

In later sections, when considering the space $H(curl,\Omega) \cap H^1(T_h)^3$, it is critical that its function only admit zero tangential jumps across interfaces.

A.1.3 Broken polynomial space

After constructing a mesh of the domain Ω, we turn to the construction of finite function spaces. In our case this consists of piecewise polynomials.

The polynomial space $\mathbb{P}_d^k$

Let $k \geq 0$ be an integer and $\alpha = (\alpha_1,\ldots,\alpha_d) \in \mathbb{N}_0^d$ be a multi-index with

$$|\alpha|_{l^1} = \sum_{i=1}^{d} \alpha_i \leq k. \tag{A.12}$$

Furthermore let $x = (x_1,\ldots,x_d)$ be a vector in $\mathbb{R}^d$ and let us use the convention

$$x^\alpha := \prod_{i=1}^{d} x_i^{\alpha_i}. \tag{A.13}$$

Then the function p_α defined as

$$p_\alpha : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \gamma_\alpha x^\alpha, \tag{A.14}$$

where $\gamma_\alpha \in \mathbb{R}$ is a coefficient, is a polynomial of d variables of total degree at most k. Hence, the set

$$\mathbb{P}_d^k := \left\{ p : \mathbb{R}^d \to \mathbb{R} \,\middle|\, \exists(\gamma_\alpha) \subset \mathbb{R} : p(x) = \sum_{|\alpha|_{l^1} \leq k} \gamma_\alpha x^\alpha \right\} \tag{A.15}$$

is the space of all polynomials of d variables with degree at most k. Its dimension is

$$\dim(\mathbb{P}_d^k) = \binom{k+d}{k} = \frac{(k+d)!}{k!d!}. \tag{A.16}$$

A.1.4 The broken polynomial space $\mathbb{P}_d^k(T_h)$

Let $K \in T_h$ be a mesh element. Then, we define $\mathbb{P}_d^k(K)$ as

$$\mathbb{P}_d^k(K) := \left\{ p|_K : K \to \mathbb{R} \middle| p \in \mathbb{P}_d^k \right\}. \tag{A.17}$$

Then, the broken polynomial space $\mathbb{P}_d^k(T_h)$ on the mesh T_h now consists of functions which are polynomials on each mesh element, that is,

$$\mathbb{P}_d^k(T_h) := \left\{ v \in L^2(\Omega) \middle| \forall K \in T_h : v|_K \in \mathbb{P}_d^k(K) \right\}. \tag{A.18}$$

It follows that

$$\dim(\mathbb{P}_d^k(T_h)) = \operatorname{card}(T_h) \times \dim(\mathbb{P}_d^k). \tag{A.19}$$

A.2 Admissible Mesh Sequences

One of the main goals of this work is to prove the convergence of dG methods, that is, to prove that, as the meshsize tends to zero, the error between the approximate solution and the exact solution also tends to zero. One of the most important concepts used to prove this is that of admissible mesh sequences. So let us consider the mesh sequence

$$T_H := (T_h)_{h \in H}, \tag{A.20}$$

where H denotes a countable subset $\mathbb{R}^+$ having 0 as only accumulation point.

In the following we consider the shape- and contact-regular mesh sequences [20, Definition 1.38]. The stated properties are necessary for the convergence analysis. For further information on this issue we refer to [20, Section 1.4.1].

A.2.1 Geometric properties

We have previously defined F_K as the set of faces composing the boundary of an element K and N_∂ as the maximum number of mesh faces composing the boundary of elements in T_h. Then, for shape- and contact-regular meshes, we can relate these quantities and the meshsize h through following lemma:

Lemma A.10. *(Bound on $\operatorname{card}(F_K)$ and N_∂, [20, Lemma 1.41]) Let T_H be a shape- and contact-regular mesh sequence. Then, for all $h \in H$ and all $K \in T_h$, $\operatorname{card}(F_K)$ and N_∂ are uniformly bounded in h.*

A.2.2 Inverse and trace inequality

We state two inequalities for the broken polynomial spaces $\mathbb{P}_d^k(T_h)$ on a shape- and contact-regular mesh, which are used for analyzing dG methods. The inverse inequality provides an upper bound on the gradient of discrete functions and the discrete trace inequality that provides an upper bound on the face values of discrete functions.

Lemma A.11. *(Inverse inequality, [20, Lemma 1.44]) Let T_H be a shape. and contact-regular mesh sequence. Then, for all $h \in H$, all $K \in T_h$ and all $v_h \in \mathbb{P}_d^k(T_h)$,*

$$\|\nabla v_h\|_{L^2(K)^d} \leq C_{inv} h_K^{-1} \|v_h\|_{L^2(K)}. \tag{A.21}$$

The constant C_{inv} depends only on d, k and the shape- and contact-regularity parameters.

Lemma A.12. *(Discrete trace inequality, [20, Lemma 1.46]) Let T_H be a shape- and contact-regular mesh sequence. Then, for all $h \in H$, all $K \in T_h$, all $F \in F_K$ and all $v_h \in \mathbb{P}_d^k(T_h)$,*

$$\|v_h\|_{L^2(F)} \leq C_{tr} h_K^{-1/2} \|v_h\|_{L^2(K)}. \tag{A.22}$$

The constant C_{tr} only depends on d, k and the shape- and contact-regularity parameters.

Lemma A.13. *(Continuous trace inequality, [20, Lemma 1.49]) Let T_H be a shape- and contact-regular mesh sequence. Then, for all $h \in H$, all $K \in T_h$. all $F \in F_K$ and all $v \in H^1(T_h)$,*

$$\|v\|_{L^2(F)}^2 \leq C_{cti} \left(2\|\nabla v\|_{L^2(K)^d} + d h_K^{-1} \|v\|_{L^2(K)} \right) \|v\|_{L^2(K)}, \tag{A.23}$$

where the constant C_{cti} depends on d and the shape- and contact-regularity parameters.

A.2.3 Polynomial approximation

In dG methods we look for the approximate solution in the piecewise polynomial space $\mathbb{P}_d^k(T_h)$. Therefore, it is of the utmost importance to guarantee the construction of the mesh sequence allows for optimal polynomial approximation. Thus we define in which sense we require that the mesh sequence admits optimal polynomial approximation through the next definition. We refer to [20, Section 1.4.4] for details.

Definition A.14. *The mesh sequence T_H has optimal polynomial approximation properties if, for all $h \in H$, all $K \in T_h$, and all polynomial degree k, there is a linear interpolation operation $I_K^k : L^2(K) \to \mathbb{P}_d^k(K)$ such that, for all $s \in \{0, \dots, k+1\}$ and all $v \in H^s(K)$, we have*

$$|v - I_K^k v|_{H^m(K)} \leq C_{app} h_K^{s-m} |v|_{H^s(K)}, \qquad \forall m \in \{0, \dots, s\}, \tag{A.24}$$

where the constant C_{app} is independent of both K and h.

We assume that the mesh-sequence T_H is admissible, that is, that it is shape- contact-regular and has optimal polynomial approximation properties in the sense of Definition A.14.

In the later error analysis we will often use the L^2-orthogonal projection onto the broken polynomial space $\mathbb{P}_d^k(T_h)$ defined as $\pi_h^{L^2} : L^2(\Omega) \to \mathbb{P}_d^k(T_h)$ such that for all $v \in L^2(\Omega)$,

$$(\pi_h^{L^2} v.\varphi_h)_{L^2(\Omega)} = (v, \varphi_h)_{L^2(\Omega)}, \qquad \forall \varphi_h \in \mathbb{P}_d^k(T_h). \tag{A.25}$$

Admissible mesh sequences provide optimality of the L^2-projection in the following sense:

Lemma A.15. *(Optimality of L^2-orthogonal projection) Let T_H be an admissible mesh sequence. Then, for all $s \in \{0,\dots,k+1\}$ and all $v \in H^s(K)$, there holds*

$$|v - \pi_h^{L^2} v|_{H^m(K)} \leq C'_{app} h_K^{s-m} |v|_{H^s(K)}, \qquad \forall m \in \{0,\dots,s\}. \tag{A.26}$$

The constant C'_{app} is independent of both K and h.

The following bound for polynomial approximations on mesh faces is a direct consequence of (A.26) and the continuous trace inequality from Lemma A.13.

Lemma A.16. *(Polynomial approximation on mesh faces) Under the assumption of Lemma A.15 with $s \geq 1$ it holds for all $h \in H$, all $K \in T_h$, and all $F \in F_K$,*

$$|v - \pi_h^{L^2} v|_{L^2(F)} \geq C''_{app} h_K^{s-1/2} |v|_{H^s(K)}, \tag{A.27}$$

with constant C''_{app} independent of both K and h.