# The Discontinuous Galerkin Method for Maxwell's Equations

Application to Bodies of Revolution and Kerr-Nonlinearities

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Dipl.-Math. Elisabeth Blank
aus Karlsruhe

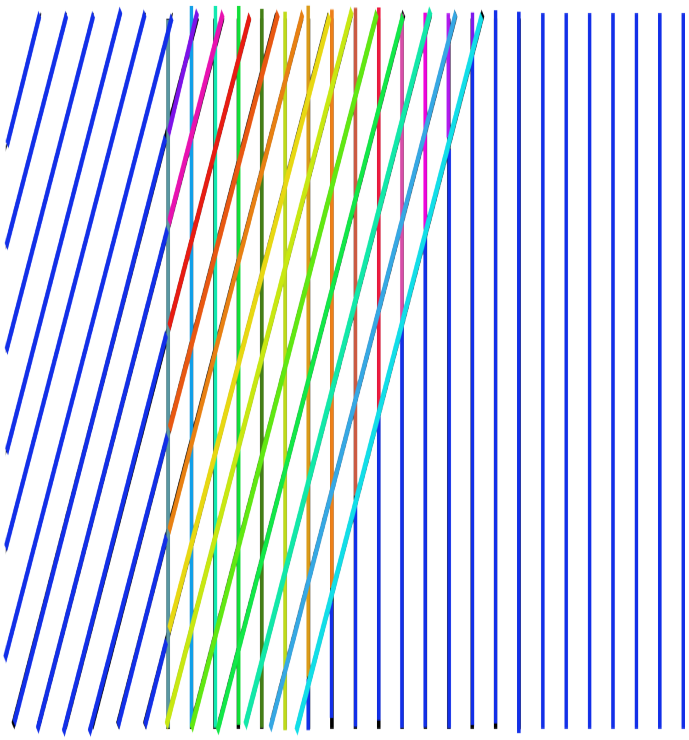
| | |
|---|---|
| Tag der mündlichen Prüfung: | 06.02.2013 |
| Referent: | Prof. Dr. W. Dörfler |
| 1. Koreferent: | Prof. Dr. K. Busch |
| 2. Koreferent: | Prof. Dr. C. Wieners |

Karlsruhe Institute of Technology

Fakultät für Mathematik
Institut für Numerische Mathematik

# The Discontinuous Galerkin Method for Maxwell's Equations

Application to Bodies of Revolution and Kerr-Nonlinearities



PhD Thesis
by

**Dipl.-Math. Elisabeth Blank**

08.01.2013

| | |
|---|---|
| Instructor: | Prof. Dr. Willy Dörfler |
| 2nd Instructor: | Prof. Dr. Kurt Busch |
| 3rd Instructor: | Prof. Dr. Christian Wieners |

# Contents

Chapter 1

# 1 Introduction

## 1.1 The Discontinuous Galerkin Method

Reed and Hill [1] applied a first form of the discontinuous Galerkin (DG) method to the neutron transport equation in 1973. Since then it has undergone a fast development and finds numerous applications to e.g. the Euler equations of gas dynamics, the shallow water equations, the equations of magneto-hydrodynamics, the compressible Navier-Stokes equations and Maxwell's equations – only to mention a few; see e.g. [2] for an overview and analysis of many of these applications. Today, the DG method can be looked upon as a suitable and efficient tool to treat physical, chemical, meteorological, biological, mathematical and many other problems.
The DG method can be used for time and/or spatial discretization; as references for the time-stepping DG method, see e.g. [3, 4, 5], which is only a small selection. In that case, the differential equations are parabolic. We will use it for the spatial discretization of the time-dependent, linear Maxwell's equations in three-dimensional rotationally symmetric geometries (in so-called Bodies of Revolution, BOR) and of the time-dependent Kerr-nonlinear Maxwell's equations, which can be formulated as hyperbolic conservation laws. For time integration we use an explicit time stepping method, the explicit low-storage Runge-Kutta method, originally introduced by Williamson in 1980 [6] and further developed by M. Carpenter and C. Kennedy [7]. The combination of the DG method for spatial discretization and of an explicit Runge-Kutta scheme for time-integration, which is also called the Runge-Kutta Discontinuous Galerkin (RKDG) method in mathematics or the Discontinuous Galerkin Time-Domain (DGTD) method in physics, has been introduced by Cockburn and Shu in several papers, [8] maybe being one of the latest. [9] gives an introduction to the DG method and contains a more complete list of references on the RKDG method; it can be downloaded online. We also mention [10], [11] and [12], which is a review paper about RKDG methods, as references on the DG method. [13] gives a review of the DG method applied to nanophotonics. We emphasize that there are numerous other good books and papers on this topic.

The DG method encounters several advantages. It is a high order accuracy method; for many cases optimal convergence rates can be shown, see e.g. [14, 15, 16, 17, 18, 19], which is only a very small selection. The DG method is a local method, which allows a high flexibility with meshes and which can thus handle complicated geometries. For linear problems, parallelization is possible. Due to its locality, discontinuous solutions can be treated as well.

A main ingredient of any DG scheme is the so-called numerical flux, which serves as a connection between the single elements in order to construct the global approximation from all locally obtained approximations. The notion of the numerical flux is taken from finite volume methods, where the numerical flux meets the same purpose, i.e. to transport the information from one local cell to another. The numerical flux plays a central role in this thesis.

## 1.2 BOR Maxwell's Equations

The electromagnetic characterization of rotationally symmetric systems (bodies of revolution, BOR) plays an important role for a large number of technical applications. Examples range from coaxial cables and cylindrical resonators in the microwave regime to lenses, tapered fibers or plasmonic nanoparticles in the optical spectrum. In numerical calculations, their symmetry can often be exploited to reduce the effective dimensionality of the system which reduces the computational or analyical effort significantly. Indeed, over the past decades most of the commonly employed numerical techniques have been extended to also treat BOR systems efficiently. Besides the Method of Moments (MoM) [20] and the Finite Element Method (FEM) [21], this is also true for the Finite–Difference Time–Domain (FDTD) method [22]. Particularly the BOR–FDTD method has proven popular and was applied to a large variety of systems, ranging from optical lenses [23] or diffractive elements [24] to plasmonic nanostructures [25]. In this thesis we apply the Runge-Kutta Discontinuous Galerkin method to BOR Maxwell's equations and demonstrate how to obtain an efficient algorithm for solving them.

## 1.3 Kerr-Nonlinear Maxwell's Equations

The physicist John Kerr discovered the optical Kerr effect in 1875. It is a nonlinear optical phenomenon which arises due to a change in the refractive index of a material in response to an incoming electric field. The nonlinear behavior of the medium is responsible for effects like self-focusing or self-modulation. In the first case, the refractive index increases with the electric field intensity and the medium acts as a focusing lens for an electromagnetic wave. In the second example, the index of refraction varies in time and intensity of the incoming pulse, leading to a phase shift which produces a shift in the frequency of the pulse. Increasing intensity leads to lower frequencies, and decreasing intensity gives higher frequencies. Near an extremum of the intensity, the frequency of the pulse behaves approximately linear.
Typically, nonlinearities are observed only if high intensities are present, as, e.g. in case of lasers. Examples of applications involving the optical Kerr effect are fast sensors for the measurement of electromagnetic fields, the fast determination of the structure of molecules, or image enhancement and image conversion in presence of ultraviolet radiation (see e.g. [26] for an overview) or bistable optical systems, see e.g. [27] and the references therein.

In this work we also apply the RKDG method to Kerr-nonlinear Maxwell's equations. In contrast to the linear BOR Maxwell's equations many aspects are different in this case. This is especially the case for the numerical flux, which is state dependent for nonlinear problems. This increases the computational effort immensely. For many relevant applications this makes simulations practically impossible. Therefore an appropriate approximation to the analytically given numerical flux is indispensable and leads so the research field of Riemann solvers. There exist numerous different Riemann solvers, like the Roe solver [28], the HLL solver [29] and its modified versions, like e.g. the HLLC or HLLE solver; for more Riemann solvers, see e.g. [30]. The HLL solver proved to give very good numerical performances; see e.g. [31], where several versions of the HLL flux are compared for the magneto-hydrodynamic equations.
In this work we give an exact numerical flux and approximative numerical fluxes. We present several linear numerical fluxes, like a Lax-Friedrichs flux, and an HLL-like flux and compare these numerical fluxes with each other with respect to efficiency and accuracy.

## 1.4 Organization of this Thesis

This thesis consists of two main parts. Part I comprises theoretical topics, Part II is about the application of the RKDG method to linear BOR and Kerr-nonlinear Maxwell's equations.

Part I starts with Maxwell's equations and a brief overview of the corresponding constitutive relations, the behavior of the electromagnetic fields at interfaces and boundary conditions. Subsection 2.4.2 contains a short introduction to uniaxial perfectly matched layers. In section 2.3 we give Maxwell's equations in the weak form, and in section 2.6 we reformulate them as a conservation law, including theoretical aspects about hyperbolic conservation laws in section 2.6.1.
In chapter 3 with introduce the Runge-Kutta Discontinuous Galerkin method. A complete DG method consists of defining an appropriate triangulation with a finite element space of discontinuous functions, and a numerical flux. The choice of a numerical flux is not unique and there are many possible ways to do it. In section 3.2 we give details about the numerical flux and its connection to finite volume methods, including examples of well-known and widely used numerical fluxes. We will choose a numerical flux that solves a so-called Riemann problem, and section 3.3 contains details about the solution of a Riemann problem which we need in order to construct a numerical flux.
Space discretization leads to a semi-discrete scheme that needs to be integrated in time. We use an explicit low-storage Runge-Kutta scheme of 4th order and with 5 stages (see [7]).
In section 3.4 we give some main convergence results of the RKDG method. This requires also the stability of the RK time stepping, which is the content of section 3.4.2. This also leads to the necessity of so-called slope limiters, especially for nonlinear problems, where discontinuous solutions can occur, leading to oscillations near discontinuities. This is the well-known Gibbs-phenomenon. A slope limiter stabilizes the DG scheme and ensures its high-order nature which is lost due to oscillations.

Part II is about the application of the RKDG method to Maxwell's equations.
In chapter 4.1 we derive BOR Maxwell's equations and introduce their weak form in section 4.1.1. We then start by applying the DG method to the two-dimensional BOR Maxwell's equations in 4.2, where the basic concepts of the DG space discretization are explained. In section 4.2.2 we present the numerical flux for BOR Maxwell's equations. We give an efficient way of computing the resulting system matrices in section 4.2.5. We conclude this chapter with several numerical tests. We proceed with the three-dimensional BOR Maxwell's equations in chapter 4.4 in the same manner and give numerical tests in section 4.5, including the introduction of uniaxial perfectly matched layers. We consider different test systems, such as traveling waves in an open system, a traveling Gaussian pulse in a fiber or a simulations in a tapered fiber.
In chapter 5 we then come to the Kerr-nonlinear Maxwell's equations. In order to complete the DG scheme, a numerical flux is needed. For this we need to solve the corresponding Riemann problem, on which we focus in section 5.2. In subsections 5.2.1 and 5.2.3 we give the corresponding Hugoniot Locus and Riemann invariants, respectively. In section 5.3.2 we present approximative numerical fluxes, and section 5.4 concludes the chapter with numerical tests, where we consider a traveling Gaussian pulse to test the performance of the RKDG scheme and the different numerical fluxes.

# Part I

# Theory

## Maxwell's Equations

## Hyperbolic Conservation Laws

## The Runge-Kutta Discontinuous Galerkin Method

Chapter 2

# 2 Macroscopic Maxwell's Equations

In 1873 Maxwell formulated four coupled equations to describe the evolution of electric and magnetic fields and thus the propagation of light. Maxwell's equations in SI units and in differential form are given as

**Maxwell's equations**

$$\frac{\partial \mathbf{D}(\mathbf{x},t)}{\partial t} - \nabla \times \mathbf{H}(\mathbf{x},t) = -\mathbf{J}(\mathbf{x},t), \tag{2.0.1a}$$

$$\frac{\partial \mathbf{B}(\mathbf{x},t)}{\partial t} + \nabla \times \mathbf{E}(\mathbf{x},t) = 0, \tag{2.0.1b}$$

$$\nabla \cdot \mathbf{D}(\mathbf{x},t) = \rho(\mathbf{x},t), \tag{2.0.1c}$$

$$\nabla \cdot \mathbf{B}(\mathbf{x},t) = 0, \tag{2.0.1d}$$

supplemented with initial, boundary and interface conditions (about which will be talked later). Here, $\mathbf{x} \in \mathbb{R}^3$ is the spatial variable, $t \in \mathbb{R}$ is the time variable, $\mathbf{E}$ is the electric field, $\mathbf{H}$ is the magnetic field, $\mathbf{B}$ is the magnetic induction, and $\mathbf{D}$ the electric displacement. $\mathbf{J}$ is the vector electric current density function, and $\rho$ the scalar charge density. For now the fields are assumed to be smooth enough with $\mathbf{E}, \mathbf{D}, \mathbf{B}, \mathbf{H}$ in some appropriate function space $\mathbb{X}$ so that Maxwell's equations (2.0.1) are well defined; this will be specified in more detail in section 2.3, where we address the weak formulation of Maxwell's equations. We also assume that interchanging time derivatives and spatial derivatives is possible. In addition, in the forthcoming, we suppress the field dependencies on the spatial variable $\mathbf{x}$ and time variable $t$ for clarity, unless it is explicitly brought out otherwise. For theory on macroscopic Maxwell's equations, see e.g. [32, 33, 34, 35], only to mention a small selection of literature.

By taking the time derivative of (2.0.1a) and (2.0.1b) and using (2.0.1d) we get the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0. \tag{2.0.2}$$

One can separate Maxwell's equations by the character of their derivatives. Then equations (2.0.1a) and (2.0.1b) are called Maxwell's curl-equations, and (2.0.1c) and (2.0.1d) are termed as divergence conditions. This designation is motivated by the fact that the divergence conditions can be looked upon as constraints on the electromagnetic fields that need to be fulfilled at all times, whereas the curl-equations are relevant for time-evolution. This can be seen follows: We apply the divergence to equations (2.0.1a) and (2.0.1b) and get

$$\nabla \cdot \frac{\partial \mathbf{D}(\mathbf{x},t)}{\partial t} = -\nabla \cdot \mathbf{J}(\mathbf{x},t) = \frac{\partial \rho}{\partial t}(\mathbf{x},t),$$

$$\nabla \cdot \frac{\partial \mathbf{B}(\mathbf{x},t)}{\partial t} = 0,$$

where we also used the continuity equation (2.0.2) and the fact that $\nabla \cdot (\nabla \times \mathbf{X}) = 0$ for any vector function $\mathbf{X}$. We thus find that the divergence of $\mathbf{D}$ and $\mathbf{B}$ is constant in time. Therefore, if the divergence condition is initially fulfilled and the current and charge density follow the continuity equation, the curl-equations suffice to describe the dynamics of the electromagnetic fields. In case a numerical scheme is applied to discretize the curl-equations, the divergence conditions do not have to be fulfilled automatically, as was pointed out in e.g. Ref. [36]. It might be necessary to design a scheme that takes the divergence constraints numerically into account, as suggested in Ref. [37], where a Discontinuous Galerkin Method is applied to Maxwell's equations using a locally divergence-free basis.

## 2.1 Constitutive Relations

Maxwell's equations alone do not suffice to determine the electromagnetic fields. Their interaction with matter is described through the constitutive laws

$$\begin{aligned} \mathbf{D}(\mathbf{x},t) &= \epsilon_0 \mathbf{E}(\mathbf{x},t) + \mathbf{P}[\mathbf{E},\mathbf{H}](\mathbf{x},t), \\ \mathbf{H}(\mathbf{x},t) &= \mu_0 \mathbf{B}(\mathbf{x},t) - \mathbf{M}[\mathbf{E},\mathbf{H}](\mathbf{x},t), \end{aligned} \tag{2.1.1}$$

where $\mathbf{P}, \mathbf{M}$ shall be vector valued functionals with $\mathbf{P} = (P_j)_j$, $\mathbf{M} = (M_j)_j$ and $P_j, M_j : \mathbb{X} \to \mathbb{R}$ for a fixed value of $t$. $\mathbf{P}$ is called the polarization and $\mathbf{M}$ the magnetization. We stress that the expression for $\mathbf{P}$ and $\mathbf{M}$ in equation (2.1.1) should be understood as a short-hand writing to illustrate the character of the relation between the electromagnetic fields. More details of the exact form of $P_j$ and its general relation to $\mathbf{E}$ can be found in, e.g., [36], [33], [32].
These constitutive equations inhibit some freedom in choosing $\mathbf{P}$ and $\mathbf{M}$. For instance, $\mathbf{D}$ and $\mathbf{H}$ can be linear in $\mathbf{E}$ and $\mathbf{B}$, then

$$\begin{aligned} \mathbf{D}(\mathbf{x},t) &= \epsilon_0 \epsilon \mathbf{E}, \\ \mathbf{H}(\mathbf{x},t) &= \mu_0 \mu \mathbf{B}, \end{aligned}$$

mirroring a medium that reacts linearly when light propagates through it. In general $\epsilon$ and $\mu$ can vary in $\mathbf{x}$ and $t$. $\mathbf{P}$ and $\mathbf{M}$ can also be nonlinear in $\mathbf{E}$ and $\mathbf{H}$, like e.g.

$$\mathbf{P}[\mathbf{E},\mathbf{H}] = \epsilon_0 \underline{\underline{\chi}}^{(3)} |\mathbf{E}|^2 \mathbf{E}, \tag{2.1.2}$$

and we set $\mathbf{M} = \mathbf{0}$. This describes a so-called Kerr-nonlinear medium. Generally, the third-order nonlinear susceptibility $\underline{\underline{\chi}}^{(3)}$ is a tensor with components $\chi^{(3)}_{jklm}$, see [35]. If $\underline{\underline{\chi}}^{(3)}$ consists of constants so that $\chi^{(3)}_{jklm} \equiv \chi^{(3)}$, we write $\underline{\underline{\chi}}^{(3)} \equiv \chi^{(3)}$ again. A fundamental attribute of the susceptibility is its frequency dependence, see e.g. Refs [35, 34].
We note one can also derive the expression (2.1.2) by making a power series ansatz, see the references we have mentioned above. In this case, $\chi^{(3)}$ must be small enough so that this power series converges.

## 2.2 Interfaces

Let us assume to have a medium with a region 1 and a region 2, creating an interface in between. In region 1 we denote the permittivity by $\epsilon_1$ and the permeability by $\mu_1$, in region 2 we have the electromagnetic parameters $\epsilon_2$ and $\mu_2$, as depicted in figure 2.1. $\hat{\mathbf{n}}$ shall be a unit normal pointing from $I$ into region 1.

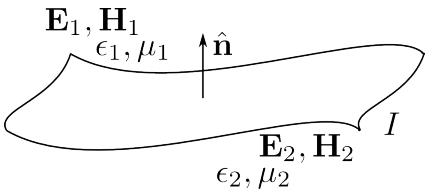



Figure 2.1: Boundary conditions at an interface between two media with different electromagnetic properties.

By $\mathbf{E}_1, \mathbf{H}_1$ we denote the limit of electromagnetic fields when approaching $I$ from region 1, and by $\mathbf{E}_2, \mathbf{H}_2$ the ones when approaching $I$ from region 2. By using Gauss' and Stokes' theorem (see e.g. Refs [38] and [39, Lemma 5.3], which takes into account more mathematical details), it can be seen that across $I$ the tangential components of $\mathbf{E}$ and the normal components of $\mathbf{H}$ need to be continuous, that is

$$\begin{aligned} \hat{\mathbf{n}} \times (\mathbf{E}_1 - \mathbf{E}_2) &= \mathbf{0}, \\ \hat{\mathbf{n}} \cdot (\mu_1 \mathbf{H}_1 - \mu_2 \mathbf{H}_2) &= \mathbf{0}, \end{aligned} \tag{2.2.1}$$

whereas for the normal components of $\mathbf{E}$ and the tangential components of $\mathbf{H}$ we have

$$\begin{aligned} \hat{\mathbf{n}} \times (\mathbf{H}_1 - \mathbf{H}_2) &= \mathbf{J}_I, \\ \hat{\mathbf{n}} \cdot (\epsilon_1 \mathbf{E}_1 - \epsilon_2 \mathbf{E}_2) &= \rho_I, \end{aligned}$$

By $\mathbf{J}_I$ we denote the value of the current density $\mathbf{J}$ on the surface $I$; $\rho_I$ has to be understood in the same manner. Thus, if $\epsilon$ and $\mu$ are discontinuous across $I$, the electromagnetic fields $\mathbf{E}, \mathbf{H}$ are not continuous across $I$.

## 2.3 Weak Formulation of Maxwell's Equations

Until now we have assumed all electromagnetic fields to be "smooth enough" so that all derivatives are defined. In this section we will specify the meaning of this. For this we will define the so-called div- and curl-spaces $H(\mathrm{div}; \Omega)$ and $H(\mathrm{curl}; \Omega)$. We will not go into details. The interested reader may choose to look into [39], for instance.

**Definition 2.1.**
*A bounded domain $\Omega \subset \mathbb{R}^n$ is called Lipschitz if its boundary $\partial\Omega$ is Lipschitz, that is, if there exists a finite number of domains $\Omega_i$, local coordinate systems $(x_i, y_i, z_i)$ and Lipschitz-continuous functions $f(x_i, y_i)$ such that $\partial\Omega$ is a subset of the union of all $\Omega_i$ and $\Omega \cap \Omega_i = \{(x_i, y_i, z_i) \in \Omega_i : \ z_i > f(x_i, y_i)\}$.*

Figure 2.2 illustrates the concept of Lipschitz continuity of a domain.

**Definition 2.2.**
*By $L^2(\Omega)^3$ we denote the three-dimensional analogue of the space of all square-integrable functions $L^2(\Omega)$, and define for $\mathbf{u} = (u_1, u_2, u_3)^T \in L^2(\Omega)^3$ and $\mathbf{v} \in L^2(\Omega)^3$ its inner product as*

$$(\mathbf{u}, \mathbf{v}) := \int_\Omega \mathbf{u} \cdot \mathbf{v} \, \mathrm{d}\Omega.$$

For later purposes, we give the definition of the Sobolev spaces $W^{k,p}(\Omega)$ and $H^k(\Omega)$. See the book "Sobolev Spaces" by Adams and Fournier [40] for more details.

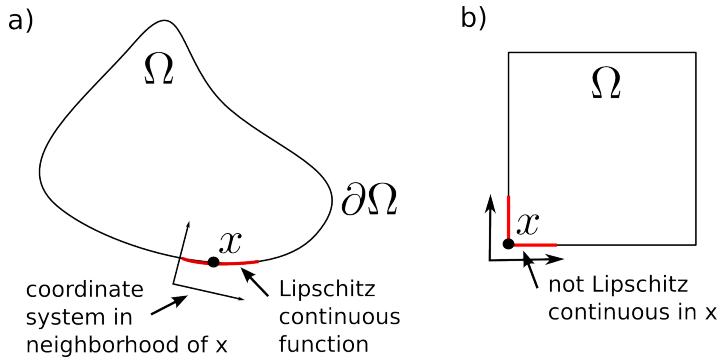



Figure 2.2: a) Example of a Lipschitz domain. b) Counter example of a non-Lipschitz domain: No neighborhood of the edges of the square with a Lipschitz continuous function can be found.

**Definition 2.3.**
*Let $\Omega \subset \mathbb{R}^n$ be open. The* Sobolev space $W^{k,p}(\Omega)$ *is defined as*

$$W^{k,p}(\Omega) = \{u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega)\ \forall |\alpha| \leq k\}$$

*with $1 \leq p \leq \infty$. $k \in \mathbb{N}$ is called the order of $W^{k,p}(\Omega)$. $W^{k,p}(\Omega)$ is equipped with the norm*

$$\|u\|_{W^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\alpha|\leq k} \|D^\alpha u\|^p_{L^p(\Omega)}\right)^{1/p}, & 1 \leq p < \infty; \\ \max_{|\alpha|\leq k} \|D^\alpha u\|_{L^\infty(\Omega)}, & p = \infty. \end{cases}$$

*$(W^{k,p}(\Omega), \|\cdot\|_{W^{k,p}(\Omega)}$ is a Banach space. One denotes $H^k(\Omega) := W^{k,2}(\Omega)$. $(H^k(\Omega), \|\cdot\|_{W^{k,2}(\Omega)})$ is a Hilbert space.*

**Definition 2.4.**
*Let $\Omega \subset \mathbb{R}^3$ be a bounded Lipschitz domain. The space of all functions with square-integrable divergence is defined as*

$$H(\mathrm{div};\Omega) := \left\{\mathbf{u} \in L^2(\Omega)^3 :\ \nabla \cdot \mathbf{u} \in L^2(\Omega)\right\}, \tag{2.3.1}$$

*endowed with the norm*

$$||\mathbf{u}||_{H(\mathrm{div};\Omega)} := \left(||\mathbf{u}||^2_{L^2(\Omega)^3} + ||\nabla \cdot \mathbf{u}||^2_{L^2(\Omega)}\right)^{1/2}.$$

*The space of all $\mathbf{u} \in H(\mathrm{div};\Omega)$ with zero trace is*

$$H_0(\mathrm{div};\Omega) := \{\mathbf{u} \in H(\mathrm{div};\Omega) :\ \mathbf{u} \cdot \hat{\mathbf{n}}|_{\partial\Omega} = 0\}, \tag{2.3.2}$$

*where $\mathbf{u} \cdot \hat{\mathbf{n}}|_{\partial\Omega}$ is continuous and well-defined for $\mathbf{u} \in H(\mathrm{div};\Omega)$. It can be shown (see [39], Theorem 3.22) that $H(\mathrm{div};\Omega)$ is the closure of $(C^\infty(\overline{\Omega}))^3$ with respect to $||\cdot||_{H(\mathrm{div};\Omega)}$ and $H_0(\mathrm{div};\Omega)$ is the closure of $(C_0^\infty(\overline{\Omega}))^3$ with respect to $||\cdot||_{H(\mathrm{div};\Omega)}$. In an analogous manner, we define the curl-space*

$$H(\mathrm{curl};\Omega) := \left\{\mathbf{u} \in L^2(\Omega)^3 :\ \nabla \times \mathbf{u} \in L^2(\Omega)\right\} \tag{2.3.3}$$

*with norm*

$$||\mathbf{u}||_{H(\mathrm{curl};\Omega)} := \left(||\mathbf{u}||^2_{L^2(\Omega)^3} + ||\nabla \times \mathbf{u}||^2_{L^2(\Omega)^3}\right)^{1/2}.$$

*The space of all* $\mathbf{u} \in H_0(\mathrm{curl};\Omega)$ *with zero trace is*

$$H_0(\mathrm{curl};\Omega) := \{\mathbf{u} \in H(\mathrm{curl};\Omega) : \ \hat{\mathbf{n}} \times \mathbf{u}|_{\partial\Omega} = 0\}\,, \tag{2.3.4}$$

*where* $\mathbf{u} \times \hat{\mathbf{n}}|_{\partial\Omega}$ *is well-defined for* $\mathbf{u} \in H_0(\mathrm{curl};\Omega)$*. It can be shown that* $H(\mathrm{curl};\Omega)$ *is the closure of* $(C^\infty(\overline{\Omega}))^3$ *with respect to the norm* $||\cdot||_{H(\mathrm{curl};\Omega)}$ *and* $H_0(\mathrm{curl};\Omega)$ *is the closure of* $(C_0^\infty(\overline{\Omega}))^3$ *with respect to* $||\cdot||_{H(\mathrm{curl};\Omega)}$*.*

P. Monk shows in [39] that these spaces are well-defined, and gives further properties. We also cite an embedding result for the space $X := H(\mathrm{curl};\Omega) \cap H_0(\mathrm{div};\Omega)$.

**Theorem 2.5.**
*Let* $\Omega \subset \mathbb{R}^3$ *be a bounded Lipschitz domain. For* $k > \frac{1}{2}$*, the space* $X$ *is continuously embedded into the Sobolev space* $H^k(\Omega)^3$*. Consequently,* $X$ *is compactly embedded in* $L^2(\Omega)^3$*. If* $\Omega$ *is even* $C^1$ *or convex, then we have* $k = 1$*. The same holds for* $H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div};\Omega)$*.*

*Proof.* See e.g. [41]. □

With these definitions we are ready to formulate Maxwell's equations (2.0.1) in the weak sense. Let us first note that either equations (2.0.1a) to (2.0.1b) can be formulated in the weak sense and equations (2.0.1c) to (2.0.1d) are to be understood pointwise (strongly), or (2.0.1a) to (2.0.1b) have to be understood pointwise and (2.0.1c) to (2.0.1d) weakly. We will give the curl-equations in the weak form, since – as we have alluded to at the end of section 2 – they are relevant for time evolution of the fields; furthermore, these are the equations we will work with in later chapters.
So let $\mathbf{E}, \mathbf{H} \in H_0(\mathrm{curl};\Omega)$ and let $\boldsymbol{\psi} \in (C_0^\infty(\overline{\Omega}))^3$ be a test function, integrate over the domain $\Omega$ and get

**Weak Maxwell's curl-equations**

$$\int_\Omega \left(\frac{\partial \mathbf{D}(\mathbf{x},t)}{\partial t} - \int_\Omega \nabla \times \mathbf{H}(\mathbf{x},t)\right) \cdot \boldsymbol{\psi}(\mathbf{x},t)\,\mathrm{d}\Omega = -\int_\Omega \mathbf{J}(\mathbf{x},t) \cdot \boldsymbol{\psi}(\mathbf{x},t)\,\mathrm{d}\Omega, \tag{2.3.5a}$$

$$\int_\Omega \left(\frac{\partial \mathbf{B}(\mathbf{x},t)}{\partial t} + \nabla \times \mathbf{E}(\mathbf{x},t)\right) \cdot \boldsymbol{\psi}(\mathbf{x},t)\,\mathrm{d}\Omega = 0, \tag{2.3.5b}$$

plus initial conditions. We have assumed to have zero boundary conditions $\hat{\mathbf{n}} \times \mathbf{E} = \mathbf{0}$, since this choice simplifies the weak formulation of Maxwell's equations, but in many cases this may be too restrictive. More about the nontrivial theory of (non-zero) traces can be found, e.g., in Refs [39, 42] and [43, Ch. 26 by L. Demkowicz].

For theory on existence and uniqueness of solutions to Maxwell's equations we refer to e.g. [39] and [44]; for an analytic solution for Kerr-nonlinear Maxwell's equations in one space dimension see [45].

## 2.4 Boundary Conditions

### 2.4.1 Perfect Electric and Magnetic Conductor

Inside a perfect electric conductor (PEC) the electric field $\mathbf{E}$ vanishes, thus we have the following boundary condition for the tangential component of $\mathbf{E}$:

$$\hat{\mathbf{n}} \times \mathbf{E} = \mathbf{0}. \tag{2.4.1}$$

Also recall that the tangential component of $\mathbf{E}$ must be continuous across material interfaces, see (2.2.1).
If we consider scattering problems, $\mathbf{E}$ is the sum of a known incident field $\mathbf{E}^{\text{in}}$ and a scattered field $\mathbf{E}^{\text{s}}$ which has to be determined. Then the PEC condition reads

$$\hat{\mathbf{n}} \times \mathbf{E}^{\text{s}} = -\hat{\mathbf{n}} \times \mathbf{E}^{\text{in}}. \tag{2.4.2}$$

The same can be formulated for the $\mathbf{H}$-field. Then one speaks of a perfect magnetic conductor (PMC).

### 2.4.2 Uniaxial Perfectly Matched Layers

Many relevant physical systems are open, not closed, which is the case for a perfect conductor. Berenger [46] found 1994 a way to numerically model open systems by introducing so-called perfectly matched layers (PML) via a split-field approach. Another formulation is the uniaxial PML (UPML) which was introduced by Gedney in 1996 [47]. Here, the PML is an artificially introduced layer to absorb traveling waves in a medium.
Figure 2.3 shall illustrate a PML for a two dimensional Cartesian system. Around the medium (beige) an artificial layer – the PML (grey)– is introduced, which allows electromagnetic waves to pass (orange), that is, the layer is "perfectly matched" to the medium itself. Inside the layer, any wave decays exponentially fast; at the outer boundary of the PML, reflection may occur (violet). Yet, if the layer is broad enough, the exponential decay prevents the medium from reflection effects. The damping of the wave inside the layer in the $i$-direction ($i = x, y$) can be controled by so-called PML parameters; in figure 2.3 these are denoted by $\sigma_x$, and $\sigma_y$. They can be adjusted as necessary.
Later it was shown that both approaches are equivalent to a third one: the stretched-coordinate PML approach by Chew and Weedon in 1994 [48], and by Teixeira and Chew in 1998 [49].

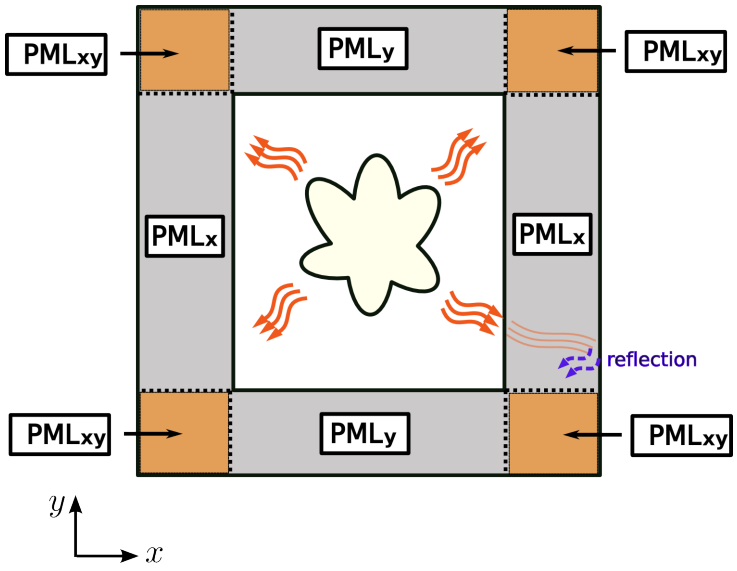



Figure 2.3: Illustration of a perfectly matched layer around a medium.

Here, we will work with the UPML approach for Maxwell's equations in bodies of revolution (BOR), which is why we will give a brief overview of the procedure and present the detailed computations of the auxiliary differential equations when we arrive at BOR Maxwell's equations. All the details of UPML can be found in the paper by Gedney [47], or e.g. in [22], [50], [51]. A mathematical discussion of PML is a topic of itself, and there

exist many works. We only mention two papers by Abarbanel and Gottlieb [52], [53], who showed that Maxwell's equations with Berenger's PML are only weakly well-posed in the sense that the norms of the Fourier transforms of the split fields $||\check{H}_x(t)||$ and $||\check{H}_y(t)||$ (see definition (2.4.3) below) are not only bounded by the norm of the corresponding initial fields, but also by the norm of the initial spatial derivatives of $\check{E}_y(t)$ and $\check{E}_z(t)$, which leads to instabilities in the solution after a small perturbation, i.e. to ill-posedness. In [53] they suggest a formulation that is well-posed in the sense that small perturbations do not lead to instabilities; one reason for well-posedness lies in the fact that the additional equations for the auxiliary variables are ordinary differential equations, which do not change the well-posedness if Maxwell's equations were well-posed in the beginning. We will work with such auxiliary variables.
In order to formulate Maxwell's equations with UPML we start by Fourier transforming them to the frequency domain. For a function $f \in L^1(\mathbb{R})$ we define its Fourier transform as

$$\check{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t) e^{-\mathrm{i}\omega t} \, \mathrm{d}t, \tag{2.4.3}$$

and its inverse Fourier transform as

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \check{f}(\omega) e^{\mathrm{i}\omega t} \, \mathrm{d}\omega.$$

For theory on the Fourier transform, see, e.g. [54]. We note that this definition can be extended to $f \in L^p(\mathbb{R})$ for $p \geq 1$, see e.g. [55].
Recalling theorem 2.5, we Fourier transform Maxwell's equations to the frequency space as

$$\begin{aligned} \nabla \times \check{H} &= -\mathrm{i}\omega\epsilon\check{E}, \\ \nabla \times \check{E} &= \mathrm{i}\omega\mu\check{H}. \end{aligned} \tag{2.4.4}$$

Here, $\omega \in \mathbb{R}$. The basic idea is to replace the material parameters $\epsilon, \mu$ by material tensors $\underline{\underline{\epsilon}} := \epsilon\underline{\underline{\Lambda}}, \underline{\underline{\mu}} := \mu\underline{\underline{\Lambda}}$, where in the Cartesian case, the tensor $\underline{\underline{\Lambda}}$ is defined as

$$\underline{\underline{\Lambda}} := \begin{pmatrix} \frac{s_y s_z}{s_x} & 0 & 0 \\ 0 & \frac{s_x s_z}{s_y} & 0 \\ 0 & 0 & \frac{s_y s_x}{s_z} \end{pmatrix}, \tag{2.4.5}$$

with

$$s_k(k) = \kappa_k(k) - \frac{\sigma_k}{\mathrm{i}\omega} \qquad (k = x, y, z),$$

where the $\sigma_k$ are the so-called PML parameters. Very often, $\kappa_k = 1$ is chosen, and we follow this choice.
In cylindrical coordinates the tensor is given as (see [56])

$$\underline{\underline{\Lambda}} := \begin{pmatrix} \frac{s_\phi s_z}{s_r} & 0 & 0 \\ 0 & \frac{s_r s_z}{s_\phi} & 0 \\ 0 & 0 & \frac{s_\phi s_r}{s_z} \end{pmatrix} \tag{2.4.6}$$

with

$$\begin{aligned} s_r(r) &= \kappa_r(r) - \frac{\sigma_r(r)}{\mathrm{i}\omega}, \\ s_z(z) &= \kappa_z(z) - \frac{\sigma_z(z)}{\mathrm{i}\omega}, \\ s_\phi(r) &= \frac{\tilde{r}(r)}{r}, \end{aligned}$$

where

$$\tilde{r}(r) = r_1 + \int_{r_1}^{r} s_r(r')dr'.$$

In $x$-direction we have $\sigma_x \neq 0, \sigma_y = \sigma_z = 0$, in $y$-direction it is $\sigma_x = \sigma_z = 0, \sigma_y \neq 0$, in $z$-direction $\sigma_x = \sigma_y = 0, \sigma_z \neq 0$, and in the corners $\sigma_k \neq 0$ $(k = x, y, z)$. In the medium, we have $\sigma_k = 0$. So, Maxwell's equations with UPML in Cartesian coordinates are given as

$$\begin{aligned} \nabla \times \check{H} &= -\mathrm{i}\omega\epsilon\underline{\underline{\Lambda}}\check{E}, \\ \nabla \times \check{E} &= \mathrm{i}\omega\mu\underline{\underline{\Lambda}}\check{H}. \end{aligned} \tag{2.4.7}$$

As an example we demonstrate the derivation of the auxiliary differential equation for the $\check{E}_x$-component. The computations for the other components are quite similar, and for BOR Maxwell's equations we will present all the details later.
Let us plug (2.4.5) into (2.4.4) so that we have

$$-\epsilon\mathrm{i}\omega\check{E}_x = \partial_y\check{H}_z - \partial_z\check{H}_y + \mathrm{i}\omega\epsilon\left(\frac{s_y s_z}{s_r} - 1\right)\check{E}_x.$$

We want to eliminate $\mathrm{i}\omega$, and we do this by first defining a so-called polarization current $\check{J}_i$ with $i \in \{x, y, z\}$ and with

$$\check{J}_x := \mathrm{i}\omega\epsilon\left(\frac{s_y s_z}{s_x} - 1\right)\check{E}_x.$$

Then we introduce a new auxiliary variables $P_i$ $(i \in \{x, y, z\})$. This results in additional equations for each $P_i$, so-called auxiliary differential equations (ADE). In the end, after reshaping the equations in a clever way, we get 12 equations in total for the electric fields $\mathbf{E}, \mathbf{H}$ and the corresponding polarizations $\mathbf{P}^{(E)}, \mathbf{P}^{(H)}$. For the $E_x$-component this looks as follows. First we encounter

$$\begin{aligned} \frac{s_y s_z}{s_x} - 1 &= \frac{\left(1 - \frac{\sigma_y}{\mathrm{i}\omega}\right)\left(1 - \frac{\sigma_z}{\mathrm{i}\omega}\right)}{1 - \frac{\sigma_x}{\mathrm{i}\omega}} - 1 = \frac{\left(\mathrm{i}\omega - \sigma_y\right)\left(1 - \frac{\sigma_z}{\mathrm{i}\omega}\right)}{\mathrm{i}\omega - \sigma_x} - 1 \\ &= \frac{1}{\mathrm{i}\omega - \sigma_x}\left(\frac{\sigma_y\sigma_z}{\mathrm{i}\omega} + \mathrm{i}\omega - \sigma_z - \sigma_y - \mathrm{i}\omega + \sigma_x\right), \end{aligned}$$

and therefore

$$\check{J}_x = \frac{\mathrm{i}\omega\epsilon}{\mathrm{i}\omega - \sigma_x}\left(\frac{\sigma_y\sigma_z}{\mathrm{i}\omega} - \sigma_z - \sigma_y + \sigma_x\right)\check{E}_x.$$

Now introduce $\check{P}_x^E := \check{J}_x + A\check{E}_x$. $A$ stands for an unknown expression which shall be determined such that $\mathrm{i}\omega$ drops out. We make the ansatz

$$(\mathrm{i}\omega - \sigma_x)\check{P}_x^E = \mathrm{i}\omega\epsilon\left(\frac{\sigma_y\sigma_z}{\mathrm{i}\omega} + \sigma_x - \sigma_z - \sigma_y\right)\check{E}_x + (\mathrm{i}\omega - \sigma_x)A\check{E}_x.$$

Thus we need $\mathrm{i}\omega A + \mathrm{i}\omega\epsilon(\sigma_x - \sigma_y - \sigma_z) = 0 \Leftrightarrow A = \epsilon(\sigma_y - \sigma_x + \sigma_z)$, and thus it follows:

$$-\mathrm{i}\omega\check{P}_x^E = -\sigma_x\check{P}_x^E - \epsilon\left(\sigma_y\sigma_z - \sigma_x\sigma_y + \sigma_x^2 - \sigma_x\sigma_z)\right)\check{E}_x.$$

Now we apply backward Fourier transform, taking into account that

$$\check{J}_x = \check{P}_x^E - A\check{E}_x = \check{P}_x^E - \epsilon(\sigma_y - \sigma_x + \sigma_z)\check{E}_x,$$

and we obtain Maxwell's equation for $E_x$ in time domain together with an auxiliary differential equation for $P_x^E$ as

$$\begin{aligned} \epsilon\partial_t E_x &= \partial_y H_z - \partial_z H_y + P_x^E - \epsilon(\sigma_y - \sigma_x + \sigma_z)E_x, \\ \partial_t P_x^E &= -\sigma_x P_x^E - \epsilon\left(\sigma_y\sigma_z - \sigma_x\sigma_y + \sigma_x^2 - \sigma_x\sigma_z)\right)E_x. \end{aligned}$$

## 2.5 Reduction to Lower Dimensions

### Reduction to Two Dimensions

If the system is homogeneous in one direction, e.g. the $z$-direction, the electromagnetic fields are constant in that direction, and the $z-$derivative drops out. Assuming $H_z = 0$ (TM polarization) or $E_z = 0$ (TE polarization), we obtain the set of equations (neglecting sources)

**TM Polarization**

$$\begin{aligned}\partial_t E_z &= (\epsilon_0\epsilon)^{-1}(\partial_x H_y - \partial_y H_x),\\ \partial_t H_x &= -(\mu_0\mu)^{-1}\partial_y E_z,\\ \partial_t H_y &= (\mu_0\mu)^{-1}\partial_x E_z.\end{aligned}$$

**TE Polarization**

$$\begin{aligned}\partial_t E_x &= (\epsilon_0\epsilon)^{-1}\partial_y H_z,\\ \partial_t E_y &= -(\epsilon_0\epsilon)^{-1}\partial_x H_z,\\ \partial_t H_z &= (\mu_0\mu)^{-1}(\partial_x E_y - \partial_y E_x).\end{aligned}$$

### Reduction to One Dimension

If the system also inhibits a translational invariance, like e.g. in the $y$-direction, this results in two equations for $E_z$ and $H_y$ (TM polarization):

$$\begin{aligned}\partial_t E_z &= (\epsilon_0\epsilon)^{-1}\partial_x H_y,\\ \partial_t H_y &= (\mu_0\mu)^{-1}\partial_x E_z.\end{aligned}$$

## 2.6 Maxwell's Equations as a Conservation Law

Maxwell's curl-equations can be reformulated in conservation form. For a deeper insight into theory about conservation laws and hyperbolic equations, see e.g. [57], [58] and [30]. Here, we give a brief overview.

### 2.6.1 Overview of the Theory of Hyperbolic Conservation Laws

We consider partial differential equations of the form

$$\mathbf{Q}\partial_t \mathbf{u} + \sum_{j=1}^{m} \frac{\partial \mathbf{F}_j(\mathbf{u})}{\partial x_j} = \mathbf{S}\mathbf{u}, \quad \mathbf{x} = (x_1, \ldots, x_m)^T \in \mathbb{R}^m, \quad t \in \mathbb{R}, \tag{2.6.1}$$

where $\partial_t \mathbf{u} = \frac{\partial \mathbf{u}}{\partial t}$ denotes the partial time derivative of the vector valued function $\mathbf{u} := (u_1, \ldots, u_n)^T$ with $\mathbf{u} : \mathbb{R}^m \times \mathbb{R} \to \Omega$, where $\Omega \subset \mathbb{R}^n$ is open, i.e. $u_i = u_i(x_1, \ldots, x_m, t)$, and it is $\mathbf{S} \in \mathbb{R}^{n\times n}$, $\mathbf{Q} \in \mathbb{R}^{n\times n}$. $\mathbf{u}$ is called the *state vector*, and $\Omega$ is called the set of states. $\mathbf{F}_j : \Omega \to \mathbb{R}^n, j = 1, \ldots, m$, is called the flux function; it is $\mathbf{F}_j = (F_{1j}, \ldots, F_{pj})^T$, where all $F_{ij}$ shall be smooth. We define the so-called *flux vector* $\mathbf{F} := (\mathbf{F}_1, \ldots, \mathbf{F}_m)$, which is a $(n \times m)$-matrix, i.e.

$$\mathbf{F} = \begin{pmatrix} F_{11} & \cdots & F_{1m} \\ \vdots & & \vdots \\ F_{n1} & \cdots & F_{nm} \end{pmatrix}. \tag{2.6.2}$$

We furthermore define the divergence of a matrix-valued field as follows:

**Definition 2.6.**
*The divergence* $\nabla \cdot \mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ *of a matrix-valued field* $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ *is defined as*

$$\nabla \cdot \mathbf{F} := \begin{pmatrix} \partial_{x_1} F_{11} + \cdots + \partial_{x_m} F_{1m} \\ \vdots \\ \partial_{x_1} F_{n1} + \cdots + \partial_{x_m} F_{nm} \end{pmatrix}, \tag{2.6.3}$$

*or, shortly,*

$$\nabla \cdot \mathbf{F} = \left( \sum_{j=1}^{m} \partial_{x_j} F_{ij} \right)_{i=1}^{n}.$$

With this we can write equation (2.6.1) as

$$\mathbf{Q}\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \mathbf{S}\mathbf{u}. \tag{2.6.4}$$

This system is said to be in *conservative form*. One also calls it a *conservation law*.

**Definition 2.7.**
*The Jacobian matrix* $\mathbf{A}_j(\mathbf{u}) \in \mathbb{R}^{n \times n}$ *of the flux vector components* $\mathbf{F}_j$ *is defined as*

$$\mathbf{A}_j(\mathbf{u}) := \left( \frac{\partial F_{ij}(\mathbf{u})}{\partial u_k} \right)_{1 \le i,k \le n}, \quad j = 1, \ldots, m.$$

**Definition 2.8.**
*The conservation law* (2.6.4) *is called hyperbolic if, for any arbitrary vector* $\mathbf{v} \in \mathbb{R}^m$ *and* $\mathbf{u} \in \Omega \subset \mathbb{R}^n$*, the matrix*

$$\mathbf{A}(\mathbf{u}; \mathbf{v}) := \sum_{j=1}^{q} v_j \mathbf{A}_j(\mathbf{u})$$

*has n real eigenvalues* $\lambda_1 \le \cdots \le \lambda_n$ *(where* $\lambda_i = \lambda_i(\mathbf{u}; \mathbf{v}), i = 1, \ldots, n$*) with corresponding n linearly independent right eigenvectors* $\mathbf{r}_1, \ldots, \mathbf{r}_n$ *(where* $\mathbf{r}_i = \mathbf{r}_i(\mathbf{u}; \mathbf{v}), i = 1, \ldots, n$*), i.e.*

$$\mathbf{A}(\mathbf{u}; \mathbf{v})\mathbf{r}_i = \lambda_i \mathbf{r}_i, \quad i = 1, \ldots, n.$$

*If all the the eigenvalues are distinct, the system is called strictly hyperbolic.*

If everything is smooth enough (!), one can rewrite (2.6.4) in non-conservative form by using the Jacobian as

$$\mathbf{Q}\partial_t \mathbf{u} + \mathbf{A}(\mathbf{u})\nabla \mathbf{u} = \mathbf{S}\mathbf{u}. \tag{2.6.5}$$

By $\nabla \mathbf{u}$ we denote the partial derivative of $\mathbf{u}$ with respect to $x$. If all entries of the matrices $\mathbf{A}(\mathbf{u})$ and $\mathbf{S}$ are constant in $\mathbf{u}$, the system is called linear with constant coefficients. If $\mathbf{A}(\mathbf{u})$ and $\mathbf{S}$ are independent of $\mathbf{u}$, but dependent on $x, t$, then one says the system is linear with variable coefficients. If only $\mathbf{S}$ depends on $\mathbf{u}$, the system is still linear, and it is quasi-linear if $\mathbf{A}(\mathbf{u})$ depends non-linearly on $\mathbf{u}$. If $\mathbf{S} = \mathbf{0}$, the system is called homogeneous.
For further reading we refer to [59], [58], [57], [30], which is only a very small selection from literature.
We start our theoretical study with the conservation law (2.6.4) with $\mathbf{S} = \mathbf{0}$ and with initial conditions, i.e. the Cauchy problem

$$\begin{cases} \mathbf{Q}\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \mathbf{0}, \\ \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^m, \end{cases}$$

where $\mathbf{u}_0 : \mathbb{R}^m \to \Omega \subset \mathbb{R}^n$ is a given function.

### 2.6.2 Integral Formulation of Conservation Laws

Conservation laws can also be expressed as integral equations, and with respect to physics this seems reasonable, since several governing equations are derived using conservation laws in integral from, as conservation of mass or momentum. Furthermore, less smoothness of solutions is required, and discontinuous functions should be allowed. When using the discontinuous Galerkin method, to which we will come in the next chapter, this is an indispensable requirement.
There exist several different integral forms which are equivalent to each other. Each one assumes to have a so-called control volume, which we shall define as $V := [t_1, t_2] \times D$, where $[t_1, t_2]$ is the time domain of the control volume, and $D$ is the spatial domain of interest. For instance, in one dimension, this would be $D = [x_l, x_r]$; see figure 2.4 for an illustration.

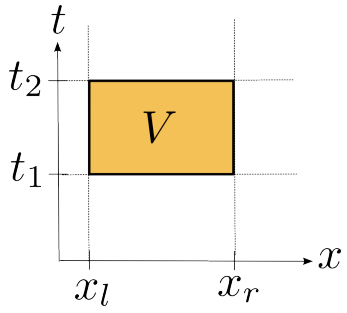



Figure 2.4: Sketch of a control volume $V$ in one dimension in $xt$-plane.

#### Integral Form No. 1

Let $\mathbf{S} = \mathbf{0}$ and $\mathbf{Q} = \mathrm{Id}$. We get a first integral form by integrating the differential equation (2.6.4) over $D$ so that

$$\partial_t \int_D \mathbf{u}(\mathbf{x}, t)\, \mathrm{d}\mathbf{x} = -\int_{\partial D} \mathbf{F}(\mathbf{u}(\mathbf{x}, t))\mathbf{n}\, \mathrm{d}\Gamma, \tag{2.6.6}$$

where $\mathbf{n}$ is the outer normal of unit length of $D$. In one dimension, we simply have

$$\int_{\partial D} \mathbf{F}(\mathbf{u}(\mathbf{x}, t))\mathbf{n}\, \mathrm{d}\Gamma = \left(\mathbf{F}(\mathbf{u}(\mathbf{x}_l, t)) - \mathbf{F}(\mathbf{u}(\mathbf{x}_r, t))\right)\mathbf{n}.$$

#### Integral Form No. 2

Starting with (2.6.6) and integrating over $[t_1, t_2]$ we get a second integral form:

$$\left(\int_D \mathbf{u}(\mathbf{x}, t_2)\mathrm{d}\mathbf{x} - \int_D \mathbf{u}(\mathbf{x}, t_1)\, \mathrm{d}\mathbf{x}\right) = -\int_{t_1}^{t_2} \int_{\partial D} \mathbf{F}(\mathbf{u}(\mathbf{x}, t))\mathbf{n}\, \mathrm{d}\mathbf{x}\, \mathrm{d}t. \tag{2.6.7}$$

From both equations one can see that changes of $\mathbf{u}$ inside the domain $D$ are only possible due to fluxes $\mathbf{F}(\mathbf{u}(\mathbf{x}, t))\mathbf{n}$ over the boundary $\partial D$.

#### Integral Form No. 3: Weak Form

We assume again $\mathbf{S} = 0$ and $\mathbf{Q} = \mathrm{Id}$. The weak form of (2.6.4) is

$$\int_D (\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u})) \cdot \mathbf{\Psi}\, \mathrm{d}\mathbf{x} = 0. \tag{2.6.8}$$

for all test functions $\mathbf{\Psi}$. On each control volume $D \subset \Omega$ we have

$$\int_D (\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u})) \cdot \mathbf{\Psi} \, \mathrm{d}\mathbf{x} = -\int_{\partial D} \mathbf{F}(\mathbf{u}(\mathbf{x},t))\mathbf{n} \cdot \mathbf{\Psi} \, \mathrm{d}\Gamma. \tag{2.6.9}$$

If we integrate over $[t_1, t_2] \times D$, we get the integral form No. 2 in the weak sense, and its right hand side reads

$$\int_{t_1}^{t_2} \int_{\partial D} \mathbf{F}(\mathbf{u}(\mathbf{x},t))\mathbf{n} \cdot \mathbf{\Psi}(\mathbf{x},t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t = (t_2 - t_1) \int_{\partial D} \left( \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \mathbf{F}(\mathbf{u}(\mathbf{x},t))\mathbf{n} \, \mathrm{d}t \right) \cdot \mathbf{\Psi}(\mathbf{x},t) \, \mathrm{d}\mathbf{x}. \tag{2.6.10}$$

### 2.6.3 Conservation Form of Maxwell's Equations

Maxwell's curl-equations read

$$\nabla \times \mathbf{E} = -\mu \partial_t \mathbf{H}, \qquad \nabla \times \mathbf{H} = \epsilon \partial_t \mathbf{E}.$$

We want to bring these into conservation form (2.6.4). For this we define the state vector $\mathbf{u} := (\mathbf{E}, \mathbf{H})^T \in \mathbb{R}^6$ and the flux vector $\mathbf{F}(\mathbf{u}) := (\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_z)^T \in \mathbb{R}^{6\times 3}$, where the components $\mathbf{F}_k$ $(k = x, y, z)$ are given as $\mathbf{F}_k = (-\hat{\mathbf{e}}_k \times \mathbf{H}, \hat{\mathbf{e}}_k \times \mathbf{E})^T \in \mathbb{R}^6$. We also define the material matrix

$$Q := \begin{pmatrix} \underline{\underline{\epsilon}} & \mathbf{0} \\ \mathbf{0} & \underline{\underline{\mu}} \end{pmatrix},$$

where we have assumed that the material tensors $\underline{\underline{\epsilon}}, \underline{\underline{\mu}}$ are $3 \times 3-$matrices, and $\epsilon$ and $\mu$ shall be constant. In general they can vary in space. Furthermore, $Q \in \mathbb{R}^{6\times 6}$ shall be invertible. We can thus write Maxwell's equations in conservation form as

$$Q \, \partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = 0. \tag{2.6.11}$$

One can show by direct computation that this is a hyperbolic system with the eigenvalues $\lambda_1 = 0$ and $\lambda_{2,3} = \pm \frac{1}{\sqrt{\epsilon\mu}}$, where each eigenvalue has multiplicity 2.

Chapter 3

# 3 The Runge-Kutta Discontinuous Galerkin Method

The DG method was first introduced by Reed and Hill for the neutron transport equation in 1973 [1]. Since then it has undergone a fast development and finds numerous applications to e.g. the Euler equations of gas dynamics, the shallow water equations, the equations of magneto-hydrodynamics, the compressible Navier-Stokes equations and Maxwell's equations – only to mention a few. The DG method can be used for time and/or spatial discretization. We will apply it to BOR Maxwell's equations for spatial discretization and later to Maxwell's equations with a Kerr-nonlinearity. We use an explicit time stepping method, the explicit low-storage Runge-Kutta method, originally developed by Williamson [6] and extended by M. Carpenter and C. Kennedy [7]. This combination, the DG method for spatial discretization and an explicit Runge-Kutta scheme for time-integration, also called Runge-Kutta Discontinuous Galerkin (RKDG) method in mathematics and Discontinuous Galerkin Time-Domain (DGTD) method in physics, has been introduced by Cockburn and Shu in several papers, [8] maybe being one of the latest. In [9], which can be downloaded online and which also gives an introduction to the DG method, a more complete list can be found. Also [10], [11] and [12] (a review paper about RKDG methods) shall be mentioned as literature about the DG method. [13] gives a review of the DG method applied to nanophotonics. Of course, there are many other good books and papers on this topic.

The DG method encounters several advantages:

- It's a high order accuracy method.
- It allows meshes with elements of any kind; thus, it can handle complicated geometries.
- It is a local method, e.g. in the linear case parallelization is possible.
- It can handle discontinuous solutions.
- It can be implemented relatively easy.

### The Key Ideas

A complete DG method consists of the following main steps:

(1) Space discretization, including the definition of a finite element space with discontinuous functions. This leads to a local scheme, where the exact solution is approximated on each element.

(2) The global approximation to the exact solution is obtained by connecting all local solutions on the local elements via the so-called *numerical flux*. The choice of a numerical flux is the main ingredient of a DG method. It has to be chosen such that the resulting scheme is consistent and convergent to the exact solution. The choice of a numerical flux is not unique and there are many possible ways to do it. We will choose a flux that solves a so-called Riemann problem.

(3) Space discretization leads to a semi-discrete scheme that needs to be integrated in time. We use an explicit low-storage Runge-Kutta scheme of 4th order and with 5 stages, see [7]. For nonlinear problems, it may be necessary to use a stability preserving RK method, like a high-order strong stability preserving Runge-Kutta scheme by Ruuth and Spiteri [60].

Sometimes a step (4) is needed: If discontinuous solutions can occur (e.g. in case of nonlinear problems) a so-called slope limiter may be required to stabilize the DG scheme due to oscillations near discontinuities, which is the well-known Gibbs-phenomenon. Furthermore, another important aspect is regaining the high-order nature of the DG method by applying a slope limiter.

## 3.1 Space Discretization with the Discontinuous Galerkin Method

As we have seen in section 2.6 Maxwell's curl-equations can be written as a conservation law, and we have seen it is a hyperbolic system of equations. For later purposes, let us recall the conservation law (2.6.4)

$$\mathbf{Q}\,\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \mathbf{S}\mathbf{u}, \tag{3.1.1}$$

$$\mathbf{u}(\mathbf{x}, t) = g(\mathbf{x}, t),\ \mathbf{x} \in \partial\Omega, \tag{3.1.2}$$

$$\mathbf{u}(\mathbf{x}, 0) = f(\mathbf{x}) \tag{3.1.3}$$

with the unknown solution $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ with $\mathbf{u} : [0, T] \times \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{Q}$ and $\mathbf{S}$ are $m \times m$-matrices, not necessarily constant, $\Omega \subset \mathbb{R}^n$ and $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^{m \times n}$. The divergence $\nabla \cdot \mathbf{F}$ was defined in 2.6.

The weak form of (3.1.1) on $\Omega$ is given as

$$\int_\Omega (\mathbf{Q}\,\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) - \mathbf{S}\mathbf{u}) \cdot \mathbf{\Psi}\, \mathrm{d}\Omega = \mathbf{0} \tag{3.1.4}$$

for all test functions $\mathbf{\Psi}$. Now we tessellate $\Omega$ by a number of $K$ conforming simplices $\Omega_k$, i.e.

$$\Omega \approx \Omega^h := \bigcup_{k=1}^{K} \Omega_k,$$

for a conforming triangulation $\mathcal{T}_h := \{\Omega_k\}$. We define the corresponding finite element space of discontinuous functions as

$$V_h := \{\mathbf{u}_h \in L^\infty(\Omega) : \mathbf{u}_h^k := \mathbf{u}_h|_{\Omega_k} \in V(\Omega_k)\ \forall\ \Omega_k \in \mathcal{T}_h\}. \tag{3.1.5}$$

$V(\Omega_k)$ is called the local space, and it should be chosen such that spurious, non-physical solutions are avoided; see e.g. [36, 61] on the topic of spurious solutions of Maxwell's equations. One possibility are $H(\mathrm{curl})$-conforming elements, as e.g. Nedelec's elements [62]. A finite element space is called conforming, if it is a proper subspace of the original continuous function space, in our case, of $H(\mathrm{curl})$. Hesthaven and Warburton chose another way [63], the *nodal* Discontinuous Galerkin method, and we follow it by letting $V(\Omega_k) = \mathcal{P}^p(\Omega_k)$, which is the space of multivariate polynomials of total degree $p \in \mathbb{N}$, as in e.g. [8] or [11].
Let $\mathbf{n}$ be the external outer normal of unity length, pointing from local element $\Omega_k$ to a neighboring element $\Omega_{k'}$. On each $\Omega_k$ we have

$$\int_{\Omega_k} \left[(\mathbf{Q}\,\partial_t \mathbf{u} - \mathbf{S}\mathbf{u}) \cdot \Psi - \mathbf{F}(\mathbf{u}) \cdot \nabla \mathbf{\Psi}\right] \mathrm{d}\Omega = -\int_{\partial\Omega_k} \mathbf{F}(\mathbf{u})\,\mathbf{n} \cdot \mathbf{\Psi}\, \mathrm{d}\Gamma. \tag{3.1.6}$$

The (continuous) Galerkin ansatz consists in requiring the residual

$$\mathcal{R}_h := \partial_t \mathbf{u}_h + \nabla \cdot \mathbf{F}(\mathbf{u}_h) - \mathbf{S}\mathbf{u}_h$$

to be orthogonal to all test functions $\boldsymbol{\Psi_h} \in V_h$, i.e.

$$\int_{\Omega_k} \mathcal{R}_h \cdot \boldsymbol{\Psi}_h \, \mathrm{d}\Omega_k = \mathbf{0}.$$

In the discontinuous Galerkin ansatz the functions may be constrained across the interfaces, so that we get (after integration by parts)

$$\int_{\Omega_k} \partial_t \mathbf{u}(\mathbf{x},t)\, \boldsymbol{\Psi} - \mathbf{F}(\mathbf{u}(\mathbf{x},t)) \cdot \nabla \boldsymbol{\Psi} \, \mathrm{d}\Omega_k = -\int_{\partial\Omega_k} \mathbf{F}(\mathbf{u}(\mathbf{x},t))\, \mathbf{n} \cdot \boldsymbol{\Psi} \, \mathrm{d}\Gamma. \tag{3.1.7}$$

We want to model the right hand side of (3.1.7) in an adequate way by a so-called numerical flux $\mathbf{F}^{\mathrm{num}}$. This numerical flux inhibits the information about how the edge values of a local cell $\Omega_k$ are connected to the edge values of a neighboring cell $\Omega_{k'}$. It is thus a function of $\mathbf{u}_h^k$ and $\mathbf{u}_h^{k'}$. That is, it is needed to reconstruct the global solution from all local solutions. So to complete any DG scheme one needs an expression for the numerical flux. Thus we set

$$\int_{\Omega_k} \left[ (\mathbf{Q}\, \partial_t \mathbf{u}_h^k - \mathbf{S}\mathbf{u}_h^k) \cdot \boldsymbol{\Psi}_h - \mathbf{F}(\mathbf{u}_h^k, \mathbf{u}_h^{k'}) \cdot \nabla \boldsymbol{\Psi}_h \right] \mathrm{d}\Omega_k = -\int_{\partial\Omega_k} \mathbf{F}^{\mathrm{num}}(\mathbf{u}_h^k, \mathbf{u}_h^{k'})\, \mathbf{n} \cdot \boldsymbol{\Psi}_h \, \mathrm{d}\Gamma \tag{3.1.8}$$

for all $\boldsymbol{\Psi_h} \in V_h$. We apply integration by parts again and get

$$\int_{\Omega_k} (\mathbf{Q}\, \partial_t \mathbf{u}_h^k + \nabla \cdot \mathbf{F}(\mathbf{u}_h^k) - \mathbf{S}\mathbf{u}_h^k) \cdot \boldsymbol{\Psi}_h \, \mathrm{d}\Omega_k = \int_{\partial\Omega_k} (\mathbf{F}(\mathbf{u}_h^k) - \mathbf{F}^{\mathrm{num}}(\mathbf{u}_h, \mathbf{u}_h^{k'}))\, \mathbf{n} \cdot \boldsymbol{\Psi}_h \, \mathrm{d}\Gamma. \tag{3.1.9}$$

In section 3.2 we will specify the expression $\mathbf{F}^{\mathrm{num}}$ of the numerical flux.

## 3.2 The Numerical Flux

The numerical flux connects the single solutions $\mathbf{u}_h^k$ on each element $\Omega_k$ and thus recovers the global approximation $\mathbf{u}_h$. Due to our discontinuous ansatz, the functions $\mathbf{u}_h^k$ can have – and generally do have – different values on the edges $e$ between two neighboring elements. The numerical flux $\mathbf{F}^{\mathrm{num}}(\mathbf{u}_h)$ connects these different edge values. It thus depends on the interior and exterior values of $\mathbf{u}_h^k(\mathbf{x}_e, t)$ of element $\Omega_k$, where $\mathbf{x}_e$ is a point on the edge $e := \partial\Omega_k \cap \partial\Omega_{k'}$. The meaning of "interior" and "exterior" is defined as follows.

$$\mathbf{u}_h^k(\mathbf{x}^{\mathrm{int}(\Omega_k)}, t) = \lim_{\mathbf{x}\to\mathbf{x}_e, \mathbf{x}\in\Omega_k} \mathbf{u}_h^k(\mathbf{x}, t),$$

$$\mathbf{u}_h^k(\mathbf{x}^{\mathrm{ext}(\Omega_k)}, t) = \begin{cases} g_h(\mathbf{x}, t), & x \in \partial\Omega, \\ \lim\limits_{\mathbf{x}\to\mathbf{x}_e, \mathbf{x}\in\Omega_{k'}} \mathbf{u}_h^k(\mathbf{x}, t), & \text{otherwise.} \end{cases}$$

By $g_h(\mathbf{x}, t)$ the discrete boundary values are meant. In what will follow, we will drop the index $k$, whenever clarity is not lost by it, and abbreviate $\mathbf{u}_h^k(\mathbf{x}^{\mathrm{int}(\Omega_k)}, t) = \mathbf{u}_h^-$ and call it the interior edge value of element $\Omega_k$, while $\mathbf{u}_h^k(\mathbf{x}^{\mathrm{ext}(\Omega_k)}, t) = \mathbf{u}_h^+$ is called the exterior edge value of $\Omega_k$. Figure 3.1 shows a sketch of this situation.

Thus the numerical flux $\mathbf{F}^{\mathrm{num}}$ depends on the exterior and interior edge values, that is $\mathbf{F}^{\mathrm{num}}(\mathbf{u}_h) = \mathbf{F}^{\mathrm{num}}(\mathbf{u}_h^-, \mathbf{u}_h^+)$. Information from one cell to another is transported in

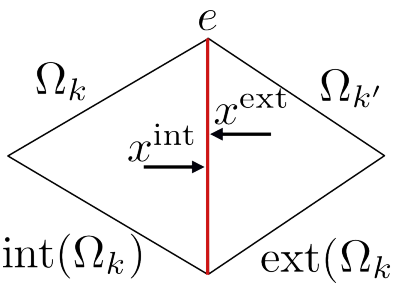



Figure 3.1: A sketch of the meaning of "interior" and "exterior".

direction of the outer unit normal $\mathbf{n}$ of $\Omega_k$, so we need the numerical flux along $\mathbf{n}$, that is, we let

$$\mathbf{F}_{\mathbf{n}}^{\text{num}}(\mathbf{u}_h^-, \mathbf{u}_h^+) := \mathbf{F}^{\text{num}}(\mathbf{u}_h^-, \mathbf{u}_h^+)\mathbf{n}, \qquad \text{and} \qquad \mathbf{F}_{\mathbf{n}}(\mathbf{u}) := \mathbf{F}(\mathbf{u})\mathbf{n}.$$

There are many possible choices of a numerical flux, see e.g. [57], [9]; in general it is not unique. In section 3.2.2 we will give some examples of numerical fluxes. In order to produce a convergent DG scheme, the numerical flux has to fulfill the following conditions [9], [8]:

1. It is locally Lipschitz.
2. It is consistent with the flux vector $\mathbf{F}$, i.e. $\mathbf{F}_{\mathbf{n}}^{\text{num}}(\mathbf{z}, \mathbf{z}) = \mathbf{F}_{\mathbf{n}}(\mathbf{z})$ for all $\mathbf{z} \in \mathbb{R}^n$.
3. It is conservative, that is, we require

   $$\mathbf{F}_{\mathbf{n}}^{\text{num}}(\mathbf{u}_{h,k}^-, \mathbf{u}_{h,k}^+) + \mathbf{F}_{\mathbf{n}}^{\text{num}}(\mathbf{u}_{h,k'}^-, \mathbf{u}_{h,k'}^+) = \mathbf{0}$$

   on all edges $e = \Omega_k \cap \Omega_{k'}$. This condition ensures that the numerical scheme is an approximation to the conservation law, and not to an arbitrary other problem. It also means that the numerical flux at the boundary between $\Omega_k$ and $\Omega_{k'}$ is the same as the one separating $\Omega_{k'}$ and $\Omega_k$.
4. If $\mathbf{F}_{\mathbf{n}}^{\text{num}}$ is a vector-valued numerical flux, the mapping $\mathbf{z} \mapsto \mathbf{F}_{\mathbf{n}}^{\text{num}}(\mathbf{z}, \cdot)$ shall be non-decreasing. We say a vector valued function is non-decreasing if all its components are non-decreasing.

These requirements are motivated from finite volume methods, where, for scalar problems, a numerical flux with these properties gives rise to a monotone scheme. Monotone schemes were shown to be stable and convergent to the exact solution of order zero, see [64], [65] and [66]. More details can also be found in [59, Prop. 4.2] or [57], giving only a small selection of references. Thus, the numerical flux in a DG method is chosen such that for piecewise constant approximations $\mathbf{u}_h$ the scheme leads to a monotone finite volume method. We will say more about monotonicity in the next section, where we give some basics about finite volume methods in order to motivate the choice and role of the numerical flux in a DG scheme.

### 3.2.1 The Numerical Flux and Finite Volume Methods

Originally, the concept of a numerical flux was inspired by finite volume methods, where the numerical flux is also a means of transporting information from one cell to another. And indeed, a DG method with polynomial order zero is nothing else than a finite volume method. In this section we want to motivate the connection between the numerical flux

$\mathbf{F}^{\text{num}}$ of a DG scheme and the numerical flux of a finite volume method. We refer to e.g. [57, 58] for substantial theory about hyperbolic equations and their numerical treatment by using finite volume methods.
For the following we let $\mathbf{S} = \mathbf{0}$ and $\mathbf{Q} = \text{Id}$ in (2.6.4). The conservation law (2.6.4) reads in integral form

$$\int_\Omega \partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) \, \mathrm{d}\mathbf{x} = 0, \tag{3.2.1}$$

or, equivalently,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \mathbf{u} \, \mathrm{d}\mathbf{x} = - \int_{\partial\Omega} \mathbf{n} \cdot \mathbf{F}(\mathbf{u}) \, \mathrm{d}\Gamma. \tag{3.2.2}$$

The expression on the right hand side gives a flux across the boundary of $\Omega$, that is, $\mathbf{u}$ changes inside $\Omega$ only due to the flux across its boundary $\partial\Omega$. This is the conservative property.
A finite volume space discretization consists of the following main steps (figure 3.2 a) shall illustrate the procedure):

1. Subdivide the domain into so-called grid cells (finite volumes) $\Omega_k$. A neighboring element is denoted by $\Omega_{k'}$. In one space dimension we write $\Omega_i = [x_{i-1/2}, x_{i+1/2}]$ and denote a neighboring element by $\Omega_j$. In this case the step size in space is denoted by $\Delta x_i := x_{i+1/2} - x_{i-1/2}$. We also define a time grid $0 = t_0 \leq t_1 \leq \cdots \leq t_M = T$ with time step size $\Delta t_l := t_{l+1} - t_l$ $(l = 1, \ldots, M)$. For a uniform one-dimensional grid in space, it is $\Delta x_i = \Delta x$ for all $i$, and for a uniform grid in time it is $\Delta t_l = \Delta t$ for all $l$.

2. The conservation law is formulated on each cell $\Omega_k$ as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega_k} \mathbf{u}(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} = - \int_{\partial\Omega_k} \mathbf{n}_k \cdot \mathbf{F}(\mathbf{u}(\cdot, t)) \, \mathrm{d}\Gamma, \tag{3.2.3}$$

   where $\mathbf{n}_k$ is the outward unit normal vector to $\Omega_k$.

3. On each $\Omega_k$ approximate the integral on the left hand side by

$$\frac{1}{\text{vol}(\Omega_k)} \int_{\Omega_k} \mathbf{u}(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \approx \mathbf{u}_k(t).$$

   At each time step $t_l$ we have $\mathbf{u}_k(t) \approx \mathbf{u}_k(t_l) =: \mathbf{u}_k^l$. The integral on the left side is the cell average of $\mathbf{u}$ over $\Omega_k$, and so, $\mathbf{u}_k^l$ is an approximation to this cell average. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega_k} \mathbf{u}(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \approx \text{vol}(\Omega_k) \frac{\mathrm{d}\mathbf{u}_k(t)}{\mathrm{d}t}.$$

   We also want to approximate the flux integral on the right hand side of (3.2.3). We first note that it holds

$$\int_{\partial\Omega_k} \mathbf{n}_k \cdot \mathbf{F}(\mathbf{u}) = \sum_{e_{k,k'} \subset \partial\Omega_k} \int_{e_{k,k'}} \mathbf{n}_k \cdot \mathbf{F}(\mathbf{u}) \, \mathrm{d}\Gamma,$$

   where $\partial\Omega_k$ is the union of all edges $e_{k,k'}$ with $e_{k,k'} = \partial\Omega_k \cap \partial\Omega_{k'}$ being an edge separating $\Omega_k$ and $\Omega_{k'}$. Since we only have information about $\mathbf{u}_k(t)$ and since in case of hyperbolic systems information travels with finite speed, we approximate the flux integral by only using the values $\mathbf{u}_k(t)$. One then introduces a function

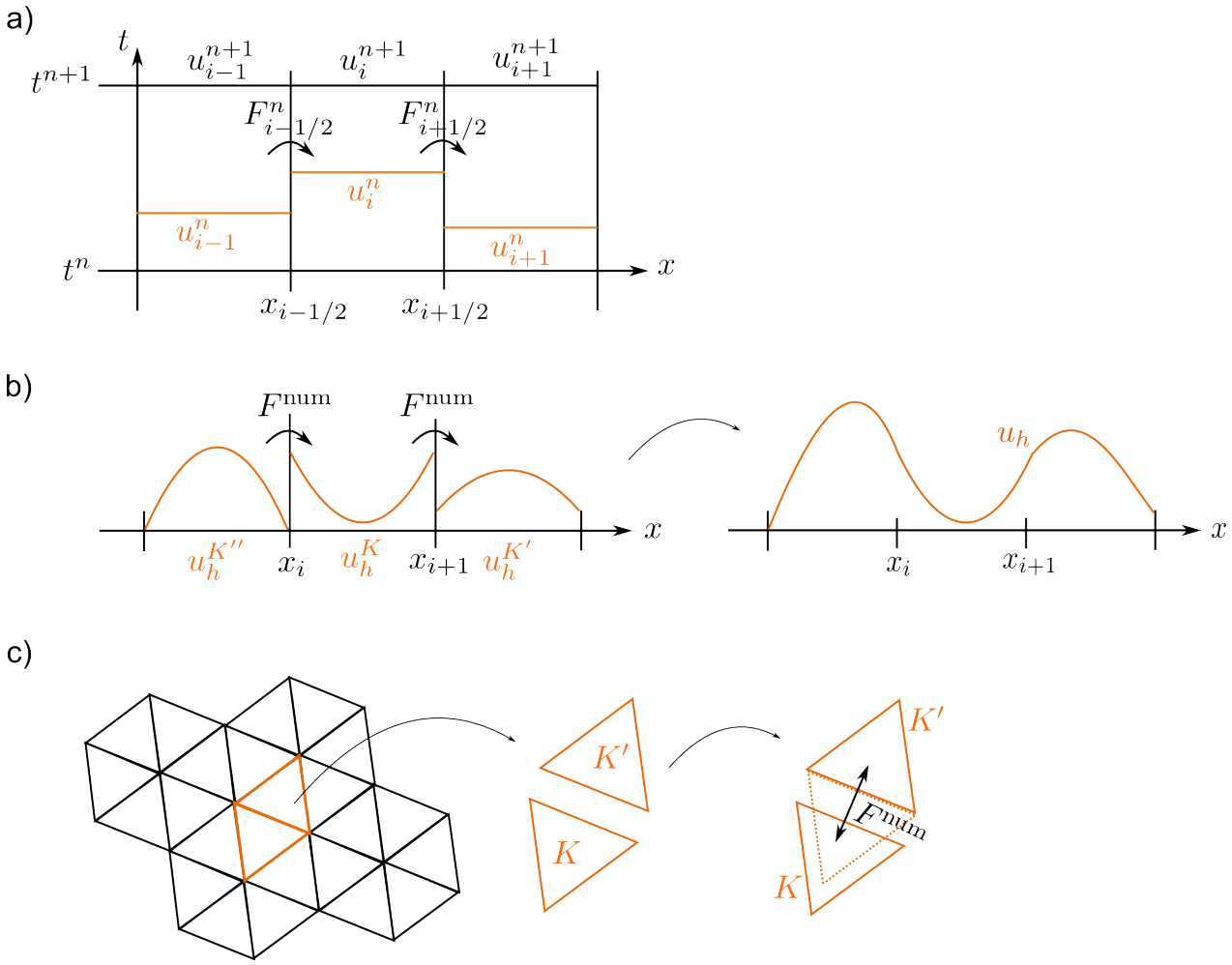



Figure 3.2: a) Finite volume discretization. b) Analogue in the DG method for one space dimension, c) and for two dimensions.

$\mathcal{F} = \mathcal{F}(\mathbf{u}_k, \mathbf{u}_{k'})$ which shall be an approximation to the flux integral in (3.2.3) so that

$$\int_{e_{k,k'}} \mathbf{n}_k \cdot \mathbf{F}(\mathbf{u}) \, \mathrm{d}\Gamma \approx |e_{k,k'}| \mathcal{F}(\mathbf{u}_k, \mathbf{u}_{k'}; \mathbf{n}_{e_{k,k'}}), \tag{3.2.4}$$

where $\mathbf{n}_{e_{k,k'}}$ is the unit normal to $e_{k,k'}$ in direction of $\Omega_{k'}$. The function $\mathcal{F}$ is called *numerical flux*, and in order to complete any finite volume method it has to be chosen appropriately. This gives a semi-discrete scheme of the form

$$\mathrm{vol}(\Omega_k) \frac{\mathrm{d}\mathbf{u}_k(t)}{\mathrm{d}t} = - \sum_{e_{k,k'} \subset \partial\Omega_k} |e_{k,k'}| \mathcal{F}(\mathbf{u}_k, \mathbf{u}_{k'}; \mathbf{n}_{e_{k,k'}})$$

4. By using a time-stepping method (e.g. an explicit Euler scheme) we integrate in time and obtain

$$\mathbf{u}_k^{l+1} = \mathbf{u}_k^l - \frac{\Delta t}{\mathrm{vol}(\Omega_k)} \sum_{e_{k,k'} \subset \partial\Omega_k} |e_{k,k'}| \mathcal{F}(\mathbf{u}_k, \mathbf{u}_{k'}; \mathbf{n}_{e_{k,k'}}). \tag{3.2.5}$$

For example, in one space dimension, it is

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{x_1}^{x_2} u \, \mathrm{d}x = F(u(x_1, t)) - F(u(x_2, t)). \tag{3.2.6}$$

The domain is the interval $\Omega := [x_1, x_2]$ with subdomains $\Omega_i$. One thus looks for methods of the form

$$u_i^{l+1} = u_i^l - \frac{\Delta t}{\Delta x}(F_{i+1/2}^l - F_{i-1/2}^l), \tag{3.2.7}$$

where

$$F_{i\pm1/2}^l \approx \frac{1}{\Delta t} \int_{t_l}^{t_{l+1}} F(u(x_{i\pm1/2}, t)) \, \mathrm{d}t.$$

For more information and theory, see e.g. [57] and [59].
With (3.2.5) we have obtained a numerical scheme in *conservation form*.

**Definition 3.1** (Conservative Scheme)**.**
*A method of the form*

$$\mathbf{u}_k^{l+1} = \mathbf{u}_k^l - \frac{\Delta t}{\mathrm{vol}(\Omega_k)} \sum_{e_{k,k'} \subset \partial\Omega_k} |e_{k,k'}| \mathcal{F}(\mathbf{u}_k, \mathbf{u}_{k'}; \mathbf{n}_{e_{k,k'}}).$$

*is called a conservative scheme for the conservation law* $\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}$. $\mathcal{F}(\mathbf{u}_k, \mathbf{u}_{k'}; \mathbf{n}_{e_{k,k'}})$ *is called numerical flux.*

For the scalar case in one space dimension, a conservative scheme is of the form (see [57])

$$u_i^{l+1} = u_i^l - \frac{\Delta t}{\Delta x}(\mathcal{F}_{i+1/2}^l - \mathcal{F}_{i-1/2}^l),$$

where $\mathcal{F}_{i+1/2}^l = \mathcal{F}_{i+1/2}^l(u_{i-a}^l, \ldots, u_{i+b}^l)$; $a, b$ are two integers, and $a = b$ is allowed.

In the last section we have mentioned monotonicity, and the next definition explains what a monotone scheme is.

**Definition 3.2** (Monotone Scheme)**.**
*A one-dimensional scalar scheme of the form*

$$u_i^{l+1} = H(u_{i-a}^l, \dots, u_{i+b}^l) = \sum_{s=-a}^{k} B_s u_{i+s}^l \tag{3.2.8}$$

*with $B_s \geq 0$ for all $s$ is called monotone. Again $a$ and $b$ are integers, and $a = b$ is possible. For a monotone scheme $H$ has to fulfill*

$$\frac{\partial H}{\partial u_s^l} \geq 0 \qquad \text{for all } s. \tag{3.2.9}$$

This can also be written as [57, Section 12.12]

$$\frac{\partial u_i^{l+1}}{\partial u_s^l} \geq 0,$$

meaning that if the value of any $u_s^l$ at time step $t_l$ is increased, then the value of $u_i^{l+1}$ at the next time step cannot decrease.
A *numerical flux* is called *monotone*, if the resulting scheme with the choice of $\mathcal{F}_{i+1/2}^l = \mathcal{F}_{i+1/2}^l(u_{i-a}^l, \dots, u_{i+b}^l)$ can be cast into the conservative form of definition 3.1. We will give examples of (monotone) fluxes in section 3.2.2.
The theory for multi-dimensional scalar conservation laws is more elaborate; yet, many results and definitions can be formulated for the system case, for instance, by considering each component of the system; see e.g. [30, Ch. 5.3.4], [57, Ch. 20], [59].

### Necessary Conditions for Convergence of a Finite Volume Method

**Definition 3.3.**
*A numerical scheme is called convergent if and only if the numerical solution converges to the exact solution if the grid is refined, i.e. as $\Delta x, \Delta t \to 0$.*

It is well known that a linear method that is consistent and stable converges to the exact solution. This gives the following necessary conditions for convergence, see e.g. [57, Ch. 4.3., Ch. 8]:

1. **Continuity:** The numerical flux shall be Lipschitz continuous.
2. **Conservation property:** It shall hold $\mathcal{F}(\mathbf{u}_k, \mathbf{u}_{k'}; \mathbf{n}) = -\mathcal{F}(\mathbf{u}_{k'}, \mathbf{u}_k; -\mathbf{n})$.
3. **Consistency:** The numerical flux shall be consistent with the physical flux, i.e.

   $$\mathcal{F}(\mathbf{u}, \mathbf{u}; \mathbf{n}) = \mathbf{F}_\mathbf{n}(\mathbf{u}, \mathbf{u}). \tag{3.2.10}$$

   This is motivated from the following: If $\mathbf{u}(x,t)$ is locally constant in $x$, i.e. it is $\mathbf{u}_k^l = \mathbf{u}_{k'}^l \equiv \tilde{\mathbf{u}}_k$ for all $k$ and for all $l$, then the exact solution is reproduced, that is,

   $$\frac{1}{\mathrm{vol}(\Omega_k)} \int_{\Omega_k} \mathbf{u}(\mathbf{x}, t)\, \mathrm{d}\mathbf{x} = \tilde{\mathbf{u}}_k.$$

   Therefore, (3.2.4) shall also hold exactly, which gives (3.2.10).
4. **Stability:** If the *CFL number* fulfills the inequality

   $$\frac{\Delta t}{\Delta x} \max_p |\lambda_p| \leq C,$$

   a finite volume method is stable. Here, $\lambda_p$ is the $p$th eigenvalue of $F'(u)$ (in the one-dimensional scalar case) or of the Jacobian $\frac{\partial \mathbf{F}}{\partial \mathbf{u}}$ (in the multi-dimensional system case); see e.g. [57, Ch. 4.4], [30, Section 5.3.3]. The constant $C$ depends on the chosen stencil. For a three-point stencil, $C = 1$, for a five-point stencil, $C = 2$.

These conditions motivate the necessary properties 3.2 of a numerical flux of a DG method (see section 3.2), which are needed to fulfill the assumptions of an important convergence result for the scalar case, the so-called Lax-Wendroff Theorem.

**Theorem 3.4** (Lax-Wendroff Theorem)**.**
*Let $(\Delta t_i)_i$ and $(\Delta x_i)_i$ be a sequence of grids with $\Delta t_i, \Delta x_i \to 0$ as $i \to \infty$. Let $u_h^{(i)}$ be the approximate solution on grid number $i$, generated by a consistent and conservative method. Furthermore,*

*(i) $u_h^{(i)}$ is uniformly bounded, i.e. $\sup_i ||u_h^{(i)}||_{L^\infty(\mathbb{R}\times\mathbb{R}_+)} \leq C_1$,*

*(ii) $u_h^{(i)}$ converges in $L^\infty_{\text{loc}}(\mathbb{R} \times \mathbb{R}_+)$ and almost everywhere to a function $u$.*

*Then $u$ is a weak solution of the conservation law. 3.8 gives the definition of a weak solution.*

*Proof.* See e.g. [57, Ch. 12.10, Th. 12.1], [59, Prop. 4.1]. □

**Remark 3.5.**

*(i) The second prerequisite can be replaced by the condition that $u$ is BV bounded. See e.g. [57], [58].*

*(ii) A finite volume method with a consistent, monotone flux is first order accurate (see e.g. [59, Lemma 4.1]).*

### 3.2.2 Examples of Monotone Numerical Fluxes

We collect results from references [12], [67], [59], [11], [57], giving a small selection of work on the choice of numerical fluxes.

1. **Lax-Friedrichs flux.**

$$\mathbf{F}_\mathbf{n}^{\text{LF}}(\mathbf{u}_h^-, \mathbf{u}_h^+) = \frac{1}{2}(\mathbf{F}_\mathbf{n}(\mathbf{u}_h^-) + \mathbf{F}_\mathbf{n}(\mathbf{u}_h^+)) + \frac{1}{2}C(\mathbf{u}_h^- - \mathbf{u}_h^+) = \{\{\mathbf{F}_\mathbf{n}\}\} + \frac{1}{2}C[\![\mathbf{u}_h]\!],$$

where $C \geq$ is an upper bound on the biggest absolute eigenvalue of the Jacobian $\frac{\partial \mathbf{F}_\mathbf{n}}{\partial \mathbf{u}}(\mathbf{u}_h)$, and

$$\{\{\mathbf{F}_\mathbf{n}\}\} = \frac{1}{2}(\mathbf{F}_\mathbf{n}(\mathbf{u}_h^-) + \mathbf{F}_\mathbf{n}(\mathbf{u}_h^+)) \quad \text{and} \quad [\![\mathbf{u}_h]\!] = \frac{1}{2}(\mathbf{u}_h^- - \mathbf{u}_h^+)$$

is the mean value and the jump, respectively. For more details on the Lax-Friedrichs flux, see e.g. [11, Ch. 2.3], [67, Ch. 2.2.2, Ch. 3.3.1], [12, Ch. 2.1, Ch. 3.1].

2. **Local Lax-Friedrichs flux.**

$$\mathbf{F}_\mathbf{n}^{\text{LLF}}(\mathbf{u}_h^-, \mathbf{u}_h^+) = \{\{\mathbf{F}_\mathbf{n}\}\} + \frac{1}{2}C[\![\mathbf{u}_h]\!],$$

where now $C = \max_i\{|\lambda_i(\mathbf{u}_h^-)|, |\lambda_i(\mathbf{u}_h^+)|\}$ is the bigger value of the largest eigenvalue of the Jacobian $\frac{\partial \mathbf{F}_\mathbf{n}}{\partial \mathbf{u}}(\mathbf{u}_h^-)$ in element $K$ and the largest eigenvalue of $\frac{\partial \mathbf{F}_\mathbf{n}}{\partial \mathbf{u}}(\mathbf{u}_h^+)$ in the neighboring element $\Omega_{k'}$. See e.g. [67, Ch. 3.3.1].

3. **Upwind schemes.**
   Let $\mathbf{F}$ be linear, $\mathbf{F_n}(\mathbf{u}) = A\mathbf{u}$, where $A$ is a constant matrix. Let us first assume $A$ is diagonal. Then we let $(\mathbf{F}_\mathbf{n}^{\text{num}})_i = \mathbf{F_n}(\mathbf{u}^+)_i$ if $A_{ii} > 0$ and $(\mathbf{F}_\mathbf{n}^{\text{num}})_i = \mathbf{F_n}(\mathbf{u}^-)_i$ if $A_{ii} < 0$. In general, let $T$ be the matrix that diagonalizes $A$, i.e. $A = T\Lambda T^{-1}$. One then defines
   $$\mathbf{F}_\mathbf{n}^{\text{num}}(\mathbf{u}_h^-, \mathbf{u}_h^+) = A\{\{\mathbf{u}_h\}\} + \frac{1}{2}|A|\,[\![\mathbf{u}_h]\!], \tag{3.2.11}$$
   where $|A| := T|\mathbf{\Lambda}|\mathbf{T}^{-1}$, $|\mathbf{\Lambda}| = \text{diag}(|\Lambda_{ii}|)$.
   For instance, if $\nabla \cdot \mathbf{F}(\mathbf{u}) = \partial_1(A_1\mathbf{u}) + \partial_2(A_2\mathbf{u})$, it is $\mathbf{F_n}(\mathbf{u}) = n_1A_1\mathbf{u} + n_2A_2\mathbf{u} =: A\mathbf{u}$. In the nonlinear case one would take an appropriate linearization $\bar{A}$ of $\mathbf{F}'_\mathbf{n}$, and defines
   $$\mathbf{F}_\mathbf{n}^{\text{num}}(\mathbf{u}_h^-, \mathbf{u}_h^+) = \{\{\mathbf{F_n}\}\} + \frac{1}{2}|\bar{A}|\,[\![\mathbf{u}]\!].$$
   For instance, if $\mathbf{F}'_\mathbf{n}$ has only real eigenvalues, one can try $\bar{A} := \{\{\mathbf{F}'_\mathbf{n}\}\}$. Another example would be Roe's linearization, see e.g. [28] for the original work by Roe, or [30] and the references therein. For the upwind scheme mentioned here, see [11, Ch. 2.4, 6.6.2].

4. **Riemann solver (Godunov flux).**
   Godunov's method uses the solution of the Riemann problem to create a numerical flux. The Godunov flux is known to be a monotone flux which produces convergent finite volume schemes of order 1, as was shown by Harten et. al. [64], Kuznetsov [65], Crandall and Majda [66]. See also LeVeque [57].
   Imagine the boundary $\Omega_k \cap \Omega_{k'}$ to be extended to a full plane in $\mathbb{R}^{n-1}$. On the side of element $\Omega_k$, we assume to have the constant value $\mathbf{u}^-$, while on the side of $\Omega_{k'}$ we have the constant value $\mathbf{u}^+$. In case of the DG method, recall that we solve the conservation law locally, and afterwards we need to restore a global solution by using the numerical flux which transports the information from cell to cell. Therefore, the value of $\mathbf{u}_h^k$ at the boundary $\Omega_k \cap \Omega_{k'}$ (which is $\mathbf{u}^-$) can be different from the value of $\mathbf{u}_h^{k'}$ at $\Omega_k \cap \Omega_{k'}$ (i.e. $\mathbf{u}^+$), and the numerical flux needs to connect these different values, thus depending on $\mathbf{u}^-$ and $\mathbf{u}^+$. One now determines the exact solution of the so-called *Riemann problem*, which is defined as
   $$\partial_t\mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \mathbf{0}, \tag{3.2.12}$$
   $$\mathbf{u}(x, t_0) = \begin{cases} \mathbf{u}^- & \text{for } x < x_0, \\ \mathbf{u}^+ & \text{for } x > x_0. \end{cases} \tag{3.2.13}$$
   Without loss of generality (due to translational invariance in $x$-direction) we have assumed to have initially a jump in $x = 0$ and $t = 0$, i.e. $x_0 = 0$, $t_0 = 0$. We denote the solution of (3.2.12) by $\tilde{\mathbf{u}}(x, t)$. Later we will see that $\tilde{\mathbf{u}}(x, t) = \tilde{\mathbf{u}}\left(\frac{x}{t}; \mathbf{u}^-, \mathbf{u}^+\right)$. The numerical flux is then defined as
   $$\mathbf{F}_\mathbf{n}^{\text{num}}(\mathbf{u}_h^-, \mathbf{u}_h^+) := \mathbf{F_n}(\tilde{\mathbf{u}}(0; \mathbf{u}^-, \mathbf{u}^+), \tag{3.2.14}$$
   i.e. one evaluates the Riemann solution at $\frac{x}{t} = 0$. This flux is called *Godunov flux*.
   It can be shown that in the linear case, i.e. $\mathbf{F}(\mathbf{u}) = A\mathbf{u}$, Godunov's flux is the upwind flux in (3.2.11), see e.g. [30, Ch. 5.4.2], [59, Ch. 2.1]. More about Godunov's method can be found in e.g. [59, Ch. 2.1], [57], [30].

In the next section we will give more details about the Godunov flux which needs the solution of the Riemann problem (3.2.12). We will show how to solve a Riemann problem and how the solution looks like. We will look at the linear and nonlinear case.

## 3.3 The Numerical Flux and the Riemann Problem

In the last section we have given the Godunov flux (3.2.14), where $\tilde{\mathbf{u}}(\frac{x}{t}; \mathbf{u}^-, \mathbf{u}^+)$ is the solution of the Riemann problem (3.2.12). For the DG scheme we choose the numerical flux in the same manner. As seen in the last section and section 3.2.1, the motivation for this choice comes from finite volume methods: If we choose piecewise constant polynomials in the DG scheme with a numerical flux that solves a corresponding Riemann problem we obtain a finite volume method that is stable and convergent of order 1.
In this section we will show how the Riemann problem can be solved in order to get the numerical flux for a DG method. We will also answer the question under which circumstances this solution is unique. We will only collect the facts and results that are important for our purposes. For all the details see e.g. [57], [30], [58], [59].

### 3.3.1 The Riemann Problem for Linear Hyperbolic Systems

Let us consider (3.2.12) for $\nabla = (\partial_x, 0, 0)$ and $\mathbf{F}(\mathbf{u}) = A\mathbf{u}$, where $\mathbf{u} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$, supplemented with initial data

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_L & \text{for } x < 0, \\ \mathbf{u}_R & \text{for } x > 0. \end{cases}$$

Since the system shall be hyperbolic, we have $n$ eigenvalues $\lambda_1 < \cdots < \lambda_n$ with $n$ linearly independent right eigenvectors $\mathbf{r}_k$ such that $A\mathbf{r}_k = \lambda_k \mathbf{r}_k$ for $k = 1, \ldots, n$. If we transform to characteristic variables by letting $\mathbf{v} = R^{-1}\mathbf{u}$, where $R$ is the matrix containing the right eigenvectors $\mathbf{r}_k$ as columns, then the conservation law decouples, and we need to solve $p$ advection problems

$$\frac{\partial v_k}{\partial t} + \lambda_k \frac{\partial v_k}{\partial x} = 0 \tag{3.3.1}$$

with initial data $\mathbf{v}_0(x) := \mathbf{v}(x,0) = R^{-1}\mathbf{u}(x,0)$. Note this is possible since the problem is linear and thus $R$ is constant. The solution can be found using the method of characteristics (see e.g. [59], [57], [30]) and it is given as

$$v_k(x,t) = v_{0,k}(x - \lambda_k t).$$

Thus, transforming back we obtain

$$\mathbf{u}(x,t) = R\mathbf{v}(x,t) = R(v_{0,k}(x - \lambda_k t))_{k=1}^n \quad \text{for} \quad k = 1, \ldots, n.$$

Since we have $n$ linearly independent right eigenvectors, we can decompose the solution $\mathbf{u}$ as

$$\mathbf{u} = \sum_{k=1}^{n} \alpha_k \mathbf{r}_k,$$

and the same can be done for $\mathbf{u}_L$ and $\mathbf{u}_R$:

$$\mathbf{u}_L = \sum_{k=1}^{n} \alpha_{kL} \mathbf{r}_k, \quad \mathbf{u}_R = \sum_{k=1}^{n} \alpha_{kR} \mathbf{r}_k.$$

Thus the $k$th advection equation (3.3.1) has initial data

$$\mathbf{v}(x,0) = \begin{cases} \alpha_{kL} & \text{for } x < 0, \\ \alpha_{kR} & \text{for } x > 0 \end{cases}$$

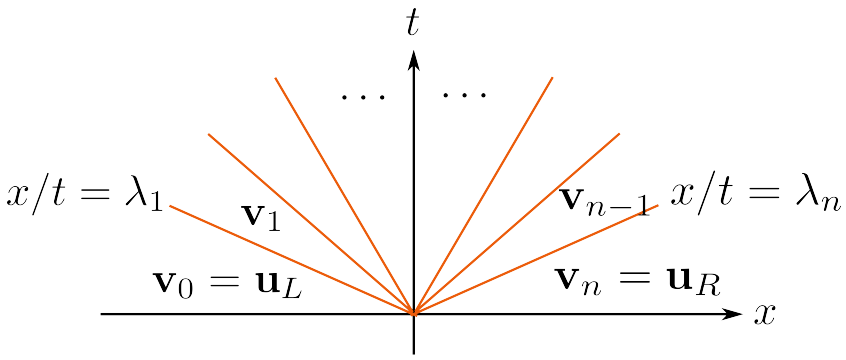



Figure 3.3: Sketch of the wave composition in $x-t-$plane of a solution of the Riemann problem.

and the solution is given as

$$\mathbf{v}_k(x,t) = \begin{cases} \alpha_{kL} & \text{for } x - \lambda_k t < 0 \;\Leftrightarrow\; \frac{x}{t} < \lambda_k, \\ \alpha_{kR} & \text{for } x + \lambda_k t > 0 \;\Leftrightarrow\; \frac{x}{t} > \lambda_k. \end{cases}$$

As a consequence the solution of the Riemann problem (3.2.12) is found to be self-similar, that means,

$$\mathbf{u}(x,t) = \tilde{\mathbf{u}}(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R).$$

It consists of $n$ waves with characteristic speeds $\lambda_k$ $(k = 1, \ldots, n)$, as shown in figure 3.3.

### 3.3.2 The Riemann Problem for Nonlinear Hyperbolic Systems

Again we consider the Riemann problem (3.2.12), but now $\mathbf{F}(\mathbf{u})$ shall be nonlinear in $\mathbf{u}$. Locally, the wave composition looks like in the linear case; yet, the eigenvalues and eigenvectors now may vary with $\mathbf{u}$, and so globally, a completely different picture can form, the characteristics being curves not straight lines.
Characteristic fields can be classified into two basic types: There exist genuinely nonlinear and linearly degenerated characteristic fields.

**Definition 3.6.**
*If the kth characteristic field fulfills*

$$\nabla\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \neq 0 \quad \forall \mathbf{u} \in \mathbb{R}^n, \tag{3.3.2}$$

*it is called genuinely nonlinear. If*

$$\nabla\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0 \quad \forall \mathbf{u} \in \mathbb{R}^n, \tag{3.3.3}$$

*the kth characteristic field is called linearly degenerate.*

If a $k$th characteristic field is genuinely nonlinear, it shall be normalized such that

$$\nabla\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 1. \tag{3.3.4}$$

For both cases the left eigenvectors $\mathbf{l}_k$ and right eigenvectors $\mathbf{r}_k$ shall also be normalized to

$$\mathbf{l}_k^T(\mathbf{u})\mathbf{r}_k(\mathbf{u}) = 1. \tag{3.3.5}$$

Let us illustrate what the solution of a nonlinear Riemann problem may look like by studying the (inviscid) Burger's equation.

**Introduction: Burger's Equation**

(Inviscid) Burger's equation reads

$$\partial_t u + \partial_x \left( \frac{u^2}{2} \right) = 0.$$

We choose the initial data

$$u(x,0) = \begin{cases} 1 & \text{for } x < 0, \\ 0 & \text{for } x > 0. \end{cases}$$

We assume to have scalar $u$, and $F(u) = u^2/2$. The eigenvalue is $\lambda(u) = F'(u) = u$, thus the characteristics are given by

$$x(\xi) = u_0(\xi)t + \xi,$$

where $\xi$ is a parameter. Thus,

$$x(\xi,t) = \begin{cases} \xi + t & \text{for } \xi < 0, \\ \xi & \text{for } \xi > 0. \end{cases}$$

Figure 3.4 illustrates the situation. We define the line left of the line $t = 0$ as $x_L := \xi_- + t$ with $\xi_- \in \{\xi : \quad \xi < 0\}$, and analogously the line right of $t = 0$ as $x_R := \xi_+ + t$, where $\xi_+ \in \{\xi : \quad \xi > 0\}$. We see that the characteristics cross – this is exactly the case if the lines $x_L$ and $x_R$ intersect, namely if

$$-\xi + t = \xi \quad \text{for} \quad \xi > 0,$$

since $\xi_- = -\xi_+$. Solving for $\xi$ gives

$$\xi = \frac{1}{2}t \quad (\xi > 0),$$

i.e. the set of intersection points $\xi_+$ is a half line with origin in zero and slope $\frac{1}{2}$. This line is called a *shock*; see figure 3.4 (right) for a visualization. For growing time the solution steepens more and more, until its slope becomes infinity. This is the point where a shock forms. In figure 3.5 we see two examples of this steepening effect: once, where the initial data is given as $u(x,0) = 1$ for $x < 0$ and $u(x,0) = 0$ for $x > 0$; and when the initial data is a hat function.
The speed of such a shock is determined by the Rankine-Hugoniot jump condition, which is given in the following definition.

**Definition 3.7** (Rankine-Hugoniot Jump Condition)**.**
*If* $\mathbf{u}$ *is a shock solution of the conservation law* $\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}$*, then* $\mathbf{u}$ *fulfills the Rankine-Hugoniot jump condition*

$$\mathbf{F}(\mathbf{u}_R) - \mathbf{F}(\mathbf{u}_L) = s(\mathbf{u}_R - \mathbf{u}_L), \tag{3.3.6}$$

*where* $s \in \mathbb{R}$ *is the shock speed.*

Of course, if the solution $\mathbf{u}$ of a conservation law has a discontinuity, it cannot be a classical solution. Therefore, we need to take weak solutions into account.

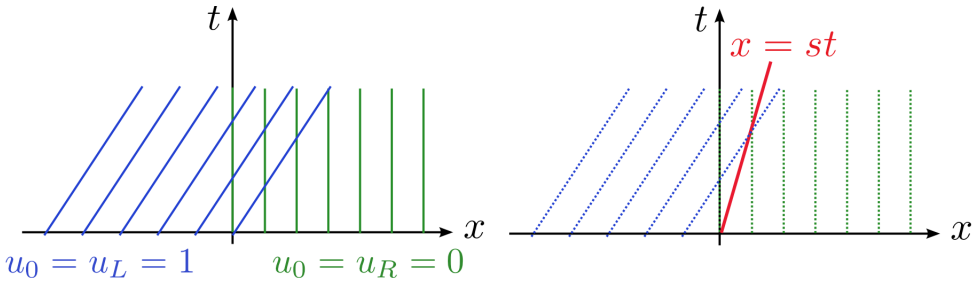



Figure 3.4: Left: Characteristics for initial data $u(x,0) = 1$ for $x < 0$, $u(x,0) = 0$ for $x > 0$. The lines cross for $x = \frac{1}{2}t$, and a shock forms (red line on the right).

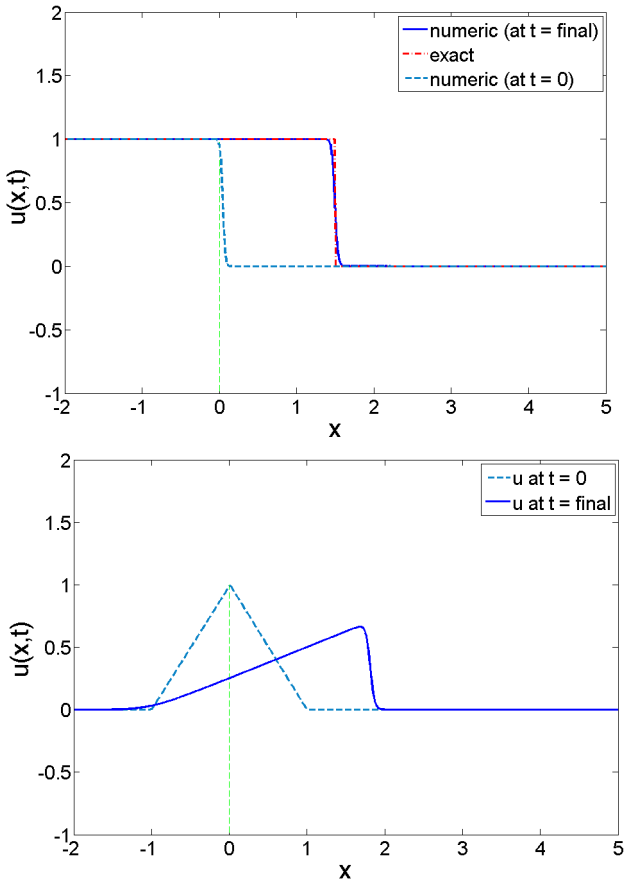



Figure 3.5: Examples of the formation of a shock for Burger's equation, where the initial data has a jump (left) and is a hat function (right). The codes for generating the plots were provided by Willy Dörfler.

**Definition 3.8** (Weak Solution)**.**
*Let $\Omega \subset \mathbb{R}^n$ be open. Consider the conservation law $\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}$ with initial data $\mathbf{u}_0 \in L^\infty_{\text{loc}}(\mathbb{R}^m)^n$. $\mathbf{u} \in L^\infty_{\text{loc}}(\mathbb{R}^m \times \mathbb{R})^n$ is called a weak solution of the conservation law if and only if it fulfills*

$$\int_{\mathbb{R}} \int_{\mathbb{R}^m} \mathbf{u} \cdot \frac{\partial \boldsymbol{\psi}}{\partial t} + \sum_{j=1}^{m} \mathbf{F}_j(\mathbf{u}) \cdot \frac{\partial \boldsymbol{\psi}}{\partial x_j} \, \mathrm{d}\mathbf{x} \, \mathrm{d}t + \int_{\mathbb{R}^m} \mathbf{u}_0(\mathbf{x}) \cdot \boldsymbol{\psi}(\mathbf{x}, 0) \, \mathrm{d}\mathbf{x} = 0 \tag{3.3.7}$$

*for all test functions $\boldsymbol{\psi} \in C^1_0(\mathbb{R}^m \times \mathbb{R})^n$ with compact support.*

If we talk about weak solutions we have also to think about uniqueness of solutions. And

indeed, there can be several weak solutions of a Riemann problem. Consider for instance Burger's equation with the following initial data:

$$\partial_t u + \partial_x \left( \frac{u^2}{2} \right) = 0, \tag{3.3.8}$$

$$u(x,0) = \begin{cases} 0 & \text{for } x < 0, \\ 1 & \text{for } x > 0, \end{cases} \tag{3.3.9}$$

i.e. $u_L = 0 < u_R = 1$. In this case the characteristics do not intersect, see figure 3.6. We cannot determine a solution via the method of characteristics in the region
$u_L t = 0 \leq x \leq t = u_R t$.

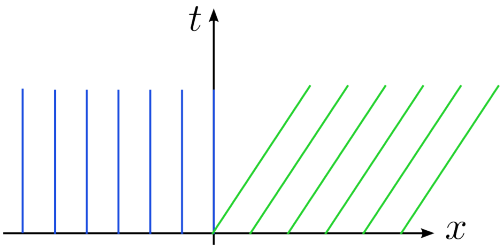



Figure 3.6: Characteristics for Burger's equation for initial data $u_0(x,0) = 0$ for $x < 0$ and $u_0(x,0) = 1$ for $x > 0$.

In this situation, there are several solutions. In fact, for $u_L \neq u_R$ there is an entire family of solutions, see [59, Ch. 2.3]. In our case, the function

$$u(x,t) = \begin{cases} 0 & \text{for } x < \frac{1}{2}t, \\ 1 & \text{for } x > \frac{1}{2}t \end{cases}$$

with the shock speed

$$s = \frac{F(u_L) - F(u_R)}{u_L - u_R} = \frac{0 - 1/2}{0 - 1} = \frac{1}{2},$$

where we used the Rankine-Hugoniot jump condition (3.3.6), is a weak solution of (3.3.8). But there is also another solution, namely

$$u(x,t) = \begin{cases} 0 & \text{for } \frac{x}{t} \leq 0 = u_L, \\ \frac{x}{t} & \text{for } u_L = 0 \leq \frac{x}{t} \leq 1 = u_R, \\ 1 & \text{for } \frac{x}{t} \geq 1 = u_R, \end{cases}$$

which is a continuous solution of (3.3.8). This is due to the fact that any function $v(x,t) = \frac{x}{t}$ for $t > 0$ is a solution of Burger's equation, since it holds:

$$\partial_t v + \partial_x \frac{v^2}{2} = \frac{-x}{t^2} + \frac{x}{t} \cdot \frac{1}{t} = 0.$$

A function of the form $u(x,t) = v(\frac{x}{t})$ is called self-similar. The solution $u(x,t) = \frac{x}{t}$ in the region $F'(u_L) = u_L = 0 \leq \frac{x}{t} \leq 1 = u_R = F'(u_R)$ is called a *rarefaction fan*; its so-called *head* is given by $\frac{x}{t} = F'(u_L) = u_L$, and its *tail* by $\frac{x}{t} = F'(u_R) = u_R$. Inside the rarefaction fan, the solution $u$ is continuous, that is, we have a continuous transition form left to right. See figure 3.7 for a sketch of this situation.
So we see there are at least two solutions of Burger's equation with initial data $u_0(x,0) = 0$ for $x < 0$ and $u_0(x,0) = 1$ for $x > 0$. In applications, as in e.g. physics, a solution has

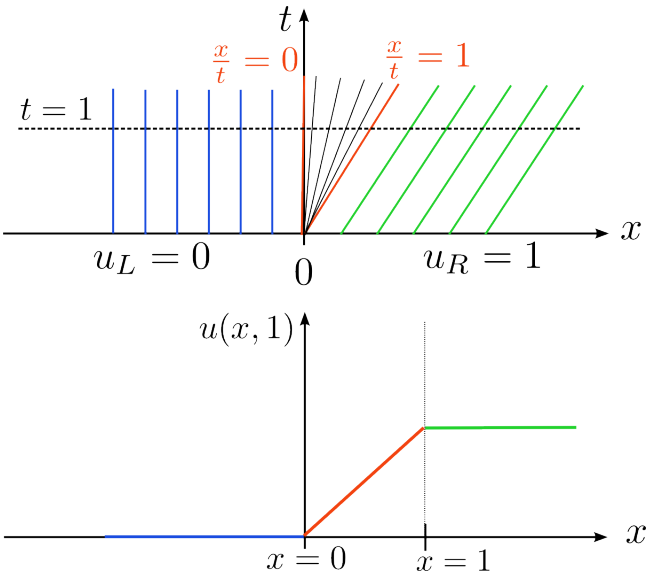



Figure 3.7: Top: A rarefaction fan for Burger's equation in $x-t$-plane with initial data $u_0(x,0) = 0$ for $x < 0$ and $u_0(x,0) = 1$ for $x > 0$. Bottom: Corresponding solution $u(x,1)$ at the chosen time $t = 1$. Between $x = 0$ and $x = 1$, $u(x,1)$ is continuous.

to be unique. So the question is if there is some criterion to ensure uniqueness and with which to choose the physically relevant solutions. This leads to so-called *entropy solutions* that fulfill some kind of entropy condition, see section 3.3.2 of this work. For details and corresponding theory we refer to e.g. [59], [57], [58]. One can show in the scalar case that, if $u_0 \in L^1(\mathbb{R}^m) \cap L^\infty(\mathbb{R}^m)$, then the conservation law has a unique entropy solution $u \in L^\infty(\mathbb{R}^m \times (0,T))$; see e.g. [59, Ch. 3.3] for a proof.

In the following sections we take a closer look at the first notions we have encountered for Burger's equation: shocks, rarefaction waves and entropy condition, amongst others. The references we use are [59], [57], [58].

### Rarefaction Waves

We are looking for piecewise smooth solutions of the Riemann problem

$$\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}, \tag{3.3.10}$$

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_L & \text{for } x < 0, \\ \mathbf{u}_R & \text{for } x > 0. \end{cases} \tag{3.3.11}$$

In the last section 3.3.1 we have seen that in the linear case, the solution $\mathbf{u}$ is self-similar. Thus, for the nonlinear case, we make the solution ansatz

$$\mathbf{u}(x,t) = \mathbf{w}\left(\frac{x}{t}\right), \tag{3.3.12}$$

where we assume that $\mathbf{w}$ is smooth. That means we make a classical ansatz. Therefore it holds

$$\partial_x \mathbf{F}(\mathbf{u}) = \frac{\partial \mathbf{F}(\mathbf{u})}{\partial \mathbf{u}} \partial_x \mathbf{u} =: A(\mathbf{u}) \partial_x \mathbf{u}.$$

Note that for discontinuous solutions we need to work in the weak sense, and in this case this equivalence does not hold, as can be seen via studying Burger's equation, which is one

of the simplest nonlinear equations. See e.g. the book by LeVeque [57, Ch. 11] for more details.
If we plug the ansatz (3.3.12) into the conservation law (in the classical sense)

$$\partial_t \mathbf{u} + A(\mathbf{u})\partial_x \mathbf{u} = \mathbf{0},$$

then we obtain

$$\partial_t \mathbf{u} = -\frac{x}{t^2}\mathbf{v}'\left(\frac{x}{t}\right),$$
$$\partial_x \mathbf{u} = \frac{1}{t}\mathbf{v}'\left(\frac{x}{t}\right),$$

and by letting $\xi := \frac{x}{t}$, we thus have

$$-\xi\mathbf{v}'(\xi) + A(\mathbf{v}(\xi))v'(\xi) = \mathbf{0} \qquad \Longleftrightarrow \qquad (A(\mathbf{v}(\xi)) - \xi\,\mathrm{Id})v'(\xi) = \mathbf{0}.$$

Hence, either $\mathbf{v}'(\xi)$ is the $\mathbf{0}$-vector, or it is an eigenvector of $A(\mathbf{v}(\xi))$ with eigenvalue $\xi$, i.e. there exists an index $k \in \{1, \dots, n\}$ such that

$$\mathbf{v}'(\xi) = \alpha(\xi)\mathbf{r}_k(\mathbf{v}(\xi)), \quad \lambda_k(\mathbf{v}(\xi)) = \xi. \tag{3.3.13}$$

Differentiating $\lambda_k(\mathbf{v}(\xi))$ with respect to $\xi$ gives

$$\nabla\lambda_k(\mathbf{v}(\xi)) \cdot \mathbf{v}'(\xi) = 1,$$

and inserting the first equation of (3.3.13), we obtain:

$$\alpha(\xi)\nabla\lambda_k(\mathbf{v}(\xi)) \cdot \mathbf{r}_k(\mathbf{v}(\xi)) = 1,$$

which can only be true if the $k$th characteristic field is genuinely nonlinear, see equation (3.3.2). If we assume the normalization (3.3.4), then $\alpha(\xi) = 1$. So we either have $\mathbf{v}'(\xi) = \mathbf{0}$ or $\mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi))$ with eigenvalue $\lambda_k(\mathbf{v}(\xi)) = \xi$. A function with these properties is called an integral curve.

**Definition 3.9** (Integral curve of a hyperbolic equation)**.**
*Let $\tilde{\mathbf{u}} : I \to \mathbb{R}^n$, $\xi \mapsto \tilde{\mathbf{u}}(\xi)$, where $I \subset \mathbb{R}$, be a smooth curve in state space. Let $\mathbf{r}_i$ be a vector field. $\tilde{\mathbf{u}}$ is called an integral curve of $\mathbf{r}_i$ :$\Leftrightarrow$ in every point $\tilde{\mathbf{u}}(\xi)$ the tangential vector $\tilde{\mathbf{u}}'(\xi)$ is an eigenvector of the Jacobi matrix $\partial\mathbf{F}/\partial\mathbf{u}(\tilde{\mathbf{u}}(\xi))$ with eigenvalue $\lambda_i(\tilde{\mathbf{u}}(\xi))$.*

Thus, having a certain set of eigenvectors $\mathbf{r}_i(\mathbf{u})$, the curve $\tilde{\mathbf{u}}(\xi)$ is an integral curve only if its tangential vector is a multiple of the eigenvector $\mathbf{r}_i(\tilde{\mathbf{u}}(\xi))$, i.e.

$$\tilde{\mathbf{u}}'(\xi) = \alpha(\xi)\mathbf{r}_i(\tilde{\mathbf{u}}(\xi)). \tag{3.3.14}$$

That is, $\tilde{\mathbf{u}}'(\xi)$ always has the same direction as $\mathbf{r}_i(\tilde{\mathbf{u}}(\xi))$. We see that the function $\mathbf{v}$ is an integral curve of the vector field $\mathbf{r}_k(\mathbf{v}(\xi))$, and $\alpha(\xi) = 1$. Hence, if we assume that the $k$th characteristic field $\mathbf{r}_k$ is genuinely nonlinear, and if we further assume that $\mathbf{u}_L$ and $\mathbf{u}_R$ lie on the same integral curve, i.e. $\mathbf{v}(\lambda_k(\mathbf{u}_L)) = \mathbf{u}_L$ and $\mathbf{v}(\lambda_k(\mathbf{u}_R)) = \mathbf{u}_R$, where $\lambda_k$ shall increase from $\mathbf{u}_L$ to $\mathbf{u}_R$, then the continuous self-similar solution of the Riemann problem (3.3.10) with the ansatz (3.3.12) looks as follows:

$$\mathbf{u}(x,t) = \begin{cases} \mathbf{u}^L & \text{for } \frac{x}{t} \le \lambda_k(\mathbf{u}_L), \\ \mathbf{v}(\frac{x}{t}) & \text{for } \lambda_k(\mathbf{u}_L) \le \frac{x}{t} \le \lambda_k(\mathbf{u}_R), \\ \mathbf{u}^R & \text{for } \frac{x}{t} \ge \lambda_k(\mathbf{u}_L). \end{cases} \tag{3.3.15}$$

This solution is called a *$k$-rarefaction wave*, and it connects the states $\mathbf{u}_L$ and $\mathbf{u}_R$ in a continuous manner. This definition also holds in the weak sense.

**Definition 3.10.**
*A self-similar weak solution of the form* (3.3.15) *is called a k-rarefaction wave.*

The following theorem summarizes the results found up until now.

**Theorem 3.11.**

*(i) If the kth characteristic field is genuinely nonlinear, then a given state* $\mathbf{u}_L$ *can be connected to a right state* $\mathbf{u}_R$ *by a k-rarefaction wave.*

*(ii) The characteristic curves of a k-rarefaction wave are straight lines along which the solution* $\mathbf{u}$ *is constant.*

*Proof.* See [59, Ch. 3.1, Th. 3.1 and Ch. 5, Th. 5.1]. □

Figure 3.8 shall illustrate a rarefaction wave.

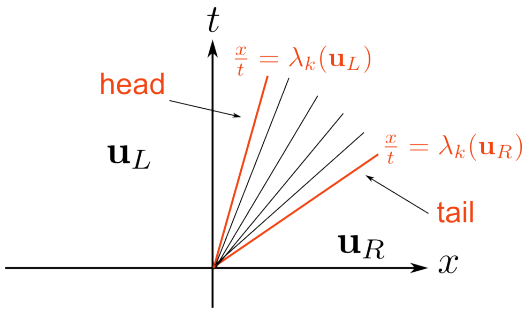



Figure 3.8: Sketch of a rarefaction fan in $x-t$-plane. If $\lambda_k(\mathbf{u}_{R,L})$ is positive, $\frac{x}{t} = \lambda_k(\mathbf{u}_L)$ is the head of the rarefaction and $\frac{x}{t} = \lambda_k(\mathbf{u}_R)$ is the tail. If $\lambda_k(\mathbf{u}_{R,L})$ is negative, it is the other way around.

### Riemann Invariants

Rarefaction waves are a special type of elementary waves associated with a certain characteristic family. Riemann invariants are another type of elementary waves.

**Definition 3.12** (Riemann Invariant)**.**
*A smooth function* $h = h(\mathbf{u})$ *is called a k-Riemann invariant if it fulfills*

$$\nabla_{\mathbf{u}} h(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0. \tag{3.3.16}$$

*That is, if* $h(\mathbf{u})$ *is a k-Riemann invariant, it is constant along any curve* $\mathbf{v} : \mathbb{R} \to \mathbb{R}^n \Leftrightarrow$

$$\frac{d}{d\xi} h(\mathbf{v}(\xi)) = 0. \tag{3.3.17}$$

Note that

$$\frac{d}{d\xi} h(\mathbf{v}(\xi)) = \nabla_{\mathbf{v}} h(\mathbf{v}(\xi)) \cdot \mathbf{v}'(\xi) = 0.$$

This holds if $\mathbf{v}$ is an integral curve, i.e. if $\mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi))$. So a $k$-Riemann invariant is constant along integral curves. Also recall the definition of linearly degenerated characteristic fields (3.3.3). We see that if the $k$-characteristic field is linearly degenerate, then $\lambda_k$ is a $k$-Riemann invariant. Locally there exist $p-1$ Riemann invariants corresponding to $\lambda_k$. Furthermore, all $k$-Riemann invariants are constant on a $k$-rarefaction wave. The proofs can be found in [59, Ch. 3.2].

**Definition 3.13.**
*Let $\Omega \subset \mathbb{R}^2$ and let $\mathbf{u} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^n$ be a smooth solution of the conservation law $\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}$ on $\Omega$. We call $\mathbf{u}$ a k-simple wave if $h(\mathbf{u})$ is constant in $\Omega$ for any k-Riemann invariant $h$.*

**Theorem 3.14.**
*If $\mathbf{u}$ is a k-simple wave, the characteristics of the kth field are straight lines, and $\mathbf{u}$ is constant along those characteristics.*

*Proof.* See [59], Theorem 3.3 in Ch. 3.2. □

**Remark 3.15.**

*(i) The simple waves of a genuinely nonlinear field are (centered) rarefaction waves. Their characteristics form a so-called rarefaction fan which is bounded by the head $\frac{x}{t} = \lambda_k(\mathbf{u}_L)$ and the tail $\frac{x}{t} = \lambda_k(\mathbf{u}_R)$ (if $\lambda_k(\mathbf{u}_{R,L})$ is positive; if it is negative, it is the other way around). See figure 3.8.*

*(ii) Recall that if the k-characteristic field is linearly degenerate, then $\lambda_k$ is a k-Riemann invariant, and thus, $\lambda_k$ is constant on any k-simple wave $\mathbf{u}$. That means, $\lambda_k(\mathbf{u}) = \lambda_k(\mathbf{u}^*) =: \lambda_k^*$ for an arbitrary point $\mathbf{u}^* = \mathbf{u}(x^*, t^*)$. Specifically, $\lambda_k^* = \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}_R)$. The characteristics are now parallel lines $x - \lambda_k^* t$, and $\mathbf{u} = \mathbf{u}(x - \lambda_k^* t)$. (For more details, see e.g. [59, Ex. 3.5].) Thus, the two states $\mathbf{u}_L$ and $\mathbf{u}_R$ cannot be connected by a continuous k-wave. The connecting wave is discontinuous. This issue will be addressed in the next section.*

Now choose $h(\mathbf{u}) = w_i(\mathbf{u})$ to be the $i$-th Riemann invariant. Then using (3.3.17) we obtain

$$\nabla_u w_i(\mathbf{u}(\xi)) \cdot \mathbf{u}'(\xi) = 0.$$

Using (3.3.14) and $\alpha(\xi) = 1$ (due to normalization) we see

$$\nabla_u w_i(\mathbf{u}(\xi)) \cdot \mathbf{v}_i(\mathbf{u}(\xi)) = 0$$

for all $\mathbf{u}(\xi)$ with varying $\xi$. From this follows

$$\nabla_u w_i \cdot \mathbf{v}_i = 0. \tag{3.3.18}$$

Thus a Riemann invariant is a function that is orthogonal to the eigenvector $\mathbf{v}_i$ in every point $\mathbf{u}$. One can use (3.3.14) and (3.3.18) to determine the $p-1$ Riemann invariants of a hyperbolic system. (3.3.14) is a system of ordinary differential equations,

$$\frac{d\mathbf{u}}{d\xi} = \mathbf{v}_i(\mathbf{u}(\xi)),$$

with $\alpha(\xi) = 1$, $\mathbf{u} =: (u_1, \ldots, u_n)^T$ and $\mathbf{v}_i := (v_i^{(1)}, \ldots, v_i^{(n)})^T$. This gives after integration the Riemann invariants

$$R_i(\mathbf{u}) = \int_\Omega \mathrm{d}\mathbf{u} - \int_I \mathbf{v}_i(\mathbf{u}(\xi))\,\mathrm{d}\xi \equiv \text{const}\,. \tag{3.3.19}$$

Evaluating $R_i(\mathbf{u})$ at any point $\mathbf{u}^*$ gives the constant. These Riemann invariants fulfill

$$\nabla_{\mathbf{u}} R_i \cdot \mathbf{v}_i = 0. \tag{3.3.20}$$

Here, by the product "$\cdot$" the following is meant: If $\nabla_u R_i$ is the Jacobian of the vector $R_i$, i.e.

$$\nabla_{\mathbf{u}} R_i = \begin{pmatrix} \frac{\partial R_{i,1}}{\partial w_1} & \dots & \frac{\partial R_{i,1}}{\partial w_n} \\ \vdots & & \vdots \\ \frac{\partial R_{i,n}}{\partial w_1} & \dots & \frac{\partial R_{i,m}}{\partial w_n} \end{pmatrix}, \tag{3.3.21}$$

then "$\cdot$" denotes the matrix-vector product, i.e.

$$\nabla_{\mathbf{u}} R_i \cdot \mathbf{v}_i = \mathbf{0}, \quad \mathbf{0} \in \mathbb{R}^n.$$

That is, the product of a row of the Jacobian, $\left(\frac{\partial R_{i,j}}{\partial w_1}, \dots, \frac{\partial R_{i,j}}{\partial w_n}\right)$ $(j = 1, \dots, n)$, which is a vector in $\mathbb{R}^{1\times n}$, with the eigenvector $\mathbf{v}_i \in \mathbb{R}^n$ has to be 0. Since the vector spaces $\mathbb{R}^n$ and $\mathbb{R}^{1\times n}$ are isomorphic, one could interpret this also as the standard scalar product between a row of $\nabla_{\mathbf{u}} R_i$ and the eigenvector $\mathbf{v}_i$, that is

$$\left(\frac{\partial R_{i,j}}{\partial w_1}, \dots, \frac{\partial R_{i,j}}{\partial w_n}\right)^T \cdot \mathbf{v}_i = 0.$$

### Shocks and Contact Discontinuities

In remark 3.15 we have mentioned the possibility of having discontinuous simple waves connecting a left and a right state. We will see that we have basically two discontinuous simple waves: a shock and a contact discontinuity.

**Definition 3.16.**
*A function*

$$\mathbf{u}(x,t) = \begin{cases} \mathbf{u}^L & \text{for } x < st, \\ \mathbf{u}^R & \text{for } x > st, \end{cases} \tag{3.3.22}$$

*where $s \in \mathbb{R}$ is the shock speed (see* (3.3.6)*), fulfilling the Rankine-Hugoniot jump condition*

$$\mathbf{F}(\mathbf{u}_R) - \mathbf{F}(\mathbf{u}_L) = s(\mathbf{u}_R - \mathbf{u}_L), \tag{3.3.23}$$

*is a weak solution of the conservation law $\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}$. It is called a discontinuity wave. The shock speed $s$ is the speed of propagation of this discontinuity.*

As in the case of rarefaction waves, we are interested in knowing how two states $\mathbf{u}_L$ and $\mathbf{u}_R$ can be connected by a discontinuous waves. This is answered in the following definitions and results, taken from [59] and [57].

**Definition 3.17** (Hugoniot Locus)**.**
*Let $\mathbf{u} \in \mathbb{R}^n$. The set*

$$\mathcal{HL} := \{\mathbf{u} : \ s(\mathbf{u}^*, \mathbf{u})(\mathbf{u} - \mathbf{u}^*) = \mathbf{F_n}(\mathbf{u}) - \mathbf{F_n}(\mathbf{u}^*)\}, \tag{3.3.24}$$

*where $s(\mathbf{u}^*, \mathbf{u}) \in \mathbb{R}$ is the shock speed from definition 3.3.22 in dependence of $\mathbf{u}^*$ and $\mathbf{u}$, is called Hugoniot Locus. A Hugoniot Locus gives the set of all points $\mathbf{u}^*$ that can be connected to an arbitrary point $\mathbf{u}$ by a discontinuity.*

**Example 3.18** (Hugoniot Locus for a two-dimensional system)**.**
*We consider the linear hyperbolic conservation law*

$$\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0},$$
$$\mathbf{u}_0 = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0 \end{cases}$$

*with* $\mathbf{u} = (u_1, u_2)^T$ *and* $\mathbf{F}(\mathbf{u}) = A\mathbf{u}$*, where* $A \in \mathbb{R}^{2,2}$*. Let* $\lambda_1$ *and* $\lambda_2$ *be the eigenvalues of* $A$ *with corresponding eigenvectors* $v_1, v_2$*. We can thus decompose* $\mathbf{u}_L$ *and* $\mathbf{u}_R$ *as*

$$\mathbf{u}_L = a_1\mathbf{v}_1 + a_2\mathbf{v}_2,$$
$$\mathbf{u}_R = b_1\mathbf{v}_1 + b_2\mathbf{v}_2$$

*with some constants* $a_i, b_i$*. For linear problems, the Hugoniot Locus is a family of straight lines with the direction of* $\mathbf{v}_1$ *(in red) and of* $\mathbf{v}_2$ *(in green) as illustrated in figure 3.9.* $\mathbf{u}_L$ *is the intersection of a line* $a_1\mathbf{v}_1$ *for a certain choice of* $a_1$ *(thick red lines) and of a line* $a_2\mathbf{v}_2$ *for a certain value of* $a_2$ *(thick green lines); the same holds for* $\mathbf{u}_R$ *for special choices of* $b_1, b_2$*. The Hugoniot Locus contains all values of* $\mathbf{u}^* = \mathbf{u}_L$ *so that the right region* $\mathbf{u}_R$ *can be reached by a discontinuity; or equivalently, it consists of all* $\mathbf{u}^* = \mathbf{u}_R$ *so that* $\mathbf{u}_R$ *and* $\mathbf{u}_L$ *are connected by a discontinuity. Thus the question is how to reach the point* $\mathbf{u}_R$ *by a shock, starting in* $\mathbf{u}_L$ *and going via* $\mathbf{u}_M$*.*

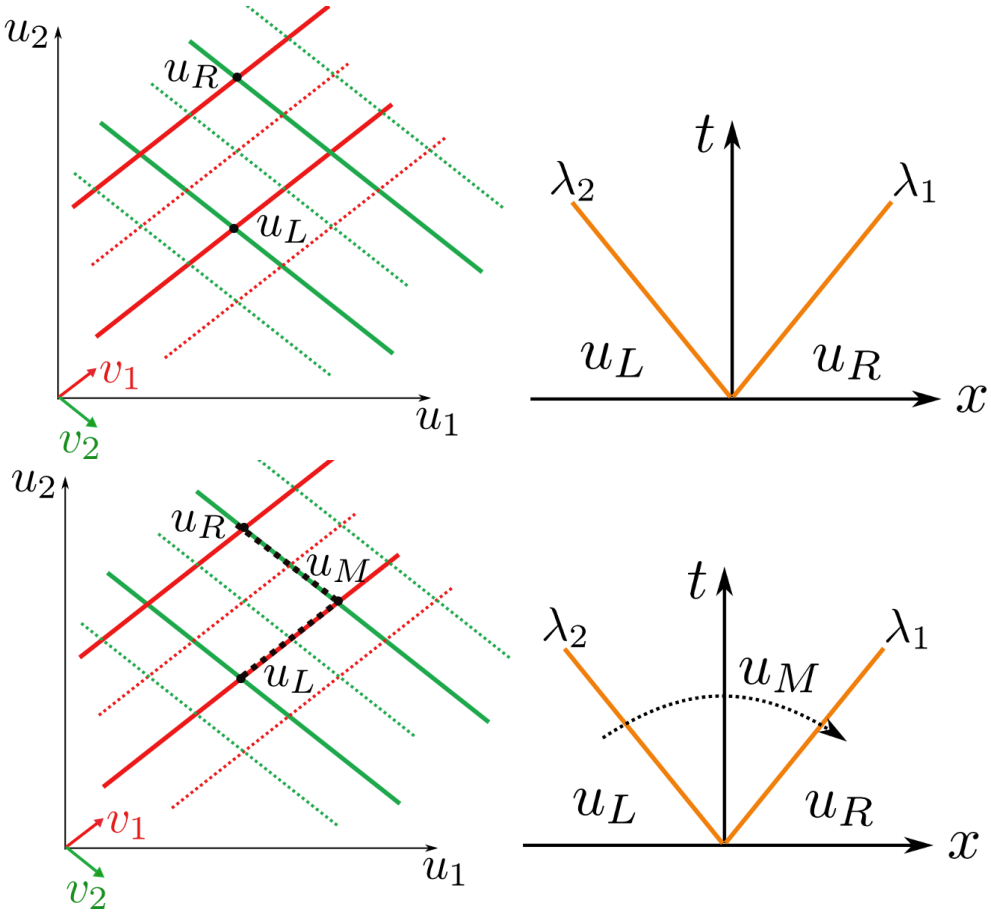



Figure 3.9: A family of straight lines with the direction of $\mathbf{v}_1$ (red) and of $\mathbf{v}_2$ (green). $\mathbf{u}_L$ is the intersection of a line $a_1\mathbf{v}_1$ for a certain choice of $a_1$ (thick red lines) and of a line $a_2\mathbf{v}_2$ for a certain value of $a_2$ (thick green lines); likewise $\mathbf{u}_R$ is the intersection point of the lines $b_1\mathbf{v}_1$ and $b_2\mathbf{v}_2$ for certain choices of $b_1, b_2$. Bottom: $\mathbf{u}_M$ is the state vector in the region between $\mathbf{u}_L$ and $\mathbf{u}_R$.

In section 5.2.1 we will determine the Hugoniot Loci for Kerr-nonlinear Maxwell's equations. For other examples corresponding to other differential equations, see e.g. the book by LeVeque [57]. The next theorem gives more information about the structure of a Hugoniot Locus.

**Theorem 3.19** (Shock Curves)**.**
*The Hugoniot Locus $\mathcal{HL}$ consists locally of $n$ smooth curves $\mathcal{L}_k(\mathbf{u}^*)$, $k = 1, \ldots, n$. Furthermore, if the $k$th characteristic field is genuinely nonlinear, the curve $\mathcal{L}_k(\mathbf{u}^*)$ is called a shock curve.*

*Proof.* See [59]. □

**Definition 3.20** (Shock Wave)**.**
*Let $\mathbf{u}^* = \mathbf{u}_L$, and suppose $\mathcal{L}_k(\mathbf{u}_L)$ is a shock curve. That is, the $k$-th characteristic field is* genuinely nonlinear*. Then, if $\mathbf{u}_R$ is a point on $\mathcal{L}_k(\mathbf{u}_L)$, the discontinuity in definition 3.3.22 is called a $k$-shock wave or only a shock. The same holds if $\mathbf{u}^* = \mathbf{u}_R$ and $\mathbf{u}_L$ is a point on $\mathcal{L}_k(\mathbf{u}_R)$.*

We will see later, when we come to the uniqueness of solutions of the Riemann problem in section 3.3.2, that one differentiates between an admissible and a non-admissible shock.
If the $k$th characteristic field is *linearly degenerate*, we have the following:

**Theorem 3.21** (Contact Discontinuity)**.**
*For a linearly degenerate characteristic field, the curve $\mathcal{L}_k(\mathbf{u}^*)$ is an integral curve of the vector field $\mathbf{r}_k$, and it holds:*

$$s(\mathbf{u}^*, \mathbf{u}) = \lambda_k(\mathbf{u}^*).$$

*In this case, letting $\mathbf{u}_L$ on $\mathcal{L}_k(\mathbf{u}_R)$ (or, equivalently, $\mathbf{u}_R$ on $\mathcal{L}_k(\mathbf{u}_L)$) the speed of shock is $s = \lambda_k^* = \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}_R)$. Then, the weak solution*

$$\mathbf{u}(x,t) = \begin{cases} \mathbf{u}^L & \text{for } x < \lambda^* t, \\ \mathbf{u}^R & \text{for } x > \lambda^* t \end{cases} \tag{3.3.25}$$

*is called a contact discontinuity.*

*Proof.* See [59]. □

#### Uniqueness of Solutions of the Riemann Problem: The Entropy Condition

In section 3.3.2 we have seen that Burger's equation has infinitely many solutions. In classical physics, this makes no sense; there exists a unique solution. Physically and mathematically the question after uniqueness is of high interest. In case of conservation laws, the so-called entropy condition and sometimes a supplementary condition on the shock speed ensure uniqueness.
Physically, entropy can roughly be understood as the total energy of a system, that is, it is the difference between the energy in the system and the energy running out of it: $\text{energy}_{\text{in}} - \text{energy}_{\text{out}}$. Entropy is a positive value, i.e. it is impossible that more energy leaves the system than is in the system itself. Thus entropy gives a physical condition, which of all the solutions of the Riemann problem that have been found are the physically relevant ones. The authors of [68] show that the entropy condition is enough to determine the shocks and that the corresponding shock speed never exceeds the speed of light in vacuum. Mathematically, there are several conditions on the shock speed available, such as the Lax entropy condition, Oleinik's condition, or the condition by Smoller and Johnson [69] for two dimensional systems. Other conditions are also possible; their choice is motivated by the physical problem at hand. E.g. Seccia [70] chooses the entropy condition $\eta > 0$ along with a reflection and transmission criterion to identify physically relevant shocks of the Riemann problem corresponding to Kerr-nonlinear Maxwell's equations. For the same problem, LaBourdonnaye [71] as well shows unique solvability, yet he takes the condition of Smoller-Johnson besides positivity of entropy.

**Definition 3.22** (Lax Entropy Condition)**.**

*(i) Let the kth characteristic field be* genuinely nonlinear*. The Lax entropy condition [57, Ch. 11.13] reads*

$$\begin{aligned} &\lambda_k(\mathbf{u}_R) < s < \lambda_{k+1}(\mathbf{u}_R), \\ &\lambda_{k-1}(\mathbf{u}_L) < s < \lambda_k(\mathbf{u}_L) \end{aligned} \tag{3.3.26}$$

*for a $k \in \{1, \dots, n\}$, with the convention $\lambda_0 := -\infty, \lambda_{n+1} := \infty$. $s$ is the speed of the discontinuity. In this case we have a shock.*

*(ii) If the kth characteristic field is* linearly degenerate*, then we have*

$$\lambda_k(\mathbf{u}_L) = s = \lambda_k(\mathbf{u}_R).$$

*Then the discontinuity is a contact discontinuity.*

**Definition 3.23** (Liu Entropy Condition)**.**
*The entropy condition by Liu is given as follows [72], [73], [74]:*

$$s(\mathbf{u}_L, \mathbf{u}_R) \leq s(\mathbf{u}_L, \mathbf{u}) \tag{3.3.27}$$

*for any point $\mathbf{u}$ on the shock curve $\mathcal{L}_k(\mathbf{u}_L)$ which lies between $\mathbf{u}_L$ and $\mathbf{u}_R$. Liu's strict condition replaces "$\leq$" by "$<$". This implies the inequality*

$$\lambda_k(\mathbf{u}_R) < s(\mathbf{u}_L, \mathbf{u}_R) < \lambda_k(\mathbf{u}_L). \tag{3.3.28}$$

For a genuinely nonlinear field Liu's condition is equivalent to the condition by Lax.

**Definition 3.24** (Admissible Shocks)**.**
*We say a shock is admissible if it fulfills one of the conditions above. A contact discontinuity is always admissible.*

In fact, this definition follows from Theorem 5.2, Ch. 5 in [59]. The following is a combination of a definition and of results found and proven in [69], [75, Th. 2.1], [76, Th. 4.5].

**Theorem 3.25** (Admissibility Condition of Smoller-Johnson)**.**
*Let $n = 2$. Assume $\lambda_1 < \lambda_2$. Let $\mathcal{L}_1(\mathbf{u}^*)$ be a shock curve with an arbitrary, but fixed point $\mathbf{u}^* \in \mathbb{R}^2$ corresponding to the* decreasing *eigenvalue $\lambda_1$, and $\mathcal{L}_2(\mathbf{u}^*)$ to the decreasing eigenvalue $\lambda_2$ (this means, $\mathbf{u}^*$ is on the left in the $x-t$-plane). By $\mathcal{L}_i^*(\mathbf{u}^*)$ we denote the shock curve corresponding to the* increasing *eigenvalue $\lambda_i$ (i.e. $\mathbf{u}^*$ is on the right in the $x-t$-plane). If the following entropy condition for shocks holds,*

$$\lambda_k(\mathbf{u}) < s(\mathbf{u}, \mathbf{u}^*) < \lambda_k(\mathbf{u}^*) \quad \text{for} \quad u \in \mathcal{L}_i(\mathbf{u}^*), \tag{3.3.29}$$

$$\lambda_k(\mathbf{u}^*) < s(\mathbf{u}^*, \mathbf{u}) < \lambda_k(\mathbf{u}) \quad \text{for} \quad u \in \mathcal{L}_i^*(\mathbf{u}^*), \tag{3.3.30}$$

*then the kth characteristic is a shock wave ($k = 1, 2$). The (additional) condition of Smoller-Johnson (L) reads*

$$(L) \quad \begin{cases} \text{for } \mathbf{u} \in \mathcal{L}_1(\mathbf{u}^*) \setminus \{\mathbf{u}^*\} \text{ (i.e. } k=1): & s(\mathbf{u}, \mathbf{u}^*) < \lambda_2(\mathbf{u}^*), \\ \text{for } \mathbf{u} \in \mathcal{L}_2(\mathbf{u}^*) \setminus \{\mathbf{u}^*\} \text{ (i.e. } k=2): & s(\mathbf{u}, \mathbf{u}^*) > \lambda_1(\mathbf{u}^*), \\ \text{for } \mathbf{u} \in \mathcal{L}_1^*(\mathbf{u}^*) \setminus \{\mathbf{u}^*\}: & s(\mathbf{u}, \mathbf{u}^*) < \lambda_2(\mathbf{u}^*), \\ \text{for } \mathbf{u} \in \mathcal{L}_2^*(\mathbf{u}^*) \setminus \{\mathbf{u}^*\}: & s(\mathbf{u}, \mathbf{u}^*) > \lambda_1(\mathbf{u}^*). \end{cases} \tag{3.3.31}$$

*I.e. a point $\mathbf{u} \in \mathbb{R}^2$ can be connected to $\mathbf{u}^*$ by an* admissible *shock wave if $(L)$ is fulfilled.*

*Proof.* See [69], [75, Th. 2.1], [76, Th. 4.5]. □

### Solution of the Riemann Problem

From the above discussions, we now can say how the solution of the Riemann problem

$$\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}, \tag{3.3.32}$$

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}^L & \text{for } x < 0, \\ \mathbf{u}^R & \text{for } x > 0 \end{cases} \tag{3.3.33}$$

looks like.

**Theorem 3.26.**

*(i) Let the kth characteristic field ($k \in 1, \ldots, p$) be* genuinely nonlinear. *For all $\mathbf{u}_L \in \mathbb{R}^n$ there exists a neighborhood $U_\epsilon(\mathbf{u}_L)$ of $\mathbf{u}_L$ so that, if $\mathbf{u}_R \in U_\epsilon(\mathbf{u}_L)$, $\mathbf{u}_L$ and $\mathbf{u}_R$ can be connected either by a k-rarefaction wave or an admissible k-shock. An appropriately chosen entropy condition decides whether it has to be shock or a rarefaction.*

*(ii) If the kth characteristic field is* linearly degenerate, *then $\mathbf{u}_L$ and $\mathbf{u}_R$ can be connected by a contact discontinuity.*

*This weak solution is unique.*

The proof of uniqueness can be found in [59], Theorem 6.1 in Ch. 6. Figure 3.10 shows a sketch of a possible composition of waves solving the Riemann problem.

**Remark 3.27.**
*In case of a non-strict hyperbolic system this result still holds if the eigenvalues form a complete set. Then each characteristic field is still either genuinely nonlinear or linearly degenerate, and there is still only one possibility for the corresponding simple wave.*

There may be also other special waves solving the Riemann problem, depending on the problem, see e.g. [59], or [70] as an example of application to nonlinear Maxwell's equations. In our applications, when we solve the Riemann problem corresponding to Kerr-nonlinear Maxwell's equations, such special waves do not occur.

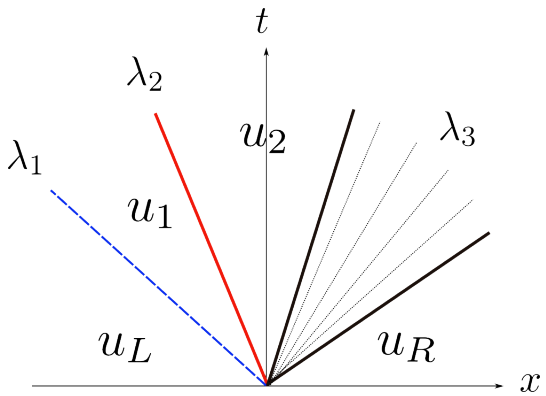



Figure 3.10: A sketch of a weak solution of the Riemann problem consisting (from left to right) of a left contact discontinuity, a left shock and a right rarefaction.

## 3.4 Convergence Theory for the RKDG Method

Shu and Cockburn showed in 1989 in [77] that a RKDG method of order $k+1$ can be devised by (a) applying a DG method with polynomials of degree $k$ to discretize in space, (b) using a total variation diminishing (TVD) explicit time discretization of order $k+1$, and (c) in case of problems with discontinuous solutions, by utilizing a so-called slope limiter to ensure the TVD property. They have shown that using a Runge-Kutta scheme of order $k+1$ results in a TVD stable method without loosing its accuracy.
In this section we first look at the convergence behavior of the space discretization via the DG method, followed by stability and convergence results for the time integration via the RK method. Both is needed to show that the RKDG method is of order $k+1$, as proven by Shu and Cockburn [77].

### 3.4.1 Convergence Theory for the DG Space Discretization

For a one-dimensional scalar hyperbolic conservation law, Johnson and Pitkäranta [15] showed in 1986 a convergence rate of $O(h^{p+1})$ for regular triangulations. Here, $p$ is the polynomial order used in the DG scheme. Richter proved in 1988 [14, Th. 3.1] the following result on the convergence rate of a DG scheme applied to a linear scalar hyperbolic conservation law in two space dimensions of the form

$$a_1\partial_x u + a_2\partial_y u = f(x,y), \quad (x,y) \in \Omega \subset \mathbb{R}^2, \tag{3.4.1}$$

where $a_1, a_2$ are constant.

**Theorem 3.28.**
*If the solution $u$ of* (3.4.1) *is smooth enough and the triangulation of $\Omega$ is uniform, then it holds*

$$||u - u_h||_{L^2(\Omega)} \leq Ch^{p+1}||u||_{H^{p+2}(\Omega)}.$$

*Proof.* See [14, Th. 3.1]. □

In case of Maxwell's equations, the following optimal $L^2$-convergence was shown:

**Theorem 3.29.**
*We consider the linear, isotropic Maxwell's curl-equations on convex domains, i.e.*

$$\epsilon\partial_t\mathbf{E} - \nabla\times\mathbf{H} = \mathbf{j} - \sigma\mathbf{E}, \tag{3.4.2}$$
$$\epsilon\partial_t\mathbf{H} + \nabla\times\mathbf{E} = 0, \tag{3.4.3}$$

*where $\sigma$ is the conductivity and $\mathbf{j}$ the current density. The set of equations* (3.4.2) *can be brought into a second-oder vector wave equation for the electric field $\mathbf{E}$ or $\mathbf{H}$ with $\mathbf{E} \in H_0(\mathrm{curl};\Omega)$ or $\mathbf{H} \in H_0(\mathrm{curl};\Omega)$, respectively, where $\Omega \subset \mathbb{R}^3$. We denote by $\mathbf{u} = \mathbf{E}$ or $\mathbf{u} = \mathbf{H}$ the exact solution of this vector valued wave equation and by $\mathbf{u}_h$ the approximate solution obtained via a DG space discretization. For smooth $\mathbf{u}$ on convex domains (see [16] for detailed regularity requirements), optimal $L^2$-convergence is obtained, that is*

$$||\mathbf{u} - \mathbf{u}_h||_{L^\infty([0,T];L^2(\Omega)^3)} \leq Ch^{p+1},$$

*where $L^\infty([0,T];L^2(\Omega)^3)$ is the so-called Bochner space with the norm*

$$||\mathbf{u}||_{L^\infty([0,T];L^2(\Omega)^3)} = \operatorname{ess\,sup}_{t\in[0,T]} ||\mathbf{u}(t)||_{L^2(\Omega)^3},$$

*and $[0,T]$ denotes the time interval.*

*Proof.* See [16, Th. 4.3]. □

The following theorem gives the error of the DG method for a multi-dimensional scalar conservation law.

**Theorem 3.30.**
*Let $\Omega \subset \mathbb{R}^n$. Consider the conservation law*

$$\begin{aligned} \partial_t u + \partial_x \mathbf{F}(u) &= \mathbf{0} && \text{on } \Omega \times (0,T), \\ u(x,0) &= \mathbf{u}_0 && \text{for } x \in \Omega, \\ u &= \gamma && \text{on } \partial\Omega \times (0,T), \end{aligned}$$

*with $u_0 \in L^\infty(\Omega)$ and $\gamma \in L^\infty(\partial\Omega \times (0,T))$. Let $\mathbf{F}(u) \in W^{k+2,\infty}(\Omega)$ (see definition 2.3). Let $L_h$ be the approximation operator to $-\nabla \cdot \mathbf{F}$ generated by the DG method, where the finite element space $V_h$ is chosen to consist of piecewise polynomial functions. Furthermore, we assume the used quadrature rules are of order $2p+1$ over the edges and of order $2p$ inside each element of a regular triangulation. Then the approximation error is of order $\Delta x^{p+1}$.*

*Proof.* See Proposition 2.1 in [17]. □

### 3.4.2 Convergence Theory for the Runge-Kutta Time Discretization

For time integration we use a Runge-Kutta scheme. In our simulations we chose the low-storage Runge Kutta method with 5 stages and of 4th order (denoted as (5,4)-RK) as presented in a paper by Carpenter and Kennedy in 1994 [7], originally introduced by Williamson in 1980 [6]. Their Runge-Kutta scheme requires only $2N$ storage and has a better accuracy and a larger stability domain than the (3,3)-RK method by Williamson. Yet, theoretical results about stability and convergence of the RKDG method are mostly formulated with respect to another version by Shu [78]. There exist several RK formulations, of which we mention here a version by Butcher [79, Ch. 23] and a version by Ruuth and Spiteri [60]. We will point out that all these formulations are connected with each other, so that theoretical results obtained for the scheme by Shu (3.4.6) also extend to the low-storage Runge-Kutta scheme (3.4.10) by Carpenter and Kennedy.

Let us consider the time-dependent equation

$$\frac{\partial \mathbf{u}_h(t)}{\mathrm{d}t} = L_h(\mathbf{u}_h(t)), \tag{3.4.4}$$

where $\mathbf{u}_h$ is an approximation in space to the exact solution $\mathbf{u}$, and $L_h$ is a discretization of some operator; in our case, $L_h(\mathbf{u}_h(t))$ denotes the discretized form of $\nabla \cdot \mathbf{F}(\mathbf{u})$ in the conservation law $\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \mathbf{0}$. Consider also a time grid $0 = t_0 < t_1 < \cdots < t_M = T$ of a time interval $[0,T]$. We denote $\Delta t_l := t_l - t_{l-1}$ with $l = 1, \ldots, M$, and let $\mathbf{u}_h^l$ be the solution approximation in time step $l$.

### Butcher's version

Butcher's version of an explicit Runge-Kutta method with $s$ stages for (3.4.4) reads

$$\begin{aligned}
&\text{for } \; l = 1, \ldots, M : \\
&\quad \mathbf{u}_h^{(0)} = \mathbf{u}_h^l, \\
&\quad \text{for } \; i = 1, \ldots, s : \\
&\qquad \mathbf{u}_h^{(i)} := L_h(\mathbf{u}_h^{(0)} + \Delta t_l \sum_{j=1}^{i-1} a_{ij} \mathbf{u}_h^{(j)}) \\
&\quad \text{end} \\
&\quad \mathbf{u}_h^{l+1} = \mathbf{u}_h^l + \Delta t_l \sum_{j=1}^{s} b_k \mathbf{u}_h^{(j)}
\end{aligned} \tag{3.4.5}$$

with the convention $\sum_{j=1}^{0} a_{ij} \mathbf{u}_h^{(j)} = 0$ so that $\mathbf{u}_h^{(1)} = L_h(\mathbf{u}_h^l)$. The $\mathbf{u}_h^{(i)}$ are recursively defined. See e.g. Butcher's book [79, Ch. 23], [80], [81] for more details on Runge-Kutta methods.

### Version of Shu and Osher

In the framework of the DG method, another formulation of the Runge-Kutta time discretization was introduced in the paper by Shu in 1988 [78]:

$$\begin{aligned}
&\text{for } \; l = 1, \ldots, M : \\
&\quad \mathbf{u}_h^{(0)} = \mathbf{u}_h^l, \\
&\quad \text{for } \; i = 1, \ldots, s : \\
&\qquad \mathbf{u}_h^{(i)} := \sum_{j=0}^{i-1} \Big( \alpha_{ij} \mathbf{u}_h^{(j)} + \Delta t_l \beta_{ij} L_h(\mathbf{u}_h^{(j)}) \Big) \\
&\quad \text{end} \\
&\quad \mathbf{u}_h^{(l+1)} = \mathbf{u}_h^{(s)} \\
&\text{end.}
\end{aligned} \tag{3.4.6}$$

Yet, in order to have optimal convergence results, their Runge-Kutta method has to be of order $s + 1$.

### Ruuth's and Spiter's version

Especially for nonlinear problems time integration can be a sensitive part of the numerical process with respect to stability. Ruuth and Spiteri give in [82] a so-called strong stability preserving Runge-Kutta method up to order 4 with 5 stages. In [60] they use the

formulation

$$\begin{aligned}
&\text{for } \; l = 1, \dots, M : \\
&\quad \mathbf{u}_h^{(0)} = \mathbf{u}_h^l, \\
&\quad \text{for } \; i = 1, \dots, s : \\
&\qquad \mathbf{u}_h^{(i)} := \mathbf{u}_h^{(0)} + \Delta t_l \sum_{j=0}^{i-1} \kappa_{ij} L_h(\mathbf{u}_h^{(j)}) \\
&\quad \text{end} \\
&\quad \mathbf{u}_h^{(l+1)} = \mathbf{u}_h^{(s)} \\
&\text{end.}
\end{aligned} \tag{3.4.7}$$

They also give a connection between the different Runge-Kutta formulations (see [60, Section 3.2], which is a relation between the coefficients in the different formulations (3.4.5), (3.4.6) and (3.4.7). The coefficients $\beta_{ij}, \alpha_{ij}$ and $\kappa_{ij}$ are recursively related by

$$\kappa_{ij} := \beta_{ij} + \sum_{k=j+1}^{i-1} \alpha_{ik} \kappa_{kj}. \tag{3.4.8}$$

And between the coefficients $b_i, a_{ij}$ and $\kappa_{ij}$ we have the relation

$$\begin{aligned}
&b_i = \kappa_{s,i-1}, \quad i = 1, \dots, s, \\
&a_{ij} = \kappa_{i-1,j-1}, \quad j = 1, \dots, i-1, \;\; i = 1, \dots, s-1.
\end{aligned} \tag{3.4.9}$$

### Williamson's version

Williamson [6] works with the following form:

$$\begin{aligned}
&\text{for } \; i = 1, \dots, s : \\
&\qquad \Delta \mathbf{u}_h^{(i)} = A_i \, \Delta \mathbf{u}_h^{(i-1)} + \Delta t_l L_h(\mathbf{u}_h^{(i-1)}) \\
&\text{end} \\
&\mathbf{u}_h^{l+1} = \mathbf{u}_h^l + B_i \, \Delta \mathbf{u}_h^{(i)}
\end{aligned} \tag{3.4.10}$$

for $l = 1, \dots, M$. The relation between the coefficients $A_i, B_i$ and $b_i, a_{ij}$ from Butcher's scheme (3.4.5) is

$$\begin{aligned}
&B_i = a_{i+1,i} \quad \text{for} \quad i \neq s, \\
&B_s = b_s, \\
&A_i = \frac{b_{i-1} - B_{i-1}}{b_i} \quad \text{for} \quad i \neq 1, \; b_i \neq 0, \\
&A_i = \frac{a_{i+1,i-1}}{B_i} \quad \text{for} \quad i \neq 1, \;\; b_i = 0.
\end{aligned} \tag{3.4.11}$$

Carpenter and Kennedy extended Williamson's RK scheme to a RK method of order 4 with 5 stages [7].
In the next sections we give an overlook of some of the theoretical results about stability and convergence of the Runge-Kutta method in the context of DG methods. This field has been widely explored by several authors. We refer to e.g. [78], [8], [67], [10], [12] (and the references therein). However, there are still open questions, for instance, concerning nonlinear multi-dimensional systems of conservation laws.

### 3.4.3 Stability of the Runge-Kutta Method for One-Dimensional Scalar Conservation Laws

We consider a nonlinear scalar conservation law in one space dimension

$$\partial_t u + \partial_x F(u) = 0.$$

In order to ensure stability of the Runge-Kutta scheme (3.4.6), the coefficients $\alpha_{ij}, \beta_{ij}$ have to be chosen appropriately, that is, we are looking for a condition on $\alpha_{ij}, \beta_{ij}$. Let us first redefine the intermediate states $u_h^{(i)}$ as

$$\begin{aligned} w_h^{(j)} &:= \alpha_{ij} u_h^{(j)} + \Delta t_m \beta_{ij} L_h(u_h^{(j)}), \quad j = 0, \dots, i-1, \\ u_h^{(i)} &= \sum_{j=0}^{i-1} w_h^{(j)}. \end{aligned}$$

The $w_h^{(j)}$ can be rewritten as

$$w_h^{(j)} = v_j + \delta L_h(v_j)$$

with $v_j =: \alpha_{ij} u_h^{(j)}$ and $\delta := \Delta t_l \frac{\beta_{ij}}{\alpha_{ij}}$, where $\alpha_{ij} = 0$ only if $\beta_{ij} = 0$; in this case, we set $\delta = 0$. This is a single forward Euler step. Let us now define what is meant by stability.

**Theorem 3.31** (Stability Property)**.**
*We say the RK method* (3.4.6) *has the stability property if and only if*

$$|u_h^{l+1}| \le |u_h^l| \qquad \text{for all } l = 1, \dots, M.$$

**Theorem 3.32** (Stability of the RK Method)**.**
*Assume the $w_h^{(j)}$ fulfill the local stability property $|w_h^{(j)}| < |v_j|$ with $|\delta| \le \delta_0$, where $\delta_0 := \max_{1 \le l \le M} \{\Delta t_l \frac{\beta_{ij}}{\alpha_{ij}}\}$ with the additional conditions*

$$\begin{aligned} &(i) \quad \text{If } \beta_{ij} \neq 0, \text{ then } \alpha_{ij} \neq 0, \\ &(ii) \quad \alpha_{ij} \ge 0, \\ &(iii) \quad \sum_{j=0}^{i-1} \alpha_{ij} = 1, \quad i = 1, \dots, s. \end{aligned} \tag{3.4.12}$$

*That is, $\alpha_{ij} = 0$ only if $\beta_{ij}$, and in this case, it is $\delta_0 = 0$. Then Runge-Kutta method* (3.4.6) *has the stability property of definition 3.31, and*

$$|u_h^l| \le |u_h^0| \qquad \text{for all } m \ge 0,$$

*where $u_h^0 = \mathbf{P}_{V_h} u_0$ is the projection of the initial data $u_0 = u(\cdot, 0)$ on the finite element space $V_h$.*

*Proof.* See [12, Ch. 2.2] and [67, Ch. 2.3.2]. □

In order to ensure $L^2$-stability, the *CFL condition* needs to be fulfilled, which reads

$$\max_k |\lambda_k| \frac{\Delta x}{\Delta t} \le \text{CFL},$$

where $\lambda_k$ is an eigenvalue of the Jacobian of $\mathbf{F}$. This condition must also be kept in the nonlinear case; see [12].

### Stability of the Intermediate Step

Theorem 3.32 says that the RK scheme is stable if the intermediate step $u_h^{(i)} \mapsto w_h^{(i)} = v_j + \delta L_h(v_j)$ is locally stable. In order to show the stability of the intermediate step, we introduce the total variation diminishing (TVD) property.

**Definition 3.33** (Total Variation Diminishing (TVD))**.**
*Let $u_h$ be the global approximate solution in $\Omega$ of $u$ in $\partial_t u + \partial_x F(u) = 0$, consisting of local approximate solutions $u_j$ on $\Omega_j$, where $\Omega = \cup_{j=1}^K \Omega_j$, $K > 0$. We assume that $\Omega_{j+1}$ is the right neighbor of $\Omega_j$. The total variation of $u_h$ is defined as*

$$|u_h|_{\mathrm{TV}} := \sum_{j=1}^{K} |u_{j+1} - u_j|.$$

*Let $u_h^l$ be the approximate solution in space at time step $t_m$. Then a time stepping method has the TVD property if and only if*

$$|u_h^{l+1}|_{\mathrm{TV}} \leq |u_h^l|_{\mathrm{TV}}$$

*for all $l \geq 1$.*

For the piecewise constant case, i.e. $p = 0$, Harten showed in [83] that a monotone scheme is TVD. Furthermore, in this case monotone schemes were shown to be stable, convergent and first order accurate, see the works by Harten et. al. [64], Kuznetsov [65], Crandall and Majda [66]. The general case, i.e. $p > 0$, is more complex. In this case one studies the total variation in the local means (TVDM). A time-stepping method has the *TVDM property* if

$$|\bar{u}_h^{l+1}|_{\mathrm{TV}} \leq |\bar{u}_h^l|_{\mathrm{TV}},$$

where $\bar{u}_h^l$ is the local means of $u_h^l$ on $\Omega_j$ at time step $t_l$, defined as

$$\bar{u}_j^l := \frac{1}{\mathrm{vol}(\Omega_j)} \int_{\Omega_j} u_h^l(x)\, \mathrm{d}x.$$

**Theorem 3.34.**
*If the Runge-Kutta method is TVD or TVDM, then the intermediate step is locally stable, giving a stable time-stepping scheme.*

*Proof.* See [67, Ch. 2.4.2]. □

The requirement of having the TVDM property leads to so-called *sign conditions* (see e.g. [67], [12]). In order to have an RK method with the TVDM property, these sign conditions must be fulfilled, which is not automatically the case. This can be ensured by a so-called *slope limiter* $\Lambda\Pi_h$. Shu and Cockburn [12] devised a generalized slope limiter such that the intermediate step $u_h^{(i)} \mapsto w_h^{(i)}$ is TVDM stable, or at least TVBM stable (i.e. the total variation is bounded in the means). See also [67, Ch. 2.4.2], [11, Ch. 5.6.2]. Furthermore, we have the following theorem on convergence of the approximation.

**Theorem 3.35** (Convergence to the Entropy Solution)**.**
*Assume that the generalized slope limiter $\Lambda\Pi_h$ is a TVDM or a TVBM slope limiter. Assume also that all the coefficients $\alpha_{ij}$ in the RK discretization are nonnegative and satisfy the condition*

$$\sum_{j=1}^{i-1} \alpha_{ij} = 1, \qquad i = 1, \ldots, k+1.$$

*Then there is a subsequence* $\{\bar{u}_{h'}\}_{h'>0}$ *of the sequence* $\{\bar{u}_h\}_{h>0}$ *generated by the RKDG scheme that converges in* $L^\infty((0,T);L^1(\Omega))$ *to a weak solution of the conservation law* $\partial_t u + \partial_x F(u) = 0$ *on* $\Omega \times [0,T]$ *with initial data* $u(x,0) = u_0(x)$*. If the generalized slope limiter* $\Lambda\Pi_h$ *is such that*

$$||\bar{v}_h - \Lambda\Pi_h(v_h)||_{L^1(\Omega)} \leq C\Delta x|\bar{v}_h|_{TV(\Omega)},$$

*then the results hold also for the sequence of the functions* $\{u_h\}_{h>0}$*.*

*Proof.* This is Theorem 2.13 in [67, Ch. 2.4.4], and the proof can be found there. □

Note that this is a general result for the scalar case, and it is also a true statement for a nonlinear conservation law.

Integrating a slope limiter into the RKDG process is easily done. One either applies the limiter in each RK stage (to ensure TVDM/TVBM stability) or after each time step, as required. A general procedure might look as follows (see [67], [12]):

(1) Compute $u_h^0 = \Lambda\Pi_h P_{V_h} u_0$, where $P_{V_h} u_0$ is the projection of the initial data $u_0$ on $V_h$.

(2) Compute $u_h^{l+1}$ of the next time step ($l = 0, \ldots, M$):

(2a) Set $u_h^{(0)} = u_h^l$.

(2b) for $j = 1, \ldots, k+1$ compute the limited intermediate steps

$$u_h^{(j)} = \Lambda\Pi_h \left( \sum_{i=0}^{j-1} \alpha_{ji} u_h^{(i)} + \beta_{ji} \Delta t_m L_h(u_h^{(i)}) \right).$$

(2c) $u_h^{l+1} = u_h^{(k+1)}$.

The authors in [11] give Matlab codes of a selection of slope limiters. There, also numerical tests on the performance can be found.

We mention another possibility of slope limiting, the so-called weighted essentially non-oscillatory (WENO) limiter which is an extension of the essentially non-oscillating (ENO) method. For theory and practical aspects we refer to [84], to the chapter "High Order ENO and WENO Schemes for Computational Fluid Dynamics" by C.-W. Shu in [85], to [86], [87], and the references therein. We note that this is only a small selection. In [88] a very simple way of implementing a WENO limiter is given.

### 3.4.4 Stability of the Runge-Kutta Method for the Multi-Dimensional System Case

The TVDM property as defined in section 3.4.3 is only fulfilled in the scalar case for the intermediate step. For multidimensional systems, Cockburn, Hou and Shu showed in 1990 in [17] that, if the intermediate step fulfills a maximum principle (given in their Lemma 2.3) and if the mesh fulfills a certain uniformity criterion (called *B-uniform*, see their definition 2.5), and if this maximum principle is enforced by an appropriate projection $\Lambda\Pi_h$ (a slope limiter), then the following can be said about stability and convergence of the Runge-Kutta method in the multidimensional case:

**Theorem 3.36.**
*Let the quadrature rule applied in the discretization process be of order* $(2k+1)$ *over the edges and of order* $2k$ *over the elements. (*$k$ *is the degree of the polynomials used in the DG space discretization.) Then*

1. *The RKDG method is of order* $k+1$ *in time and space if* $\Delta t = O(h)$.

2. *If a certain CFL condition is fulfilled (the choice of the CFL depends on the triangulation), the approximation generated by the RKDG method fulfills the maximum principle from Lemma 2.3 in [17].*

3. *If the BV-norm of* $\bar{\mathbf{u}}_h$ *is bounded, the approximate solution converges to a weak solution of the conservation law* $\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \mathbf{0}$, *where* $\mathbf{u} = (u_1, \dots, u_n)^T$ *and* $\mathbf{x} \in \Omega \subset \mathbb{R}^n$, $t \in [0, T]$.

*Proof.* This was proven in Theorem 2.10 in [17]. □

The construction of such a slope limiter can be found in, e.g., [67, Ch. 3.2.2, Ch. 3.3.8] by Cockburn (the same is found in [12]), although the authors remark that there is no rigorous proof that this slope limiter indeed ensures stability of the RK method. Yet, their numerical tests indicate it, and additionally it is easily implemented; therefore it is used in practice.
The necessity of a B-uniform mesh can be dropped, as Wierse showed in [89] for two-dimensional scalar conservation laws; in this paper the author presents a slope limiter for general regular meshes.

### Slope Limiting in the One-dimensional System Case

For this section we assume to have $\mathbf{u} = (u_1, \dots, u_n)^T$, and we let $\mathbf{F} = (F_1, \dots, F_m)$, where $F_i = (f_1, \dots, f_n)^T$. For a scalar system of conservation laws, it turns out to be more efficient to transform to characteristic variables, then to apply the slope limiter componentwise and to transform back (see [67, Ch. 3.3.9]).
Let $\bar{A}_j := \frac{\partial \mathbf{F}}{\partial \mathbf{u}}|_{\mathbf{u}=\bar{\mathbf{u}}_j}$ be the Jacobian of $\mathbf{F}$, evaluated at the mean value $\bar{\mathbf{u}}_j$ of $\mathbf{u}$ on $\Omega_j$. The right eigenvectors of $\bar{A}_j$ shall be $\mathbf{r}_j^{(l)}$ $(l = 1, \dots, n)$, and $\mathbf{l}_j^{(l)}$ the left ones. They shall be normalized such that $\mathbf{r}_j^{(l)} \cdot \mathbf{l}_j^{(k)} = \delta_{lk}$. We define

$$\bar{R}_j := (\mathbf{r}_j^{(1)}, \dots, \mathbf{r}_j^{(m)}) \in \mathbb{R}^{n \times n}.$$

Due to the normalization of the right and left eigenvectors, it is

$$\bar{R}_j^{-1} = (\mathbf{l}_j^{(1)}, \dots, \mathbf{l}_j^{(m)})^T.$$

(1) Compute $\bar{R}_j$ and $\bar{R}_j^{-1}$.

(2) Transform to the characteristic variables $\mathbf{v} = \bar{R}_j^{-1}\mathbf{u}$.

(3) Apply a slope limiter to each component of $\mathbf{v}$.

(4) Transform back via $\mathbf{u} = \bar{R}_j \mathbf{v}$.

### 3.4.5 Convergence of the RKDG Method for the Linear One-dimensional Scalar Case

Assume $F(u) = au$, where $a = \text{const}$. In [67, Th. 2.2] the author proves the following result.

**Theorem 3.37.**
*Assume the initial function $u_0$ is an element of $H^{k+2}(\Omega)$. Define the error function $e := u - u_h$. Then it holds*

$$||e(T)||_{L^2(\Omega)} \leq C|u_0|_{H^{k+2}(\Omega)}(\Delta x)^{k+1},$$

*where $C$ depends on $k, |a|$, and $T$.*

### 3.4.6 Convergence of the RKDG Method for the Nonlinear One-dimensional Scalar Case

For piecewise-constant $u_h$, we have the $L^1$-error estimate:

**Theorem 3.38.**
$||u(T) - u_h(T)||_{L^1(\Omega)} \leq ||u_0 - u_h(0)||_{L^1(\Omega)} + C|u_0|_{TV(\Omega)}\sqrt{T\Delta x}$.

*Proof.* See [67, Th. 2.4]. □

Unfortunately, there are not yet error estimates for $k > 0$. However, there is a result in case the flux $F$ is concave or convex.

**Theorem 3.39.**
*Let $F$ be a strictly convex or concave flux. Then, for any $k \geq 0$, if the numerical solution given by the DG method converges, it converges to the entropy solution.*

*Proof.* See [90] and [67, Th. 2.5]. □

Thus, assuming to have a convex or concave flux and recalling section 3.4.4, if the RKDG scheme is devised as required in theorem 3.36 and if the numerical solution converges, it converges with order $k + 1$. Especially, if the exact solution is smooth, we can expect convergence with $O(\Delta x^{k+1})$, in the linear and nonlinear case, respectively.

# Part II

# Application

## Rotationally Symmetric and Kerr-Nonlinear Maxwell's Equations

# 4 Application: Rotationally Symmetric Maxwell's Equations

Bodies of revolution (BOR) are objects that are rotationally symmetric around a certain axis. Examples include cylinders of any kind, toroidal resonators which can be used as sensors or filters in biology, chemistry or physics (to mention only a few areas). They also range from antennas to tapered fibers as they are used in near-field scanning optical microscopy below the diffraction limit. In general, the objects of interest in optics are very small with a size ranging from $nm$ to $\mu m$. This complicates numerical simulations: The objects need to be resolved accurately, resulting in the need of flexible meshes with many elements of different sizes. This leads to long computation times, especially in three dimensions. It is therefore desirable to reduce computational costs whenever possible.
In case of BOR, the idea is to use the symmetry to reduce the computational effort. After introducing cylindrical coordinates, the azimuthal dependence is represented by a Fourier series. This results in an infinite set of equations with reduced dimension $d-1$ which have to be solved. Numerically, a (small) finite number of equations should suffice to express the azimuthal dependence of the fields sufficiently accurately. Here lies the reduction of the computational cost.

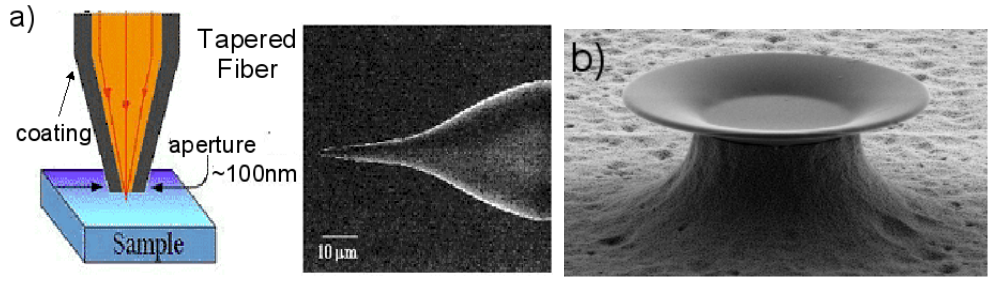



Figure 4.1: a) A tapered fiber as it is used in optical microscopy [The Awshalom Group of the University of Santa Barbara]. b) A toroidal resonator as applied in e.g. biosensing or filtering [Grossmann et al., Appl. Phys. Lett. 96 (2010)].

This chapter is organized as follows: In section 4.1 we derive BOR Maxwell's equations. We will see that introducing cylindrical coordinates leads to equations with a singularity. This problem can be overcome by introducing the weak form. In sections 4.2 and 4.4 we then apply the DG method to discretize BOR Maxwell's equations in one and two space dimensions. A main ingredient of any DG scheme is the numerical flux, and we will derive it for the BOR case. Spatial discretization leads to a semi-discrete scheme with local matrices that need to be updated element-wise. We will present an efficient way of computing those matrices by using orthogonal polynomials. In sections 4.3 and 4.5 we will come to numerical tests and results in one and two space dimensions. As a basic test system to explore stability and convergence of the scheme we consider a homogeneous cavity. After that we introduce perfectly matched layers and sources. We will present results of simulations of a traveling Gaussian pulse through a tapered silicon fiber with and without a sample made of glass. Most of the results can also be found in the paper [91].

## 4.1 Maxwell's Equations for Bodies of Revolution in Time Domain

In the succeeding sections we formulate Maxwell's equations in bodies of revolution, which we introduced at the beginning of this chapter. Then we apply the DG method in order to discretize them in space. We present here results from [91] in more details, considering also the one-dimensional case in order to introduce the basics of the approach.

Our starting point are Maxwell's curl-equations in time domain in three dimensions without any sources, that is

$$\begin{aligned}\nabla \times \mathbf{E} &= -\mu \partial_t \mathbf{H},\\ \nabla \times \mathbf{H} &= \epsilon \partial_t \mathbf{E},\end{aligned}$$

where $\mathbf{x} \in \mathbf{R}^3$ is the spatial variable, $t \geq 0$ is the time variable, $\mu$ is the permeability, $\epsilon$ the permittivity, $\mathbf{E}$ is the electric field, and $\mathbf{H}$ is the magnetic field.
We introduce cylindrical coordinates $\mathbf{x} \mapsto (r, \varphi, z)$ with respect to the standard basis $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\varphi, \hat{\mathbf{e}}_z$ so that

$$\nabla \times \mathbf{E} = \begin{pmatrix} \frac{1}{r}\partial_\varphi E_z - \partial_z E_\varphi \\ \partial_z E_r - \partial_r E_z \\ \frac{1}{r}(\partial_r (r E_\varphi) - \partial_\varphi E_r) \end{pmatrix}.$$

$\nabla \times \mathbf{H}$ is transformed in an analogous manner. We thus obtain the Maxwell's equations in cylindrical coordinates as

$$\begin{aligned}
&\epsilon \partial_t E_r - \frac{1}{r}\partial_\varphi H_z + \partial_z H_\varphi = 0,\\
&\epsilon \partial_t E_\varphi + \partial_r H_z - \partial_z H_r = 0,\\
&\epsilon \partial_t E_z - \frac{1}{r}(\partial_r (r H_\varphi) - \partial_\varphi H_r) = 0,\\
&\mu \partial_t H_r + \frac{1}{r}\partial_\varphi E_z - \partial_z E_\varphi = 0,\\
&\mu \partial_t H_\varphi - \partial_r E_z + \partial_z E_r = 0,\\
&\mu \partial_t H_z + \frac{1}{r}(\partial_r (r E_\varphi) - \partial_\varphi E_r) = 0.
\end{aligned} \tag{4.1.1}$$

We initially encounter the problem of finding $\frac{1}{r}$-terms in the equations, which are singular at the symmetry axis $r = 0$. This is a problem introduced by the chosen coordinate system. The fields are differentiable on the axis of rotation. In the FDTD and FEM setting several solutions have been suggested that overcome this problem, some of them are quite elaborate (see e.g. Ref. [92]). Here, we will give another possibility to solve the singularity problem in the next section.

### 4.1.1 Weak Form

Here, we will give another possibility to solve the singularity problem, namely by switching to a weak formulation [40, 39]. To do so, we start with Maxwell's curl-equations in coordinate-independent form. To establish the weak form we then multiply by a smooth test function $\psi$ and integrate over the domain of interest$\mathcal{B} \subset \mathbb{R}^3$. For theory concerning

weak formulations see e.g. [40, 39]. We start with Maxwell's curl-equations (2.0.1) in coordinate-independent form. To establish the weak form we multiply them by a smooth test function $\psi$ and integrate over the domain of interest $\mathcal{B} \subset \mathbb{R}^3$. For instance, for Faraday's law equation this reads

$$\int_{\mathcal{B}} \epsilon \partial_t \mathbf{E} \cdot \psi \, \mathrm{d}\mathbf{x} = \int_{\mathcal{B}} \nabla \times \mathbf{H} \cdot \psi \, \mathrm{d}\mathbf{x},$$

where we assume $\mathbf{H}, \mathbf{E}$ are smooth enough so that all derivatives are defined (see Ref. [40] for related theory). Now we introduce cylindrical coordinates with the corresponding transformation mapping $\Pi : \mathcal{D} \to \mathcal{B}$, and get the weak equations as

$$\begin{aligned} &\int_{\mathcal{D}} \epsilon \partial_t \mathbf{E} \cdot \Psi \, r \, \mathrm{d}(r, \varphi, z) \\ &= -\int_{\mathcal{D}} \begin{pmatrix} -\frac{1}{r}\partial_\varphi H_z + \partial_z H_\varphi \\ \partial_r H_z - \partial_z H_r \\ -\frac{1}{r}(\partial_r (r H_\varphi) - \partial_\varphi H_r) \end{pmatrix} \cdot \Psi \, r \, \mathrm{d}(r, \varphi, z). \end{aligned} \tag{4.1.2}$$

The equation for the $\mathbf{H}$-field is brought into the weak form in an analogous manner. Here, $\det(J_\Pi) = r$, where $J_\Pi$ is the corresponding Jacobian of the mapping $\Pi$. We realize the singularity in $r = 0$ drops out this way, in contrast to the set of equations (4.1.1). When we say we solve equation (4.1.1) in the weak sense we mean the set of equations (4.1.2) *with respect to the measure* $r \, \mathrm{d}(r, \varphi, z)$. We will proceed with the weak form and later use a Galerkin ansatz for the discretization of $\Omega := \{(r, z) : (r, \varphi, z) \in \mathcal{D}\}$.

### 4.1.2 Fourier Series Ansatz

Due to the periodicity in $\varphi$-direction in rotational symmetries (as BOR), we can make a Fourier ansatz for the $\varphi$-variable of the fields as

$$\begin{aligned} \mathbf{E}(\mathbf{r}, t) &= \sum_{m=0}^{\infty} e^{\mathrm{i}m\varphi} \, \mathbf{E}^{(m)}(r, z, t), \\ \mathbf{H}(\mathbf{r}, t) &= \sum_{m=0}^{\infty} e^{\mathrm{i}m\varphi} \, \mathbf{H}^{(m)}(r, z, t). \end{aligned} \tag{4.1.3}$$

If we plug this ansatz into the equations (4.1.2), the $\varphi$-derivative obviously drops out, and we get an infinite set of equations in the weak sense as

**BOR Maxwell's Equations**

$$\begin{aligned} &\epsilon \, \partial_t E_r^{(m)} - \frac{\mathrm{i}m}{r} H_z^{(m)} + \partial_z H_\varphi^{(m)} = 0, \\ &\epsilon \, \partial_t E_\varphi^{(m)} + \partial_r H_z^{(m)} - \partial_z H_r^{(m)} = 0, \\ &\epsilon \, \partial_t E_z^{(m)} - \frac{1}{r} H_\varphi^{(m)} - r \partial_r H_\varphi^{(m)} + \frac{\mathrm{i}m}{r} H_r^{(m)} = 0, \\ &\mu \, \partial_t H_r^{(m)} + \frac{\mathrm{i}m}{r} E_z^{(m)} - \partial_z E_\varphi^{(m)} = 0, \\ &\mu \, \partial_t H_\varphi^{(m)} - \partial_r E_z^{(m)} + \partial_z E_r^{(m)} = 0, \\ &\mu \, \partial_t H_z^{(m)} + \frac{1}{r} E_\varphi^{(m)} + r \partial_r E_\varphi^{(m)} - \frac{\mathrm{i}m}{r} E_r^{(m)} = 0. \end{aligned} \tag{4.1.4}$$

Note that $m \in \mathbb{N}_0$ is fixed. We see that the set of six three-dimensional equations (4.1.1) is reduced to an infinite set of decoupled two-dimensional equations (4.1.4). We will call this set of equations *BOR Maxwell's Equations*. Numerically, the infinite series in the $\varphi$-ansatz (4.1.3) is approximated by a finite series, choosing $m$ appropriately. A thorough theory about the justification of the $\varphi-$ansatz, its approximation, and its convergence, including an error analysis, can be found in Ref. [93]. Physically, the choice of $m$ depends on the excitation of the system, and it is often sufficient to only solve for a single or very few values of $m$. At last, we remark the Fourier series ansatz automatically fulfills periodic boundary conditions in $\varphi$.

### 4.1.3 Boundary Conditions

In order to assure well-posedness, it is required that all electromagnetic fields are continuous on the rotational axis, see [93, Prop. 2.3]. This leads to the following conditions in $r = 0$ (see also [94] and [95]):

The limit for $r \to 0$ has to be $\varphi$-independent and unique. For the $E_z$-field we thus require

$$\lim_{r\to 0} E_z(r, \varphi, z, t) = E_z(r = 0, z, t), \text{ and } E_z(r = 0, z, t) \text{ is unique.}$$

Furthermore, it has to hold $\partial_r E_z(r = 0, z, t) = 0$. At the rotational axis $r = 0$, there are three different cases:

(i) For $m = 0$ : $\mathbf{E}^{(0)}$ is polarized in $z$-direction, and $E_\varphi^{(0)}(r = 0, z) = 0$.

(ii) For $m = 1$ : $E_k^{(1)}(r = 0, z)$ (for $k = r, z$) is purely radial, that means,

$$E_r^{(1)}(r = 0, z) = E_\varphi^{(1)}(r = 0, z) \text{ for all } z \text{ and } E_z^{(1)}(r = 0, z) = 0.$$

(iii) For $m > 1$ : all fields are 0 in $r = 0$.

These boundary conditions are included in the weak formulation of BOR Maxwell's equations (4.1.4); for theory, see e.g. [93].

### 4.1.4 BOR Maxwell's Equations as a Conservation Law

In order to apply a discontinuous Galerkin discretization, we rewrite BOR Maxwell's equations (4.1.4) as a system of conservation laws as

$$Q\,\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \frac{1}{r} B\mathbf{u} \tag{4.1.5}$$

with the state vector $\mathbf{u} = (\mathbf{E}^{(m)}, \mathbf{H}^{(m)})^T \in \mathbb{R}^6$, and $m \in \mathbb{N}_0$ is fixed; note that $\mathbf{u}$ depends on $\mathbf{r} = (r, z)$ and $t$. Also, $\nabla = (\partial_r, \partial_z)^T$. $Q$ denotes the material matrix and it is given as

$$Q := \begin{pmatrix} \underline{\underline{\epsilon}} & \mathbf{0} \\ \mathbf{0} & \underline{\underline{\mu}} \end{pmatrix}.$$

$B$ is a $6 \times 6$-matrix with non-zero entries

$$B_{1,6} = \mathrm{i}m,\ B_{3,4} = -\mathrm{i}m,\ B_{3,5} = 1,$$
$$B_{4,3} = -\mathrm{i}m,\ B_{6,1} = \mathrm{i}m,\ B_{6,2} = -1.$$

$\mathbf{F}(\mathbf{u}) \in \mathbb{R}^{6,2}$ is the so-called flux vector, and it is given by

$$\mathbf{F}(\mathbf{u}) := (A_r\mathbf{u}, A_z\mathbf{u}),$$

where the $6 \times 6$-matrices $A_r$ and $A_z$ only have the following non-zero entries:

$$(A_r)_{2,6} = 1,\ (A_r)_{3,5} = -1,\ (A_r)_{5,3} = -1,\ (A_r)_{6,2} = 1;$$
$$(A_z)_{1,5} = 1,\ (A_z)_{2,4} = -1,\ (A_z)_{4,2} = -1,\ (A_z)_{5,1} = 1.$$

If the material tensors $\underline{\underline{\epsilon}}$ and $\underline{\underline{\mu}}$ are symmetric, then so are the matrices $Q, A_r$ and $A_z$. Therefore, (4.1.5) represents a hyperbolic system; recall the basic notions about hyperbolic equations in section 2.6.1 of this thesis. Also, we give [58, Ch. 5] as an additional reference. In the special case of isotropic media, i.e., $\underline{\underline{\epsilon}} = \mathrm{diag}(\epsilon, \epsilon, \epsilon)$ and $\underline{\underline{\mu}} = \mathrm{diag}(\mu, \mu, \mu)$, we can calculate the eigenvalues explicitly and find them to be $\lambda_1 = 0$ (with double multiplicity) and $\lambda_{2,3} = \pm\frac{1}{\sqrt{\epsilon\mu}}$ (each with double multiplicity), identical to the Cartesian case.

## 4.2 The Runge-Kutta Discontinuous Galerkin Method Applied to 2D-1D-BOR Maxwell's Equations

In this section we apply the Runge-Kutta Discontinuous Galerkin (RKDG) method to BOR Maxwell's equations in two space dimensions. Our aim is to illustrate the RKDG process, especially also in view of practice. We start by reducing a problem in two space dimensions with space variables $(r, \varphi)$ to a problem in one space dimension by applying the $\varphi$-ansatz (4.1.3) and call the resulting set of equations *1D-BOR Maxwell's equations*. Then, the electromagnetic fields are solely dependent on $r$. The system will be discretized in space with the DG method and integrated in time with a low-storage (4,5)-Runge-Kutta method, as were introduced in sections 3.1 and 3.4.2.
As a physical motivation for a two-dimensional problem we consider an infinite waveguide in three space dimensions, as illustrated in figure 4.2.

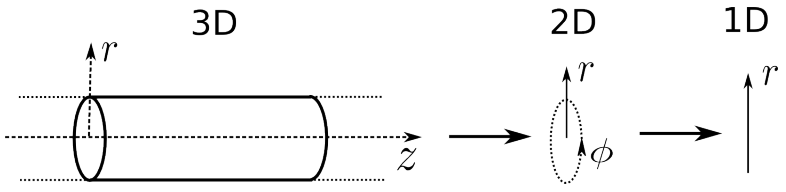



Figure 4.2: An infinite waveguide in three dimensions. The electromagnetic fields are invariant in $z-$direction, and thus the dimension is reduced to two. After the $\varphi$-ansatz the computational dimension is one instead of two.

In this case the electromagnetic fields are invariant in $z$-direction, that is, it is $\mathbf{E}(r, z_1, t) = \mathbf{E}(r, z_2, t)$ for all $z_1, z_2$ on the $z$-axis. Thus, in (4.1.4) all $z$-dependencies and all $z$-

derivatives drop out, and we get 1D-BOR Maxwell's equations as

$$\begin{aligned}
&\epsilon\,\partial_t E_r - \frac{\mathrm{i}m}{r} H_z = 0,\\
&\epsilon\,\partial_t E_\varphi + \partial_r H_z = 0,\\
&\epsilon\,\partial_t E_z - \frac{1}{r} H_\varphi - \partial_r H_\varphi + \frac{\mathrm{i}m}{r} H_r = 0,\\
&\mu\,\partial_t H_r + \frac{\mathrm{i}m}{r} E_z = 0,\\
&\mu\,\partial_t H_\varphi - \partial_r E_z = 0,\\
&\mu\,\partial_t H_z + \frac{1}{r} E_\varphi + \partial_r E_\varphi - \mathrm{i}m E_r = 0.
\end{aligned} \tag{4.2.1}$$

For clarity we have dropped the superscript $(m)$ of the electromagnetic fields, which we will maintain in the following sections.
The following steps need to be taken if we want to apply the RKDG method to the set of equations (4.2.1):

1. We discretize in space by applying the DG method.
   a) Discretize the space with a discontinuous Galerkin ansatz and choose an appropriate finite element space.
   b) Choose a numerical flux so that the resulting DG method is consistent and stable. In our case, this will be an upwind flux.
   c) The basis of the finite element space must be chosen such that the matrices in the resulting semi-discrete scheme are well conditioned.
2. Integrate the semi-discrete scheme in time by using a low-storage Runge-Kutta method (see section 3.4.2).

### 4.2.1 DG Space Discretization

We apply the DG method to discretize (4.2.1) in space. We thus consider $\mathbf{u} = (\mathbf{E}, \mathbf{H})^T \in \mathbb{R}^6$, where $\mathbf{u}$ depends on $(r, t)$ with $r \in \Omega \subset \mathbb{R}$ and $t \geq 0$. As discussed in section 3.1, we divide the interval $\Omega$ in $K$ elements; this renders subintervals $\Omega_k := [r_L^k, r_R^k]$ so that $\Omega = \bigcup_{k=1}^K \Omega_k$. We define the finite element space of discontinuous functions

$$V_h := \{\mathbf{u}_h \in L^\infty(\Omega)^6 : \mathbf{u}_h|_{\Omega_k} \in V(\Omega_k),\ k = 1, \ldots, K\}.$$

$V(\Omega_k)$ is called the local approximation space, and we choose $V(\Omega_k) = \mathcal{P}^p$ with $p \in \mathbb{N}$ to be the space of one-dimensional polynomials of degree at most $p$. We approximate $\mathbf{u}$ by $\mathbf{u}_h \in V_h$, and we use Lagrange interpolation to represent $\mathbf{u}_h|_{\Omega_k}$. This gives the so-called *nodal representation* of the fields (see [11]) as

$$\mathbf{u}_h(r, t) = \sum_{i=1}^{N_p} \mathbf{u}_h^k(r_i^k, t) l_i^k(r), \tag{4.2.2}$$

where $l_i^k$ is the one-dimensional Lagrange polynomial on $\Omega_k$, $N_p = p + 1$ is the number of nodes and $r_i^k$ are suitably chosen interpolation points on $\Omega_k$. For our purposes we will choose Gauss-Lobatto grid points as in [11]. The test functions $\boldsymbol{\psi}$ are also approximated by $\boldsymbol{\psi}_h \in V_h$:

$$\boldsymbol{\psi}(r, t) \approx \boldsymbol{\psi}_h(r, t) = \sum_{i=1}^{N_p} \boldsymbol{\psi}_h^k(r_i^k, t) l_i^k(r). \tag{4.2.3}$$

The Galerkin ansatz requires the residual

$$\mathbf{R}_h := \mathbf{Q}\,\partial_t \mathbf{u}_h + \nabla \cdot \mathbf{F}(\mathbf{u}_h) - \frac{1}{r} B\mathbf{u}_h$$

to be orthogonal to all test functions $\boldsymbol{\psi}_h \in V_h$ with respect to the measure $r\,\mathrm{d}r$ (in two dimensions this would be $r\,\mathrm{d}\mathbf{r}$ with $\mathbf{r} = (r, z)$), i.e.

$$\int_{\Omega_k} \mathbf{R}_h \cdot \boldsymbol{\psi}_h \, r\,\mathrm{d}r = \mathbf{0}. \tag{4.2.4}$$

In section 3.2 we introduced the numerical flux, which is a main ingredient of any DG scheme and which provides a way of enabling the coupling between neighboring elements. Let us repeat an important observation about the hyperbolic system (4.1.5), which can also be motivated from finite volume methods; see (3.2.2) in section 3.2.1. If we integrate (4.1.5) over each $\Omega_k$ and apply integration by parts (Green's Theorem) to the divergence part $\nabla \cdot \mathbf{F}$, i.e.

$$\int_{\Omega_k} (\nabla \cdot \mathbf{F})\, r\,\mathrm{d}r = \int_{\partial\Omega_k} (\mathbf{F}\mathbf{n})\, r\,\mathrm{d}r - \int_{\Omega_k} \mathbf{F}\,\mathbf{e}_r \; r\,\mathrm{d}r,$$

$\mathbf{e}_r$ being the standard basis vector in $r$-direction, this leads to the expression

$$\int_{\Omega_k} (\mathbf{Q}\,\partial_t \mathbf{u} - \frac{1}{r} B\mathbf{u})\, r\,\mathrm{d}r + \int_{\partial\Omega_k} \mathbf{F}(\mathbf{u})\mathbf{n}\, r\,\mathrm{d}r - \int_{\Omega_k} \mathbf{F}(\mathbf{u})\,\mathbf{e}_r\, r\,\mathrm{d}r = \mathbf{0},$$

or, equivalently,

$$\int_{\Omega_k} (\mathbf{Q}\,\partial_t \mathbf{u} - \frac{1}{r} B\mathbf{u})\, r\,\mathrm{d}r = -\int_{\partial\Omega_k} \mathbf{F}(\mathbf{u})\mathbf{n}\, r\,\mathrm{d}r + \int_{\Omega_k} \mathbf{F}(\mathbf{u})\,\mathbf{e}_r\, r\,\mathrm{d}r,$$

meaning that $\mathbf{u}$ changes inside $\Omega_k$ only due to the flux across the boundary (first term on the right hand side) and the consumption inside $\Omega_k$ (second term). Turning back to the approximation $\mathbf{u}_h^k$ of $\mathbf{u}$ on $\Omega_k$, we see that within the discontinuous Galerkin formulation we have

$$\int_{\Omega_k} (\mathbf{Q}\,\partial_t \mathbf{u}_h^k - \frac{1}{r} B\mathbf{u}_h^k) \cdot \boldsymbol{\psi}_h\, r\,\mathrm{d}r = -\int_{\partial\Omega_k} \mathbf{F}^{\mathrm{num}}(\mathbf{u}_h^k)\mathbf{n} \cdot \boldsymbol{\psi}_h\, r\,\mathrm{d}r + \int_{\Omega_k} \mathbf{F}(\mathbf{u}_h^k) \cdot \nabla\boldsymbol{\psi}_h\, r\,\mathrm{d}r$$

with the numerical flux $\mathbf{F}^{\mathrm{num}}$, which is an approximation to the flux vector $\mathbf{F}$ and which has to fulfill the requirements from section 3.2, including continuity on $\partial\Omega_k$. We apply integration by parts to the second term on the right hand side and obtain

$$\begin{aligned}
\int_{\Omega_k} (\mathbf{Q}\,\partial_t \mathbf{u}_h^k + \nabla \cdot \mathbf{F}(\mathbf{u}_h^k) - \frac{1}{r} B\mathbf{u}_h^k) \cdot \boldsymbol{\psi}_h\, r\,\mathrm{d}r = \\
&- \int_{\partial\Omega_k} \mathbf{F}^{\mathrm{num}}(\mathbf{u}_h^k)\mathbf{n} \cdot \boldsymbol{\psi}_h r\,\mathrm{d}\, r + \int_{\partial\Omega_k} \mathbf{F}(\mathbf{u}_h^k)\mathbf{n} \cdot \nabla\boldsymbol{\psi}_h \; r\,\mathrm{d}r \\
&= \int_{\partial\Omega_k} (\mathbf{F}(\mathbf{u}_h^k) - \mathbf{F}^{\mathrm{num}}(\mathbf{u}_h^k))\mathbf{n} \cdot \boldsymbol{\psi}_h\, r\,\mathrm{d}r.
\end{aligned}$$

Thus, using a discontinuous Galerkin ansatz, instead of the residual expression in (4.2.4) we have

$$\int_{\Omega_k} \mathbf{R}_h \cdot \boldsymbol{\psi}_h\, r\,\mathrm{d}\mathbf{r} = -\int_{\partial\Omega_k} (\mathbf{F}(\mathbf{u}_h^k) - \mathbf{F}^{\mathrm{num}}(\mathbf{u}_h))\mathbf{n} \cdot \boldsymbol{\psi}_h\, r\mathrm{d}\mathbf{r}, \tag{4.2.5}$$

where $\mathbf{n}$ denotes the outer normal vector of the boundary $\partial\Omega_k$. Recall that the choice of the numerical flux $\mathbf{F}^{\mathrm{num}}$ plays a crucial role for the stability and accuracy of the scheme,

see 3.2 and section 3.2.1.
For the one-dimensional BOR Maxwell's equations (4.2.1) we shall choose a flux $\mathbf{F}^{\text{num}} = \mathbf{F}^*$, and equation (4.2.5) becomes

$$
\begin{aligned}
&\int_{\Omega_k} \epsilon \partial_t E_r \cdot \psi \; r \,\mathrm{d}r - \int_{\Omega_k} \frac{\mathrm{i}m}{r} H_z \cdot \psi \; r \,\mathrm{d}r = 0,\\
&\int_{\Omega_k} \epsilon \partial_t E_\varphi \cdot \psi \; r \,\mathrm{d}r + \int_{\Omega_k} \partial_r H_z \cdot \psi \; r \,\mathrm{d}r = -\left[(F_{E_\varphi} - F^*_{E_\varphi})\psi\right]_{r_L^k}^{r_R^k},\\
&\int_{\Omega_k} \epsilon \partial_t E_z \cdot \psi \; r \,\mathrm{d}r + \int_{\Omega_k} -\frac{1}{r} H_\varphi - \partial_r H_\varphi + \frac{\mathrm{i}m}{r} H_r) \cdot \psi \; r \,\mathrm{d}r = -\left[(F_{E_z} - F^*_{E_z})\psi\right]_{r_L^k}^{r_R^k},\\
&\int_{\Omega_k} \mu \partial_t H_r \cdot \psi \; r \,\mathrm{d}r + \int_{\Omega_k} \frac{\mathrm{i}m}{r} E_z \cdot \psi \; r \,\mathrm{d}r = 0\\
&\int_{\Omega_k} \mu \partial_t H_\varphi \cdot \psi \; r \,\mathrm{d}r - \int_{\Omega_k} \partial_r E_z \cdot \psi \; r \,\mathrm{d}r = \left[(F_{H_\varphi} - F^*_{H_\varphi})\psi\right]_{r_L^k}^{r_R^k},\\
&\int_{\Omega_k} \mu \partial_t H_z \cdot \psi \; r \,\mathrm{d}r + \int_{\Omega_k} \frac{1}{r} E_\varphi + \partial_r E_\varphi - \mathrm{i}m E_r \cdot \psi \; r \,\mathrm{d}r = \left[(F_{H_z} - F^*_{H_z})\psi\right]_{r_L^k}^{r_R^k},
\end{aligned}
$$

where $\mathbf{F}^*_{\mathbf{E}} := (0, F^*_{E_\varphi}, F^*_{E_z}) \in \mathbb{R}^3, \mathbf{F}^*_{\mathbf{H}} := (0, F^*_{H_\varphi}, F^*_{H_z}) \in \mathbb{R}^3$ are the components of the numerical flux vector $\mathbf{F}^* := (\mathbf{F}^*_{\mathbf{E}}, \mathbf{F}^*_{\mathbf{H}}) \in \mathbb{R}^{3,2}$.
Plugging in the ansatzes (4.2.2) and (4.2.3) we get a first version of the semi-discrete scheme of (4.2.1):

$$
\begin{aligned}
&\epsilon \partial_t \mathbf{E}_r^k \int_{\Omega_k} l_i^k l_j^k \; r \,\mathrm{d}r - \mathrm{i}m \mathbf{H}_z^k \int_{\Omega_k} l_i^k l_j^k \,\mathrm{d}r = 0\\
&\epsilon \partial_t \mathbf{E}_\varphi^k \int_{\Omega_k} l_i^k l_j^k \; r \,\mathrm{d}r + \mathbf{H}_z^k \int_{\Omega_k} l_i^k \left(d_r l_j^k\right) r \,\mathrm{d}r = -\left[r \left(\mathbf{F}_{E_\varphi} - \mathbf{F}^*_{E_\varphi}\right) l_j^k\right]_{r_L^k}^{r_R^k},\\
&\epsilon \partial_t \mathbf{E}_z^k \int_{\Omega_k} l_i^k l_j^k \; r \,\mathrm{d}r - \mathbf{H}_\varphi^k \int_{\Omega_k} \left(l_i^k l_j^k + l_i^k \left(d_r l_j^k\right) r\right) \mathrm{d}r + \mathrm{i}m \mathbf{H}_r^k \int_{\Omega_k} l_i^k l_j^k \,\mathrm{d}r = -\left[r \left(\mathbf{F}_{E_z} - \mathbf{F}^*_{E_z}\right) l_j^k\right]_{r_L^k}^{r_R^k},\\
&\mu \partial_t \mathbf{H}_r^k \int_{\Omega_k} l_i^k l_j^k \; r \,\mathrm{d}r + \mathrm{i}m \mathbf{E}_z^k \int_{\Omega_k} l_i^k l_j^k \,\mathrm{d}r = 0,\\
&\mu \partial_t \mathbf{H}_\varphi^k \int_{\Omega_k} l_i^k l_j^k \; r \,\mathrm{d}r - \mathbf{E}_z^k \int_{\Omega_k} l_i^k (d_r l_j^k) \; r \,\mathrm{d}r = -\left[r \left(\mathbf{F}_{H_\varphi} - \mathbf{F}^*_{H_\varphi}\right) l_j^k\right]_{r_L^k}^{r_R^k},\\
&\mu \partial_t \mathbf{H}_z^k \int_{\Omega_k} l_i^k l_j^k \; r \,\mathrm{d}r + \mathbf{E}_\varphi^k \int_{\Omega_k} \left(l_i^k l_j^k + l_i^k \left(d_r l_j^k\right) r\right) \mathrm{d}r - \mathrm{i}m \mathbf{E}_r^k \int_{\Omega_k} l_i^k l_j^k \; r \,\mathrm{d}r = \left[r \left(\mathbf{F}_{H_z} - \mathbf{F}^*_{H_z}\right) l_j^k\right]_{r_L^k}^{r_R^k},
\end{aligned}
\tag{4.2.6}
$$

where $\mathbf{H}_n^k$ and $\mathbf{E}_n^k$ $(n = r, \varphi, z)$ are the time-dependent coefficient vectors, i.e.

$$(\mathbf{H}_n^k)_i = H_n^k(r_i^k, t) \text{ and } (\mathbf{E}_n^k)_i = E_n^k(r_i^k, t).$$

We have dropped the index $h$ for clarity, but kept $k$ to emphasize the locality of the scheme; furthermore we suppressed the $r$-dependency of the Lagrange polynomials $l_i^k = l_i^k(r)$ and abbreviated the total derivative as $\frac{d}{dr} =: d_r$.

### 4.2.2 A Numerical Flux for BOR Maxwell's Equations

In the semi-discrete scheme (4.2.6) an explicit expression for the numerical flux on the right hand sides is still required. As we mentioned in section 3.2, the numerical flux $\mathbf{F}^{\text{num}}(\mathbf{u}_h)$ connects the different edge values of $\mathbf{u}_h^k$ of a local cell $\Omega_k$, and thus the global approximation $\mathbf{u}_h$ is obtained. In section 3.3 we discussed how to get a numerical flux by

solving a Riemann problem. It was shown that in the finite volume setting with piecewise constant approximations, this choice of a numerical flux produces stable and convergent schemes of order 1, as was shown by Harten et. al. [64], Kuznetsov [65], Crandall and Majda [66]. See also LeVeque [57]. Thus, if we choose piecewise constant polynomials in the DG scheme with a numerical flux that has been chosen to solve a corresponding Riemann problem we obtain a finite volume method that is stable and convergent. This motivates the choice of numerical flux as the solution of a Riemann problem. Furthermore, we have seen in section 3.2 that for DG schemes applied to general hyperbolic equations an optimal convergence rate of $O(h^{p+1})$ for the $L^2$-error was shown in Ref. [14], and for Maxwell's equations (in Cartesian coordinates) in [16]. On the basis of numerical tests we will show in sections 4.3 and 4.5 to get such a convergence rate also for our scheme.
In [96] a numerical flux for dispersive and lossy Maxwell's equations is determined by solving a Riemann problem. In an analogous manner we can compute a numerical flux for the BOR case.

As we are interested in the transport of information across each edge of a local element in direction of the outer normal $\mathbf{n} = (n_r, 0, 0)^T$, pointing to a neighboring cell, we consider $\mathbf{F_n} = \mathbf{Fn}$. In our case it holds

$$\mathbf{F_n} = (-\mathbf{n} \times \mathbf{H}, \mathbf{n} \times \mathbf{E})^T =: (\mathbf{F_E}, \mathbf{F_H})^T.$$

Recalling (4.2.5), we need an expression for $\mathbf{F_n} - \mathbf{F_n^*}$. To do so we solve the corresponding Riemann problem. The eigenvalues of the system matrix of equation (4.1.5) contain important information on the solution of the Riemann problem. They were given as $\lambda_1 = 0$ and $\lambda_{2,3} = \pm\frac{1}{\sqrt{\epsilon\mu}}$ (all three of double multiplicity). The solution of the Riemann problem consists of three waves: The waves associated with the eigenvalues $\lambda_{2,3}$ are shocks with the shock speed $\lambda_L := -1/\sqrt{\epsilon_L\mu_L} = \lambda_2(\epsilon_L, \mu_L)$ and $\lambda_R := 1/\sqrt{\epsilon_R\mu_R} = \lambda_3(\epsilon_R, \mu_R)$, and the wave with the speed $\lambda_1$ is a contact discontinuity with speed 0. Figure 4.3 shows a sketch of the wave composition in the $x-t$-plane. Across each wave the Rankine-Hugoniot jump condition holds, that is

$$\begin{aligned} \text{left to 1}: &\quad \mathbf{F_n}(\mathbf{u}_L) - \mathbf{F_n^{(1)}}(\mathbf{u}_1) = \lambda_L Q_L(\mathbf{u}_L - \mathbf{u}_1), \\ \text{1 to 2}: &\quad \mathbf{F_n^{(1)}}(\mathbf{u}_1) - \mathbf{F_n^{(2)}}(\mathbf{u}_2) = \lambda_1 Q_1(\mathbf{u}_1 - \mathbf{u}_2) = \mathbf{0}, \\ \text{1 to right}: &\quad \mathbf{F_n^{(2)}}(\mathbf{u}_2) - \mathbf{F_n}(\mathbf{u}_R) = \lambda_R Q_R(\mathbf{u}_2 - \mathbf{u}_R), \end{aligned} \tag{4.2.7}$$

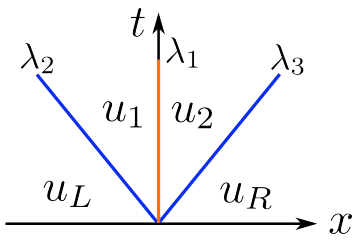



Figure 4.3: Solution of the Riemann problem for Maxwell's equations in the $x-t$-plane.

where $\mathbf{F_n^{(1)}}$ and $\mathbf{F_n^{(2)}}$ are the numerical fluxes in region 1 and 2, respectively; $\mathbf{u}_1$ is the state vector in the 1-region and $\mathbf{u}_2$ the state vector in the 2-region (recall $\mathbf{u} = (\mathbf{E}, \mathbf{H})^T$ for Maxwell's equations); both vectors are unknown. $\mathbf{u}_L$ and $\mathbf{u}_R$ are the given edge values in the interior and exterior of the local cell. Furthermore, we assume $Q_i = \text{diag}(\epsilon_i, \epsilon_i, \epsilon_i, \mu_i, \mu_i, \mu_i)$, where $i = R, L$. From the relations (4.2.7) we get the numerical flux by solving for $\mathbf{F_n}(\mathbf{u}_L) - \mathbf{F_n^{(1)}}(\mathbf{u}_1)$ (recall the expression on the right hand side

of (4.2.5)). The computations are quite similar to those in [11] or [97]. We will give an expression for the numerical flux for the two-dimensional BOR Maxwell's equation (4.1.4) before we come to the one for the one-dimensional case (4.2.1). For the proof we need the following definitions.

**Definition 4.1.**
*(i) We define*

1. *the jump of a scalar as* $[\![f]\!] := f^- - f^+ = f_L - f_R$
2. *and for a vector as* $[\![\mathbf{f}]\!] := \mathbf{n}^- \cdot \mathbf{f}^- + \mathbf{n}^+ \cdot \mathbf{f}^+ = \mathbf{n}_L \cdot \mathbf{f}_L + \mathbf{n}_R \cdot \mathbf{f}_R$.

*From a different point of view, this jump is the difference* $f_L - f_R$ *of the left and right edge value of an element in direction of the outer normal.*

*(ii) We also define the average of a scalar or vector as*

$$\{\{f\}\} := \frac{f^- + f^+}{2} = \frac{f_L + f_R}{2}.$$

**Lemma 4.2.**
*The numerical flux for the two-dimensional BOR Maxwell's equations* (4.1.4) *is given as*

$$\begin{aligned}\mathbf{G_E} &:= (\mathbf{F_E} - \mathbf{F_E^*})\mathbf{n} = -[\mathbf{n} \times \mathbf{H} - (\mathbf{n} \times \mathbf{H}^*)] = -\frac{1}{2\{\{Z\}\}}\mathbf{n} \times [Z_{\mathrm{R}}[\![\mathbf{H}]\!] - \alpha \mathbf{n} \times [\![\mathbf{E}]\!]] \\ \mathbf{G_H} &:= (\mathbf{F_H} - \mathbf{F_H^*})\mathbf{n} = [\mathbf{n} \times \mathbf{E} - (\mathbf{n} \times \mathbf{E}^*)] = \frac{1}{2\{\{Y\}\}}\mathbf{n} \times [Y_{\mathrm{R}}[\![\mathbf{E}]\!] + \alpha \mathbf{n} \times [\![\mathbf{H}]\!]],\end{aligned} \tag{4.2.8}$$

*where "*$\mathrm{R}$*" (or "*$+$*") shall denote the exterior of the local cell, "*$\mathrm{L}$*" ("*$-$*") the interior, thus* $\mathbf{E_L}$ *is the value of the* $\mathbf{E}$*-field in the interior of* $\Omega_k$*,* $\mathbf{E_R}$ *the value in the exterior, and* $[\![\mathbf{E}]\!] := \mathbf{E_L} - \mathbf{E_R}$ *are the field differences at the faces of the elements.* $[\![\mathbf{H}]\!]$ *is defined in an analogous manner. Furthermore, in two dimensions the normal vector is given as* $\mathbf{n} = (n_r, 0, n_z)^T$*;* $Z_{\mathrm{L,R}} = \sqrt{\mu_{\mathrm{L,R}}/\epsilon_{\mathrm{L,R}}}$ *is the local impedance, and* $Y_{\mathrm{L,R}} = 1/Z_{\mathrm{L,R}}$ *is the local conductance, and* $\alpha \in [0, 1]$*. If* $\alpha = 1$ *we have an upwind flux, if* $\alpha = 0$ *it is a central flux. We note this flux looks like the Cartesian flux as given in Ref. [11].*

*Proof.* Recall section 4.1.4 and equation (4.1.5):

$$Q\,\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \frac{1}{r} B\mathbf{u}.$$

We define the matrix $\mathcal{A} := n_r A_r + n_z A_z$; it holds

$$\mathcal{A}\mathbf{u} = \begin{pmatrix}(\mathcal{A}\mathbf{u})_E \\ (\mathcal{A}\mathbf{u})_H\end{pmatrix} = \begin{pmatrix} n_z H_\varphi \\ n_r H_z - n_z r H_r \\ -n_r H_\varphi \\ -n_z E_\varphi \\ n_z E_r - n_r E_z \\ n_r E_\varphi \end{pmatrix} = \begin{pmatrix} -\mathbf{n} \times \mathbf{H} \\ \mathbf{n} \times \mathbf{E} \end{pmatrix}.$$

We observe $\mathbf{F_n} = \mathcal{A}\mathbf{u}$. For the time being we abbreviate $\mathbf{F_n}(\mathbf{u}_L) = \mathbf{F}_L$, $\mathbf{F_n}(\mathbf{u}_R) = \mathbf{F}_R$, $\mathbf{F_n^{(1)}}(\mathbf{u}_1) = \mathbf{F}^{(1)}$ and $\mathbf{F_n^{(2)}}(\mathbf{u}_2) = \mathbf{F}^{(2)}$. The Rankine-Hugoniot jump conditions (4.2.7) give

$$\begin{aligned}&(1) \quad \mathbf{F}_L - \mathbf{F}^{(1)} = \lambda_L Q_L(\mathbf{u}_L - \mathbf{u}_1), \\ &(2) \quad \mathbf{F}^{(1)} = \mathbf{F}^{(2)}, \text{ and } \mathbf{u}_1 = \mathbf{u}_2 \quad (\text{since } \lambda_1 = 0), \\ &(3) \quad \mathbf{F}^{(2)} - \mathbf{F}_R = \lambda_R Q_R(\mathbf{u}_2 - \mathbf{u}_R).\end{aligned}$$

In order to obtain an expression for $\mathbf{F_E} - \mathbf{F_E^*}$, that is, for the numerical flux in the star region (consisting of region 1 and 2), we solve the Rankine-Hugoniot conditions for $\mathbf{F}_{\mathbf{E}_L} - \mathbf{F}_{\mathbf{E}}^{(1)}$.
Remember that $\mathbf{F}_L = \mathcal{A}\mathbf{u}_L$, $\mathbf{F}_R = \mathcal{A}\mathbf{u}_R$ and $\mathbf{F}^{(1)} = \mathcal{A}\mathbf{u}_1$. Thus, we have

$$(\mathcal{A}\mathbf{u}_L) - (\mathcal{A}\mathbf{u})^{(1)} = \lambda_L Q_L(\mathbf{u}_L - \mathbf{u}_1),$$
$$(\mathcal{A}\mathbf{u})^{(1)} - (\mathcal{A}\mathbf{u}_R) = \lambda_R Q_R(\mathbf{u}_1 - \mathbf{u}_R).$$

We solve the second equation for $\mathbf{u}_1$,

$$\mathbf{u}_1 = \frac{1}{\lambda_R} Q_R^{-1}\left[(\mathcal{A}\mathbf{u})^{(1)} - (\mathcal{A}\mathbf{u}_R)\right] + \mathbf{u}_R$$

and plug this into the first equation:

$$(\mathcal{A}\mathbf{u}_L) - (\mathcal{A}\mathbf{u})^{(1)} = \lambda_L Q_L(\mathbf{u}_L - \frac{1}{\lambda_R} Q_R^{-1}[(\mathcal{A}\mathbf{u})^{(1)} - (\mathcal{A}\mathbf{u}_R)] - \mathbf{u}_R).$$

Since $Q$ is diagonal with constant entries $\epsilon$ and $\mu$, we can rearrange in terms in the following manner to obtain

$$\begin{aligned} &- (\mathcal{A}\mathbf{u})^{(1)} + \frac{\lambda_L}{\lambda_R} Q_L Q_R^{-1}(\mathcal{A}\mathbf{u})^{(1)} = \lambda_L Q_L(\mathbf{u}_L - \mathbf{u}_R) + \frac{\lambda_L}{\lambda_R} Q_L Q_R^{-1}(\mathcal{A}\mathbf{u}_R) - (\mathcal{A}\mathbf{u}) \quad | \cdot \lambda_R Q_R \\ \Leftrightarrow\ &(\lambda_L Q_L - \lambda_R Q_R)(\mathcal{A}\mathbf{u})^{(1)} = \lambda_L \lambda_R Q_L Q_R(\mathbf{u}_L - \mathbf{u}_R) + \lambda_L Q_L(\mathcal{A}\mathbf{u}_R) - \lambda_R Q_R(\mathcal{A}\mathbf{u}_L). \end{aligned}$$

Furthermore, we note that

$$\lambda_L Q_L = -\operatorname{diag}(Z_\mathrm{L}, Z_\mathrm{L}, Z_\mathrm{L}, Y_\mathrm{L}, Y_\mathrm{L}, Y_\mathrm{L}),$$

$$\lambda_R Q_R = \operatorname{diag}(Z_\mathrm{R}, Z_\mathrm{R}, Z_\mathrm{R}, Y_\mathrm{R}, Y_\mathrm{R}, Y_\mathrm{R}),$$
and thus
$$(\lambda_L Q_L - \lambda_R Q_R)^{-1} = -\operatorname{diag}\left(\frac{1}{2\{\{Z\}\}}, \frac{1}{2\{\{Z\}\}}, \frac{1}{2\{\{Z\}\}}, \frac{1}{2\{\{Y\}\}}, \frac{1}{2\{\{Y\}\}}, \frac{1}{2\{\{Y\}\}}\right),$$

$$\lambda_L Q_L \lambda_R Q_R = -\operatorname{diag}(Z_\mathrm{L}Z_\mathrm{R}, Z_\mathrm{L}Z_\mathrm{R}, Z_\mathrm{L}Z_\mathrm{R}, Y_\mathrm{L}Y_\mathrm{R}, Y_\mathrm{L}Y_\mathrm{R}, Y_\mathrm{L}Y_\mathrm{R}).$$

Noting $Z_\mathrm{L}Z_\mathrm{R}(\mathbf{E}_L - \mathbf{E}_R) = [\![\mathbf{E}]\!]$ and $Y_\mathrm{L}Y_\mathrm{R}(\mathbf{H}_L - \mathbf{H}_R) = [\![\mathbf{H}]\!]$, it follows:

$$(\mathcal{A}\mathbf{u})^{(1)} = \begin{pmatrix} \mathbf{F}_\mathbf{E}^{(1)} \\ \mathbf{F}_\mathbf{H}^{(1)} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\{\{Z\}\}}\left[-[\![\mathbf{E}]\!] - Z_L(\mathcal{A}\mathbf{u}_R)_E - Z_R(\mathcal{A}\mathbf{u}_L)_E\right] \\ -\frac{1}{2\{\{Y\}\}}\left[-[\![\mathbf{H}]\!] - Y_L(\mathcal{A}\mathbf{u}_R)_E - Y_R(\mathcal{A}\mathbf{u}_L)_H\right]. \end{pmatrix}$$

We observe

$$Z_L(\mathcal{A}\mathbf{u}_R)_E + Z_R(\mathcal{A}\mathbf{u}_L)_E = Z_L(-\mathbf{n}\times\mathbf{H}_R) + Z_R(-\mathbf{n}\times\mathbf{H}_L) = -\mathbf{n}\times 2\{\{Z\mathbf{H}\}\},$$

so that

$$\begin{aligned} \begin{pmatrix} \mathbf{F}_\mathbf{E}^{(1)} \\ \mathbf{F}_\mathbf{H}^{(1)} \end{pmatrix} &= \begin{pmatrix} \frac{1}{2\{\{Z\}\}}\left[[\![\mathbf{E}]\!] - 2\mathbf{n}\times\{\{Z\mathbf{H}\}\}\right] \\ \frac{1}{2\{\{Y\}\}}\left[[\![\mathbf{H}]\!] + 2\mathbf{n}\times\{\{Y\mathbf{E}\}\}\right] \end{pmatrix} = \begin{pmatrix} \frac{1}{2\{\{Z\}\}}\left[[\![\mathbf{E}]\!] - 2\mathbf{n}\times\{\{Z\mathbf{H}\}\}\right] \\ \frac{1}{2\{\{Y\}\}}\left[[\![\mathbf{H}]\!] + 2\mathbf{n}\times\{\{Y\mathbf{E}\}\}\right] \end{pmatrix} \\ &= \begin{pmatrix} -\mathbf{n}\times\frac{1}{2\{\{Z\}\}}\left[2\{\{Z\mathbf{H}\}\} + \mathbf{n}\times[\![\mathbf{E}]\!]\right] \\ \mathbf{n}\times\frac{1}{2\{\{Y\}\}}\left[2\{\{Y\mathbf{E}\}\} - \mathbf{n}\times[\![\mathbf{H}]\!]\right] \end{pmatrix}. \end{aligned}$$

At last we have

$$\begin{aligned}\mathbf{F}_{\mathbf{E}_L} - \mathbf{F}_{\mathbf{E}}^{(1)} &= (-\mathbf{n}\times\mathbf{H}_L) - \mathbf{n}\times\frac{1}{2\{\{Z\}\}}[-2\{\{Z\mathbf{H}\}\} - \mathbf{n}\times[\![\mathbf{E}]\!]] \\ &= -\mathbf{n}\times\frac{1}{2\{\{Z\}\}}(2\{\{Z\}\}\mathbf{H}_L - 2\{\{Z\mathbf{H}\}\} - \mathbf{n}\times[\![\mathbf{E}]\!]) \\ &= -\frac{1}{2\{\{Z\}\}}\mathbf{n}\times(Z_R[\![\mathbf{H}]\!] - \mathbf{n}\times[\![\mathbf{E}]\!]).\end{aligned}$$

An analog computation for $\mathbf{F}_{\mathbf{H}_L} - \mathbf{F}_{\mathbf{H}}^{(1)}$ gives the final result. □

### 4.2.3 A Numerical Flux for 1D-BOR Maxwell's Equations

For the 1D-BOR case, the unit normal is $\mathbf{n} = (n_r, 0, 0)^T$, and thus the numerical flux expression for (4.2.1) becomes

$$\begin{aligned}\mathbf{G}_{\mathbf{E}} &= -[\mathbf{n}\times\mathbf{H} - (\mathbf{n}\times\mathbf{H})^*] = -\frac{1}{2\{\{Z\}\}}\begin{pmatrix}0\\ -Z_R n_r[\![H_z]\!] + \alpha[\![E_\varphi]\!]\\ Z_R n_r[\![H_\varphi]\!] + \alpha[\![E_z]\!]\end{pmatrix}, \\ \mathbf{G}_{\mathbf{H}} &= [\mathbf{n}\times\mathbf{E} - (\mathbf{n}\times\mathbf{E})^*] = \frac{1}{2\{\{Y\}\}}\begin{pmatrix}0\\ -Y_R n_r[\![E_z]\!] - [\![H_\varphi]\!]\\ Y_R n_r[\![E_\varphi]\!] - [\![H_z]\!]\end{pmatrix}.\end{aligned} \tag{4.2.9}$$

### 4.2.4 Semi-Discrete Scheme

In the following, we denote the discrete version of the numerical flux (4.2.9) as $\mathbb{G}_E$ and $\mathbb{G}_H$. The components are $\mathbb{G}_E = (\mathbb{G}_{E_r}, \mathbb{G}_{E_\varphi}, \mathbb{G}_{E_z})^T$ (analogously for $\mathbb{G}_H$). By defining the local matrices

$$\begin{aligned}(M_r^k)_{ij} &:= \int_{\Omega_k} l_i^k(r) l_j^k(r)\, r\,\mathrm{d}r, \\ (S_r^k)_{ij} &:= \int_{\Omega_k} r\, l_i^k(r)(d_r l_j^k(r))\,\mathrm{d}r, \\ (M^k)_{ij} &:= \int_{\Omega_k} l_i^k(r) l_j^k(r)\, r\,\mathrm{d}r, \\ \mathcal{F}_{ij}^k &:= \int_{\partial\Omega_k} l_i^k(r)\, l_j^k(r)\, r\,\mathrm{d}r,\end{aligned} \tag{4.2.10}$$

we can rewrite (4.2.6) as

$$\begin{aligned}&\epsilon M_r^k \partial_t \mathbf{E}_r^k - \mathrm{i}m M^k \mathbf{H}_z^k = 0, \\ &\epsilon M_r^k \partial_t \mathbf{E}_\varphi^k + \left(S_r^k \mathbf{H}_z^k - \mathcal{F}^k \mathbb{G}_{E_\varphi}\right) = 0, \\ &\epsilon M_r^k \partial_t \mathbf{E}_z^k - (M_r^k + S_r^k)\mathbf{H}_\varphi^k - \mathcal{F}^k \mathbb{G}_{E_z} + \mathrm{i}m M^k \mathbf{H}_r^k = 0. \\ &\mu M_r^k \partial_t \mathbf{H}_r^k + \mathrm{i}m M^k \mathbf{E}_z^k = 0, \\ &\mu M_r^k \partial_t \mathbf{H}_\varphi^k - \left(S_r^k \mathbf{E}_z^k - \mathcal{F}^k \mathbb{G}_{H_\varphi}\right) = 0, \\ &\mu M_r^k \partial_t \mathbf{H}_z^k + (M_r^k + S_r^k)\mathbf{E}_\varphi^k - \mathcal{F}^k \mathbb{G}_{H_z} - \mathrm{i}m M^k \mathbf{E}_r^k = 0.\end{aligned} \tag{4.2.11}$$

We call $M_r^k$ the *BOR mass matrix*, $S_r^k$ the *BOR stiffness matrix* and $M^k$ the *local mass matrix*; $\mathcal{F}^k$ is the face matrix (see e.g. [11, Ch. 3]). With the definitions

$$\mathbb{H}^k := \begin{pmatrix} \mathbf{H}_r^k \\ \mathbf{H}_\varphi^k \\ \mathbf{H}_z^k \end{pmatrix}, \quad \mathbb{E}^k := \begin{pmatrix} \mathbf{E}_r^k \\ \mathbf{E}_\varphi^k \\ \mathbf{E}_z^k \end{pmatrix},$$

$$\mathbb{L}_{\mathrm{BOR}} := \begin{pmatrix} 0 & 0 & im(M_r^k)^{-1}M^k \\ 0 & 0 & -(M_r^k)^{-1}S_r^k \\ -im(M_r^k)^{-1}M^k & (M_r^k)^{-1}(M^k + S_r^k) & 0 \end{pmatrix}$$

and

$$(M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_E = ((M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_{E_r}, (M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_{E_\varphi}, (M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_{E_z})^T$$

(provided $(M_r^k)^{-1}$ is invertible which we assume in our numerical setting, as we will see), we can write the semi-discrete scheme (4.2.11) in short vector-matrix notation as

**Semi-Discrete Scheme in 1D**

$$\begin{aligned} \partial_t \mathbb{E}^k &= \frac{1}{\epsilon}\left(\mathbb{L}_{\mathrm{BOR}}\mathbb{H}^k + (M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_H\right), \\ \partial_t \mathbb{H}^k &= \frac{1}{\mu}\left(-\mathbb{L}_{\mathrm{BOR}}\mathbb{E}^k + (M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_E\right). \end{aligned} \tag{4.2.12}$$

This semi-discrete scheme is integrated in time with a low-storage Runge-Kutta method of order 4 with 5 stages, as was discussed in section 3.4.2.
In principle, one can evaluate the matrices (4.2.10) by using a suitable quadrature rule for each element $\Omega_k$. Recalling section 3.4.2 this quadrature rule would have to be of order $2p+1$ (where $p$ is the polynomial order) in order to achieve a convergence rate of $p+1$ for the RKDG scheme. From those matrices, one can then easily pre-compute and store the matrices (4.2.10) which are required to set up the semi-discrete scheme (4.2.11). However, such a quadrature-based approach requires the storage of four matrices for each element. This is not only demanding in terms of memory, it also negates much of the advantage that makes the discontinuous Galerkin time-domain approach so attractive for implementation. In the next section, we will present an alternative procedure to calculate the local matrices (4.2.10), which at least partially overcomes the shortcomings of a quadrature-based approach.

### 4.2.5 Efficient Computation of the Local Matrices

A significant advantage of the RKDG approach in Cartesian coordinates is that all local matrices (4.2.10) can be expressed in terms of a few global template matrices. Here, we will demonstrate how to achieve something similar despite the explicit $r$-dependence.

#### Transformation to a Reference Element

All operations are carried out on a reference element $I$, not on the physical domain $\Omega_k$. The transformation between $I := [-1, 1]$ and $\Omega_k$ is given by the map $\Psi : I \to \Omega^k$ so that we can express $r \in \Omega_k$ by

$$r(x) = r_L^k + \frac{1}{2}(r_R^k - r_L^k)(1+x) =: \Psi(x) \text{ for } x \in I. \tag{4.2.13}$$

It is

$$\det(J_\Psi) = \frac{1}{2}(r_R^k - r_L^k) =: \frac{h^k}{2},$$

where by $J_\Psi$ we mean the Jacobian of the affine mapping $\Psi$.

### Choosing a Basis

As already mentioned earlier, the nodal representation of $\mathbf{u} = (\mathbf{E}, \mathbf{H})^T$ is expressed as

$$\mathbf{u}_h^k(r(x), t) = \sum_{i=1}^{N_p} \mathbf{u}_h^k(r_i^k, t) l_i(x), \tag{4.2.14}$$

where the $r_i^k$ are suitable grid points on $\Omega_k$, $x \in I$, and $l_i$ are Lagrange polynomials defined on $I$.
We can express the fields also by a so-called *modal representation* (see [11]), which uses another basis $\{B_n\}_{n=1}^{N_p}$ of $V_h$, to be determined afterwards:

$$\mathbf{u}_h^k(x, t) = \sum_{i=1}^{N_p} \tilde{\mathbf{u}}_i^k(t) B_i(x), \tag{4.2.15}$$

where the $\tilde{\mathbf{u}}_i^k$ are the expansion coefficients. We want to find a transformation matrix that exhibits the change of bases from $\{l_n\}_{n=1}^{N_p}$ to $\{B_n\}_{n=1}^{N_p}$. As a special case we can choose $x = x_i$ in (4.2.14), where $x_i$ are suitable grid points on $I$, and obtain by combining (4.2.14) with (4.2.15):

$$\mathbf{u}_h^k(x_i, t) = \sum_{j=1}^{N_p} \tilde{\mathbf{u}}_j^k B_j(x_i) = \sum_{j=1}^{N_p} \mathbf{u}_h^k(r_j^k, t) l_j(x_i). \tag{4.2.16}$$

A basic property of the Lagrange polynomials is the fact, that in the grid points they are exactly one, that is, $l_j(x_i) = \delta_{ij}$, where $\delta_{ij}$ is the Kronecker delta. Thus, $l_j(x_i)$ is only non-vanishing for $i = j$, and we get

$$\sum_{j=1}^{N_p} \tilde{\mathbf{u}}_j^k B_j(x_i) = \mathbf{u}_h^k(r_i^k, t), \tag{4.2.17}$$

or written in matrix-vector notation:

$$\mathbf{u}_h = V \tilde{\mathbf{u}}_h,$$

where $\mathbf{u}_h := (\mathbf{u}_h^k(r_i^k, t))_{i=1}^{N_p}$, $\tilde{\mathbf{u}}_h := (\tilde{\mathbf{u}}_h^{(i)})_{i=1}^{N_p}$, and $V_{ij} := B_j(x_i)$ is the so-called *generalized Vandermonde matrix*. $V$ describes the transformation between the two bases, as we can see in the following relation:

$$B_i(x) = \sum_{j=1}^{N_p} B_i(x_j) l_j(x) = \sum_{j=1}^{N_p} V_{ji} l_j(x) = \sum_{j=1}^{N_p} (V_{ij})^T l_j(x), \tag{4.2.18}$$

or, equivalently, $\mathbf{B} = V^T \mathbf{l}$, where $\mathbf{B} = (B_i)_{i=1}^{N_p}, \mathbf{l} = (l_i)_{i=1}^{N_p}$. Since $V$ is a transformation matrix of two bases, it is invertible, and it holds $\mathbf{l} = (V^T)^{-1}\mathbf{B}$. Let us briefly collect this result in the following lemma.

**Lemma 4.3.**
*Let $\{l_n\}_n$ be the basis of Lagrange polynomials of $V_h$ and $\{B_n\}_n$ another basis of polynomials. The transformation between the two bases is performed by the generalized Vandermonde matrix $V$ via*

$$V^T \mathbf{l}(x) = \mathbf{B}(x)$$

*with $V_{ij} := B_j(x_i)$, or in sum notation: $l_i(x) = \sum_{k=1}^{N_p} (V_{ki})^{-T} B_k(x)$.*

We are left with the open question how to choose the basis $\{B_n\}_n$. To answer this question, let us look at the local mass matrix $M^k$, defined by (4.2.10) as

$$(M^k)_{ij} := \int_{\Omega_k} l_i^k(r) l_j^k(r) \, \mathrm{d}r.$$

Transforming to $I$, we get:

$$(M^k)_{ij} = J^k \int_I l_i(x) l_j(x) \, \mathrm{d}x =: J^k M_{ij}, \tag{4.2.19}$$

where $M$ is the global mass matrix, independent of the element $\Omega_k$. Now we change to the basis $\{B_n\}_n$:

$$\begin{aligned} M_{ij} &= \int_I l_i(x) l_j(x) \, \mathrm{d}x = \int_I \sum_{k=1}^{N_p} (V_{ik})^{-1} B_k(x) \sum_{m=1}^{N_p} (V_{jm})^{-1} B_m(x) \, \mathrm{d}x \\ &= \sum_{k=1}^{N_p} \sum_{m=1}^{N_p} (V_{ki})^{-T} (V_{mj})^{-T} \int_I B_k(x) B_m(x) \, \mathrm{d}x. \end{aligned} \tag{4.2.20}$$

It is well known that choosing $\{B_n\}_n$ to be a monomial basis leads to an ill-conditioned mass matrix $M$, which also is true for the other local matrices in (4.2.10). We need to choose another basis. From (4.2.20) we see that, if we choose an orthonormal basis, it holds $M = (VV^T)^{-1}$. $V$ will be invertible also numerically, and $M$ is well-conditioned. We follow [11, Ch. 3.1] (and the references therein) in order to obtain an orthogonal basis. Applying the Gram-Schmidt process to the monomial basis with respect to a weighted scalar product gives an orthonormal basis of Jacobi polynomials. For basic properties of the Jacobi polynomials in one dimension, see e.g. [98]. We thus have (see [11])

$$B_n = \hat{P}_{n-1}^{(\alpha,\beta)} = \frac{P_{n-1}^{(\alpha,\beta)}}{\sqrt{c_{n-1}}}, \tag{4.2.21}$$

where the $P_n^{(\alpha,\beta)}$ are the one-dimensional Jacobi polynomials which are orthogonal with respect to the weight function $w(x) = (1-x)^\alpha (1+x)^\beta$ with $\alpha, \beta > -1$ (see e.g. [98]) and $c_n := \frac{2}{2n+1}$.
We are now able to show how to compute the local matrices (4.2.10).

### Computation of the Mass Matrices

**Lemma 4.4** (Mass Matrix)**.**
*The local mass matrix $M^k$ can be computed as $M^k = J^k M$, where $J^k = \det(J_\Psi)$ is the determinant of the transformation from $I$ to $\Omega_k$. $M$ is the global mass matrix, which is independent of each element $\Omega_k$; considering implementation, it thus has to be stored only once. It can be computed efficiently and stably by using orthogonal Jacobi-polynomials by*

$$M = (VV^T)^{-1},$$

*where $V$ is the Vandermonde matrix given as*

$$V_{ij} = P_{j-1}^{(0,0)}(x_i).$$

*Here, $P_n^{(0,0)}$ are the one-dimensional Jacobi polynomials with $\alpha = \beta = 0$, which are orthogonal with respect to the weight function $w(x) \equiv 1$. The $x_i$ are Gauss-Lobatto quadrature points on $I$; see e.g. [81, Ch. 8], [80, Ch. 10].*

*Proof.* We only give the main aspects of the proof; all the details can be found in [11]. In (4.2.19) and (4.2.20) we have already seen that it is $M^k = J^k M$ and $M = (VV^T)^{-1}$. Recalling (4.2.20) again, we also encounter that

$$M_{ij} = \int_I l_i(x) l_j(x)\,\mathrm{d}x = \sum_{k=1}^{N_p}\sum_{m=1}^{N_p} (V_{ki})^{-T}(V_{mj})^{-T} \int_I \hat{P}_{k-1}^{(\alpha,\beta)}(x)\hat{P}_{m-1}^{(\alpha,\beta)}(x)\,\mathrm{d}x.$$

The Jacobi polynomials have to be orthogonal with respect to the weight function $w^{(0,0)}(x) = 1$, i.e. it is $\alpha = \beta = 0$, and thus $V_{ij} = \hat{P}_{j-1}^{(0,0)}(x_i)$. □

**Lemma 4.5** (The BOR Mass Matrix)**.**
*The BOR mass matrix $M_r^k$ can be computed as*

$$M_r^k = J^k(r_L^k M + J^k M_r),$$

*where*

$$(M_r)_{ij} := \int_{-1}^{1} (1+x) l_i(x) l_j(x)\,\mathrm{d}x$$

*is the global BOR mass matrix, and it can be computed similarly to the mass matrix by using Jacobi polynomials, but this time with $\alpha = 0, \beta = 1$, that is*

$$M_r = (V_1 V_1^T)^{-1}, \;\; (V_1)_{ij} := P_{j-1}^{(0,1)}(x_i).$$

*Proof.* Transforming to the reference element gives

$$\begin{aligned}(M_r^k)_{ij} &= \int_{\Omega_k} r l_i^k(r) l_j^k(r)\,\mathrm{d}r = J^k \left( \int_{-1}^{1} [r_L^k + \frac{r_R^k - r_L^k}{2}(1+x)]\, l_i(x) l_j(x)\,\mathrm{d}x \right) \\ &= J^k(r_L^k M_{ij} + J^k (M_r)_{ij}).\end{aligned}$$

We need to choose Jacobi polynomials that are orthogonal with respect to the weight function $w^{(0,1)}(x) = 1 + x$, i.e. we have $\alpha = 0, \beta = 1$. Then we proceed as in case of the mass matrix so that we get the Vandermonde matrix $(V_1)_{ij} = \hat{P}_{j-1}^{(0,1)}(x_i)$. □

### Computation of the BOR Stiffness Matrix

**Lemma 4.6.**
*The BOR stiffness matrix can also be computed as the composition of template matrices, i.e.*

$$S_r^k = r_L^k S + J^k S_r,$$

*where $S$ is the stiffness matrix,*

$$S_{ij} = \int_I l_i\,(d_x l_j)\,\mathrm{d}x,$$

*and $S_r$ the BOR stiffness matrix from* (4.2.10)*. Furthermore it holds $S = MD_r$, with $D_r$ being the differentiation matrix as given in [11, Ch. 3.2], that is*

$$(D_r)_{ij} = \left.\frac{dl_j(x)}{dx}\right|_{x_i}.$$

*$D_r$ can also be computed by using the Vandermonde matrix, i.e. $D_r = V_x V^{-1}$ with*

$$(V_x)_{ij} = \left.\frac{\mathrm{d}\hat{P}_{j-1}^{(\alpha,\beta)}}{\mathrm{dx}}\right|_{\mathbf{x}_i}.$$

*The matrix $S_r$ is the global BOR stiffness matrix,*

$$(S_r)_{ij} := \int_{-1}^{1} (1+x) l_i(x) \left(d_x l_j(x)\right) \mathrm{d}x,$$

*and analogously, it is $S_r = M_r D_r$.*

**Remark 4.7.**

1. *We use $r$ as a subscript in $M_r, S_r$ and $D_r$ in analogy with the derivative $\mathrm{d}_r$ in BOR Maxwell's equations, although to set up $D_r$ the $x-$derivatives of $l_i$ are taken and the orthonormal Jacobi polynomials $\hat{P}_n^{(\alpha,\beta)}$ are evaluated in Gauss-Lobatto grid points $x_i$ on $I$.*
2. *In the semi-discrete equations* (4.2.11)*, we need to multiply by $(M_r^k)^{-1}$ and thus obtain an ordinary differential equation in time* (4.2.12)*. This means we do not have to compute the global BOR stiffness matrix $S_r$ at all, since*

$$(M_r^k)^{-1} S_r^k = r_L^k (M_r^k)^{-1}(MD_r) + (M_r^k)^{-1}(J^k M_r) D_r = r_L^k (M_r^k)^{-1}(MD_r) + D_r.$$

*Proof.* As before, transforming to $I = [-1, 1]$ gives

$$\begin{aligned} S_r^k &= \int_{\Omega_k} l_i^k(r)\, d_r l_j^k(r)\, r \,\mathrm{d}r \\ &= J^k \Big( \int_{-1}^{1} l_i(x) \frac{2}{r_R^k - r_L^k} d_x l_j(x) \left[ r_L^k + \frac{r_R^k - r_L^k}{2}(1+x) \right] \mathrm{d}x \Big) \\ &= J^k \frac{1}{J^k} (x_l^k S_{ij} + J^k (S_r)_{ij}). \end{aligned}$$

The factor $\frac{1}{J^k} = \frac{2}{r_R^k - r_L^k}$ comes from the transformation of the derivative $d_r$ to $d_x$.
What remains to be shown is the relation $S_r = M_r D_r$. The statement $S = MD_r$ is proven in [11, Ch. 3.2]. We proceed in an analogous fashion. It is

$$(S_r)_{ij} = \int_{-1}^{1} (1+x) l_i(x)\, d_x l_j(x)\, \mathrm{d}x.$$

Expanding $d_x l_j(x)$ in terms of Lagrange polynomials as

$$d_x l_j(x) = \sum_{n=1}^{N_p} d_x l_j(x_n)\, l_n(x)$$

and inserting this expansion together with the Jacobi polynomials in the matrix $S_r$ gives

$$\begin{aligned} (S_r)_{ij} &= \int_{-1}^{1} (1+x) l_i(x)\, d_x l_j(x)\, \mathrm{d}x = \sum_{n=1}^{N_p} (d_x l_j(x_n)) \int_{-1}^{1} (1+x)\, l_i(x)\, l_n(x)\, \mathrm{d}x \\ &= \sum_{n=1}^{N_p} (M_r)_{in} (D_r)_{nj}. \end{aligned}$$

□

### Computation of the Face Matrix

If we recall the semi-discrete scheme (4.2.12), there remains the computation of the expressions $\mathcal{F}^k\mathbb{G}_E$ and $\mathcal{F}^k\mathbb{G}_H$, that is (note that $\partial\Omega_k = \{r_L^k, r_R^k\}$),

$$\int_{\partial\Omega_k} (\mathbf{F}_E - \mathbf{F}_E^*)(r)\, l_j^k(r)\, r\,\mathrm{d}r = \Big[ r\,(\mathbf{F}_E - \mathbf{F}_E^*)(r)\, l_j^k(r) \Big]_{r_L^k}^{r_R^k}. \tag{4.2.22}$$

The right hand side can be computed directly by evaluating the expression in the left and right boundary values $r_L^k, r_R^k$ of the element $\Omega_k = [r_L^k, r_R^k]$. The same is true for $\mathbf{F}_H - \mathbf{F}_H^*$. In the one-dimensional case the face matrix is thus easily computable at low cost. In addition, we want to give some details of the construction of the face matrix with respect to implementation issues.
On each element $\Omega_k$ we have $N_p = p + 1$ nodes $r_1^k, \ldots, r_{N_p}^k$, where $p$ is the polynomial order, and in total we have $K$ elements. We collect all grid points in an $N_p \times K$-array

$$\mathbf{r} := \begin{pmatrix} r_1^1 & r_1^2 & \cdots & r_1^K \\ \vdots & \vdots & \vdots & \vdots \\ r_{N_p}^1 & r_{N_p}^2 & \cdots & r_{N_p}^K \end{pmatrix}. \tag{4.2.23}$$

Thus the edge nodes on element number $k$ are $r_L^k = r_{N_p}^k$ and $r_R^k = r_{N_p}^k$. Let us denote by $\mathbf{G}_E^{(k)}$ the numerical flux on element number $k$. The we get for $j = 1, \ldots, N_p$:

$$\begin{aligned} \Big[ r\, \mathbf{G}_E^{(k)}(r) l_j^k(r) \Big]_{r_L^1}^{r_R^1} &= \Big[ r\, l_j^k(r) \Big]_{r_1^k}^{r_{N_p}^k} \mathbf{G}_E^{(k)}(r_{N_p}^k - r_1^k) \\ &= (r_{N_p}^k\, l_j^k(r_{N_p}^k) - r_1^k\, l_j^k(r_1^k))\; \mathbf{G}_E^{(k)}(r_{N_p}^k - r_1^k), \end{aligned}$$

where by $\mathbf{G}_E^{(k)}(r_{N_p}^k - r_1^k)$ we mean the numerical flux (4.2.9) for 1D-BOR Maxwell's equations, evaluated at the difference $r_{N_p}^k - r_1^k$ of the edge points on element $\Omega_k$. Recalling $l_j^k(r_i) = \delta_{ij}$, we thus obtain in matrix-vector notation

$$\Big( r_{N_p}^k\, l_j^k(r_{N_p}^k) - r_1^k\, l_j^k(r_1^k) \Big)_{j=1}^{N_p} = \begin{pmatrix} -1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} r_1^k \\ r_{N_p}^k \end{pmatrix} =: \mathcal{L}^k \begin{pmatrix} r_1^k \\ r_{N_p}^k \end{pmatrix}. \tag{4.2.24}$$

The matrix $\mathcal{L}^k$ in (4.2.24) is called Lift matrix and has dimension $N_p \times 2$. Each component of the discretized numerical flux vector $\mathbb{G}_E$ in (4.2.12) is a $2 \times K$-array. We can define a $2 \times K$-vector $\mathbf{F}_e$ containing all edge points of all $K$ elements, i.e.

$$\mathbf{F}_e := \begin{pmatrix} r_1^1 & r_1^2 & \cdots & r_1^K \\ r_{N_p}^1 & r_{N_p}^2 & \cdots & r_{N_p}^K \end{pmatrix}.$$

We thus obtain

$$\mathcal{F}^k \mathbf{G}_E = \mathcal{L}^k \mathbf{F}_\mathbf{e}.$$

For more implementation issues and codes in Matlab, we refer to [11].

## 4.3 Numerical Tests

Now that we have an RKDG scheme for 1D-BOR Maxwell's equations (4.2.1), we test the performance of the method. We used some of the Matlab-codes in the book by Hesthaven and Warburton [11], some were modified for our case.

### 4.3.1 Homogeneous Waveguide

Let us first assume to have a homogeneous waveguide, i.e. $\epsilon = 1$, $\mu = 1$. The exact solution of this system is known (e.g. [99, Ch. 9.5.2], [100]). The cylinder shall have a radius of $R = 1$ and a length of $L = 1$, and we let $\Omega = [0, 1]$. We impose PEC boundary conditions on $\partial\Omega$ (see (2.4.1) in section 2.4.1). In TM mode the exact solutions read

$$\begin{aligned}
&E_z(r, \varphi, t) = J_m(\gamma_{mn} r) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn} t},\\
&E_r = E_\varphi = 0,\\
&H_z = 0,\\
&H_r(r, \varphi, t) = \frac{m\epsilon\omega_{mn}}{r\gamma_{mn}^2} J_m(\gamma_{mn} r) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn} t},\\
&H_\varphi(r, \varphi, t) = \frac{i\epsilon\omega_{mn}}{\gamma_{mn}} J_m'(\gamma_{mn} r) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn} t}.
\end{aligned}$$

Here, $J_m$ is the $m$th Bessel function of the first kind (see e.g. [98, Ch. 9]), $m = 0, 1, 2, \ldots$, $\gamma_{mn}$ the $n$ (possibly infinitely many) zeros of $J_m$, $n = 0, 1, 2, \ldots$, and $\omega_{mn}$ are the corresponding frequencies, $\omega_{mn} = c\,\gamma_{mn}$, where $c = \frac{1}{\sqrt{\epsilon\mu}}$ is the speed of light. For numerical simulations presented here we chose $m = 1$ and $n = 1$. By setting $t = 0$ in the exact solutions we get the initial field values.
Figure 4.4 shows the error plot with respect to the $L^2$-norm for this situation. For the component vectors $\mathbf{E}_z$ and $\mathbf{E}_z^h$ of the exact and of the approximate solution, respectively, we compute the error

$$||\mathbf{E}_z^h - \mathbf{E}_z||_{L^2} = \frac{||\mathbf{E}_z^h - \mathbf{E}_z||_2}{||\mathbf{E}_z^h||_2},$$

where $||\cdot||_2$ denotes the Euclidean norm. We can see a convergence behavior as we would expect from the Cartesian case (see Theorems 3.28 and 3.29). The stagnation in the slope beginning with $p = 5$ comes from time integration, where we used a time step size $0.3\,\mathrm{d}t$ (the magnitude of $\mathrm{d}t$ depends on the mesh step size). To check this we have chosen a smaller time step size $0.01\,\mathrm{d}t$ and have rerun the simulation. In this case the error decreases further, as expected, see figure 4.5.

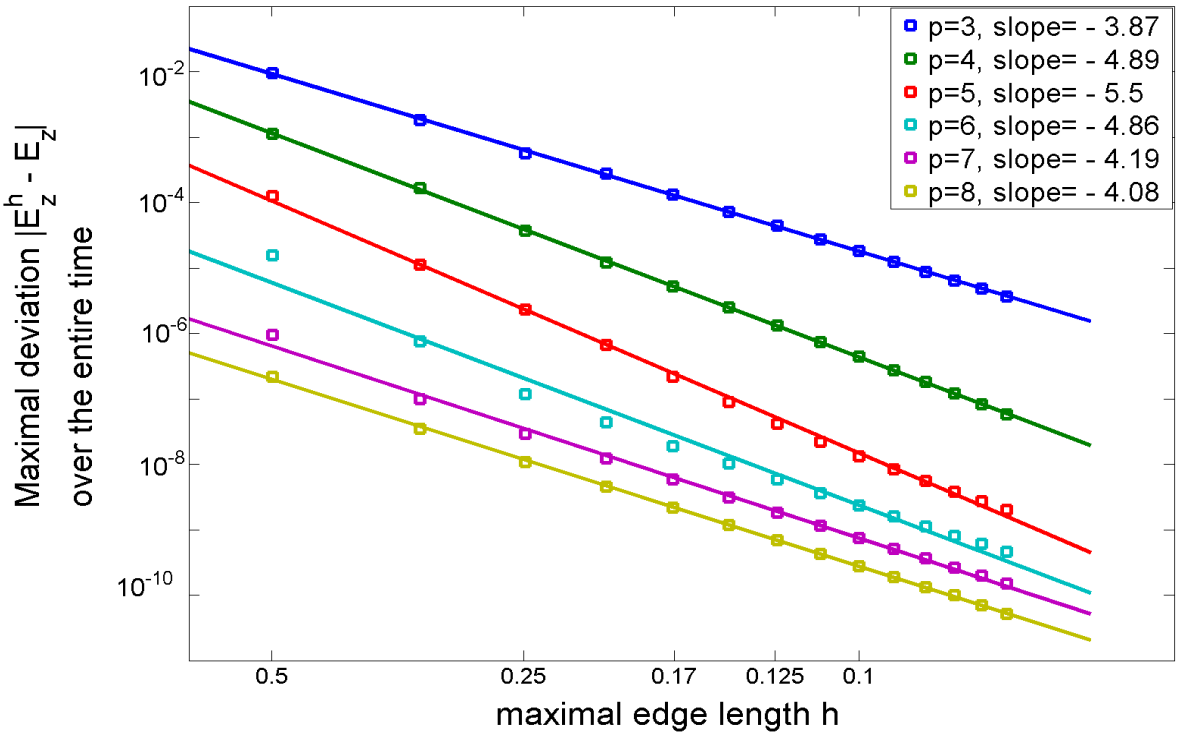



Figure 4.4: Error plot for $E_z$ in a homogeneous medium with polynomial order $p \in \{3, \dots, 8\}$, with number of elements $K = 2, \dots, 15$ and $m = 1$, in logarithmic scaling. We plot the $L^2$-error of the deviation $|E_z^h - E_z|$ of the exact solution $E_z$ to the approximation $E_z^h$ over the entire time.

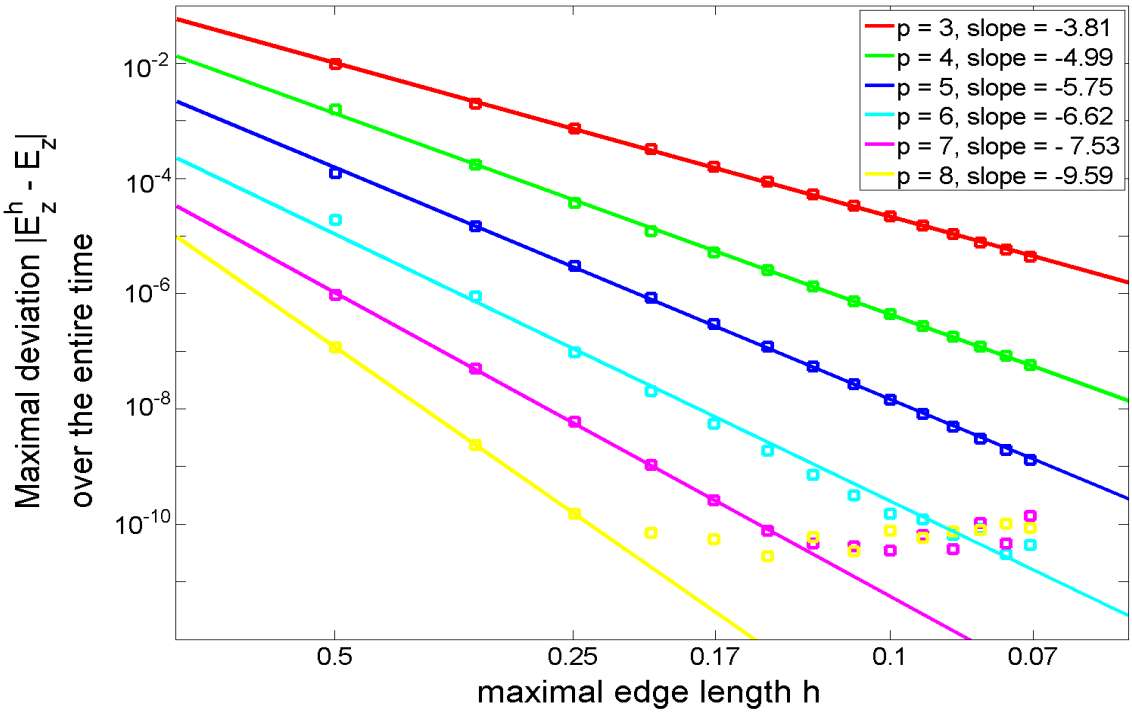



Figure 4.5: Error plot as in figure 4.4, but with a smaller time step size.

### 4.3.2 Coaxial Cable

As a next test system we look at a coaxial cable, as illustrated in figure 4.6: We have an inner radius $a$ and a total radius $b$, so that $\Omega = [0, a] \cup (a, b]$. The inner ring is filled with air, the outer ring has material parameters $\epsilon$ and $\mu$. We impose zero boundary conditions on $E_z$, that is, $E_z(r = a, \varphi, t) = 0$, $E_z(r = b, \varphi, t) = 0$ ([99, Ch. 9]). In our tests we chose $a = 0.5$ and $b = 1.0$.

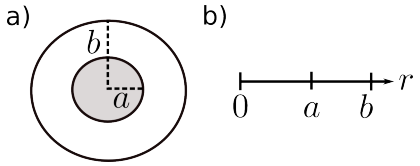



Figure 4.6: A sketch of a coaxial cable. a) From the front: The shaded gray part is filled with air, the white ring has material parameters $\epsilon$ and $\mu$; b) after dimension reduction we get the interval $[0, b]$.

The exact solution is (see e.g. [99, Ch. 9])

$$\begin{aligned}
&E_z(r, \varphi, t) = (A_1 J_m(\gamma_{mn} r) + B_1 Y_m(\gamma_{mn} r)) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn} t},\\
&E_r = E_\varphi = 0,\\
&H_z = H_r = 0,\\
&H_\varphi(r, \varphi, t) = -\frac{\mathrm{i}\epsilon\omega_{mn}}{\gamma_{mn}} (A_1 J_m'(\gamma_{mn} r) + B_1 Y_m'(\gamma_{mn} r)) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn} t},
\end{aligned}$$

where the $J_m$ are the Bessel functions of the first kind and the $Y_m$ are the Bessel functions of the second kind. $A_1$, $B_1$ and $\gamma_{mn}$ are unknowns, which need to be determined by applying the boundary conditions $E_z(r = a, \varphi, t) = 0$, $E_z(r = b, \varphi, t) = 0$. This leads to two equations for three unknowns – and thus to a one-parameter family of solutions – namely

$$\begin{aligned}
A_1 J_m(\gamma_{mn} b) + B_1 Y_m(\gamma_{mn} b) = 0,\\
A_1 J_m(\gamma_{mn} a) + B_1 Y_m(\gamma_{mn} a) = 0.
\end{aligned}$$

Solving for $A_1$ gives

$$\begin{aligned}
A_1 = -B_1 \frac{Y_m(\gamma_{mn} b)}{J_m(\gamma_{mn} b)},\\
A_1 = -B_1 \frac{Y_m(\gamma_{mn} a)}{J_m(\gamma_{mn} a)}.
\end{aligned}$$

Equating both equations and solving for $B_1$ renders

$$B_1 \left(Y_m(\gamma_{mn} b) J_m(\gamma_{mn} a) - Y_m(\gamma_{mn} a) J_m(\gamma_{mn} b)\right) = 0.$$

Since $B_1 = 0$ would lead to $A_1 = 0$, we have $B_1 \neq 0$, and thus $\gamma_{mn}$ is determined as the zero of the function

$$f(\gamma_{mn}) := Y_m(\gamma_{mn} b) J_m(\gamma_{mn} a) - Y_m(\gamma_{mn} a) J_m(\gamma_{mn} b).$$

Finally we obtain

$$\begin{aligned}
A_1 = -B_1 \frac{Y_m(\gamma_{mn} a)}{J_m(\gamma_{mn} a)},\\
B_1 \in \mathbb{R} \text{ arbitrary (but fixed).}
\end{aligned}$$

Again, it is $\omega_{mn} = c\,\gamma_{mn}$, where $c = \frac{1}{\sqrt{\epsilon\mu}}$. For this test system we observe the same convergence behavior as in the homogeneous system. We do not show an error plot for this case, since the next test system is an inhomogeneous waveguide, where the solution of a coaxial cable is needed as well; and, as we will see, we have p-convergence in that case as well.

### 4.3.3 Inhomogeneous Waveguide

As a third basic test case we look at an inhomogeneous waveguide. We assume to have a medium consisting of a region 1 and 2 with $\Omega = \Omega_1 \cup \Omega_2 = [0, a] \cup (a, b]$; in region 1 we denote the material parameters by $\epsilon_1, \mu_1$; in region 2 we have correspondingly $\epsilon_2, \mu_2$.
The exact solution is composed of the solution in region 1 and region 2 (see [99, Ch. 9.5.1 and 9.5.4]). The solution in region 1 is the one of the homogeneous case (see above), which we index by a superscript 1. The solution in region 2 is given to be the solution for the case of a coaxial cable. This gives

$$
\begin{aligned}
E_z^{(1)}(r,\varphi,t) &= A_1 J_m(\gamma_{mn}^{(1)} r) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn}t},\\
E_z^{(2)}(r,\varphi,t) &= \left(A_2 J_m(\gamma_{mn}^{(2)} r) + B_2 Y_m(\gamma_{mn}^{(2)} r)\right) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn}t},\\
H_r^{(1)}(r,\varphi,t) &= \frac{m\omega_1\epsilon_1}{(\gamma_{mn}^{(1)})^2 r} E_z^{(1)}(r,\varphi,t),\\
H_r^{(2)}(r,\varphi,t) &= \frac{m\omega_2\epsilon_2}{(\gamma_{mn}^{(2)})^2 r} E_z^{(2)}(r,\varphi,t),\\
H_\varphi^{(1)}(r,\varphi,t) &= A_1 \frac{\mathrm{i}\omega_1\epsilon_1}{\gamma_{mn}^{(1)}} J_m'(\gamma_{mn}^{(1)} r) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn}t},\\
H_\varphi^{(2)}(r,\varphi,t) &= \frac{\mathrm{i}\omega_2\epsilon_2}{\gamma_{mn}^{(2)}} \left(A_2 J_m'(\gamma_{mn}^{(2)} r) + B_2 Y_m'(\gamma_{mn}^{(2)} r)\right) e^{\mathrm{i}m\varphi} e^{-\mathrm{i}\omega_{mn}t}.
\end{aligned}
$$

Here, $\gamma_1$ and $\gamma_2$ are unknown and need to be determined, along with the coefficients $A_1, A_2, B_2$. The tangential of the electric field must vanish at $r = b$, and at $r = a$ the tangential components of the electromagnetic fields must be continuous (see section 2.2), leading to a continuity condition on $E_z$ and $H_\varphi$ (see [99, Ch. 9.5.1, Ch. 9.5.4]), namely at the interface $r = a$ we demand

$$
E_z^{(2)}(r=a) - E_z^{(1)}(r=a) = 0, \qquad H_\varphi^{(1)}(r=a) - H_\varphi^{(2)}(r=a) = 0.
$$

In addition, we have the boundary conditions $E_z^{(1)}(r=0) = 0$, $E_z^{(2)}(r=b) = 0$. Collecting this, we get a linear system of equations for the coefficient vector $\mathbf{b} := (A_1, A_2, B_2)^T$,

$$
A\,\mathbf{b} := \begin{pmatrix}
-J_m(\gamma_{mn}^{(1)} a) & J_m(\gamma_{mn}^{(2)} a) & Y_m(\gamma_{mn}^{(2)} a)\\
\frac{\epsilon_1}{\mu_1} J_m'(\gamma_{mn}^{(1)} a) & -\frac{\epsilon_2}{\mu_2} J_m'(\gamma_{mn}^{(2)} a) & -\frac{\epsilon_2}{\mu_2} Y_m'(\gamma_{mn}^{(2)} a)\\
0 & J_m(\gamma_{mn}^{(2)} b) & Y_m(\gamma_{mn}^{(2)} b)
\end{pmatrix}
\begin{pmatrix} A_1 \\ A_2 \\ B_2 \end{pmatrix} = 0.
$$

Since $\gamma_{mn}^{(1)}$ and $\gamma_{mn}^{(2)}$ are unknown as well, we need more information to solve this system uniquely. In both regions the frequency of the fields has to be similar, that is, $\omega_{mn}^{(1)} = \omega_{mn}^{(2)}$. This gives the condition

$$
\gamma_{mn}^{(2)} = \sqrt{\frac{\epsilon_2\mu_2}{\epsilon_1\mu_1}}\,\gamma_{mn}^{(1)}. \tag{4.3.1}
$$

We only have $\mathbf{b} \neq 0$ if $\det(A) = 0$; from this we can determine $\gamma_{mn}^{(1)}$ (see also [99, Ch. 9.5.1 and problem 9.8]): It is a zero of the function

$$\begin{aligned} f(\gamma_{mn}^{(1)}) = &- J_m(\gamma_{mn}^{(1)}a)Y_m(\gamma_{mn}^{(2)}b)\frac{\epsilon_2}{\gamma_{mn}^{(2)}}J'_m(\gamma_{mn}^{(2)}a) + Y_m(\gamma_{mn}^{(2)}b)\frac{\epsilon_1}{\gamma_1}J'_m(\gamma_{mn}^{(1)}a)J_m(\gamma_{mn}^{(2)}a) \\ &+ J_m(\gamma_{mn}^{(1)}a)J_m(\gamma_{mn}^{(2)}b)\frac{\epsilon_2}{\gamma_{mn}^{(2)}}Y'_m(\gamma_{mn}^{(2)}b) - J_m(\gamma_{mn}^{(2)}b)\frac{\epsilon_1}{\gamma_{mn}^{(1)}}J'_m(\gamma_{mn}^{(1)}a)Y_m(\gamma_{mn}^{(2)}a). \end{aligned}$$

If $\gamma_{mn}^{(1)}$ is known, we can compute $\gamma_{mn}^{(2)}$ from (4.3.1). Thus we have determined all unknowns and can construct a continuous solution on $[0, b]$.
Again we are interested in the question whether we can hope to get $p$-convergence. Indeed, the results suggest this convergence behavior. Figure 4.7 shows for the $E_z$-field, where we chose – without any physical motivation – $\epsilon_1 = 1, \epsilon_2 = 1.5, \mu_1 = 1, \mu_2 = 5$. Note this is a consequence due to the fact that we chose $r = a$ as a grid point so that the interface $r = a$ is resolved exactly.

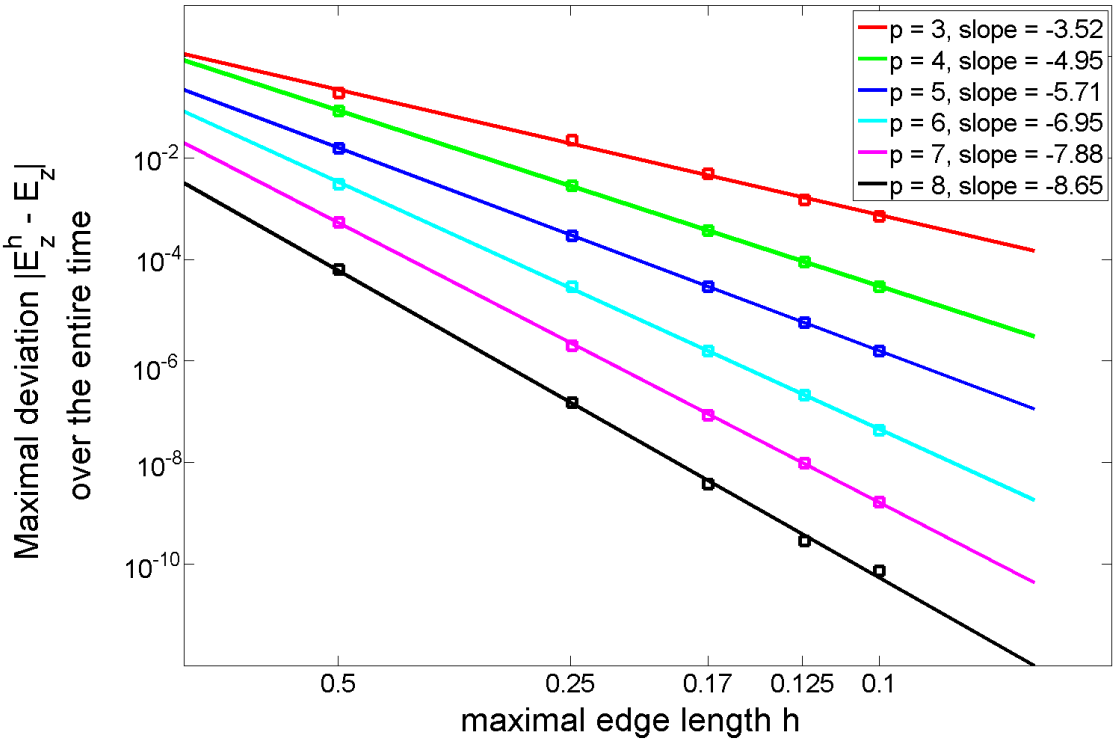



Figure 4.7: Error plot of the $E_z$-field in an inhomogeneous medium with $\epsilon_1 = 1, \epsilon_2 = 1.5, \mu_1 = 1, \mu_2 = 5$ and $m = 1, p \in \{3, \ldots, 8\}$. The number of elements is $K \in \{2, 4, 6, 8, 10\}$, the time step size is chosen to be 0.01 dt, where dt depends on the mesh size of the grid.

## 4.4 The Runge-Kutta Discontinuous Galerkin Method Applied to 3D-2D BOR Maxwell's Equations

In this section we apply the DG method to BOR Maxwell's equations in three space dimensions. After the $\varphi$-ansatz in section 4.1.2 the computational cost reduces to two dimensions, therefore we call the resulting set of equations *2D-BOR Maxwell's equations*. They read in the weak sense, after the $\varphi$-ansatz 4.1.1:

$$
\begin{aligned}
&\epsilon\partial_t E_r - \frac{\mathrm{i}m}{r} H_z + \partial_z H_\varphi = 0,\\
&\epsilon\partial_t E_\varphi + \partial_r H_z - \partial_z H_r = 0,\\
&\epsilon\partial_t E_z - \frac{1}{r}(\partial_r(rH_\varphi) - \mathrm{i}mH_r) = 0,\\
&\mu\partial_t H_r + \frac{\mathrm{i}m}{r} E_z - \partial_z E_\varphi = 0,\\
&\mu\partial_t H_\varphi - \partial_r E_z + \partial_z E_r = 0,\\
&\mu\partial_t H_z + \frac{1}{r}(\partial_r(rE_\varphi) - \mathrm{i}mE_r) = 0,
\end{aligned}
\tag{4.4.1}
$$

Recall that when we speak of BOR Maxwell's equations in the weak sense, this is meant with respect to the measure $r\,\mathrm{d}\mathbf{r}$, where in two dimensions, $\mathbf{r} = (r, z)$. Again, $m \in \mathbb{N}$ is fixed. The fields depend on $(r, z, t) \in \Omega \times \mathbb{R}$ with $\Omega \subseteq \mathbb{R}^2$. The approach is very similar to the one in one dimension, see section 4.2. Yet, there are some differences which are not minor, especially concerning the computation of the local matrices. We now have a two-dimensional geometry; we need to find multi-dimensional analogues of the Lagrange polynomials, and the natural question arises whether the resulting local matrices can be computed as efficiently as the ones in one dimension. Indeed, this can be accomplished by defining appropriate two-dimensional orthonormal polynomials.
Thus, as in one space dimension, we have to carry out the following steps during the DG discretization process:

1. Discretization of $\Omega$ and definition of a finite element space.
2. Choice of a numerical flux that gives rise to a stable and convergent DG method. We have already given a numerical flux in (4.2.8).
3. Efficient computation of the local matrices.

### 4.4.1 DG space discretization

As in one dimension, we divide $\Omega$ into $K$ elements so that $\Omega = \bigcup_{k=1}^{K} \Omega_k$. Our finite element space of discontinuous functions is

$$V_h := \{\mathbf{u}_h \in (L^\infty(\Omega))^6 : \mathbf{u}_h|_{\Omega_k} \in V(\Omega_k),\ k = 1, \dots, K\}.$$

We choose $V(\Omega_k) = \mathcal{P}^p(\Omega_k)$ as the space of multivariate polynomials of total degree $p \in \mathbb{N}$. We approximate $\mathbf{u}$ by $\mathbf{u}_h \in V_h$, and use Lagrange interpolation to represent $\mathbf{u}_h$. Thus, we obtain the nodal representation (cf. Ref. [11]) of the fields as

$$\mathbf{u}_h(r, z, t) = \sum_{i=1}^{N_p} \mathbf{u}_h^k(r_i^k, z_i^k, t) l_i^k(r, z), \tag{4.4.2}$$

where $l_i^k(r, z)$ are two-dimensional Lagrange polynomials on $\Omega_k$, $N_p = (p + 1)(p + 2)/2$ is the number of nodes and $\mathbf{r}_i^k := (r_i^k, z_i^k)$ are suitably chosen interpolation points on $\Omega_k$.
In section 4.2.1 we have seen that the Galerkin ansatz requires the residual

$$\mathcal{R}_h := \partial_t \mathbf{u}_h + \nabla \cdot \mathbf{F}(\mathbf{u}_h)$$

to be orthogonal to all test functions $\boldsymbol{\psi}_h \in V_h$ with respect to the measure $r\,\mathrm{d}\mathbf{r}$. As in (4.2.5), in the discontinuous Galerkin approach we have instead

$$\int_{\Omega_k} \mathcal{R}_h\, \boldsymbol{\psi}_h\, r\,\mathrm{d}\mathbf{r} = -\int_{\partial\Omega_k} (\mathbf{F}(\mathbf{u}_h^k) - \mathbf{F}^{\mathrm{num}}(\mathbf{u}_h^k))\, \mathbf{n}\, \boldsymbol{\psi}_h\, r\,\mathrm{d}\mathbf{r},$$

where $\mathbf{n}$ denotes the outer normal vector of the boundary $\partial\Omega_k$ and $\mathbf{F}^{\text{num}}$ is the numerical flux.
Plugging in the ansatz (4.4.2) into BOR Maxwell's equations (4.4.1) and approximating the test functions $\psi$ in $V_h$ as well, we get the semi-discrete scheme

$$
\begin{aligned}
&\epsilon\partial_t \mathbf{E}_r^k \int_{\Omega_k} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r} - \mathrm{i}m\mathbf{H}_z^k \int_{\Omega_k} l_i^k\, l_j^k\,\mathrm{d}\mathbf{r} + \mathbf{H}_\varphi^k \int_{\Omega_k} l_i^k\, \partial_z l_j^k\, r\,\mathrm{d}\mathbf{r} = \int_{\partial\Omega_k} (\mathbf{F}_{E_r} - \mathbf{F}^*_{E_r})\mathbf{n} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r},\\
&\epsilon\partial_t \mathbf{E}_\varphi^k \int_{\Omega_k} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r} + \mathbf{H}_z^k \int_{\Omega_k} l_i^k\, \partial_r l_j^k\, r\,\mathrm{d}\mathbf{r} - \mathbf{H}_r^k \int_{\Omega_k} l_i^k\, \partial_z l_j^k\, r\,\mathrm{d}\mathbf{r} = \int_{\partial\Omega_k} (\mathbf{F}_{E_\varphi} - \mathbf{F}^*_{E_\varphi})\mathbf{n} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r},\\
&\epsilon\partial_t \mathbf{E}_z^k \int_{\Omega_k} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r} - \mathbf{H}_\varphi^k \left( \int_{\Omega_k} l_i^k\, l_j^k\,\mathrm{d}\mathbf{r} + \int_{\Omega_k} l_i^k\, \partial_r l_j^k\, r\,\mathrm{d}\mathbf{r} \right) - \mathrm{i}m\mathbf{H}_r^k \int_{\Omega_k} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r}\\
&\qquad\qquad = \int_{\partial\Omega_k} (\mathbf{F}_{E_z} - \mathbf{F}^*_{E_z})\mathbf{n} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r},\\
&\mu\partial_t \mathbf{H}_r^k \int_{\Omega_k} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r} + \mathrm{i}m\mathbf{E}_z^k \int_{\Omega_k} l_i^k\, l_j^k\,\mathrm{d}\mathbf{r} - \mathbf{E}_\varphi^k \int_{\Omega_k} l_i^k\, \partial_z l_j^k\, r\,\mathrm{d}\mathbf{r} = \int_{\partial\Omega_k} (\mathbf{F}_{H_r} - \mathbf{F}^*_{H_r})\mathbf{n} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r},\\
&\mu\partial_t \mathbf{H}_\varphi^k \int_{\Omega_k} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r} - \mathbf{E}_z^k \int_{\Omega_k} l_i^k\, \partial_r l_j^k\, r\,\mathrm{d}\mathbf{r} + \mathbf{E}_r^k \int_{\Omega_k} l_i^k\, \partial_z l_j^k\, r\,\mathrm{d}\mathbf{r} = \int_{\partial\Omega_k} (\mathbf{F}_{H_\varphi} - F^*_{\mathbf{H}_\varphi})\mathbf{n} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r},\\
&\mu\partial_t \mathbf{H}_z^k \int_{\Omega_k} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r} + \mathbf{E}_\varphi^k \left( \int_{\Omega_k} l_i^k\, l_j^k\,\mathrm{d}\mathbf{r} + \int_{\Omega_k} l_i^k\, \partial_r l_j^k\, r\,\mathrm{d}\mathbf{r} \right) - \mathrm{i}m\mathbf{E}_r^k \int_{\Omega_k} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r}\\
&\qquad\qquad = \int_{\partial\Omega_k} (\mathbf{F}_{H_z} - \mathbf{F}^*_{H_z})\mathbf{n} l_i^k\, l_j^k\, r\,\mathrm{d}\mathbf{r}.
\end{aligned}
\tag{4.4.3}
$$

Again, $\mathbf{E}_r^k$ are the coefficient vectors of the fields, which are time-dependent. By $\mathbf{F}^*$ we denote the numerical flux $\mathbf{F}^{\text{num}}$, as in the one-dimensional case, which is specified in the next section.

### 4.4.2 A Numerical Flux

We already computed the numerical flux in (4.2.8), and we repeat it here:

$$
\begin{aligned}
\mathbf{G}_\mathbf{E} &:= (\mathbf{F}_\mathbf{E} - \mathbf{F}^*_\mathbf{E})\mathbf{n} = -[\mathbf{n}\times\mathbf{H} - (\mathbf{n}\times\mathbf{H}^*)] = -\frac{1}{2\{\{Z\}\}}\mathbf{n}\times[Z_\mathrm{R}[\![\mathbf{H}]\!] - \alpha\mathbf{n}\times[\![\mathbf{E}]\!]]\\
\mathbf{G}_\mathbf{H} &:= (\mathbf{F}_\mathbf{H} - \mathbf{F}^*_\mathbf{H})\mathbf{n} = [\mathbf{n}\times\mathbf{E} - (\mathbf{n}\times\mathbf{E}^*)] = \frac{1}{2\{\{Y\}\}}\mathbf{n}\times[Y_\mathrm{R}[\![\mathbf{E}]\!] + \alpha\mathbf{n}\times[\![\mathbf{H}]\!]],
\end{aligned}
$$

with $\mathbf{n} = (n_r, 0, n_z)^T$ and $\alpha \in [0,1]$. Written out componentwise, this is

$$\begin{aligned}
\mathbf{G}_{E_r} &= -\frac{1}{2\{\{Z\}\}}\big[-Z_R n_z [\![H_\varphi]\!] - \alpha(n_r(\mathbf{n}\cdot[\![\mathbf{E}]\!]) - [\![E_r]\!])\big],\\
\mathbf{G}_{E_\varphi} &= -\frac{1}{2\{\{Z\}\}}\big[Z_R(n_z[\![H_r]\!] - n_r[\![H_z]\!]) + \alpha[\![E_\varphi]\!]\big],\\
\mathbf{G}_{E_z} &= -\frac{1}{2\{\{Z\}\}}\big[Z_R n_r[\![H_\varphi]\!] - \alpha(n_z(\mathbf{n}\cdot[\![\mathbf{E}]\!]) - [\![E_z]\!])\big],\\
\mathbf{G}_{H_r} &= \frac{1}{2\{\{Y\}\}}\big[-Y_R n_z[\![E_\varphi]\!] + \alpha(n_r(\mathbf{n}\cdot[\![\mathbf{H}]\!]) - [\![H_r]\!])\big],\\
\mathbf{G}_{H_\varphi} &= \frac{1}{2\{\{Y\}\}}\big[Y_R(n_z[\![E_r]\!] - n_r[\![E_z]\!]) - \alpha[\![H_\varphi]\!]\big],\\
\mathbf{G}_{H_Z} &= \frac{1}{2\{\{Y\}\}}\big[Y_R n_r[\![E_\varphi]\!] + \alpha(n_z(\mathbf{n}\cdot[\![\mathbf{H}]\!]) - [\![H_z]\!])\big],
\end{aligned} \tag{4.4.4}$$

where $\mathbb{G}_E := (\mathbf{G}_{E_r}, \mathbf{G}_{E_\varphi}, \mathbf{G}_{E_z})^T$ and $\mathbb{G}_H := (\mathbf{G}_{H_r}, \mathbf{G}_{H_\varphi}, \mathbf{G}_{H_z})^T$. We used the vector identity $\mathbf{n}\times(\mathbf{n}\times\mathbf{V}) = \mathbf{n}(\mathbf{n}\cdot\mathbf{V}) - (\mathbf{n}\cdot\mathbf{n})\mathbf{V} = \mathbf{n}(\mathbf{n}\cdot\mathbf{V}) - \mathbf{V}$, since $\mathbf{n}\cdot\mathbf{n} = \mathbf{1}$, where $\mathbf{V}$ is an arbitrary vector.

### 4.4.3 Semi-Discrete Scheme

By defining the local matrices

$$\begin{aligned}
(M_r^k)_{ij} &:= \int_{\Omega_k} l_i^k(\mathbf{r})\, l_j^k(\mathbf{r})\; r\,\mathrm{d}\mathbf{r},\\
(S_r^k)_{ij} &:= \int_{\Omega_k} l_i^k(\mathbf{r})\, \partial_r l_j^k(\mathbf{r})\; r\,\mathrm{d}\mathbf{r},\\
(S_z^k)_{ij} &:= \int_{\Omega_k} l_i^k(\mathbf{r})\, \partial_z l_j^k(\mathbf{r})\; r\,\mathrm{d}\mathbf{r},\\
(M^k)_{ij} &:= \int_{\Omega_k} l_i^k(\mathbf{r})\, l_j^k(\mathbf{r})\; r\,\mathrm{d}\mathbf{r},\\
\mathcal{F}_{ij}^k &:= \int_{\partial\Omega_k} l_i^k(\mathbf{r})\, l_j^k(r,z)\; r\,\mathrm{d}\mathbf{r}.
\end{aligned} \tag{4.4.5}$$

we bring BOR Maxwell's equations (4.4.1) into the semi-discrete form as

$$\begin{aligned}
&\epsilon M_r^k \partial_t \mathbf{E}_r - \mathrm{i}m M^k \mathbf{H}_z + S_z^k \mathbf{H}_\varphi - \mathcal{F}^k \mathbf{G}_{E_r} = 0,\\
&\epsilon M_r^k \partial_t \mathbf{E}_\varphi + S_r^k \mathbf{H}_z - S_z^k \mathbf{H}_r - \mathcal{F}^k \mathbf{G}_{E_\varphi} = 0,\\
&\epsilon M_r^k \partial_t \mathbf{E}_z - (M^k + S_r^k)\mathbf{H}_\varphi - \mathrm{i}m M^k \mathbf{H}_r - \mathcal{F}^k \mathbf{G}_{E_z} = 0,\\
&\mu M_r^k \partial_t \mathbf{H}_r + \mathrm{i}m M^k \mathbf{E}_z - S_z^k \mathbf{E}_\varphi - \mathcal{F}^k \mathbf{G}_{H_r} = 0,\\
&\mu M_r^k \partial_t \mathbf{H}_\varphi - S_r^k \mathbf{E}_z + S_z^k \mathbf{E}_r - \mathcal{F}^k \mathbf{G}_{H_\varphi}) = 0,\\
&\mu M_r^k \partial_t \mathbf{H}_z + (M^k + S_r^k)\mathbf{E}_\varphi - \mathrm{i}m M^k \mathbf{E}_r - \mathcal{F}^k \mathbf{G}_{H_z} = 0,
\end{aligned} \tag{4.4.6}$$

At the end, as in the one-dimensional case we define

$$\mathbb{H}^k := \begin{pmatrix} \mathbf{H}_r^k \\ \mathbf{H}_\varphi^k \\ \mathbf{H}_z^k \end{pmatrix}, \quad \mathbb{E}^k := \begin{pmatrix} \mathbf{E}_r^k \\ \mathbf{E}_\varphi^k \\ \mathbf{E}_z^k \end{pmatrix},$$

$$\mathbb{L}_{\mathrm{BOR}} := \begin{pmatrix} 0 & -(M_r^k)^{-1} S_z^k & \mathrm{i}m (M_r^k)^{-1} M^k \\ (M_r^k)^{-1} S_z^k & 0 & -(M_r^k)^{-1} S_r^k \\ -\mathrm{i}m (M_r^k)^{-1} M^k & (M_r^k)^{-1}(M^k + S_r^k) & 0 \end{pmatrix}$$

and define

$$(M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_E := ((M_r^k)^{-1}\mathcal{F}^k\mathbf{G}_{E_r}, (M_r^k)^{-1}\mathcal{F}^k\mathbf{G}_{E_\varphi}, (M_r^k)^{-1}\mathcal{F}^k\mathbf{G}_{E_z})^T,$$
$$(M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_E = ((M_r^k)^{-1}\mathcal{F}^k\mathbf{G}_{H_r}, (M_r^k)^{-1}\mathcal{F}^k\mathbf{G}_{H_\varphi}, (M_r^k)^{-1}\mathcal{F}^k\mathbf{G}_{H_z})^T,$$

We can thus write the semi-discrete scheme (4.4.6) in short vector-matrix notation as

**Semi-Discrete Scheme in 2D**

$$\begin{aligned}\epsilon\partial_t\mathbb{E}^k &= \left(\mathbb{L}_{\mathrm{BOR}}\mathbb{H}^k + (M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_H\right),\\ \mu\partial_t\mathbb{H}^k &= \left(-\mathbb{L}_{\mathrm{BOR}}\mathbb{E}^k + (M_r^k)^{-1}\mathcal{F}^k\mathbb{G}_E\right).\end{aligned} \tag{4.4.7}$$

### Transformation to a Reference Element

Again all operations are performed on a reference element, which in two dimensions is given as

$$I := \{\mathbf{x} = (x, y) : x, y > -1, x + y \leq 0\}.$$

**Lemma 4.8.**

*(i) For straight-sided triangles the transformation of $I$ to $\Omega_k$ is given by the affine mapping*

$$\begin{aligned}&\Psi^k : I \to \Omega_k,\\ &\mathbf{x} \mapsto \mathbf{r} = \Psi^k(\mathbf{x}) := \mathbf{v}_1 + \frac{1}{2}(x+1)(\mathbf{v}_2 - \mathbf{v}_1) + \frac{1}{2}(y+1)(\mathbf{v}_3 - \mathbf{v}_1)\end{aligned}$$

*Here, $\mathbf{v_1} = (v_{11}, v_{12}), \mathbf{v_2} = (v_{21}, v_{22}), \mathbf{v_3} = (v_{31}, v_{32})$ denote the vertices of the triangle $\Omega_k$ as depicted in figure 4.8. In addition, we introduce the edge vectors $\mathbf{e}_1 = \mathbf{v}_2 - \mathbf{v}_1$, $\mathbf{e}_2 = \mathbf{v}_3 - \mathbf{v}_2$ and $\mathbf{e}_3 = \mathbf{v}_1 - \mathbf{v}_3$.*

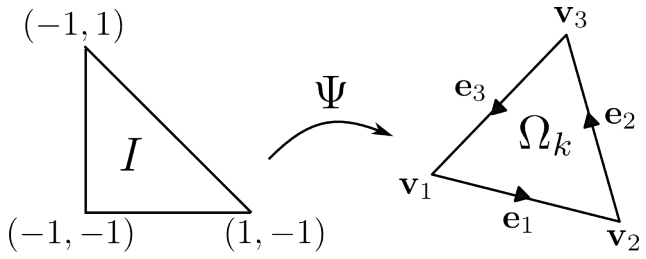



Figure 4.8: Transformation of $\Omega_k$ to the reference element $I$.

*The Jacobi matrix of $\Psi^k$ is given as*

$$J_{\Psi^k} := \nabla_{\mathbf{r}}\Psi^k = \frac{\partial \mathbf{r}}{\partial \mathbf{x}} = (\mathbf{r}_x \quad \mathbf{r}_y) = \begin{pmatrix} r_x & r_y \\ z_x & z_y \end{pmatrix} = \frac{1}{2}\begin{pmatrix} e_{11} & -e_{31} \\ e_{12} & -e_{32} \end{pmatrix}, \tag{4.4.8}$$

*and we abbreviate its determinant as*

$$\det(J_\Psi^k) =: J^k = \frac{1}{2}\,\mathrm{vol}(\Omega_k).$$

*Since $\Psi^k$ is an affine mapping, the inverse Jacobi matrix reads*

$$(\nabla\Psi^k)^{-1} = \nabla_{\mathbf{x}}(\Psi^k)^{-1} = \frac{\partial \mathbf{x}}{\partial \mathbf{r}} = (\mathbf{x}_r \quad \mathbf{x}_z) = \begin{pmatrix} x_r & x_z \\ y_r & y_z \end{pmatrix} = \frac{1}{J^k}\begin{pmatrix} -e_{32} & e_{31} \\ -e_{12} & e_{11} \end{pmatrix}.$$

*Thus,*

$$r_x = \frac{y_z}{J}, \quad r_y = -\frac{x_z}{J}, \quad z_x = -\frac{y_r}{J}, \quad z_y = \frac{x_r}{J}.$$

*(ii) The derivatives transform as*

$$\begin{aligned} \partial_r &\mapsto \frac{1}{2J}\big[-e_{32}\partial_x - e_{11}\partial_y\big], \\ \partial_z &\mapsto \frac{1}{2J}\big[e_{31}\partial_x + e_{11}\partial_y\big]. \end{aligned} \tag{4.4.9}$$

*Proof.* (i) The last statement of (i) can be found in [11]. The rest can be easily verified by direct computation.

(ii) Since $\Psi^k$ is an affine mapping, $\nabla_{\mathbf{r}} = \partial\mathbf{x}/\partial\mathbf{r}$ and $\nabla_{\mathbf{x}} = \partial\mathbf{r}/\partial\mathbf{x}$ are inverse to each other, and it follows:

$$\begin{pmatrix} \partial_r \\ \partial_z \end{pmatrix} \mapsto \frac{\partial \mathbf{r}}{\partial \mathbf{x}}^T \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} = \frac{1}{2}\begin{pmatrix} r_x\partial_x + z_x\partial_y \\ r_y\partial_x + z_y\partial_y \end{pmatrix},$$

and by using (i) we get the statement in (ii). □

### 4.4.4 Efficient Computation of the Local Matrices

It would be desirable to efficiently compute the local matrices as in the one-dimensional case as the composition of template matrices, that can be set up by using two-dimensional orthogonal polynomials. Yet, there is another good reason: It is a non-trivial task to express multi-dimensional Lagrange polynomials explicitly. As we will present in this section, we can indeed compute the local matrices as efficiently as in one space dimension. Besides a higher demand in computing the local matrices, it is also more challenging to find the $N_p$ grid points in higher dimensions, and already in two space dimensions this can be quite elaborate. A naive idea would be to use a tensor product of the one-dimensional grid points to get the grid points on $I = \{(x, y) : x, y > -1, x+y \leq 0\}$. But as the authors in [11] point out this leads to ill-conditioned schemes. In [101] Warburton presents another way of getting a set of two-dimensional grid points by a method he called *blend & warp*. The main idea is to find a function $f$ that maps an equidistant grid $\{r_i^{\text{equ}} : i = 1, \cdots, N_p\}$ to an arbitrary grid $\{r_i : i = 1, \cdots, N_p\}$ so that $f(r_i^{\text{equ}}) = r_i$. He chooses a grid consisting of Legendre-Gauss-Lobatto nodes, and we will use this set of grid points as well.

#### Orthogonal Polynomials

As we have briefly reported in section 4.2.5, the family of one-dimensional Jacobi polynomials $\{P_n^{(\alpha,\beta)}\}_{n\geq 0}$ on $[-1, 1]$, for $\alpha, \beta > -1$, is constructed by the Gram-Schmidt orthonormalization procedure with respect to the weighted $L^2$-scalar product on $(-1, 1)$ using $w(x) = (1 - x)^\alpha(1 + x)^\beta$ as a weight, see e.g. [11, App. A]. In Ref. [102, Ch. 3.3,

Class IV and Class V] a family of two-dimensional Jacobi polynomials $\{P_{n,k}^{(\alpha,\beta,\gamma)}\}_{0\le k\le n}$ on the simplex

$$S := \{(\xi,\eta) \in \mathbb{R}^2 \,:\, 0 < \eta < \xi < 1\}$$

was derived, for $\alpha, \beta, \gamma > -1$, that is $L^2(S)$-orthogonal with respect to the weight function

$$w_S(\xi,\eta) = (1-\xi)^\alpha (\xi-\eta)^\beta \eta^\gamma, \tag{4.4.10}$$

see Ref. [102, (3.10)]. The two-dimensional polynomials are obtained from the one-dimensional polynomials via

$$P_{n,k}^{(\alpha,\beta,\gamma)}(\xi,\eta) = P_{n-k}^{(\alpha,\beta+\gamma+2k+1)}(2\xi-1)\xi^k P_k^{(\beta,\gamma)}\Big(\frac{2\eta}{\xi}-1\Big).$$

Since all our later operations are carried out on $I$, we define a transformation from $I$ to $S$ by

$$\Psi_{I\to S} : I \to S, \quad \Psi_I \to S(x,y) := \frac{1}{2}(1-y, 1+x)^T. \tag{4.4.11}$$

Thus we obtain the two-variable analogues of the Jacobi polynomials on $I$ by

$$\widetilde{P}_{n,k}^{(\alpha,\beta,\gamma)}(x,y) = P_{n,k}^{(\alpha,\beta,\gamma)} \circ \Psi_{I\to S}(x,y) = P_{n-k}^{(\alpha,\beta+\gamma+2k+1)}(-y)\frac{(1-y)^k}{2^k} P_k^{(\beta,\gamma)}\Big(2\frac{1+x}{1-y}-1\Big) \tag{4.4.12}$$

with $n = 0,\dots,N, k = 0,\dots,n$ and $m = (n-k) + (N+1)k + 1 - \frac{k}{2}(k-1)$. These are orthogonal with respect to the weight function

$$w_I(x,y) = \frac{1}{2^{\alpha+\beta+\gamma}}(1+y)^\alpha (x+y)^\beta (1+x)^\gamma. \tag{4.4.13}$$

We normalize the orthogonal polynomials $\widetilde{P}_{n,k}^{(\alpha,\beta,\gamma)}$ such that their weighted $L^2(I)$-norm is unity. By doing so we get the normalized set $\{\widehat{P}_{n,k}^{(\alpha,\beta,\gamma)}\}_{0\le k\le n}$ of the two-variable polynomials (4.4.12). We remark that for the choice $\alpha = \beta = \gamma = 0$ we recover the Dubiner polynomials, see e.g. [103], [102], [104], [11, Ch. A.1].

### Computation of the Mass Matrices

**Lemma 4.9** (Mass matrix)**.**
*The local mass matrix can be computed as $M^k = J^k M$ with $J^k$ being the Jacobi determinant on element $k$, where we assume to have straight-sided triangles, and the global mass matrix $M$ is given through*

$$M = (VV^T)^{-1} \;\; \textit{with} \;\; V_{ij} = \hat{P}_{j-1}^{(0,0,0)}(\mathbf{x_i}),$$

*where $P_m^{(0,0,0)}(\mathbf{x}_i)$ are the orthogonal polynomials in (4.4.12). Thus, the computational effort for $M^k$ reduces to computing a global template matrix $M$ once.*

*Proof.* We transform the local mass matrix to the reference element $I$ by

$$(M^k)_{ij} := \int_{\Omega_k} l_i^k(\mathbf{r}) l_j^k(\mathbf{r})\, \mathbf{dr} = J^k \int_I l_i(\mathbf{x}) l_j(\mathbf{x})\, \mathrm{d}\mathbf{x} =: J^k M_{ij}.$$

As was shown in [11, Ch. 6.2], it holds $M = (VV^T)^{-1}$. □

**Lemma 4.10** (BOR mass matrix)**.**
*The local BOR mass matrix can be computed as a composition of template matrices as*

$$M_r^k = J_k(v_{11}M + e_{11}M_r^{(1)} - e_{31}M_r^{(2)}),$$

*where*

$$(M_r^{(1)})_{ij} := \int_I \frac{1}{2}(1+x)l_i(\boldsymbol{x})l_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$
$$(M_r^{(2)})_{ij} := \int_I \frac{1}{2}(1+y)l_i(\boldsymbol{x})l_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

*This can be seen as follows:*

$$\begin{aligned}(M_r^k)_{ij} &= \int_{\Omega_k} l_i^k(\mathbf{r})l_j^k(\mathbf{r})\,r\,\mathrm{d}\mathbf{r} = J^k\int_I \left[v_{11} + (1+x)\frac{e_{11}}{2} - (1+y)\frac{e_{31}}{2}\right] l_i(\mathbf{x})l_j(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &= J^k\left[v_{11}\int_I l_i(\mathbf{x})l_j(\mathbf{x})\,\mathrm{d}\mathbf{x} + \frac{e_{11}}{2}\int_I (1+x)l_i(\mathbf{x})l_j(\mathbf{x})\,\mathrm{d}\mathbf{x} - \frac{e_{31}}{2}\int_I (1+y)l_i(\mathbf{x})l_j(\mathbf{x})\,\mathrm{d}\mathbf{x}\right]\\ &= J^k\left[v_{11}M_{ij} + \frac{e_{11}}{2}(M_r^1)_{ij} - \frac{e_{31}}{2}(M_r^2)_{ij}\right].\end{aligned}$$

*Analogously to the computation of the mass matrix via the generalized Vandermonde matrix $V$, the matrices $M_r^{(1)}$ and $M_r^{(2)}$ can also be set up by the usage of BOR Vandermonde matrices, that is*

$$M_r^{(1)} = (V_r^{(1)}(V_r^{(1)})^T)^{-1} \;\; \textit{with} \;\; (V_r^{(1)})_{ij} := \hat{P}_{j-1}^{(0,0,1)}(\mathbf{x_i}),$$
$$M_r^{(2)} = (V_r^{(2)}(V_r^{(2)})^T)^{-1} \;\; \textit{with} \;\; (V_r^{(2)})_{ij} := \hat{P}_{j-1}^{(1,0,0)}(\mathbf{x_i}).$$

### Computation of the BOR Stiffness Matrices

At last, the BOR stiffness matrices in (4.4.5) can be computed as

$$S_r^k = J^k M_r(x_r D_x + y_r D_y),$$
$$S_z^k = J^k M_r(x_z D_x + y_z D_y),$$

where $D_x, D_y$ are the differentiation matrices, given as

$$(D_x)_{ij} := \partial_x l_j(\mathbf{x}_i),$$
$$(D_y)_{ij} := \partial_y l_j(\mathbf{x}_i),$$

where $\mathbf{x}_i = (x_i, y_i)$, exactly as in [11, Kap. 3.2]. They can be computed in the same way via $D_m = V_m V^{-1}$, $m \in \{x, y\}$ and

$$(V_m)_{ij} = \partial_m \widehat{P}_j^{(\alpha,\beta,\gamma)}(\mathbf{a}_i).$$

*Proof.* We transform on $I$ and recall the transformation of the derivatives (4.4.9) to get

$$\begin{aligned}(S_r^k)_{ij} &= \int_{\Omega_k} l_j^k(\mathbf{r})(\partial_r l_i^k(\mathbf{r}))\,r\,\mathrm{d}\mathbf{r}\\ &= J^k\Big[\int_I \left(v_{11} + (1+x)\frac{e_{11}}{2} - (1+y)\frac{e_{31}}{2}\right)\frac{1}{2J^k}\left(-e_{31}\partial_x - e_{12}\partial_y\right) l_i\; l_j\,\mathrm{d}\mathbf{x}\Big]\\ &= J^k\Big[-\frac{v_{11}e_{32}}{2J^k}\int_I l_i(\mathbf{x})\;\partial_x l_j(\mathbf{x})\,\mathrm{d}\mathbf{x} - \frac{v_{11}e_{12}}{2J^k}\int_I l_i(\mathbf{x})\;\partial_y l_j(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &- \frac{e_{11}e_{32}}{2J^k}\int_I \frac{1+x}{2}\;l_i(\mathbf{x})\partial_x l_j(\mathbf{x})\,\mathrm{d}\mathbf{x} - \frac{e_{11}e_{12}}{2J^k}\int_I \frac{1+x}{2}\;l_i(\mathbf{x})\;\partial_y l_j(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &+ \frac{e_{31}e_{32}}{2J^k}\int_I \frac{1+y}{2}\;l_i(\mathbf{x})\partial_x l_j(\mathbf{x})\,\mathrm{d}\mathbf{x} + \frac{e_{31}e_{12}}{2J^k}\int_I \frac{1+y}{2}\;l_i(\mathbf{x})\;\partial_y l_j(\mathbf{x})\,\mathrm{d}\mathbf{x}\Big].\end{aligned}$$

We define the matrices

$$
\begin{aligned}
(S^{(1)})_{ij} &:= \int_I l_i(\mathbf{x})\ \partial_x l_j(\mathbf{x}) d\boldsymbol{x},\\
(S^{(2)})_{ij} &:= \int_I l_i(\mathbf{x})\ \partial_y l_j(\mathbf{x}) d\boldsymbol{x},\\
(S^{(3)})_{ij} &:= \int_I \frac{1+x}{2} l_i(\mathbf{x})\ \partial_x l_j(\mathbf{x})\,\mathrm{d}\boldsymbol{x},\\
(S^{(4)})_{ij} &:= \int_I \frac{1+x}{2} l_i(\mathbf{x})\ \partial_y l_j(\mathbf{x})\,\mathrm{d}\boldsymbol{x},\\
(S^{(5)})_{ij} &:= \int_I \frac{1+y}{2} l_i(\mathbf{x})\ \partial_x l_j(\mathbf{x})\,\mathrm{d}\boldsymbol{x},\\
(S^{(6)})_{ij} &:= \int_I \frac{1+y}{2} l_i(\mathbf{x})\ \partial_y l_j(\mathbf{x})\,\mathrm{d}\boldsymbol{x}.
\end{aligned}
$$

Analogously to the one-dimensional case, one shows

$$
\begin{aligned}
S^{(1)} &= MD_x,\\
S^{(2)} &= MD_y,\\
S^{(3)} &= M_r^{(1)} D_x,\\
S^{(4)} &= M_r^{(1)} D_y,\\
S^{(5)} &= M_r^{(2)} D_x,\\
S^{(6)} &= M_r^{(2)} D_y.
\end{aligned}
$$

At last we define

$$
\begin{aligned}
S_r &:= v_{11} MD_x + e_{11} M_r^{(1)} D_x - e_{31} M_r^{(2)} D_x = M_r D_x,\\
S_z &:= v_{11} MD_y + e_{11} M_r^{(1)} D_y - e_{31} M_r^{(2)} D_y = M_r D_y,
\end{aligned}
$$

and get altogether

$$
\begin{aligned}
(S_r^k)_{ij} &= \int_{\Omega_k} l_i^k(\mathbf{r})\,\partial_r l_j^k(\mathbf{r})\, r\,\mathrm{d}\mathbf{r}\\
&= J_k\Big(v_{11}\big(x_r S_{ij}^{(1)} + y_r S_{ij}^{(2)}\big) + 2J_k x_z\big(y_r S_{ij}^{(3)} + x_r S_{ij}^{(4)}\big) - 2J_k x_z\big(x_r S_{ij}^{(5)} + y_r S_{ij}^{(6)}\big)\Big)\\
&= J^k (M_r^k)_{ij}\big(x_r (D_x)_{ij} + y_r (D_y)_{ij}\big),\\
(S_z^k)_{ij} &= \int_{\Omega_k} l_i^k(\mathbf{r})\,\partial_z l_j^k(\mathbf{r})\, r\,\mathrm{d}\mathbf{r}\\
&= J_k\Big(v_{11}\big(x_z S_{ij}^{(1)} + y_z S_{ij}^{(2)}\big) + 2J_k y_z\big(x_z S_{ij}^{(3)} + y_z S_{ij}^{(4)}\big) - 2J_k x_z\big(x_z S_{ij}^{(5)} + y_z S_{ij}^{(6)}\big)\Big)\\
&= J^k (M_r^k)_{ij}\big(x_z (D_x)_{ij} + y_z (D_y)_{ij}\big).
\end{aligned}
$$

□

**Remark 4.11.**
*We realize that*

$$
\begin{aligned}
(M_r^k)^{-1} S_r^k &= x_r D_x + y_r D_y,\\
(M_r^k)^{-1} S_z^k &= x_z D_x + y_z D_y.
\end{aligned}
$$

*That means the BOR stiffness matrices do not have to be computed explicitly. Only the matrices $D_x, D_y$ are needed which are already available (see [11]).*

### Computation of the Face Matrix

It remains to evaluate the flux expression on the right hand side of the semi-discrete scheme (4.4.3), that is, we need to compute integrals of the form

$$\mathcal{E}^k := \int_{\partial\Omega_k} {\mathbf{G}_E}^k(\mathbf{r})\; l_i^k(\mathbf{r})\, r\,\mathrm{d}\mathbf{r},$$

where ${\mathbf{G}_E}^k$ is one of the flux terms in (4.4.4) on $\Omega_k$.
In our case $\Omega_k$ is a triangle, and thus we can decompose its boundary $\partial\Omega_k$ into its three edges $e_1$, $e_2$, and $e_3$, so that we have

$$\mathcal{E}^k = \int_{\partial\Omega_k} {\mathbf{G}_E}^k(\mathbf{r})\; l_i^k(\mathbf{r})\, r\,\mathrm{d}\mathbf{r} = \sum_{m=1}^{3} \int_{e_m} {\mathbf{G}_E}^k(\mathbf{r})\; l_i^k(\mathbf{r})\, r\,\mathrm{d}\mathbf{r}, \tag{4.4.15}$$

Inserting the expansion of the fields into basis functions $l_j^k$, we find

$$\int_{e_m} \mathbf{G}^k(\mathbf{r})\; l_i^k(\mathbf{r})\, r\,\mathrm{d}\mathbf{r} = \sum_{j=1}^{N_p} \mathbf{G}_j^k \int_{e_m} l_i^k(\mathbf{r})\, l_j^k(\mathbf{r})\, r\,\mathrm{d}\mathbf{r}. \tag{4.4.16}$$

Recall the face matrix defined in (4.4.5), i.e.

$$(\mathcal{F}^k)_{ij} = \int_{\partial\Omega_k} l_i^k(\mathbf{r})\, l_j^k(\mathbf{r})\; r\,\mathrm{d}\mathbf{r}.$$

In (4.4.16) we see we can split the face matrix integral into three integrals over the three edges $e_1$, $e_2$, and $e_3$; thus, by letting

$$M_{ij}^{(k,e_m)} := \int_{e_m} l_i^k(\mathbf{r})\, l_j^k(\mathbf{r})\, r\,\mathrm{d}\mathbf{r},$$

expression (4.4.16) can be written as a matrix-vector product of the discrete flux vector $\mathbf{G}^k$ and $M^{(k,e_m)}$; the discretized version of the flux expression in (4.4.3) is thus given as $\mathcal{F}^k\mathbf{G}^k = \sum_{m=1}^{3} M^{(k,e_m)}\mathbf{G}^k$. The matrix $M^{(k,e_m)}$ is very similar to the BOR mass matrix, with the distinction that the integration only runs over a certain edge of the element instead over its volume.
We can use one-dimensional orthogonal polynomials to compute this surface integral in an efficient way. As in the previous computations of the local matrices, we transform to a reference element; in one dimension, this is $I = [-1, 1]$. We collect all results and details of how to determine $\mathcal{E}^k$ in the next lemma.

**Lemma 4.12** (Surface integral)**.**
*The integral $\mathcal{E}^k$ can be approximated by*

$$\mathcal{E}^k \approx \sum_{m=1}^{3} \sum_{j=1}^{p+1} \mathbf{G}_j^k J_k^1 \int_{e_m} l_j^k(\mathbf{r})\, l_i^k(\mathbf{r})\; r\,\mathrm{d}\mathbf{r},$$

*where in one dimension, $N_p = p + 1$. Here, $J_k^1 = \frac{1}{2}\,\mathrm{vol}(\partial\Omega_k)$. The edge matrices*

$$M_{ij}^{e_m} = \int_{e_m} l_j^k(\mathbf{r})\, l_i^k(\mathbf{r})\; r\,\mathrm{d}\mathbf{r},\ (m = 1, 2, 3)$$

*can be composed into template matrices as*

$$M^{(k,e_m)} = \frac{|\mathbf{e}_m|}{2}\left(v_{m1} M^{1D} + \frac{e_{m1}}{2} M_r^{1D}\right),$$

*where*

$$M^{1D} := (V^{1D}(V^{1D})^T)^{-1},$$
$$M_r^{1D} := (V_r^{1D}(V_r^{1D})^T)^{-1}.$$

*The matrices $V^{1D}$ and $V_r^{1D}$ are the generalized one-dimensional Vandermonde matrices*

$$(V^{1D})_{i'j'} = P_{j'-1}^{(0,0)}(\tau_{i'}),$$
$$(V_r^{1D})_{i'j'} = P_{j'-1}^{(0,1)}(\tau_{i'}).$$

*The indices $i', j'$ result from a one-to-one correspondence between the indices $i, j \in \{1, \dots, N_p\}$ for the nodes $\mathbf{x}_i$ on $e_m$ to indices $i', j' \in \{1, \dots, p+1\}$ for the nodes $\tau_{i'}$ in $[-1, 1]$ with $\mathbf{x}_i = \boldsymbol{\gamma}_m(\tau_{i'})$. $\boldsymbol{\gamma}_m(\tau)$ exploits the transformation of the edges onto the reference interval $[-1, 1]$,*

$$\boldsymbol{\gamma}_m(\tau) = \mathbf{v}_m + \frac{1+\tau}{2}\mathbf{e}_m.$$

**Remark 4.13.**
*In the code we write the face matrix $\mathcal{F}^k$ in vector notation as $\mathbf{F} = [M^{k,e_1}, M^{k,e_2}, M^{k,e_2}]^T$. Also, we have the notation $J_k^1 \hat{=} \text{Fscale}$, $\mathcal{F}^k \hat{=} \text{LIFT}$. The face matrix is also called lift matrix (see [11]).*

*Proof.* The edges $\mathbf{e}_m$ are mapped onto the reference interval $[-1, 1]$ by the transformation

$$\boldsymbol{\gamma}_m(\tau) = (\gamma_{m1}, \gamma_{m2})^T = \mathbf{v}_m + \frac{1+\tau}{2}\mathbf{e}_m.$$

$M_{ij}^{(k,e_m)}$ thus becomes

$$M_{ij}^{(k,e_m)} = \frac{|\mathbf{e}_m|}{2}\int_{-1}^{1} \gamma_{m1}(\tau)\, l_i^k(\boldsymbol{\gamma}_m(\tau)) l_j^k(\boldsymbol{\gamma}_m(\tau))\, \mathrm{d}\tau.$$

The interpolation points are chosen such that there are $p+1$ points on each edge with identical distribution. Thus we can replace the element-specific two-dimensional Lagrange polynomials $l_i^k(\mathbf{r})$ by corresponding one-dimensional Lagrange polynomials $l_{i'}(\tau)$ on the reference interval, i.e.

$$M_{ij}^{(k,e_m)} = \frac{|\mathbf{e}_m|}{2}\int_{-1}^{1} (\gamma_{m1} + \frac{1+\tau}{2}e_{m1})\, l_{i'}(\tau) l_{j'}(\tau)\, \mathrm{d}\tau.$$

Here we consider a one-to-one correspondence between the indices $i, j \in \{1, \dots, N_p\}$ and the indices $i', j' \in \{1, \dots, p+1\}$ as explained in the lemma.
Similarly to the procedure of the BOR mass matrix, we can express the Lagrange polynomials in terms of Jacobi polynomials and exploit the orthogonality relations to find

$$M^{(k,e_m)} = \frac{|\mathbf{e}_m|}{2}\left(v_{m1}M^{1D} + \frac{e_{m1}}{2}M_r^{1D}\right).$$

□

## 4.5 Numerical Tests

From sections 3.4.1 and 4.2.1 we know that DG schemes applied to general hyperbolic equations have an optimal convergence rate of $O(h^{p+1})$, as in the one-dimensional case. We will demonstrate in this section that we find $p$-convergence also for our scheme applied

the two-dimensional BOR equations.
We will start with a first basic test by looking at a two-dimensional homogeneous cavity. Here, we check the p-convergence behavior as predicted in theory, see theorem 3.29.
For many realistic cases it is necessary to simulate open systems. Unfortunately, it is non-trivial to formulate and implement exact open boundary conditions. As a well-known alternative, one can add an absorbing layer around the computational domain. If this layer is designed in a way to absorb outgoing radiation without any reflections at its interface, it is called a perfectly matched layer (PML). We therefore extend our test systems by uniaxial perfectly matched layers (UPML), as introduced in section 2.4.2,

Another important extension is the total field/scattered field (TF/SF) approach which allows – within the discontinuous Galerkin approach – a relatively easy way to add sources to the system. As sources we will consider an incoming traveling wave and a traveling Gaussian pulse.

As a first basic test system including PML we will look at a homogeneous cavity with PML in $z-$direction in order to check the performance of the PML. As a second test, we consider a glass fiber with PML and insert a traveling wave or a Gaussian pulse, respectively, as an incoming wave. As a last test we look at a tapered fiber with a traveling Gaussian pulse as an incoming wave.

In numerical simulations with PML and TF/SF we have to regard several error sources which can influence the performance of the scheme. We basically expect two error sources: an error coming from the approximation to the exact Gaussian pulse, which is a convolution integral; and an error coming from the PML, like reflection effects. To judge whether our results are reasonable we orient on what is known from finite difference time-domain methods (see e.g. [22]). We therefore carry out the following tests on the performance of the PML and the approximation to the Gaussian pulse:

1. First basic test: Increasing the polynomial order $p$ should lead to a better spatial resolution.

2. Error coming from the source: If we increase the temporal width of the Gaussian pulse we should obtain a better approximation to the convolution integral.

3. Error coming from the PML: Ideally the PML absorbs the fields completely. Inside the layer, the fields decay exponentially fast, so that there occurs no reflection back into the medium. In numerics, there will be a reflection error. To keep this error small, first the PML should be wide enough (so that the exponential decay can unfold), secondly, we can vary the PML parameters such that the reflection gets minimal. From FDTD we expect the reflection decreases down to a certain minimal value (around $10^{-4}$, if not even better) and then increases again for increasing parameter values (using log-scaling). The values of the parameters corresponding to this reflection minimum are then optimal.

Before we come to the single test systems, we introduce the UPML and the TF/SF approach for BOR Maxwell's equations.

### 4.5.1 Uniaxial Perfectly Matched Layers

In order to simulate real physical systems, we need to be able to treat open systems numerically. As already introduced in section 2.4.2, this can be performed by using a perfectly matched layer (PML). Specifically, we use unixial PML (UPML) here. For our case of

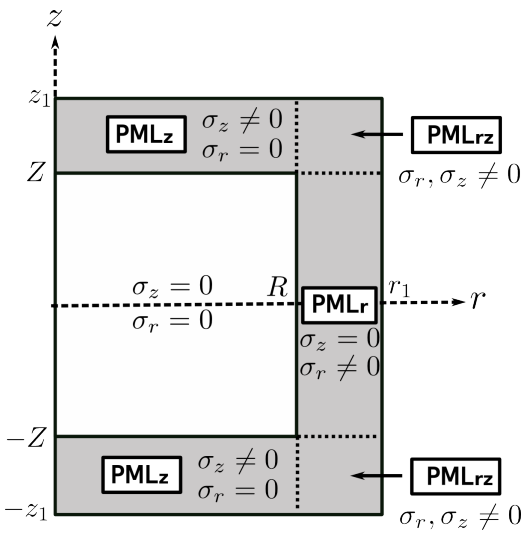



Figure 4.9: Sketch of a BOR–PML-region (gray) with PML-width $r_1 - R = z_1 - Z$. The physical region is $[0, R] \times [-Z, Z]$.

BOR Maxwell's equations, we need UPML in cylinder coordinates, which can be found in the paper by Teixeira and Chew [56].

We start by Fourier-transforming Maxwell's curl-equations to the frequency domain, i.e. (with the convention $\partial_t \leftrightarrow -\mathrm{i}\omega$)

$$\begin{aligned} \nabla \times \check{H} &= -\mathrm{i}\omega\epsilon\check{E}, \\ \nabla \times \check{E} &= \mathrm{i}\omega\mu\check{H}. \end{aligned} \tag{4.5.1}$$

By $\check{f}$ we denote the Fourier transform of a function $f$ as defined in section 2.4.2. We choose the UPML region as shown in figure 4.9. We have in $r$-direction the PML region $[R, r_1]$ and in $z-$direction the PML regions $[Z, z_1]$ and $[-Z, -z_1]$, respectively.

**Lemma 4.14** (BOR Maxwell's Equations with UPML)**.**
*BOR Maxwell's equations with UPML in cylinder coordinates read*

$$\begin{aligned} -\mathrm{i}\omega\epsilon\underline{\underline{\Lambda}}\check{E} &= \nabla \times \check{H}, \\ \mathrm{i}\omega\mu\underline{\underline{\Lambda}}\check{H} &= \nabla \times \check{E}, \end{aligned}$$

*where the tensor* $\underline{\underline{\Lambda}}$ *is given as*

$$\underline{\underline{\Lambda}} = \begin{pmatrix} \frac{s_\phi s_z}{s_r} & 0 & 0 \\ 0 & \frac{s_r s_z}{s_\phi} & 0 \\ 0 & 0 & \frac{s_\phi s_r}{s_z} \end{pmatrix}.$$

*Here,*

$$\begin{aligned} &s_r(r) = \kappa_r(r) - \frac{\sigma_r(r)}{\mathrm{i}\omega}, \qquad s_z(z) = \kappa_z(z) - \frac{\sigma_z(z)}{\mathrm{i}\omega}, \qquad s_\phi(r) = \frac{\tilde{r}(r)}{r}, \\ &\tilde{r}(r) = r_1 + \int_{r_1}^{r} s_r(r')dr'. \end{aligned}$$

For a proof, see [56]. We call $\sigma_r$ and $\sigma_z$ the *PML parameters*.

### Choice of the PML Parameters

Throughout we choose $\kappa_r(r) = \kappa_z(z) = 1$ so that

$$\sigma_r(r) = \sigma_r = \text{const}, \qquad \sigma_z(z) = \sigma_z = \text{const}\,.$$

Thus, we get

$$\begin{aligned}
&\tilde{r}(r) = r\left(1 - \frac{\sigma_r}{\mathrm{i}\omega}\right) + \frac{\sigma_r}{\mathrm{i}\omega} r_1 \text{ and} \\
&s_r = 1 - \frac{\sigma_r}{\mathrm{i}\omega}, \qquad s_z = 1 - \frac{\sigma_z}{\mathrm{i}\omega}, \qquad s_\phi = \left(1 - \frac{\sigma_r}{\mathrm{i}\omega}\right) + \frac{r_1\sigma_r}{r\,\mathrm{i}\omega}.
\end{aligned}$$

For PML in $r$-direction, it is $\sigma_r \neq 0$, $\sigma_z = 0$; for PML in $z$-direction, it is $\sigma_r = 0$, $\sigma_z \neq 0$, and in the corners $\sigma_r \neq 0$, $\sigma_z \neq 0$. In the medium, we have $\sigma_r = \sigma_z = 0$. See figure 4.9 for an illustration.

### UPML for BOR Maxwell's Equations

Thus, in frequency domain, Maxwell's equations (4.1.4) with UPMLs can be written component-wise as

$$\begin{aligned}
-\epsilon\mathrm{i}\omega\check{E}_r &= -\partial_z\check{H}_\varphi + \frac{\mathrm{i}m}{r}\check{H}_z + \frac{1}{r}\check{J}_r^{(E)}, \\
-\epsilon\mathrm{i}\omega\check{E}_\varphi &= -\partial_r\check{H}_z + \partial_z\check{H}_r + \frac{1}{r}\check{J}_\varphi^{(E)}, \\
-\epsilon\mathrm{i}\omega\check{E}_z &= \partial_r\check{H}_\varphi + \frac{1}{r}\check{H}_\varphi - \frac{\mathrm{i}m}{r}\check{H}_r + \frac{1}{r}\check{J}_z^{(E)}, \\
-\mu\mathrm{i}\omega\check{H}_r &= \partial_z\check{E}_\varphi - \frac{\mathrm{i}m}{r}\check{E}_z + \frac{1}{r}\check{J}_r^{(H)}, \\
-\mu\mathrm{i}\omega\check{H}_\varphi &= \partial_r\check{E}_z - \partial_z\check{E}_r + \frac{1}{r}\check{J}_\varphi^{(H)}, \\
-\mu\mathrm{i}\omega\check{H}_z &= -\partial_r\check{E}_\varphi - \frac{1}{r}\check{E}_\varphi - \frac{\mathrm{i}m}{r}\check{E}_r + \frac{1}{r}\check{J}_z^{(H)},
\end{aligned} \tag{4.5.2}$$

with the polarization currents $\check{\mathbf{J}}^{(E)}$ and $\check{\mathbf{J}}^{(H)}$, which are introduced such that $\mathrm{i}\omega$ is eliminated. Then we introduce a new variable $P_i$ with $i \in \{r, \varphi, z\}$, which results in an additional equation for $P_i$, a so-called *auxiliary differential equation* (ADE). In the end, we get 12 equations in total for the electromagnetic fields $\mathbf{E}, \mathbf{H}$ and the corresponding polarizations $\mathbf{P}^{(E)}, \mathbf{P}^{(H)}$, as stated in the next lemma.

**Lemma 4.15** (BOR Maxwell's Equations with UPML)**.**

$$
\begin{aligned}
r\epsilon\partial_t E_r &= -r\partial_z H_\varphi + \mathrm{i}mH_z + P_r^{(E)} + \epsilon(r_1\sigma_r - r\sigma_z)E_r,\\
\partial_t P_r^{(E)} &= -\sigma_r P_r^{(E)} - \epsilon\sigma_r r_1(\sigma_r - \sigma_z)E_r,\\
r\epsilon\partial_t E_\varphi &= -r\partial_r H_z + r\partial_z H_r + P_\varphi^{(E)} - \epsilon(r_1\sigma_r + r\sigma_z)E_\varphi,\\
r\partial_t P_\varphi^{(E)} &= -r\sigma_r P_\varphi^{(E)} + r_1\sigma_r P_\varphi^{(E)} - \epsilon\sigma_r r_1((r_1 - r)\sigma_r - r\sigma_z)E_\varphi,\\
r\epsilon\partial_t E_z &= H_\varphi + \partial_r H_\varphi - \mathrm{i}mH_r + P_z^{(E)} + \epsilon((r_1 - 2r)\sigma_r + r\sigma_z)E_z,\\
\partial_t P_z^{(E)} &= -\sigma_z P_z^{(E)} + \epsilon\sigma_r^2(r_1 - r)E_z + \epsilon\sigma_z((r_1 - 2r)\sigma_r - r\sigma_z)E_z,\\
r\mu\partial_t H_r &= r\partial_z E_\varphi - \mathrm{i}mE_z + P_r^{(H)} + \mu(r_1\sigma_r - r\sigma_z)H_r,\\
\partial_t P_r^{(H)} &= -\sigma_r P_r^{(H)} - \mu\sigma_r r_1(\sigma_r - \sigma_z)H_r,\\
r\mu\partial_t H_\varphi &= r\partial_r E_z - r\partial_z E_r + P_\varphi^{(H)} - \mu(r_1\sigma_r + r\sigma_z)H_\varphi,\\
r\partial_t P_\varphi^{(H)} &= -r\sigma_r P_\varphi^{(H)} + r_1\sigma_r P_\varphi^{(H)} - \mu\sigma_r r_1((r_1 - r)\sigma_r + r\sigma_z)H_\varphi,\\
r\mu\partial_t H_z &= -E_\varphi - \partial_r E_\varphi + \mathrm{i}mE_r + P_z^{(H)} + \mu((r_1 - 2r)\sigma_r + r\sigma_z)H_z,\\
\partial_t P_z^{(H)} &= -\sigma_z P_z^{(H)} + \mu\sigma_r^2(r_1 - r)H_z + \mu\sigma_z((r_1 - 2r)\sigma_r - r\sigma_z)H_z,
\end{aligned}
\tag{4.5.3}
$$

*where we have one auxiliary differential equation for each component within the PML. It should be noted that these auxiliary differential equations do not contain spatial derivatives, which means that no modification of the numerical flux is needed [105, 106].*

*Proof.* We only demonstrate the computations for the **E**-field. The principle is the same for the **H**-field. The equations differ in a sign (a small but very important difference).

We begin with the equation for $E_r$:

Define

$$\check{J}_r := \mathrm{i}\omega\epsilon r\left(\frac{s_\phi s_z}{s_r} - 1\right)\check{E}_r.$$

Using the relation

$$
\begin{aligned}
\frac{s_\phi s_z}{s_r} - 1 &= \frac{\left(1 - \frac{\sigma_r}{\mathrm{i}\omega} + \frac{\sigma_r r_1}{r\,\mathrm{i}\omega}\right)\left(1 - \frac{\sigma_z}{\mathrm{i}\omega}\right)}{1 - \frac{\sigma_r}{\mathrm{i}\omega}} - 1 = \frac{\left(\mathrm{i}\omega - \sigma_r + \frac{\sigma_r r_1}{r}\right)\left(1 - \frac{\sigma_z}{\mathrm{i}\omega}\right)}{\mathrm{i}\omega - \sigma_r} - 1\\
&= \frac{1}{r(\mathrm{i}\omega - \sigma_r)}\left(\frac{\sigma_r\sigma_z}{\mathrm{i}\omega}(r - r_1) - r\sigma_z + r_1\sigma_r\right),
\end{aligned}
$$

$\check{J}_r$ becomes

$$\check{J}_r = \frac{\mathrm{i}\omega\epsilon}{\mathrm{i}\omega - \sigma_r}\left(\frac{\sigma_r\sigma_z}{\mathrm{i}\omega}(r - r_1) - r\sigma_z + r_1\sigma_r\right)\check{E}_r.$$

Now introduce the new variable

$$\check{P}_r^E := \check{J}_r + x\check{E}_r.$$

$x$ stands for an unknown expression which shall be determined such that $\mathrm{i}\omega$ drops out. Making the ansatz

$$(\mathrm{i}\omega - \sigma_r)\check{P}_r^E = \mathrm{i}\omega\epsilon\left(\frac{\sigma_r\sigma_z}{\mathrm{i}\omega}(r - r_1) - r\sigma_z + r_1\sigma_r\right)\check{E}_r + (\mathrm{i}\omega - \sigma_r)x\check{E}_r,$$

we come to

$$\mathrm{i}\omega x - \mathrm{i}\omega\epsilon r\sigma_z + \mathrm{i}\omega\epsilon r_1\sigma_r = 0 \Leftrightarrow x = \epsilon(r\sigma_z - r_1\sigma_r),$$

and it follows:

$$-\mathrm{i}\omega\check{P}_r^E = -\sigma_r\check{P}_r^E + \epsilon\sigma_r r_1(\sigma_z - \sigma_r)\check{E}_r.$$

After Fourier transformation we obtain the equation for $E_r$ and the ADE for $P_r^{(E)}$.

We resume with the equation for $E_\varphi$:

As before we define

$$\check{J}_\varphi := \mathrm{i}\omega\epsilon r\left(\frac{s_r s_z}{s_\phi} - 1\right)\check{E}_\varphi$$

and use the relation

$$\frac{s_r s_z}{s_\phi} - 1 = \frac{(\mathrm{i}\omega - \sigma_r)(1 - \frac{\sigma_z}{\mathrm{i}\omega})}{\mathrm{i}\omega - \sigma_r + \sigma_r\frac{r_1}{r}} - 1 = \frac{-\sigma_z + \frac{\sigma_r\sigma_z}{\mathrm{i}\omega} - \sigma_r\frac{r_1}{r}}{\mathrm{i}\omega - \sigma_r + \sigma_r\frac{r_1}{r}},$$

leading to

$$\check{J}_\varphi = \frac{r\ \mathrm{i}\omega\epsilon}{r\ \mathrm{i}\omega - \sigma_r + \sigma_r r_1}\left(r\frac{\sigma_r\sigma_z}{\mathrm{i}\omega} - r\sigma_z - r_1\sigma_r\right)\check{E}_\varphi.$$

Define

$$\check{P}_\varphi^E := \check{J}_\varphi + x\check{E}_\varphi.$$

Then we have

$$(r\ \mathrm{i}\omega - \sigma_r + \sigma_r r_1)\check{P}_\varphi^E = r\ \mathrm{i}\omega\epsilon\left(r\frac{\sigma_r\sigma_z}{\mathrm{i}\omega} - r\sigma_z - r_1\sigma_r\right)\check{E}_\varphi + (r\ \mathrm{i}\omega - \sigma_r + \sigma_r r_1)x\check{E}_\varphi,$$

and we obtain

$$r\mathrm{i}\omega x - r^2\ \mathrm{i}\omega\epsilon\sigma_z - r\ \mathrm{i}\omega\epsilon r_1\sigma_r = 0 \Leftrightarrow x = \epsilon(r\sigma_z + r_1\sigma_r).$$

Consequently,

$$-r\ \mathrm{i}\omega\check{P}_\varphi^E = -r\sigma_r\check{P}_\varphi^E + \sigma_r r_1 P_\varphi^E - \epsilon\sigma_r r_1[r(\sigma_z - \sigma_r) + \sigma_r r_1]\check{E}_\varphi.$$

Finally, we present the details for the equation for $E_z$:

Define

$$\check{J}_z := \mathrm{i}\omega\epsilon r\left(\frac{s_r s_\phi}{s_z} - 1\right)\check{E}_z.$$

As before, it is

$$\frac{s_r s_\phi}{s_z} - 1 = \frac{-2r\sigma_r + \sigma_r r_1 + \frac{\sigma_r^2}{\mathrm{i}\omega}(r - r_1) + r\sigma_z}{r(\mathrm{i}\omega - \sigma_z)},$$

therefore we come to

$$\check{J}_z = \frac{\mathrm{i}\omega\epsilon}{\mathrm{i}\omega - \sigma_z}\left(\sigma_r(r_1 - 2r) + \frac{\sigma_r^2}{\mathrm{i}\omega}(r - r_1) + r\sigma_z\right)\check{E}_z.$$

We introduce

$$\check{P}_z^E := \check{J}_z + x\check{E}_z,$$

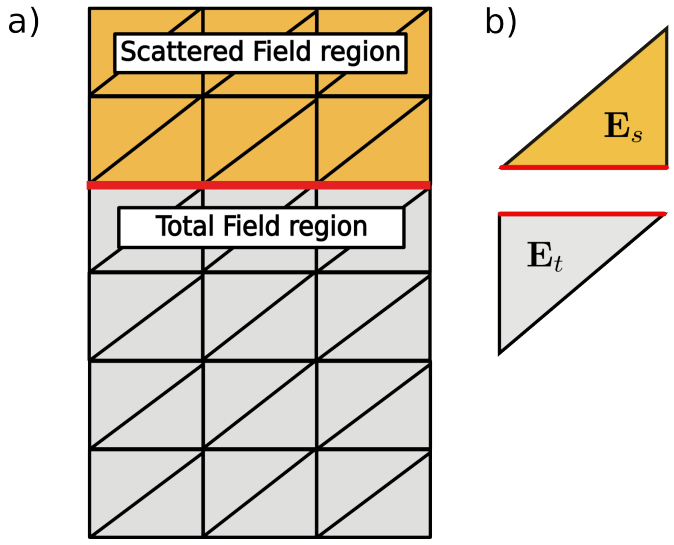



Figure 4.10: a) Splitting of a region into Scattered Field and Total Field region. (b) Two triangles at the interface between Total Field and Scattered Field.

resulting in

$$(\mathrm{i}\omega - \sigma_z)\check{P}_z^E = \mathrm{i}\omega\epsilon\left(\sigma_r(r_1 - 2r) + \frac{\sigma_r^2}{\mathrm{i}\omega}(r - r_1) + r\sigma_z\right)\check{E}_z + (\mathrm{i}\omega - \sigma_z)x\check{E}_z,$$

and finally $x = \epsilon[\sigma_r(2r - r_1) - r\sigma_z]$, so that

$$-\mathrm{i}\omega\check{P}_z^E = -\sigma_z\check{P}_z^E - \epsilon\sigma_r^2(r - r_1)\check{E}_z + \epsilon\sigma_z[\sigma_r(2r - r_1) - r\sigma_z]\check{E}_z.$$

□

### 4.5.2 Sources and the Total Field/Scattered Field Approach

The interesting questions and phenomena in optics arise when incoming light meets the physical system. Depending on the strength, the polarization, the frequency and direction of light the system will react accordingly.
As a second step towards more physical systems we thus introduce sources into our model. A very popular way for *linear* problems is the so-called total field/scattered field (TF/SF) approach; see e.g. the review article [13] and the references therein for many details. Within the TF/SF approach the total electromagnetic fields can be split into a scattered and incident part as

$$\begin{aligned}\mathbf{E}_{\text{total}}(\mathbf{x}, t) &= \mathbf{E}_{\text{in}}(\mathbf{x}, t) + \mathbf{E}_{\text{scat}}(\mathbf{x}, t),\\ \mathbf{H}_{\text{total}}(\mathbf{x}, t) &= \mathbf{H}_{\text{in}}(\mathbf{x}, t) + \mathbf{H}_{\text{scat}}(\mathbf{x}, t).\end{aligned} \tag{4.5.4}$$

It is exactly the incident field which represents a given mathematical expression of the incoming light. Given the total field, the scattered field can be determined; given the scattered field, we can compute the total field. Figure 4.10 shall illustrate the situation.
Since we work with linear Maxwell's equation, we solve in each region for the scattered and the total field. In conservative form (2.6.11) the equations to solve read

$$\begin{aligned}&\mathbf{Q}\partial_t\mathbf{u}_{\text{total}} + \nabla\cdot\mathbf{F}(\mathbf{u}_{\text{total}}) = 0 \quad \text{in the TF region,}\\ &\mathbf{Q}\partial_t\mathbf{u}_{\text{scat}} + \nabla\cdot\mathbf{F}(\mathbf{u}_{\text{scat}}) = 0 \quad \text{in the SF region.}\end{aligned}$$

If we are in the TF or SF region itself, the equations are the familiar Maxwell's equations. There we have the known field differences

$$\Delta\mathbf{E}_{\text{total}}(\mathbf{x},t) = \mathbf{E}^{\text{ext}}_{\text{total}}(\mathbf{x},t) - \mathbf{E}^{\text{int}}_{\text{total}}(\mathbf{x},t),$$
$$\Delta\mathbf{E}_{\text{scat}}(\mathbf{x},t) = \mathbf{E}^{\text{ext}}_{\text{in}}(\mathbf{x},t) - \mathbf{E}^{\text{int}}_{\text{scat}}(\mathbf{x},t).$$

With "int" we mean the interior of the local cell and with "ext" the exterior, i.e. the neighboring cell. But on the interface between total field and scattered field, this is not true anymore. Although we may consider Maxwell's equations for the total field in the TF region (or the other way around if we are in the SF region), at the interface we suddenly switch between the total and scattered field. We therefore need to adjust the field differences in the numerical flux.
Due to (4.5.4) we can express the total field by means of the scattered and incident field, that is

$$\mathbf{E}_{\text{total}} = \mathbf{E}_{\text{scat}} + \mathbf{E}_{\text{in}},$$

and we can express the scattered field by means of the total and incident field, i.e.

$$\mathbf{E}_{\text{scat}} = \mathbf{E}_{\text{total}} - \mathbf{E}_{\text{in}}.$$

The adjustment of the numerical flux at the interface can be accomplished as follows. Imagine to be in a cell in the TF region. There, the field difference is given as

$$\Delta\mathbf{E}_{\text{total}} = \mathbf{E}^{\text{ext}}_{\text{total}} - \mathbf{E}^{\text{int}}_{\text{total}}.$$

In a neighboring cell in the SF region we only have $\mathbf{E}^{\text{ext}}_{\text{scat}}$. In order to get an equality, we need to add the incident field value to get the total field as

$$\mathbf{E}^{\text{ext}}_{\text{total}} = \mathbf{E}^{\text{ext}}_{\text{scat}} + \mathbf{E}^{\text{ext}}_{\text{in}}.$$

Now we look from the other side. Imagine to be in a cell in the SF region. At the interface from SF to TF region we need to adjust the expression

$$\mathbf{E}^{\text{ext}}_{\text{scat}} = \mathbf{E}^{\text{ext}}_{\text{total}} - \mathbf{E}^{\text{ext}}_{\text{in}}.$$

Collecting everything, the field differences at the interface therefore are

$$\Delta\mathbf{E}_{\text{total}} = \mathbf{E}^{\text{ext}}_{\text{scat}} + \mathbf{E}^{\text{ext}}_{\text{in}} - \mathbf{E}^{\text{int}}_{\text{total}} = \mathbf{E}^{\text{ext}}_{\text{scat}} - \mathbf{E}^{\text{int}}_{\text{total}} + \mathbf{E}_{\text{in}},$$
$$\Delta\mathbf{E}_{\text{scat}} = \mathbf{E}^{\text{ext}}_{\text{total}} - \mathbf{E}^{\text{ext}}_{\text{in}} - \mathbf{E}^{\text{int}}_{\text{scat}} = \mathbf{E}^{\text{ext}}_{\text{total}} - \mathbf{E}^{\text{int}}_{\text{scat}} - \mathbf{E}_{\text{in}}.$$

So the only adjustment consists in adding or substracting the incident field to the total/scattered field differences.

In our simulations we considered traveling waves and a traveling Gaussian pulse in $z$-direction, respectively. We remark that the Gaussian pulse we use is not a solution to Maxwell's equations. It is an approximation to the true solution which is – after Fourier transformation – a convolution integral [32], [33]. This integral might be computed exactly (in the sense that the integral is approximated by some quadrature formula, e.g.), but the effort involved might lead to inefficiency with respect to computational time and memory. Additionally, there is an error coming from the approximation of the integral as well. Indeed, the approximation we use is "relatively good", and it becomes better for increasing values of the Gaussian pulse width $\sigma$. We look at this in the next section in more detail.

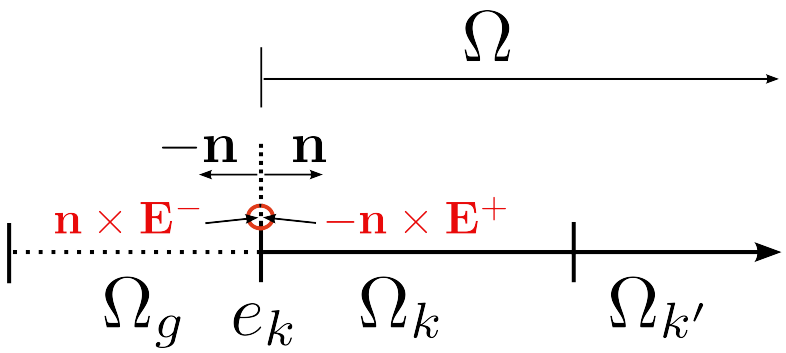



Figure 4.11: A sketch for the relation $\mathbf{n} \times \mathbf{E}^+ = -\mathbf{n} \times \mathbf{E}^-$ in case of perfectly electric conducting boundary conditions $\mathbf{n} \times \mathbf{E} = 0$.

### 4.5.3 Homogeneous Cavity as a Test System

As a first basic test we consider a homogeneous cylindrical resonator of radius $R = 1$ and length $L = 1$, i.e. $\Omega = [0, 1]^2$. We choose TM mode with $m = 1$ as an initial condition and let the fields evolve for 10 periods of oscillation. For this case the BOR equations (4.4.1) reduce to

$$\begin{aligned}
-\frac{1}{r\mu}(\mathrm{i}mE_z - r\partial_z E_\varphi) &= \partial_t H_r, \\
-\frac{1}{\mu}(\partial_z E_r - \partial_r E_z) &= \partial_t H_\varphi, \\
-\frac{1}{\epsilon}\partial_z H_\varphi &= \partial_t E_r, \\
\frac{1}{\epsilon}\partial_z H_r &= \partial_t E_\varphi, \\
\frac{1}{r\epsilon}(\partial_r(rH_\varphi) - \mathrm{i}mH_r) &= \partial_t E_z
\end{aligned}$$

in the weak sense (with respect to the measure $r\,\mathrm{d}\mathbf{r}$). We choose perfectly electric conducting (PEC) boundary conditions, i.e. $\mathbf{n} \times \mathbf{E} = 0$. At a boundary element we therefore require

$$\mathbf{n} \times \mathbf{E}^+ = -\mathbf{n} \times \mathbf{E}^-,$$

or, equivalently, $[\![\mathbf{E}]\!] = -2\mathbf{E}^-$. Figure 4.11 shall motivate an explanation of this relation in one dimension. Let $\Omega_k$ be a boundary element and $\Omega_{k'}$ its neighboring element. Imagine to have a so-called ghost cell $\Omega_g$ left to $\Omega_k$; that is, formally, $\Omega_g \notin \Omega$. At the edge $e_k := \partial\Omega_k \cap \Omega_g$, we require $\mathbf{n} \times \mathbf{E} = 0$, and $\mathbf{n}$ is the outer normal at $e_k$ pointing from $\Omega_k$ to $\Omega_g$; $-\mathbf{n}$ is the outer normal at $e_k$ pointing in the opposite direction, i.e. from $\Omega_g$ to $\Omega_k$. Therefore, in order to obtain $\mathbf{n} \times \mathbf{E} = 0$ on $e_k$, the value on $e_k$ coming from $\Omega_g$, i.e. $\mathbf{n} \times \mathbf{E}^-$, and the value on $e_k$ coming from $\Omega_k$, that is, $-\mathbf{n} \times \mathbf{E}^+$, must be the same, which means: $\mathbf{n} \times \mathbf{E}^+ = -\mathbf{n} \times \mathbf{E}^-$. For the $\mathbf{H}$-field it is $mathbfn \times \mathbf{H}^+ = \mathbf{n} \times \mathbf{H}^-$, i.e. $[\![\mathbf{H}]\!] = 0$.

For this system the exact solutions are known:

$$
\begin{aligned}
E_z(r,\varphi,z,t) &= J_m(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}\cos\left(\frac{\kappa\pi z}{L}\right)e^{-\mathrm{i}\omega_{mn}t},\\
E_r(r,\varphi,z,t) &= -\frac{\kappa\pi}{L\,\gamma_{mn}}J_m'(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}\sin\left(\frac{\kappa\pi z}{L}\right)e^{-\mathrm{i}\omega_{mn}t},\\
E_\varphi(r,\varphi,z,t) &= -\frac{\mathrm{i}m\kappa\pi}{r\,L\,\gamma_{mn}^2}J_m(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}\sin\left(\frac{\kappa\pi z}{L}\right)e^{-\mathrm{i}\omega_{mn}t},\\
H_r(r,\varphi,z,t) &= \frac{m\epsilon\omega_{mn}}{r\,\gamma_{mn}^2}J_m(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}\cos\left(\frac{\kappa\pi z}{L}\right)e^{-\mathrm{i}\omega_{mn}t},\\
H_\varphi(r,\varphi,z,t) &= \frac{\mathrm{i}\epsilon\omega_{mn}}{\gamma_{mn}}J_m'(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}\cos\left(\frac{\kappa\pi z}{L}\right)e^{-\mathrm{i}\omega_{mn}t}.
\end{aligned}
$$

Here, the $J_m$ are the Bessel functions of the first kind, $\gamma_{mn}$ its $n$th zero, $\kappa \in \mathbb{N}$ is the oscillation number, $\omega_{mn} = c\,\sqrt{\left(\frac{\gamma_{mn}}{R}\right)^2 + \left(\frac{\kappa\pi}{L}\right)^2}$ the frequency and $c = \frac{1}{\sqrt{\epsilon\mu}}$ the speed of light.

**Remark 4.16.**
*We remark the exact solutions are determined from the relations*

$$
\begin{aligned}
\mathbf{E_t} &= \frac{1}{\gamma^2}(\nabla_\mathbf{t}(\partial_z E_z) - \frac{\mathrm{i}\omega}{c}(\hat{\mathbf{e}}_\mathbf{z}\times\nabla_\mathbf{t})H_z),\\
\mathbf{H_t} &= \frac{1}{\gamma^2}(\nabla_\mathbf{t}(\partial_z H_z) + (\mathrm{i}\omega\epsilon)(\hat{\mathbf{e}}_\mathbf{z}\times\nabla_\mathbf{t})E_z).
\end{aligned}
$$

*The index "$\mathbf{t}$" denotes the transversal part of the electric fields, i.e.* $\mathbf{E_t} = (E_r, E_\varphi)$*, and* $\nabla_\mathbf{t}$ *is the transversal derivative with*

$$
\nabla_\mathbf{t} := \begin{pmatrix} \partial_r \\ \frac{1}{r}\partial_\varphi \end{pmatrix}.
$$

For a test run we chose – without physical motivation – $\kappa = 1,\ \epsilon = 1,\ \mu = 1,\ m = 1,\ n = 1$. At each time step, we record the $L^2$-error of the approximate solution $E_z$ and the analytical solution $E_z^h$, i.e.

$$
||\mathbf{E_z^h} - \mathbf{E_z}||_{L^2} = \frac{||\mathbf{E_z^h} - \mathbf{E_z}||_2}{||\mathbf{E_z^h}||_2}.
$$

In figure 4.12 we plot the $L^2$-error over the entire time in logarithmic scale, for increasing polynomial order $p$ and decreasing maximal edge length $h$. We observe an error behavior of $O(h^{p+1})$, as we know from the Cartesian case.

### 4.5.4 Homogeneous Cavity with PML in $z$-direction

We consider the test system as depicted in figure 4.13. Here, we include PML in negative $z$-direction with the following parameter settings:

- $\sigma_r = 0, \quad \sigma_z = 7.$
- The PML width shall be 0.5, i.e. we define the PML region in $z$-direction as $\mathrm{PML}_z := [0, 1.5] \times [-1, -1.5]$.

Furthermore we inject a traveling wave or a Gaussian pulse in $z$-direction with wave number $k = -2\pi$ at the interface between SF and TF region, which is located at $Z = 1$. We thus consider the SF region $\Omega_{\mathrm{SF}} := [0, 1]\times[1, 1.5]$ and the TF region $\Omega_{\mathrm{TF}} := [0, 1]\times[-1.5, 1]$. The entire region is $\Omega = [0, 1] \times [-1.5, 1.5]$. We thus have a cylinder of length 3. For our

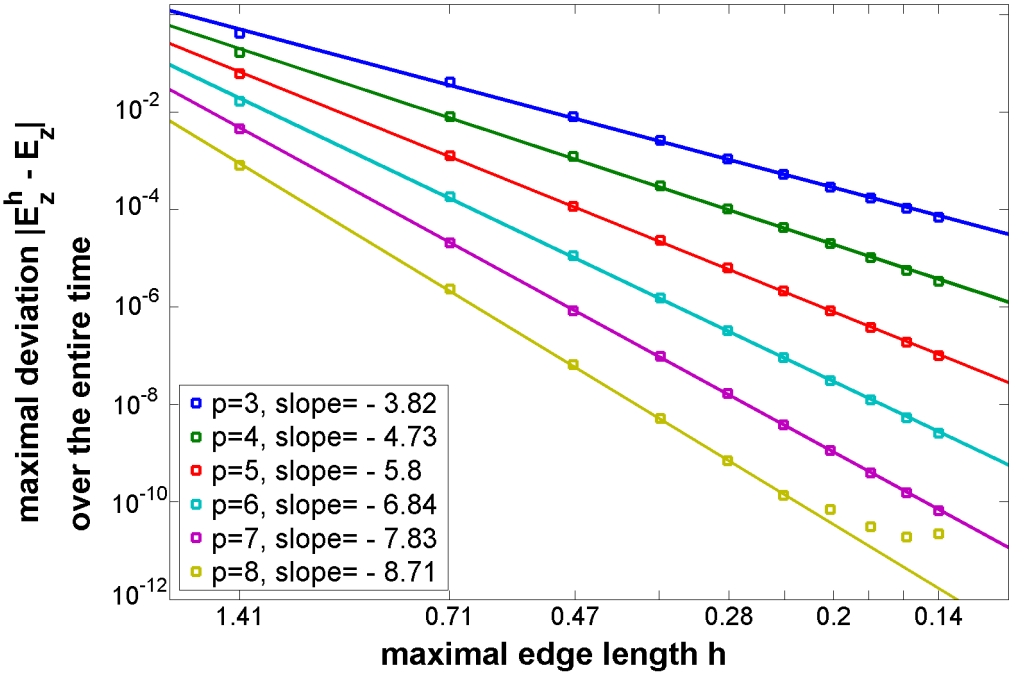



Figure 4.12: Error plot for $E_z$ in a homogeneous medium with $\epsilon = 1, \mu = 1,\ m = 1$ for increasing polynomial order $p \in \{3, \dots, 9\}$ and decreasing edge length $h$.

simulations we chose the polynomial order $p = 5$, and we impose PEC boundary conditions on the outer cylinder wall.
As mentioned previously, we take a traveling wave or a traveling Gaussian pulse as source, which are given in the following.

#### Traveling waves

We consider a closed system with PEC boundary conditions and initialize traveling waves in $z-$direction into the system. These are given as (see e.g. [100], [99], [107], and section 4.5.3)

$$
\begin{aligned}
E_z(r,\varphi,z,t) &= J_m(\gamma_{mn}\, r) e^{\mathrm{i}m\varphi} e^{\pm \mathrm{i}kz} e^{-\mathrm{i}\omega_{mn} t},\\
E_r(r,\varphi,z,t) &= \pm \frac{\mathrm{i}k}{\gamma_{mn}} J_m'(\gamma_{mn}\, r) e^{\mathrm{i}m\varphi} e^{\pm \mathrm{i}kz} e^{-\mathrm{i}\omega_{mn} t},\\
E_\varphi(r,\varphi,z,t) &= \pm \frac{\mathrm{i}mk}{r\,\gamma_{mn}^2} J_m(\gamma_{mn}\, r) e^{\mathrm{i}m\varphi} e^{\pm \mathrm{i}kz} e^{-\mathrm{i}\omega_{mn} t},\\
H_r(r,\varphi,z,t) &= \frac{m\epsilon\omega_{mn}}{r\,\gamma_{mn}^2} J_m(\gamma_{mn}\, r) e^{\mathrm{i}m\varphi} e^{\pm \mathrm{i}kz} e^{-\mathrm{i}\omega_{mn} t},\\
H_\varphi(r,\varphi,z,t) &= \frac{\mathrm{i}\epsilon\omega_{mn}}{\gamma_{mn}} J_m'(\gamma_{mn}\, r) e^{\mathrm{i}m\varphi} e^{\pm \mathrm{i}kz} e^{-\mathrm{i}\omega_{mn} t},\\
H_z(r,\varphi,z,t) &= 0.
\end{aligned}
$$

Here, $k$ is the wave number (which can be chosen arbitrarily), $\omega_{mn} = c\,\sqrt{\left(\frac{\gamma_{mn}}{R}\right)^2 + k^2}$ is the frequency and $c = \frac{1}{\sqrt{\epsilon\mu}}$ is the speed of light. We chose $k = \frac{2\pi}{\lambda} = -2\pi$ with the wavelength $\lambda = 1$. Traveling waves serve as a first test on the performance of the PML and the sources within the TF/SF framework.

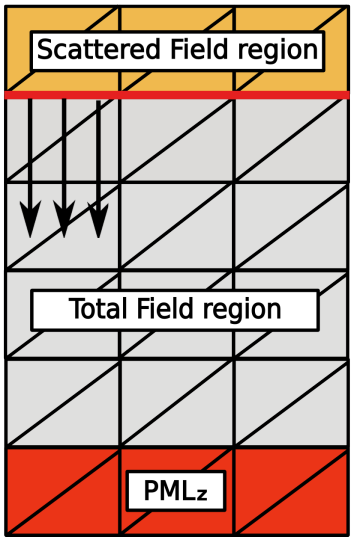



Figure 4.13: A sketch of a test system with scattered field (SF) and total field (TF) region and PML in $z$-direction. A traveling wave is launched into the system at the TF/SF edge as indicated by the black arrows.

### A Gaussian Pulse traveling in $z$-direction

We aim at simulating the wave propagation in a glass fiber. For this we multiply the exact solution known from the homogeneous test case with

$$A(t) := e^{-\mathrm{i}\omega(t_0-t)}e^{-(t_0-t)^2/(2\sigma^2)}.$$

$A(t)$ consists of an oscillating part with frequency $\omega$ and a Gaussian pulse envelope, where $\sigma$ is its width and $t_0$ is its center. For a motivation of this choice, see [13, Ch. A.1.3]. The frequency is given as $\omega = k_0 c$, where $k_0 = 2\pi\lambda$ is the free-space wave number and $\lambda$ is the wave number. We have chosen $\lambda = 1.5$ in our simulations. $\sigma$ should be a multiple of $t_0$, i.e. $\sigma = at_0$ for some $a \in \mathbb{N}$. By multiplying the exact solution of the homogeneous case from the previous subsection with $A(t)$ instead of $e^{-\mathrm{i}\omega_{mn}t}$, we obtain

$$\begin{aligned}
E_z(r,\varphi,z,t) &= J_m(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}e^{\mathrm{i}kz}e^{-\mathrm{i}\omega(t_0-t)}e^{-(t_0-t)^2/(2\sigma^2)},\\
E_r(r,\varphi,z,t) &= \frac{\mathrm{i}k}{\gamma_{mn}}J_m'(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}e^{\mathrm{i}kz}e^{-\mathrm{i}\omega(t_0-t)}e^{-(t_0-t)^2/(2\sigma^2)},\\
E_\varphi(r,\varphi,z,t) &= \frac{\mathrm{i}mk}{r\,\gamma_{mn}^2}J_m(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}e^{\mathrm{i}kz}e^{-\mathrm{i}\omega(t_0-t)}e^{-(t_0-t)^2/(2\sigma^2)},\\
H_r(r,\varphi,z,t) &= \frac{m\epsilon\omega}{r\,\gamma_{mn}^2}J_m(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}e^{\mathrm{i}kz}e^{-\mathrm{i}\omega(t_0-t)}e^{-(t_0-t)^2/(2\sigma^2)},\\
H_\varphi(r,\varphi,z,t) &= \frac{\mathrm{i}\epsilon\omega}{\gamma_{mn}}J_m'(\gamma_{mn}\, r)e^{\mathrm{i}m\varphi}e^{\mathrm{i}kz}e^{-\mathrm{i}\omega(t_0-t)}e^{-(t_0-t)^2/(2\sigma^2)},\\
H_z(r,\varphi,z,t) &= 0.
\end{aligned} \tag{4.5.5}$$

In order to check the performance of the PML and sources, we consider the following tests, as already alluded to in the introduction of this section:

1. Increasing the polynomial order $p$ must lead to a better spatial resolution. We can see this in figure 4.14.

2. Error coming from the source: For bigger $t_0$ and $\sigma$ the error coming from the Gaussian pulse source must decrease (see e.g. [13, Ch. A.1.3]). We chose successively

$$t_0 = 1, \sigma = 5; \quad t_0 = 2, \sigma = 10; \quad t_0 = 3, \sigma = 15; \quad t_0 = 4, \sigma = 20.$$

The results are shown in figure 4.15. The first bump at the very beginning of the plot comes from the error made in the source approximation. Indeed we can see that this bump gets smaller and smaller for increasing $\sigma$.

3. Error coming from the PML: As a measure for the reflection we define

$$R := \frac{\max\limits_{\mathbf{x}\in\Omega_s, t>10\sigma} E_z(\mathbf{x},t)}{\max\limits_{r\in[0,1],\, t>0} E_z(r,Z,t)},$$

that is, $R$ is the ratio of the maximum of the $E_z$-field in the SF region for times bigger than $10\sigma$ to the maximum of the $E_z$-field at the TF/SF edge (in each time step); we choose $t > 10\sigma$ since we are interested in the error coming from the reflection alone, not from the source itself (this was examined in point 2). The source error dominates and it would distort $R$. In figure 4.15 the second bump gives the maximum value of the reflected field. We observe the correct translation of the center $t_0$ by approximately a factor of the length of the cylinder, taking into account the wavelength. $R$ can be understood as an approximation to the reflection coefficient, which is the ratio of the electric field strength of the reflected wave to that of the incident wave. The results are shown in figure 4.16. Indeed we see that $R$ decreases down to a minimal value around $10^{-4}$ and then increases again. This is the same behavior known from FDTD (see e.g. [22]).

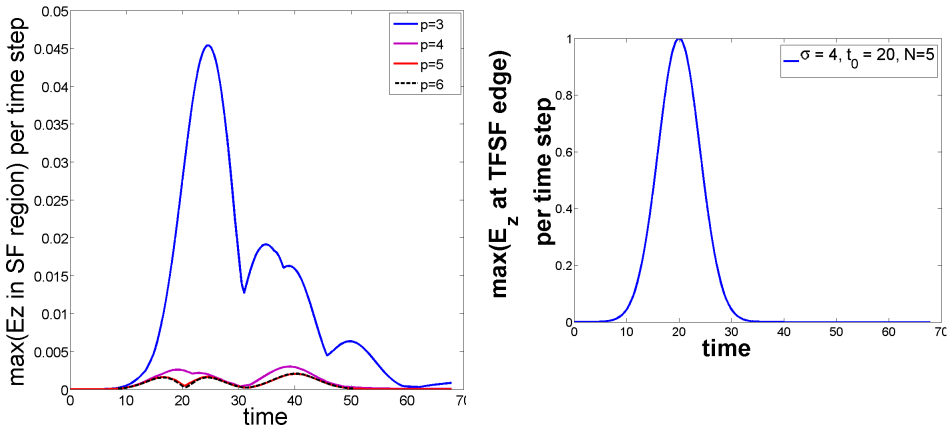



Figure 4.14: The spatial resolution of the fields gets better if the polynomial order $p$ is increased. Here, $t_0 = 1, \sigma = 5$, the simulation time is 30 units.

### 4.5.5 A Glass Fiber with PML

In order to simulate the propagation of electromagnetic waves in a glass fiber, we successively extend the systems from the previous sections in the subsequent manner, namely:

(1) We add PML in $z$-direction, and we consider a finite system with PEC boundary conditions.

(2) As a next step, we add PML in $r$- and $z$-direction, simulating an open system with decaying solutions for $r \to \infty$. This system can be looked upon as a model for a dielectric covered conducting rod (see [99, p. 524 ff.]).

(3) We simulate wave propagation in a half glass fiber.

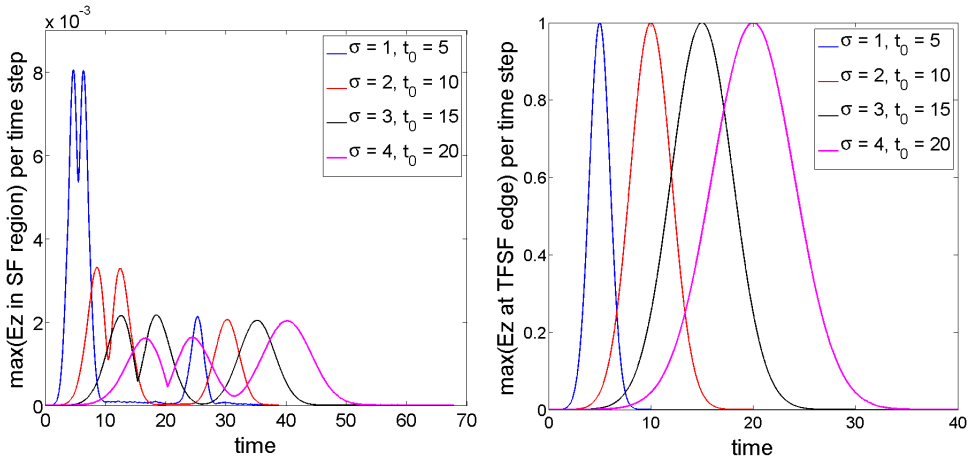



Figure 4.15: The error coming from the source decreases for increasing values of $t_0$ and $\sigma$.

(4) As a last step we consider a tapered fiber, that is, we add a sample beneath the half glass fiber. The system is composed of the regions as shown in figures 4.17 and 4.18, that is

- The entire area is $\Omega = [0, b] \times [-c, c]$, where $b, c$ are positive numbers. We denote by $b_{\text{PML}}$ the PML width. The scattered field and total field regions are

$$\Omega_{SF} = [0, b] \times [c - b_{\text{PML}}, c],$$
$$\Omega_{TF} := [0, b] \times [-c, c - b_{\text{PML}}].$$

- The PML region in $z$-direction is defined as

$$\text{PML}_z := ([0, b] \times [c - b_{\text{PML}}, c]) \cup ([0, b] \times [-c + b_{\text{PML}}, -c]),$$

where $\sigma_z \neq 0$. In $r$-direction we have the PML region

$$\text{PML}_r := [b - b_{\text{PML}}, b] \times [-c, c],$$

with $\sigma_r \neq 0$, and in $rz$-direction, i.e. in the corners, we let

$$\text{PML}_{rz} := \text{PML}_z \cap \text{PML}_r,$$

where $\sigma_r, \sigma_z$ are both nonzero.

- The fiber shall have a width of $2a\,\mu m$, the fiber region (green) is

$$\Omega_1 := [0, a] \times [-c, c] \text{ with } 0 < a < b$$

and the material parameters $\epsilon_1, \mu_1$. The rest of the region is the medium $\Omega_2$ with $\epsilon_2 = \mu_2 = 1$.

In the following we present all the details of these four systems we are considering.

### (1) Fiber with PML in $z$-direction, finite system with PEC boundary conditions

We look at a fiber with a core of radius $a$ and a cladding, as is visualized in figure 4.18(a). At $r = b$ we impose PEC boundary conditions. As a source we initialize the analytic solutions of an inhomogeneous medium. For the one-dimensional case, these are given in

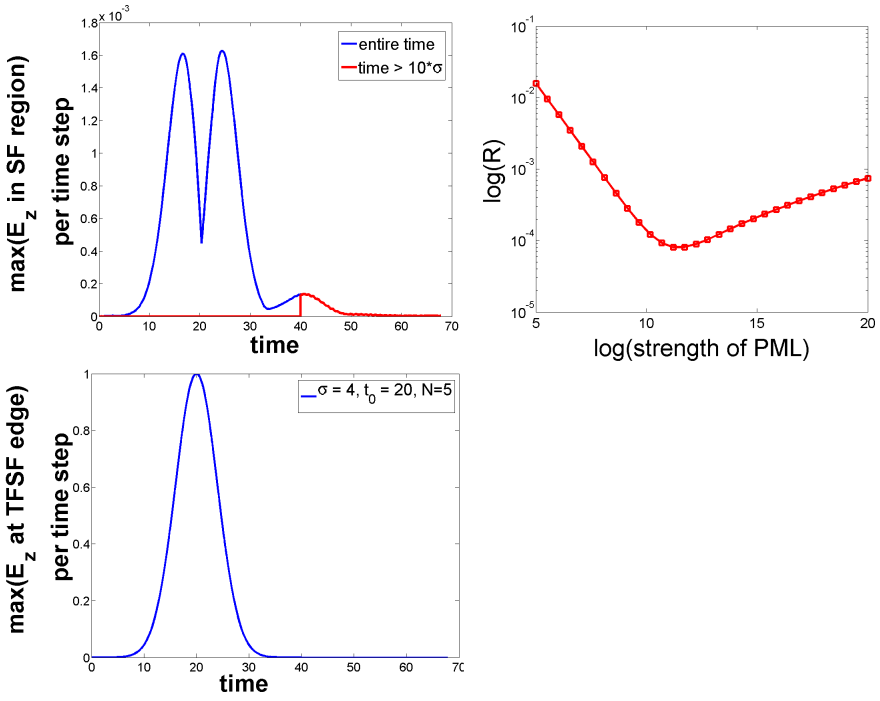



Figure 4.16: Error coming from PML in negative $z$-direction: behavior of $R$ as a measure for the reflected field strength for varying PML strength.

4.3.3. In two dimensions, the exact solution in TM mode in $\Omega$ is composed of the exact solution in the fiber region $\Omega_1$ and the medium $\Omega_2$ in the following manner:

For $0 \leq r \leq a$ :

$$
\begin{aligned}
E_z^{(1)}(r,\varphi,z,t) &= A_1 J_m(\gamma_{mn}^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
E_r^{(1)}(r,\varphi,z,t) &= A_1 \frac{\mathrm{i}k}{\gamma_{mn}^{(1)}} J_m'(\gamma_{mn}^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
E_\varphi^{(1)}(r,\varphi,z,t) &= A_1 \frac{-mk}{r(\gamma_{mn}^{(1)})^2} J_m(\gamma_{mn}^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
H_r^{(1)}(r,\varphi,z,t) &= A_1 \frac{m\epsilon_1\omega}{r(\gamma_{mn}^{(1)})^2} J_m(\gamma_{mn}^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
H_\varphi^{(1)}(r,\varphi,z,t) &= A_1 \frac{\mathrm{i}\epsilon_1\omega}{\gamma_{mn}^{(1)}} J_m'(\gamma_{mn}^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
H_z^{(1)}(r,\varphi,z,t) &= 0.
\end{aligned}
$$

For $a \leq r \leq b$ :

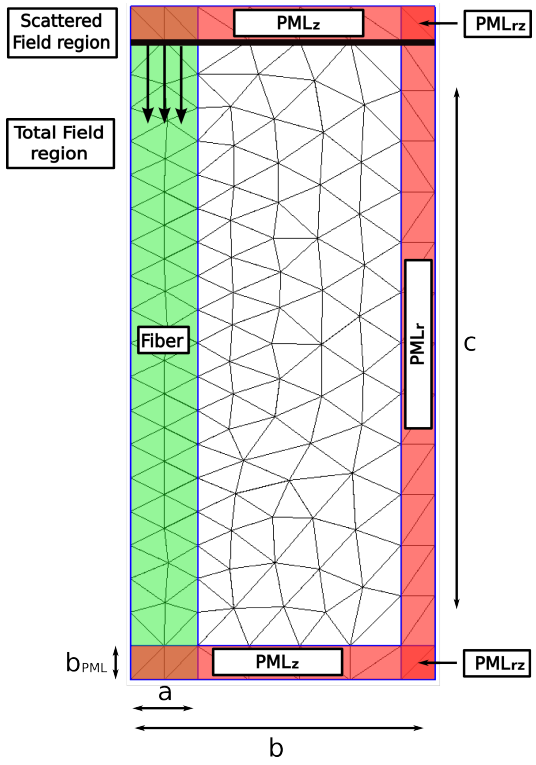



Figure 4.17: A system with a fiber of radius $a$, PML in $r$- and $z$-direction, scattered field (SF) and total field (TF) region. A Gaussian pulse is launched at the edge between TF and SF region with the traveling direction as indicated by the arrows. The mesh was generated with NetGen 4.9.9 [108].

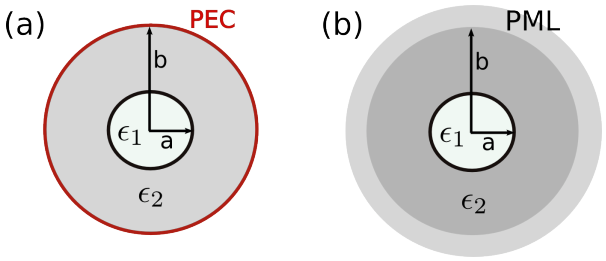



Figure 4.18: (a) Fiber with PEC boundary conditions. (b) Fiber with a PML.

$$
\begin{aligned}
E_z^{(2)}(r,\varphi,z,t) &= \big(A_2 J_m(\gamma_{mn}^{(2)} r) + B_2 Y_m(\gamma_{mn}^{(2)} r)\big) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
E_r^{(2)}(r,\varphi,z,t) &= \frac{-\mathrm{i}k}{\gamma_{mn}^{(2)}} \big(A_2 J_m'(\gamma_{mn}^{(2)} r) + B_2 Y_m'(\gamma_{mn}^{(2)} r)\big) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
E_\varphi^{(2)}(r,\varphi,z,t) &= \frac{mk}{r\,(\gamma_{mn}^{(2)})^2} \big(A_2 J_m(\gamma_{mn}^{(2)} r) + B_2 Y_m(\gamma_{mn}^{(2)} r)\big) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
H_r^{(2)}(r,\varphi,z,t) &= \frac{-m\epsilon_2\omega}{r\,(\gamma_{mn}^{(2)})^2} \big(A_2 J_m(\gamma_{mn}^{(2)} r) + B_2 Y_m(\gamma_{mn}^{(2)} r)\big) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
H_\varphi^{(2)}(r,\varphi,z,t) &= \frac{-\mathrm{i}\epsilon_2\omega}{\gamma_{mn}^{(2)}} \big(A_2 J_m'(\gamma_{mn}^{(2)} r) + B_2 Y_m'(\gamma_{mn}^{(2)} r)\big) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)},\\
H_z^{(2)}(r,\varphi,z,t) &= 0.
\end{aligned}
$$

The $Y_m$ are the Bessel functions of the second kind, see e.g. [98]. The wave number $k$ has to be determined from the dispersion relations

$$
\begin{aligned}
k^2 + (\gamma_{mn}^{(1)})^2 &= \mu_1\epsilon_1 k_0^2,\\
k^2 - (\gamma_{mn}^{(2)})^2 &= \mu_2\epsilon_2 k_0^2,
\end{aligned}
$$

where $k_0$ is the wave number in vacuum, i.e. $k_0 = \frac{\omega}{c}$. Note that now the frequency $\omega$ is a given quantity, whereas in the case of traveling waves in a homogeneous cavity we had $\omega = \omega_{mn}$, dependent on the zeros $\gamma_{mn}$ of the Bessel functions. Now it is $k = k_{mn}$, but in the forthcoming we drop the index to avoid index overflow.

The coefficients $A_1, A_2, B_2$ and the unknowns $\gamma_{mn}^{(1)}, \gamma_{mn}^{(2)}, k$ have to be determined, which can be achieved by applying boundary conditions and continuity conditions at the interface at $r = a$. That is, the tangential field components have to vanish, i.e.

$$
E_z^{(1)}(r=0) = 0, \qquad E_z^{(2)}(r=b) = 0,
$$

and the continuity condition at $r = a$ gives

$$
E_z^{(1)}(r=a) = E_z^{(2)}(r=a), \qquad H_\varphi^{(1)}(r=a) = H_\varphi^{(2)}(r=a).
$$

We will present the detailed computations for the next test system, where we include PML in $r$-direction. We refer to e.g. [99] and [107] for details.

### (2) Fiber with PML in $r$- and $z$-direction

We add a PML in $r$-direction to simulate an open fiber system as shown in figure 4.18(b) and figure 4.17. The exact solution of this system, again multiplied by a traveling Gaussian pulse, is launched as a source into the fiber. In the second region, the medium, the solution has to decay exponentially, since we are considering an open system. It consists of the modified Bessel functions of the first kind, denoted by $K_m$; see e.g. [98]. In the simulations we made the following settings:

$$
\begin{aligned}
&\Omega = [0,4.5]\times[-5,5], \qquad \Omega_{SF} = [0,4.5]\times[4.5,5], \qquad \Omega_{TF} := [0,4.5]\times[-5,4.5],\\
&\mathrm{PML}_z = \big([0,4.5]\times[4.5,5]\big) \cup \big([0,4.5]\times[-4.5,-5]\big),\\
&\mathrm{PML}_r := [4,4.5]\times[-5,5],\\
&\mathrm{PML}_{rz} := \mathrm{PML}_z \cap \mathrm{PML}_r,\\
&\Omega_1 := [0,1]\times[-5,5], \ \text{ i.e. } \ a = 1, b = 4.5, c = 5, \ \ b_{\mathrm{PML}} = 0.5.
\end{aligned}
$$

The exact solution in TM mode in $\Omega$ is composed of the exact solution in the fiber $\Omega_1$ and the medium $\Omega_2$ as follows (we drop the subscript $mn$ in $\gamma_{mn}$ for clarity):

For $0 \leq r \leq a$ :

$$
\begin{aligned}
E_z^{(1)}(r,\varphi,z,t) &= A_1 J_m(\gamma^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
E_r^{(1)}(r,\varphi,z,t) &= A_1 \frac{\mathrm{i}k}{\gamma^{(1)}} J_m'(\gamma^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
E_\varphi^{(1)}(r,\varphi,z,t) &= A_1 \frac{-mk}{r\,(\gamma^{(1)})^2} J_m(\gamma^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
H_r^{(1)}(r,\varphi,z,t) &= A_1 \frac{m\epsilon_1\omega}{r\,(\gamma^{(1)})^2} J_m(\gamma^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
H_\varphi^{(1)}(r,\varphi,z,t) &= A_1 \frac{\mathrm{i}\epsilon_1\omega}{\gamma^{(1)}} J_m'(\gamma^{(1)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
H_z^{(1)}(r,\varphi,z,t) &= 0,
\end{aligned}
\tag{4.5.6}
$$

For $a \leq r \leq \infty$ :

$$
\begin{aligned}
E_z^{(2)}(r,\varphi,z,t) &= A_2 K_m(\gamma^{(2)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
E_r^{(1)}(r,\varphi,z,t) &= \frac{-\mathrm{i}k}{\gamma^{(2)}} A_2 K_m'(\gamma_2 r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
E_\varphi^{(1)}(r,\varphi,z,t) &= \frac{mk}{r\,(\gamma^{(2)})^2} A_2 K_m(\gamma^{(2)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
H_r^{(1)}(r,\varphi,z,t) &= \frac{-m\epsilon_2\omega}{r\,)\gamma^{(2)})^2} A_2 K_m(\gamma^{(2)} r) + e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
H_\varphi^{(1)}(r,\varphi,z,t) &= \frac{-\mathrm{i}\epsilon_2\omega}{\gamma^{(2)}} A_2 K_m'(\gamma^{(2)} r) e^{\mathrm{i}m\varphi} e^{\mathrm{i}kz} e^{-\mathrm{i}\omega(t_0-t)} e^{-(t_0-t)^2/(2\sigma^2)}, \\
H_z^{(1)}(r,\varphi,z,t) &= 0,
\end{aligned}
\tag{4.5.7}
$$

with

$$k^2 + (\gamma^{(1)})^2 = \mu_1\epsilon_1 k_0^2, \tag{4.5.8}$$

$$k^2 - (\gamma^{(2)})^2 = \mu_2\epsilon_2 k_0^2. \tag{4.5.9}$$

This solution ansatz and relations (4.5.6) or (4.5.7) can be found in [99] and [107].
At the interface $r = a$ the following continuity conditions have to hold:

$$
\begin{aligned}
E_z^{(1)}(r=a) &= E_z^{(2)}(r=a), \\
H_\varphi^{(1)}(r=a) &= H_\varphi^{(2)}(r=a),
\end{aligned}
$$

This leads to a linear system of equations for the unknowns $A_1, A_2, \gamma^{(1)}, \gamma^{(2)}$ which only has a nontrivial solution if the determinant of the corresponding system matrix vanishes. This shows that $A_1$ can be chosen arbitrarily, and

$$A_2 = \frac{J_m(\gamma^{(1)} a)}{K_m(\gamma^{(2)} a)}, \tag{4.5.10}$$

which is a consequence of the continuity condition on $E_z$ at $r = a$. Using the second continuity condition on $H_\varphi$ renders

$$\epsilon_1\gamma^{(2)} J_m'(\gamma^{(1)} a) K_m(\gamma^{(2)} a) + \epsilon_2\gamma^{(1)} J_m(\gamma^{(1)} a) K_m'(\gamma^{(2)} a) = 0.$$

Thus, $\gamma^{(1)}$ is the zero of the function

$$f(\gamma^{(1)}) := \epsilon_1\gamma^{(2)}J'_m(\gamma^{(1)}a)K_m(\gamma^{(2)}a) + \epsilon_2\gamma^{(1)}J_m(\gamma^{(1)}a)K'_m(\gamma^{(2)}a).$$

Furthermore, considering (4.5.8), solving for $k^2$ and substracting both equations gives

$$(\gamma^{(1)})^2 + (\gamma^{(2)})^2 = (\mu_1\epsilon_1 - \mu_2\epsilon_2)k_0^2.$$

$\gamma^{(1)}$ can be found numerically with any zero finding routine, and $k$ is given through (4.5.8). We have thus determined all unknowns.

**Remark 4.17.**
*The derivatives of the Bessel functions fulfill the relations [109]*

$$J'_m(x) = 0.5(J_{m-1}(x) - J_{m+1}(x)),$$
$$K'_m(x) = -0.5(K_{m-1}(x) + K_{m+1}(x)).$$

*This avoids the explicit computation of the derivatives, e.g. via some difference method.*

**Remark 4.18.**
*Given $E_z^{(1)}$ and $H_z^{(1)}$ – which is 0 in TM mode – the rest of the electromagnetic fields in $\Omega_1$ can be determined by (see e.g. [107])*

$$E_r^{(1)} = \frac{i}{(\gamma^{(1)})^2}(k\partial_r E_z^{(1)} + \frac{\omega}{r}\partial_\varphi H_z^{(1)}),$$
$$E_\varphi^{(1)} = \frac{i}{(\gamma^{(1)})^2}(\frac{k}{r}\partial_\varphi E_z^{(1)} - \omega\partial_r H_z^{(1)}),$$
$$H_r^{(1)} = \frac{i}{(\gamma^{(1)})^2}(k\partial_r H_z^{(1)} - \frac{\epsilon_1\mu_1\omega}{r}\partial_\varphi E_z^{(1)}),$$
$$H_\varphi^{(1)} = \frac{i}{(\gamma^{(1)})^2}(\frac{k}{r}\partial_\varphi H_z^{(1)} + \epsilon_1\mu_1\omega\partial_r E_z^{(1)}).$$

*The fields in $\Omega_2$ are obtained by replacing $(\gamma^{(1)})^2$ by $-(\gamma^{(2)})^2$.*

### (3) A half glass fiber with PML in $r$- and $z$-direction

We change the system described in (2) as depicted in figure 4.19, so we now consider a half fiber (here without the sample), ending at the line through $z = 0$. The exact solution and the source remain the same. We choose the (not physically motivated) material values $\epsilon_1 = 2.0, \epsilon_2 = 1.0, \mu_1 = \mu_2 = 1.0$, and we repeat the simulations. Some snapshots of the waves are shown in figure 4.20.

### (4) A tapered fiber

At last we extend system (3) to include a sample. We simulate a semi-infinite dielectric fiber system as shown in figures 4.18 (b) and 4.19. The fiber has radius $r = 1\,\mu m$, and permittivity $\epsilon_1 = 1.527$ in $z = [0, \infty]$. Below the fiber, at the distance of $1\mu$m, we put a sphere with radius $r_s = 1\,\mu$m; to emphasize the effect of the sample we let $\epsilon_s = 12$. The computational domain was chosen to be $4.5\,\mu\text{m} \times 10\,\mu\text{m}$ in size and is surrounded by a PML of width $0.5\,\mu m$. The system is excited by an injection of a pulse within the fiber, traveling from the top downwards. The pulse consists of the exact solution of this system (which can be found in e.g. [98], [107]) multiplied by a Gaussian pulse with carrier frequency of $\nu_0 = \frac{2\pi}{1.5}$ and a Gaussian envelope of width $\sigma = 8$. Outside the fiber the solution decays exponentially. Some snapshots of the traveling waves are shown in figure 4.21. For mesh generating we used NetGen 4.9.9 [108].

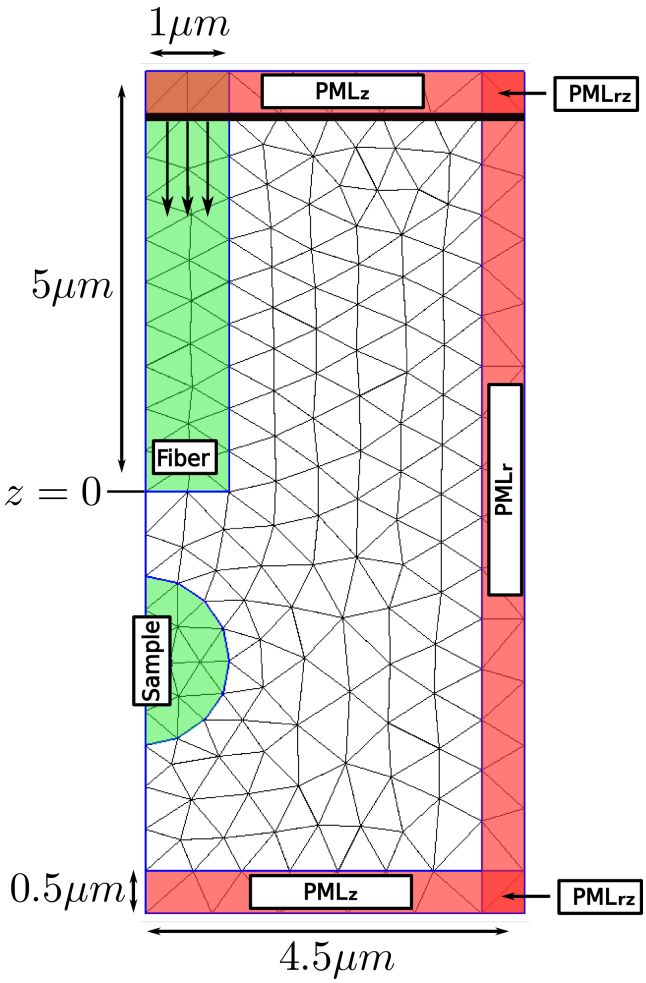



Figure 4.19: System with a fiber of radius 1 $\mu$m and length 5 $\mu$m, ending in $z = 0$, and a sample of radius 1 $\mu$m. We have PMLs in $r$- and $z$-direction with width 0.5 $\mu$m. A Gaussian pulse is launched at the edge at $z = 4.5$, as indicated by the arrows.

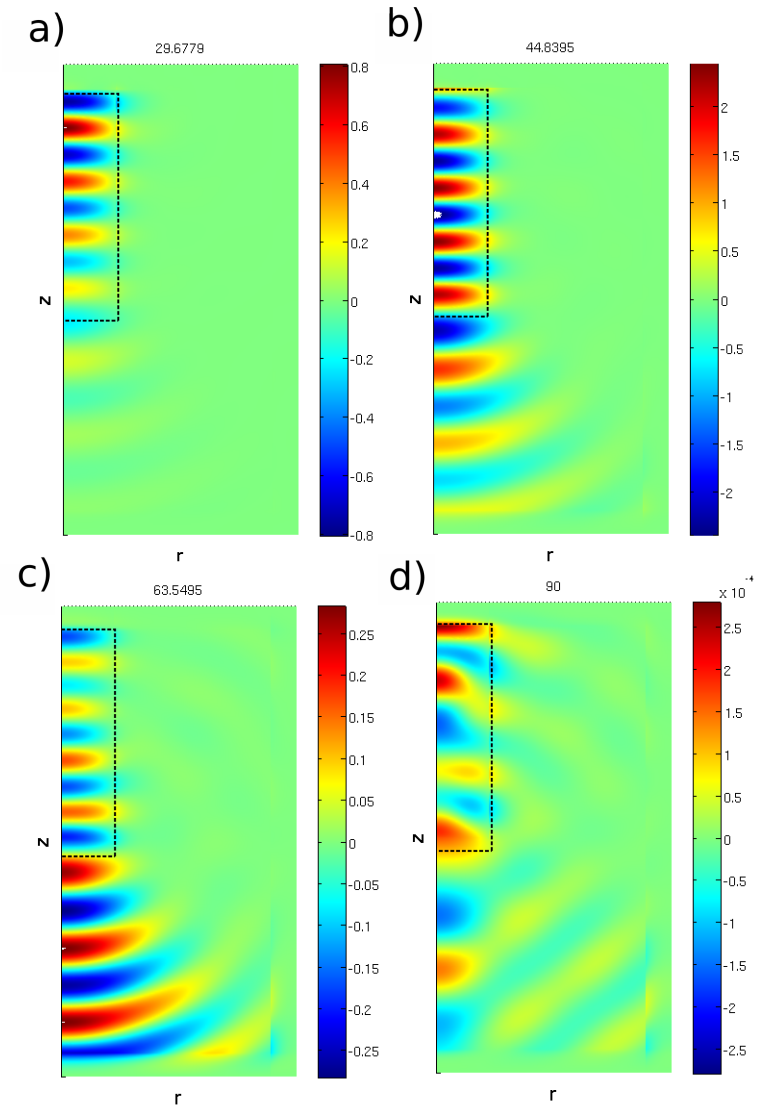



Figure 4.20: Simulation plots of $E_r$ traveling in $z$-direction along a fiber, ending in $z = 0$. The system is surrounded by PML. In the fiber we let $\epsilon_1 = 2.0$ and outside $\epsilon_2 = 1$. (a) At time unit 29.6779, (b) at time unit 44.8395, (c) at time unit 63.5495, (d) at final time 90.0.

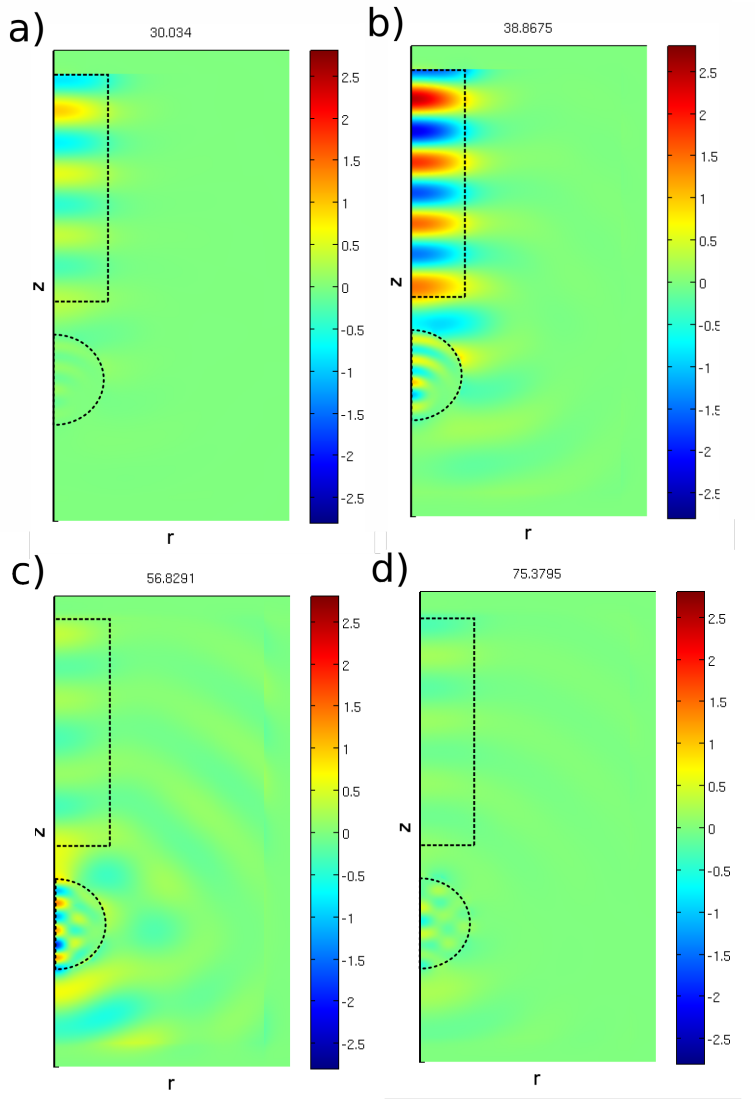



Figure 4.21: Simulation plots of $E_r$ traveling in negative $z$-direction along a fiber, ending in $z = 0$. The pulse is then scattered by a sphere with $\epsilon_s = 12$ and the outgoing radiation is absorbed by surrounding PMLs. (a) At time unit 30.0341, (b) at time unit 38.8765, (c) at time unit 56.8291, (d) at 75.3795.

## 4.6 Summary

We presented an efficient implementation of the Runge-Kutta discontinuous Galerkin method for the solution of Maxwell's equations in axi-symmetric systems (bodies of revolution, BOR) in two and three space dimensions. In contrast to a naive, quadrature-based approach, we showed how all elementary matrices can be constructed by exploiting the set of orthogonal Jacobi polynomials. Above all, we demonstrated how the stiffness matrices can be directly constructed from two global template matrices.

While our approach still requires to pre-compute and invert the BOR mass matrix $M_r^k$ for each element, it still reduces the required memory by at least a factor of two when compared to the quadrature-based approach. Since BOR systems are effectively two-dimensional, in most cases this memory requirement does not lead to significant limitations in terms of applicability. For cases where memory is very scarce, our method leads to reductions by roughly a factor of four at the price of some additional matrix-vector products.

Finally, in a set of numerical experiments, we demonstrated that our implementation yields optimal p-convergence and is a promising method for solving the time-dependent Maxwell equations in BOR systems:

1. We observed higher order convergence of our scheme for a homogeneous cavity.
2. We successively added PML and sources to our systems and showed simulations of electromagnetic waves traveling along a half fiber and scattering by a sphere, which was placed beneath the fiber in a distance of $1\mu$m. Our plots show the radiation is absorbed by the surrounding PML.

Chapter 5

# 5 Application: Kerr-Nonlinear Maxwell's Equations

The optical Kerr effect is a nonlinear optical phenomenon, which was discovered in 1875 and named after the physicist John Kerr. The Kerr effect arises due to a change in the refractive index of a material in response to an incoming electric field. The nonlinear behavior of the medium is responsible for effects like self-focusing or self-modulation. In the first case, the refractive index increases with the electric field intensity and the medium acts as a focusing lens for an electromagnetic wave. In the second example, the index of refraction varies in time and intensity of the incoming pulse, leading to a phase shift which produces a shift in the frequency of the pulse. Increasing intensity leads to lower frequencies, and decreasing intensity gives higher frequencies. Near an extremum of the intensity, the frequency of the pulse behaves approximately linearly.
Typically, nonlinearities are observed only if high intensities are present, as, e.g. in case of lasers. Examples of applications involving the optical Kerr effect are fast sensors for the measurement of electromagnetic fields, the fast determination of the structure of molecules, or image enhancement and image conversion in presence of ultraviolet radiation (see e.g. [26]).

In this section we will apply the RKDG method to Kerr-nonlinear Maxwell's equations. In contrast to the *linear* BOR Maxwell's equations many aspects are different for a nonlinear problem. This is especially the case for the numerical flux, which is given via a functional and which is state dependent for nonlinear problems. For BOR Maxwell's equations we chose the numerical flux to be the solution of a corresponding Riemann problem. This resulted in an upwind flux. Due to the linear behavior of the equations, the functional globally determines the flux and is state independent. For Kerr-nonlinear Maxwell's equations we will proceed in a likewise manner. Yet, one crucial difference and difficulty arises due to the nonlinear nature of the equations. In every point of the domain we have a different Riemann problem with a corresponding solution and thus the functional determining the numerical flux is state dependent and not globally given. This increases the computational effort immensely. For many relevant applications this makes simulations nearly impossible. Already for one-dimensional, very small systems with e.g. only ten elements, it takes minutes to run a simulation. Therefore an appropriate approximation to the analytical numerical flux is indispensable and leads to the research field of Riemann solvers. We chose several numerical fluxes, a Lax-Friedrichs flux, a Richtmyer flux, another linear flux and an HLL-like flux, which we compare with each othinside the layer, the fields decay exponentially faster with respect to efficiency and accuracy.

## 5.1 Kerr-Nonlinear Maxwell's Equations

The basis for the computation of a numerical flux for the DG method applied to Kerr-nonlinear Maxwell's equations are results from a paper by LaBourdonnaye [71]. The author presents the analytic solution of the Riemann problem corresponding to the Kerr-nonlinear Maxwell's equations. He shows uniqueness of this solution under the assumption of positivity of entropy (see e.g. [70]), and the entropy condition of Smoller-Johnson (see

section 3.3.2). In this section we cite the main results which are important for our purposes. For details and some theory, we refer to [71].

We consider the Kerr-nonlinear Maxwell's curl-equations of the form

$$\begin{aligned} \partial_t \mathbf{D}(\mathbf{E}) - \nabla \times \mathbf{H} &= 0, \\ \partial_t \mathbf{B} + \nabla \times \mathbf{E} &= 0, \end{aligned} \tag{5.1.1}$$

where the following constitutive relations shall hold:

$$\begin{aligned} \mathbf{B}(\mathbf{H}) &= \mu_0 \mathbf{H}, \\ \mathbf{D}(\mathbf{E}) &= \mathcal{E}'\left(\frac{1}{2}|\mathbf{E}|^2\right)\mathbf{E}. \end{aligned} \tag{5.1.2}$$

The function $\mathcal{E}'$ is defined as

$$\begin{aligned} \mathcal{E}(x) &:= \epsilon_0 x + \chi x^2, \\ \mathcal{E}'(x) &= \epsilon_0 + 2\chi x, \\ \mathcal{E}''(x) &= 2\chi. \end{aligned} \tag{5.1.3}$$

As a special choice we consider a Kerr-nonlinear medium so that

$$\mathcal{E}'\left(\frac{1}{2}|\mathbf{E}|^2\right) = \epsilon_0 + \chi|\mathbf{E}|^2. \tag{5.1.4}$$

$\mathcal{E}$ shall fulfill the following conditions:

(i) $\mathcal{E}'\left(\frac{x^2}{2}\right) \geq 0$ for all $x \in \mathbb{R}$.

(ii) $\mathcal{E}''\left(\frac{x^2}{2}\right) > 0$.

This leads to the requirement $\chi > 0$. For $\chi = 0$ we obtain the linear Maxwell's equations. With these assumptions, the energy, that is, the corresponding Hamiltonian, of the system is convex, implying that $\mathbf{D}(\mathbf{E})$ is invertible; the Hamiltonian is given in [71] and later in (5.2.20) of section 5.2.2. Thus the Kerr-nonlinear Maxwell's equations (5.1.1) can be rewritten as

$$\begin{aligned} \partial_t \mathbf{D}(\mathbf{E}) - \frac{1}{\mu_0}\nabla \times \mathbf{H} &= 0, \\ \partial_t \mathbf{B} + \left(\frac{\partial \mathbf{D}(\mathbf{E})}{\partial \mathbf{E}}\right)^{-1} \nabla \times \mathbf{D}(\mathbf{E}) &= 0. \end{aligned} \tag{5.1.5}$$

For the forthcoming we define the Jacobian of $\mathbf{D}(\mathbf{E})$ as $J(\mathbf{E}) := \frac{\partial \mathbf{D}(\mathbf{E})}{\partial \mathbf{E}}$ and often abbreviate $J(\mathbf{E}) = J$.

**Definition 5.1.**
*The Kronecker product $\mathbf{V} \otimes \mathbf{W}$ of two matrices $\mathbf{V} \in \mathbb{R}^{m \times n}$ and $\mathbf{W} \in \mathbb{R}^{r \times s}$ is defined as*

$$\begin{pmatrix} v_{11}\mathbf{W} & \cdots & v_{1n}\mathbf{W} \\ \vdots & \ddots & \vdots \\ v_{m1}\mathbf{W} & \cdots & v_{mn}\mathbf{W} \end{pmatrix}. \tag{5.1.6}$$

*The resulting matrix has dimension $mr \times ns$.*

With this definition we can give an explicit expression of the matrix $J$ as [71]

$$J(\mathbf{E}) = \frac{\partial \mathbf{D}(\mathbf{E})}{\partial \mathbf{E}} = \mathcal{E}'\left(\frac{1}{2}|\mathbf{E}|^2\right)\mathrm{Id} + \mathcal{E}''\left(\frac{1}{2}|\mathbf{E}|^2\right)(\mathbf{E}\otimes\mathbf{E}), \tag{5.1.7}$$

where

$$\mathbf{E}\otimes\mathbf{E} = (E_x\mathbf{E}, E_y\mathbf{E}, E_z\mathbf{E}) = \begin{pmatrix} E_x^2 & E_xE_y & E_xE_z \\ E_xE_y & E_y^2 & E_yE_z \\ E_xE_z & E_yE_z & E_z^2 \end{pmatrix} \tag{5.1.8}$$

for $\mathbf{E} = (E_x, E_y, E_z)$. Note that $J$ is symmetric. The inverse $J^{-1}$ enables a connection between $\partial_{\mathbf{E}}\mathbf{D}(\mathbf{E})$ and $\partial_{\mathbf{D}(\mathbf{E})}\mathbf{E}$, namely

$$J^{-1}\partial_{\mathbf{E}}\mathbf{D}(\mathbf{E}) = \partial_{\mathbf{D}(\mathbf{E})}\mathbf{E}. \tag{5.1.9}$$

The same is true with respect to the time derivative, i.e.

$$J^{-1}\partial_t\mathbf{D}(\mathbf{E}) = \partial_t\mathbf{E}. \tag{5.1.10}$$

*Proof.* All fields depend on $(x,t)$. Taking the partial derivative of $\mathbf{D}(x,t)$ with respect to $t$, we obtain

$$\begin{aligned}
\partial_t\mathbf{D} &= \partial_t(\mathcal{E}')\,\mathbf{E} + \mathcal{E}'\partial_t\mathbf{E} \\
&= \mathcal{E}''(E_y\partial_tE_y + E_z\partial_tE_z)\mathbf{E} + \mathcal{E}'\partial_t\mathbf{E} \\
&= \begin{pmatrix} \mathcal{E}''(E_y^2\partial_tE_y + E_yE_z\partial_tE_z) + \mathcal{E}'\partial_tE_y \\ \mathcal{E}''(E_yE_z\partial_tE_y + E_z^2\partial_tE_z) + \mathcal{E}'\partial_tE_z \end{pmatrix} \\
&= \left(\begin{pmatrix} \mathcal{E}''E_y^2 & \mathcal{E}''E_yE_z \\ \mathcal{E}''E_yE_z & \mathcal{E}''E_z^2 \end{pmatrix} + \mathcal{E}'\,\mathrm{Id}\right)\partial_t\mathbf{E} \\
&= \left(\mathcal{E}''(\mathbf{E}\otimes\mathbf{E}) + \mathcal{E}'\,\mathrm{Id}\right)\partial_t\mathbf{E} \\
&= J\partial_t\mathbf{E}.
\end{aligned}$$

By Id we mean the $2\times 2$ identity matrix. □

In view to the DG method, where the domain of interest is discretized and a corresponding mesh is generated, one often works quadrilaterals (in two space dimensions), or with polyhedra (in three dimensions). In any case, the numerical flux transports information from one cell to another across their shared faces along the unit normal $\mathbf{n}$ pointing into the neighboring cell. Imagining a face extended to be a half space, the same is true for any other face of two arbitrary neighboring cells. Due to rotational invariance we therefore consider propagation only in $x$-direction so that $\mathbf{n} = (n_x, 0, 0)^T$, leading to $\partial_tD_x = \partial_tB_x = 0$. Plugging this into (5.1.1), we can rewrite the Kerr-nonlinear Maxwell's equations as

$$\begin{aligned}
&\partial_tD_y + \partial_xH_z = 0, \\
&\partial_tD_z + \partial_x(-H_y) = 0, \\
&\partial_tB_y + \partial_x(-E_z) = 0, \\
&\partial_tB_z + \partial_xE_y = 0.
\end{aligned} \tag{5.1.11}$$

We thus have $\mathbf{D} := (D_y, D_z)^T, \mathbf{E} := (E_y, E_z)^T, \mathbf{B} := (B_z, -B_y)^T$ and $\mathbf{H} := (H_z, -H_y)^T$. The matrix $J$ in (5.1.7) then becomes

$$J = \mathcal{E}'\,\mathrm{Id} + \mathcal{E}''(\mathbf{E}\otimes\mathbf{E}) = \begin{pmatrix} \mathcal{E}' + \mathcal{E}''E_y^2 & \mathcal{E}''E_yE_z \\ \mathcal{E}''E_yE_z & \mathcal{E}' + \mathcal{E}''E_z^2. \end{pmatrix}, \tag{5.1.12}$$

The inverse of $J$ is determined to be

$$J^{-1} = \frac{1}{\det(J)} \begin{pmatrix} \mathcal{E}' + \mathcal{E}'' E_z^2 & -\mathcal{E}'' E_y E_z \\ -\mathcal{E}'' E_y E_z & \mathcal{E}' + \mathcal{E}'' E_y^2 \end{pmatrix}, \tag{5.1.13}$$

and the determinant of $J$ in (5.1.12) becomes

$$\det(J) = \mathcal{E}'(\mathcal{E}' + \mathcal{E}''|\mathbf{E}|^2) = (\epsilon_0 + \chi|\mathbf{E}|^2)(\epsilon_0 + 3\chi|\mathbf{E}|^2). \tag{5.1.14}$$

For implementation issues we reformulate (5.1.1) as

$$\begin{aligned} J^{-1}\partial_t \mathbf{E} - \nabla \times \mathbf{H} &= 0, \\ \partial_t(\mu_0 \mathbf{H}) + \nabla \times \mathbf{E} &= 0, \end{aligned}$$

which, by using (5.1.10), is equivalent to

$$\begin{aligned} \partial_t \mathbf{E} - J\nabla \times \mathbf{H} &= 0, \\ \partial_t \mathbf{H} + \frac{1}{\mu_0} \nabla \times \mathbf{E} &= 0. \end{aligned} \tag{5.1.15}$$

The one-dimensional Kerr-nonlinear Maxwell's equations are obtained by setting, for instance, $E_y = 0$ and $H_z = 0$. This case has been studied in the diploma thesis [110]. The matrix $J^{-1}$ is then given as

$$J^{-1} = \begin{pmatrix} \frac{1}{\epsilon_0 + \chi E_z^2} & 0 \\ 0 & \frac{1}{\epsilon_0 + 3\chi E_z^2} \end{pmatrix}.$$

An analytical solution of the resulting Kerr-Maxwell's equations in one dimension can be found in [45]. Analogously one could set $E_z = 0$ and $H_y = 0$ and get an analogous formula by substituting $E_z \to E_y$ and $H_y \to -H_z$.

### Conservative Form of Kerr-Nonlinear Maxwell's Equations

By defining the state vector $\mathbf{u} := (D_y, D_z, \mu_0 H_y, \mu_0 H_z)^T$ and the flux vector $\mathbf{F}(\mathbf{u}) := (H_z, -H_y, -E_z, E_y)^T$ we can write system (5.1.11) in conservative form as

$$\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = 0.$$

In [71] it is shown that system (5.1.5) is a quasilinear hyperbolic system, which is symmetrizable. Its eigenstructure looks as stated in the following lemma.

**Lemma 5.2.**
*The eigenvalues of the Jacobian $\mathbf{F}'(\mathbf{u}) \in \mathbb{R}^{4\times 4}$ are given as*

$$\lambda_1^{\pm} = \pm \frac{1}{\sqrt{\mathcal{E}'}}, \tag{5.1.16}$$

$$\lambda_2^{\pm} = \pm \frac{1}{\sqrt{\mathcal{E}' + \mathcal{E}''|\mathbf{E}|^2}}, \tag{5.1.17}$$

*where we abbreviated $\mathcal{E}' = \mathcal{E}'\left(\frac{1}{2}|\mathbf{E}|^2\right)$. The corresponding eigenvectors read*

$$\mathbf{v}_1^{\pm} = \begin{pmatrix} \tilde{\mathbf{E}} \\ \lambda_1^{\pm} \tilde{\mathbf{E}} \end{pmatrix}, \qquad \mathbf{v}_2^{\pm} = \begin{pmatrix} \mathbf{E} \\ \lambda_2^{\pm} \mathbf{E} \end{pmatrix} \tag{5.1.18}$$

*with $\tilde{\mathbf{E}} := (-E_z, E_y)^T$, i.e. $\tilde{\mathbf{E}} \perp \mathbf{E}$. It holds*

$$\lambda_1^- \le \lambda_2^- \le 0 \le \lambda_2^+ \le \lambda_1^+.$$

### 5.1.1 Characteristic Fields

**Lemma 5.3.**

*(1) The 1-characteristic field, corresponding to the eigenvalues $\lambda_1^{\pm}$, is linearly degenerate.*

*(2) The 2-characteristic field, corresponding to the eigenvalues $\lambda_2^{\pm}$, is genuinely nonlinear as long as $\mathbf{E} \neq \mathbf{0}$..*

*Proof.* (1) Recalling definition 3.6, we need to show

$$\nabla\lambda_1(\mathbf{u})^{\pm} \cdot \mathbf{v}_1(\mathbf{u})^{\pm} = 0 \qquad \forall\, \mathbf{u} \in \mathbb{R}^6.$$

Recall that $\lambda_1^{\pm} = (\mathcal{E}')^{-1/2}$. We then encounter

$$\nabla\lambda_1(\mathbf{u})^{\pm} = (J^{-1}\partial_{\mathbf{E}}\lambda_1^{\pm},\ \partial_{\mathbf{B}}\lambda_1^{\pm}) = (J^{-1}\partial_{\mathbf{E}}\lambda_1^{\pm}, \mathbf{0}) = J^{-1}\left(\pm\left(\frac{1}{\sqrt{\mathcal{E}'}}\right)' \mathbf{E}, \mathbf{0}\right),$$

where we applied the chain rule to the expression

$$\partial_{\mathbf{E}}\lambda_1^{\pm} = \partial_{\mathbf{E}}\left(\frac{1}{\sqrt{\mathcal{E}'(|\mathbf{E}|^2/2}}\right),$$

and $\left(\frac{1}{\sqrt{\mathcal{E}'}}\right)'$ means the total derivative

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{\sqrt{\mathcal{E}'(x)}}\right).$$

Since $\mathbf{E} \perp \tilde{\mathbf{E}}$ i.e. $\mathbf{E} \cdot \tilde{\mathbf{E}} = 0$, the claim follows.

(2) According to definition 3.6, we require to show

$$\nabla\lambda_2(\mathbf{u})^{\pm} \cdot \mathbf{v}_2(\mathbf{u})^{\pm} \neq 0 \quad \forall\, \mathbf{u} \in \mathbb{R}^6.$$

Again, it is

$$\nabla\lambda_2(\mathbf{u})^{\pm} = (J^{-1}\partial_{\mathbf{E}}\lambda_2^{\pm}(|\mathbf{E}|^2), \mathbf{0}),$$

where $\lambda_2^{\pm} = (\mathcal{E}' + \mathcal{E}''|\mathbf{E}|^2)^{-1/2}$. Using the chain rule we obtain

$$\partial_{\mathbf{E}}\lambda_2^{\pm} = 2\mathcal{E}''\, h(|\mathbf{E}|^2)\, \mathbf{E},$$

with $h(|\mathbf{E}|^2) := -\frac{1}{2}(\mathcal{E}' + \mathcal{E}''|\mathbf{E}|^2)^{-3/2}$. Denoting the first component of $\mathbf{v}_2^{\pm} = (\mathbf{E}, \lambda_2^{\pm}\mathbf{E})^T$ by $\mathbf{v}_{2,1}^{\pm}$, we observe

$$\partial_{\mathbf{E}}\lambda_2^{\pm} \cdot \mathbf{v}_{2,1}^{\pm} = 2\mathcal{E}''\, h(|\mathbf{E}|^2)\, \mathbf{E} \cdot \mathbf{E},$$

which only becomes zero if $\mathbf{E} = \mathbf{0}$. This ends the proof. □

For the following, whenever we say the 2-characteristic field is genuinely nonlinear, we exclude $\mathbf{E} = \mathbf{0}$.

## 5.2 The Kerr-Nonlinear Riemann Problem and its Solution

When we applied the DG method to BOR Maxwell's equations, we chose the numerical flux to be the solution of the corresponding Riemann problem. We make the same choice for Kerr-nonlinear Maxwell's equations. In this section we present the solution structure of the Kerr-nonlinear Riemann problem as given in [71]. In section 3.3 we saw that (Theorem 3.26)

(1) if the $k$th characteristic field is *genuinely nonlinear*, we either have a $k$-shock or a $k$-rarefaction wave;

(2) for a *genuinely nonlinear* $k$-field, points on a Hugoniot Locus correspond to a $k$-shock, and points on a Riemann invariant correspond to a $k$-rarefaction. An entropy condition and a second condition on the shock speed give the parts on the Hugoniot Locus to which an admissible shock corresponds;

(3) for a $k$-shock, the Rankine-Hugoniot jump conditions must hold;

(4) if the $k$-th field is *linearly degenerate*, we have a contact discontinuity. It can be determined by either applying the Rankine-Hugoniot condition or by using the corresponding $k$-Riemann invariant

(5) the solution of a Riemann problem consists of finitely many waves which are either a shock or a rarefaction, respectively, or a contact discontinuity.

In our case,

(1) the 2-field is genuinely nonlinear, so we either have a 2-shock or a 2-rarefaction. We determine a 2-shock via the Rankine-Hugoniot condition and select the admissible ones by the entropy condition $\eta > 0$ and the condition of Smoller-Johnson (3.3.31), as in [71]. The 2-rarefaction wave is determined via the 2-Riemann invariants;

(2) the 1-field is linearly degenerate, therefore we have a contact discontinuity. It can be determined by either applying the Rankine-Hugoniot condition or by using the corresponding 1-Riemann invariant.

In this section we start by determining the Hugoniot Loci and the Riemann invariants for the Kerr-nonlinear Riemann problem. Afterwards we give its unique solution.

### 5.2.1 Hugoniot Locus

In order to determine the Hugoniot Loci we use the approach in [57, Ch. 13.7]. In definition 3.17 of section 3.3.2 we defined the Hugoniot Locus as the set

$$\mathcal{HL} = \{\mathbf{u} : \ s(\mathbf{u}^*, \mathbf{u})(\mathbf{u} - \mathbf{u}^*) = \mathbf{F_n}(\mathbf{u}) - \mathbf{F_n}(\mathbf{u}^*)\}, \tag{5.2.1}$$

where $\mathbf{u} \in \mathbb{R}^n$ and $s = s(\mathbf{u}^*, \mathbf{u}) \in \mathbb{R}$ is the shock speed from definition 3.3.22. In our case, $n = 4$. A Hugoniot Locus gives the set of all points $\mathbf{u}^*$ that can be connected to an arbitrary point $\mathbf{u}$ by a discontinuity. We aim at finding all states $\mathbf{u}^*$ that can be connected to a left state $\mathbf{u}_L$ (or a right state $\mathbf{u}_R$). With view to the Riemann problem where one initially considers the left and right states $\mathbf{u}_L$ and $\mathbf{u}_R$, we are interested in the question in which case $\mathbf{u}_L$ and $\mathbf{u}_R$ are connected by a discontinuity or, more precisely, by an admissible shock (see definition 3.24). Then either $\mathbf{u}^* = \mathbf{u}_R$ or $\mathbf{u}^* = \mathbf{u}_L$, depending on whether one is interested in the connection to $\mathbf{u}_L$ or $\mathbf{u}_R$.

So let $\mathbf{u}^*$ be arbitrary and fixed. We consider $\mathbf{n} = (1,0,0)^T$ so that $\mathbf{F}_\mathbf{n}(\mathbf{u}) = \mathbf{F}(\mathbf{u})$. By using (5.2.1) we obtain via the Rankine-Hugoniot jump condition (3.3.23)

$$s\,(D_y^* - D_y) = H_z^* - H_z, \tag{5.2.2a}$$
$$s\,(D_z^* - D_z) = H_y - H_y^*, \tag{5.2.2b}$$
$$s\,(B_z^* - B_z) = E_y^* - E_y, \tag{5.2.2c}$$
$$s\,(B_y - B_y^*) = E_z^* - E_z. \tag{5.2.2d}$$

Additionally, we have the constitutive relations (5.1.2)

$$\begin{aligned} D_y &= (\epsilon_0 + \chi|\mathbf{E}|^2)E_y, \\ D_z &= (\epsilon_0 + \chi|\mathbf{E}|^2)E_z. \end{aligned} \tag{5.2.3}$$

Recall that $|\mathbf{E}|^2 = E_y^2 + E_z^2$. We assume to have $\chi \neq 0$. Furthermore, at this point we let $\mathbf{u} \neq \mathbf{u}^*$. We see that we have six equations (5.2.2a) to (5.2.2d), together with equations (5.2.3), for seven unknowns in total, namely $D_y, D_z, E_y, E_z, B_y, B_z$ and the shock speed $s$. Thus we will get a one-parameter family of solutions. One of the unknowns should be chosen to be a parameter in such a way that all expressions become relatively simple. In the diploma thesis [110], where the one-dimensional Kerr-nonlinear Riemann was studied, the $E$-field was chosen to be the parameter, and we follow by either taking $E_y$ or $E_z$ to be the parameter. We choose $E_y$ in the subsequent, but $E_z$ is equally well possible. We then solve equations (5.2.2a) to (5.2.3) for the remaining unknowns.
We define

$$\alpha = \alpha(E_z) := \epsilon_0 + \chi(E_y^2 + E_z^2). \tag{5.2.4}$$

It is important to note that $\alpha$ depends on $E_z$. We combine (5.2.2a) with (5.2.2c) and (5.2.2b) with (5.2.2d) and obtain

$$s^2\,(D_y^* - D_y) = E_y^* - E_y \quad \Longleftrightarrow \quad s^2 = \frac{E_y^* - E_y}{D_y^* - \alpha E_y}, \tag{5.2.5}$$

$$s^2\,(D_z^* - D_z) = E_z^* - E_z \quad \Longleftrightarrow \quad s^2 = \frac{E_z^* - E_z}{D_z^* - \alpha E_z}. \tag{5.2.6}$$

Note that we have assumed $E_y^* \neq E_y$ and $E_z^* \neq E_z$ for the moment. We equate both equations for $s^2$ and get after some rearrangement the equation

$$(-\chi E_y^*)E_z^3 + (E_z^* E_y \chi)E_z^2 + \left[D_y^* - E_y^*(\epsilon_0 + \chi E_y^2)\right] E_z + \gamma = 0, \tag{5.2.7}$$

where we introduced

$$\begin{aligned} \gamma &:= \beta D_z^* + (\epsilon_0 + \chi E_y^2)E_z^* E_y - D_y^* E_z^*, \\ \beta &:= E_y^* - E_y. \end{aligned}$$

As a first step we need to distinguish the following cases:

**Case 1:**
$E_y^* = 0, E_z^* \neq 0$. Then (5.2.7) becomes

$$\begin{aligned} & \chi E_z^* E_y E_z^2 + \chi E_z^* E_y^3 - \chi (E_z^*)^3 E_y = 0 \\ \Longleftrightarrow \quad & E_z^2 + E_y^2 - (E_z^*)^2 = 0 \\ \Longleftrightarrow \quad & E_z^2 = (E_z^*)^2 - E_y^2. \end{aligned} \tag{5.2.8}$$

The solution is given as

$$E_z^{(1,2)} = \pm\sqrt{(E_z^*)^2 - E_y^2}. \tag{5.2.9}$$

Since all values have to be real, we require $(E_z^*)^2 - E_y^2 \geq 0$, which leads to the condition $|E_z^*| \leq |E_y|$.

**Case 2:**
$E_y^* \neq 0, E_z^* = 0$. Equation (5.2.7) reads

$$\begin{aligned} &-\chi E_y^* E_z^3 + \left[(\epsilon_0 + \chi(E_y^*)^2)E_y^* - \beta(\epsilon_0 + \chi E_y^2) - E_y(\epsilon_0 + \chi E_y^2)\right] E_z = 0 \\ \iff \quad &\left(-E_z^2 + (E_y^*)^2 - E_y^2\right) E_z = 0, \end{aligned}$$

and it follows that $E_z = 0$ or $E_z^2 = (E_y^*)^2 - E_y^2$. Since $E_z$ has to be real, we need $|E_y^*| \geq |E_y|$. The condition $|E_y^*| \geq |E_y|$ may lead to selective discontinuities; yet, $E_z$ must be continuous, and therefore $E_z = 0$ is the solution.

**Case 3:**
$E_y^* = 0, E_z^* = 0$. The left hand side of (5.2.7) is zero, and $E_z$ can be chosen arbitrarily, but it mus depend continuously on $E_y$. Especially, $E_z = E_y$ or $E_z = 0$ are possible. In this case, the speed $s$ of the shock is equal to the eigenvalue $\lambda_1^{\pm}$.

**Case 4:**
$E_y^* \neq 0, E_z^* \neq 0$. Equation (5.2.7) is a third order equation in $E_z$, which can be analytically solved by applying the Cardano formulas which were published in 1545 (see e.g. [111, 112]).

The Cardano formulas are formulated for the equation

$$z^3 + pz + q = 0. \tag{5.2.10}$$

A general third order equation of the form

$$Ax^3 + Bx^2 + Cx + D = 0, \qquad A \neq 0,$$

can be transformed to the form (5.2.10) by first dividing through $A$ and obtaining the normal form

$$x^3 + ax^2 + bx + c = 0$$

with $a = B/A,\ b = C/A, d = D/A$. With the substitution $x := z - a/3$ one obtains the reduced form

$$z^3 + pz + q = 0$$

with

$$p := b - \frac{a^2}{3}, \tag{5.2.11}$$

$$q := \frac{2a^3}{27} - \frac{ab}{3} + c. \tag{5.2.12}$$

Depending on the discriminant $\mathcal{D}$ which is defined as

$$\mathcal{D} := \left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3,$$

one has to distinguish the following cases:

(1) $\mathcal{D} > 0$: There is one real solution and two imaginary solutions.

(2) $\mathcal{D} = 0$: There are only real solutions.

(3) $\mathcal{D} < 0$: There are three different real solutions.

Since the electromagnetic fields are physical quantities, only real solutions are of interest. These real solutions are:

(1) $\mathcal{D} > 0$: Letting $a^{\pm} := -\frac{q}{2} \pm \sqrt{D}$, the solution is $z^{+} = v_1 + v_2$, where $v_{1,2} = \operatorname{sign}(a^{\pm})\sqrt[3]{a^{\pm}}$.

(2) $\mathcal{D} = 0$: If $p \neq 0 \neq q$, the solutions are $z_1^{(0)} = \frac{3q}{p}$ and $z_2^{(0)} = -\frac{3q}{2p}$. If $p = q = 0$, the only solution is $z^{(0)} = 0$.

(3) $\mathcal{D} < 0$:

$$
\begin{aligned}
z_1^{-} &= \sqrt{-\frac{4}{3}p} \cdot \cos\left(\frac{1}{3}\arccos\left(-\frac{q}{2}\sqrt{-\frac{27}{p^3}}\right)\right), \\
z_2^{-} &= -\sqrt{-\frac{4}{3}p} \cdot \cos\left(\frac{1}{3}\arccos\left(-\frac{q}{2}\sqrt{-\frac{27}{p^3}}\right) + \frac{\pi}{3}\right), \\
z_3^{-} &= -\sqrt{-\frac{4}{3}p} \cdot \cos\left(\frac{1}{3}\arccos\left(-\frac{q}{2}\sqrt{-\frac{27}{p^3}}\right) - \frac{\pi}{3}\right).
\end{aligned}
$$

Note that in this case $p < 0$.

For our case we divide equation (5.2.7) by $-\chi E_y^*$ and define

$$
\begin{aligned}
a &:= -\frac{E_z^* E_y}{E_y^*}, \\
b &:= -E_y^2 - (E_y^*)^2 - (E_z^*)^2, \\
c &:= -\frac{\gamma}{\chi E_y^*}.
\end{aligned}
$$

Next, we make the substitution $E_z = z - \frac{a}{3}$, which leads to

$$
z^3 + pz + q = 0,
$$

where $p$ and $q$ are given in (5.2.11). Following 5.2.1 we obtain $E_z$ by $E_z = z - \frac{a}{3}$, where in detail

$$
\begin{cases}
\mathcal{D} > 0: & E_z^{+} := z^{+} - \frac{a}{3}, \\
\mathcal{D} = 0: & E_{z,i}^{(0)} := z_i^{(0)} - \frac{a}{3}\ (i = 1, 2), \ \text{ if } p \neq 0 \neq q; \\
 & E_z^{(0)} = 0, \text{ if } \ p = 0 = q, \\
\mathcal{D} < 0: & E_{z,i}^{-} := z_i^{-} - \frac{a}{3} \quad (i = 1, 2, 3).
\end{cases}
$$

For the case $\mathbf{u}^* = \mathbf{u}$, especially if $E_y^* = E_y$, equation (5.2.7) reduces to

$$
-E_z^3 + E_z^* E_z^2 + (E_z^*)^2 E_z - (E_z^*)^3 = 0,
$$

and $a = -E_z^*, b = -(E_z^*)^2, c = (E_z^*)^3$. Also, we then have $p = -\frac{4}{3}(E_z^*)^2$ and $q = \frac{16}{27}(E_z^*)^3$. For $E_z^* \neq 0$ we are in the case $\mathcal{D} = 0$ with $p \neq 0 \neq q$. For $E_z^* = 0$ we would have $\mathcal{D} = 0$ and $p = q = 0$, and then $E_z = 0$.

Having determined $E_z$, $D_y$ and $D_z$ are also known. $B_y$ and $B_z$ are obtained from equations (5.2.2c) and (5.2.2d) as

$$B_y^+ = B_y^* + \sqrt{(E_z^* - E_z)(D_z^* - D_z)}, \tag{5.2.13a}$$

$$B_y^- = B_y^* - \sqrt{(E_z^* - E_z)(D_z^* - D_z)}, \tag{5.2.13b}$$

$$B_z^+ = B_z^* - \sqrt{(E_y^* - E_y)(D_y^* - D_y)}, \tag{5.2.13c}$$

$$B_z^- = B_y^* + \sqrt{(E_y^* - E_y)(D_y^* - D_y)}, \tag{5.2.13d}$$

and by using (5.2.5) the shock speed is given by

$$s^{\pm} = \pm \left( \frac{E_z^* - E_z}{D_z^* - D_z} \right)^{\frac{1}{2}}, \tag{5.2.14}$$

or, equivalently,

$$s^{\pm} = \pm \left( \frac{E_y^* - E_y}{D_y^* - D_y} \right)^{\frac{1}{2}}. \tag{5.2.15}$$

We obtain equations (5.2.13a) and (5.2.13b) by using (5.2.14), and equations (5.2.13c) and (5.2.13d) by taking (5.2.15). The superscript "+" in (5.2.13a) and (5.2.13c) shall emphasize that the corresponding shock speed is $s^+$, and $s^-$ corresponds to equations (5.2.13b) and (5.2.13d) with superscript "-".
Figure 5.2 shows two exemplary plots of all possible solutions for all the cases $\mathcal{D} < 0$, $\mathcal{D} = 0$ and $\mathcal{D} > 0$ for (1) $E_y^* = 1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = 2$ and (2) $E_y^* = 1$, $E_z^* = 1$, $B_y^* = 2$, $B_z^* = -2$, $\chi = 0.1$, $\epsilon_0 = 1$, where we have the notations as displayed in table 5.2.1. For instance, for $E_y \in [-|E_y^*|, |E_y^*|]$ it is $\mathcal{D} < 0$, and we have three real solutions on the plus or minus branch, respectively; if $\mathcal{D} > 0$, for instance for $|E_y| = 2$, we have one real solution on the plus or minus branch, respectively. It is important to emphasize that we have a plus branch and a minus branch for $B_y$ and $B_z$, respectively. Figure 5.3 shows a more detailed analysis of figure 5.2. In figures 5.4 and 5.5 all plus branches $B_y^+$ and $B_z^+$ are marked red and all minus branches $B_y^-$ and $B_z^-$ are marked blue.
The discriminant $\mathcal{D}$ varies smoothly in $E_y$ and $E_z$, see figure 5.6 for a visualization. Also, the tangential components of the electric field have to be continuous across interfaces (see section 2.2), i.e. $E_z$ is continuous. Furthermore, the Hugoniot Locus consists of a set of points leading at least to a locally smooth curve. We therefore define the subsequent selection rules in order to select the correct parts of the complete set of solutions. For all cases we let $E_y \in \mathbb{R}$ vary as a parameter, fix arbitrary $\mathbf{E}^*, \mathbf{B}^*$ and define the regions

$$\begin{aligned} \mathrm{I} &:= (-\infty, -|E_y^*|), \\ \mathrm{II} &:= [-|E_y^*|, |E_y^*|], \\ \mathrm{III} &:= (-\infty, -|E_y^*|), \end{aligned}$$

Figure 5.1 visualizes the situation. If $E_y$ is such that $\mathcal{D} < 0$, $E_z$ is given as follows:

**Case 1:** $E_y^* < 0$, $E_z^* > 0$ or $E_y^* > 0$, $E_z^* < 0$.

$$\begin{aligned} E_z|_{\mathrm{I}} &= E_{z,1}^-, \\ E_z|_{\mathrm{II}} &= E_{z,2}^-, \\ E_z|_{\mathrm{III}} &= E_{z,3}^-. \end{aligned}$$

**Case 2:** $E_y^* > 0$, $E_z^* > 0$ or $E_y^* < 0$, $E_z^* < 0$.

$$E_z|_{\mathrm{I}} = E_{z,3}^-,$$
$$E_z|_{\mathrm{II}} = E_{z,2}^-,$$
$$E_z|_{\mathrm{III}} = E_{z,1}^-.$$

**Case 3:** $E_z^* = 0$ and $\mathcal{D} = 0$. The solution is $E_z = 0$.

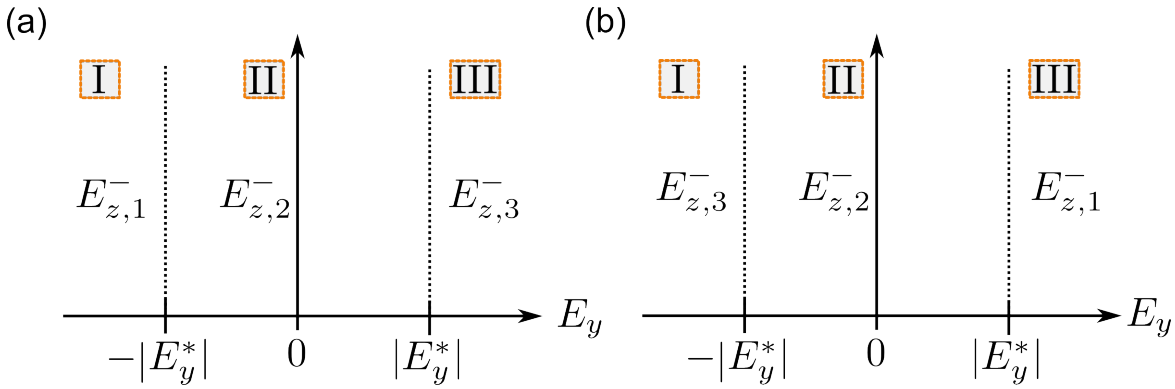



Figure 5.1: (a) Solution $E_z$ for the case $E_y^* < 0$, $E_z^* > 0$ or $E_y^* > 0$, $E_z^* < 0$. (b) Solution $E_z$ for the case $E_y^* > 0$, $E_z^* > 0$ or $E_y^* < 0$, $E_z^* < 0$.

For $\mathcal{D} = 0$ with $p \neq 0 \neq q$, we first realize that $z_1^{(0)} = -\frac{1}{2} z_2^{(0)}$. In the linear case the Hugoniot Locus consists of a family of straight lines with the directions of the eigenvectors (see the linear example in section 3.3.2); in the nonlinear case these directions are *locally* given by the eigenvectors of the system, and locally the winding curves are straight lines, as in the linear case. Therefore we can choose either $z_1^{(0)}$ or $z_2^{(0)}$, since their directions are the same. In our tests we chose $z_2^{(0)}$. Figures 5.9 and 5.8 each show a branch of the Hugoniot Locus which is obtained by applying the selection rules to the complete solution of the system. Figure 5.7 shows a plot of $E_z$ after applying the selection rules and the resulting plus and minus branches of $B_y$ and $B_z$, respectively. At the bottom of figure 5.9 a sketch of the Hugoniot Loci for the Kerr-nonlinear Riemann problem is shown, including several branches. Compare this picture with the linear case, see figure 3.9.

| $\mathcal{D} > 0$ | $\mathcal{D} = 0$ | $\mathcal{D} < 0$ |
|---|---|---|
| superscript "+" | superscript "(0)" if $p \neq 0 \neq q$<br>superscript "(00)" if $p = q = 0$ | superscript "-" |

| minus branches $B_y^-$, $B_z^-$ | plus branches $B_y^+$, $B_z^+$ |
|---|---|
| subscript "m" | subscript "p" |

Table 5.1: Notation for the solutions in figure 5.2. The numbers $1, 2, 3$ in the subscript denote the solution number.

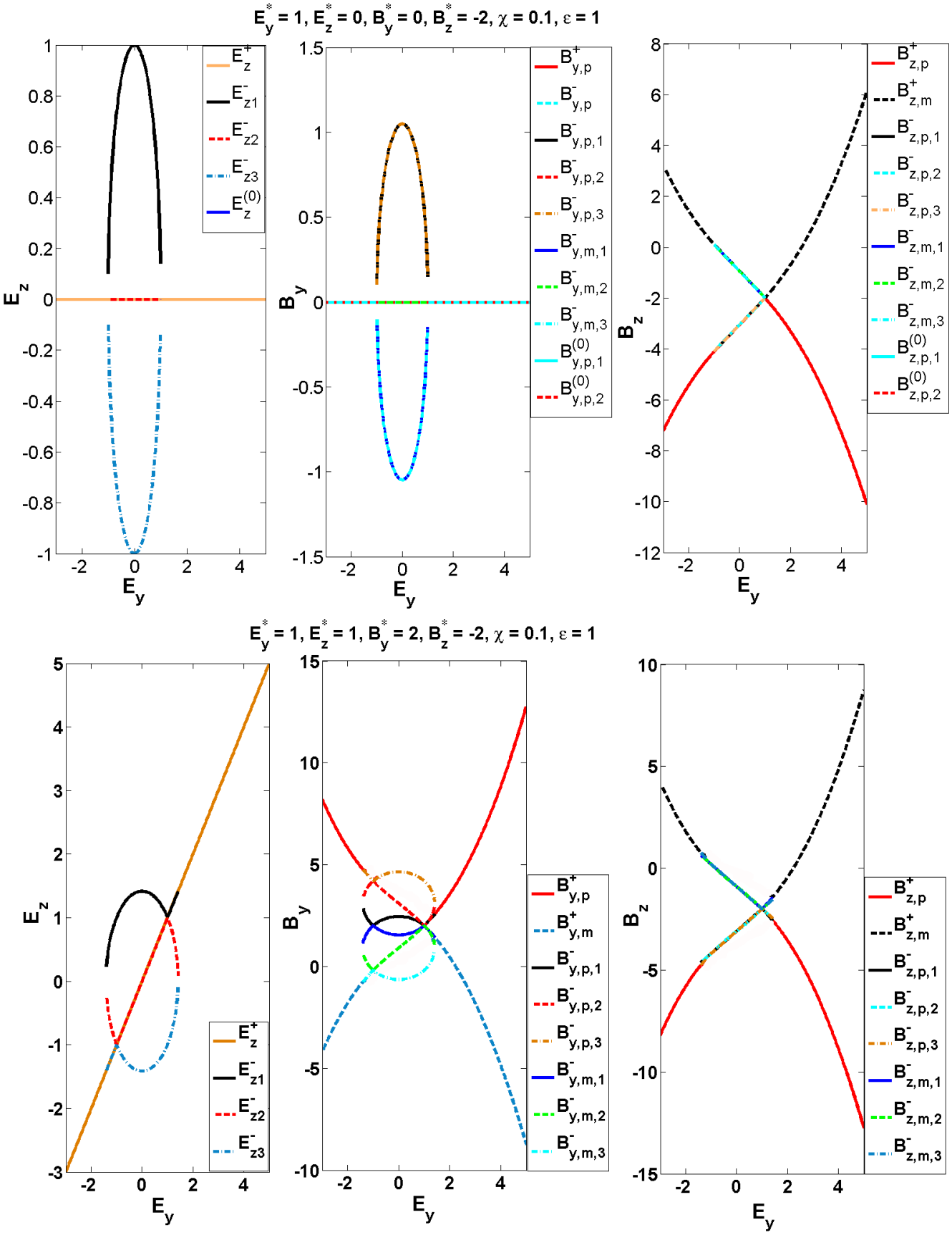



Figure 5.2: Plots of all the solutions given by the Cardano formulas. We have chosen $E_y \in [-|E_y^*| - 2, E_y^* + 4]$ with (Top) $E_y^* = 1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = 2$, $\chi = 0.1$, $\epsilon_0 = 1$. (Bottom) $E_y^* = 1$, $E_z^* = 1$, $B_y^* = 2$, $B_z^* = -2$, $\chi = 0.1$, $\epsilon_0 = 1$. The empty spaces in the plots for of $E_z$ and $B_y$ are due to a deficiency of the plotting program. They should be closed so that ellipses are obtained.

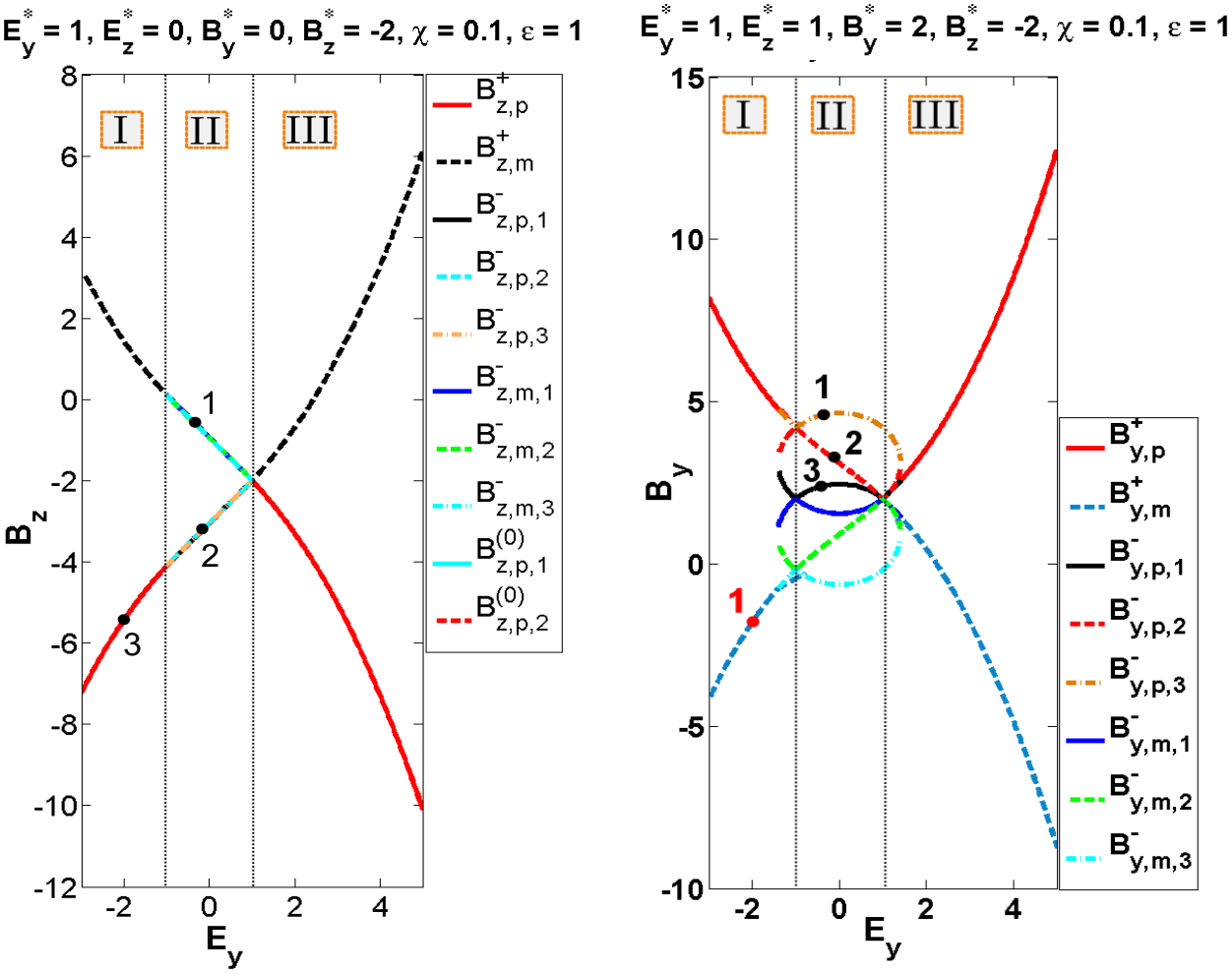



Figure 5.3: Details of the plots of figure 5.2. The left picture shows all solutions given via the Cardano formulas for the $B_z$ component of the Hugoniot Locus. In region II, where $\mathcal{D} < 0$, we have three real solutions in point 1 on the plus branch which are identical, and three real solutions in point 2 on the minus branch which are identical as well. In point 3 in region I, where $\mathcal{D} > 0$, we have exactly one real solution. On the right, we see the same for the $B_y$ component: In points 1 to 3 in region II on the plus branch, it is $\mathcal{D} < 0$, and there are three real solutions which are not identical. In region I, in the red point 1, it is $\mathcal{D} > 0$ and there is exactly one real solution.

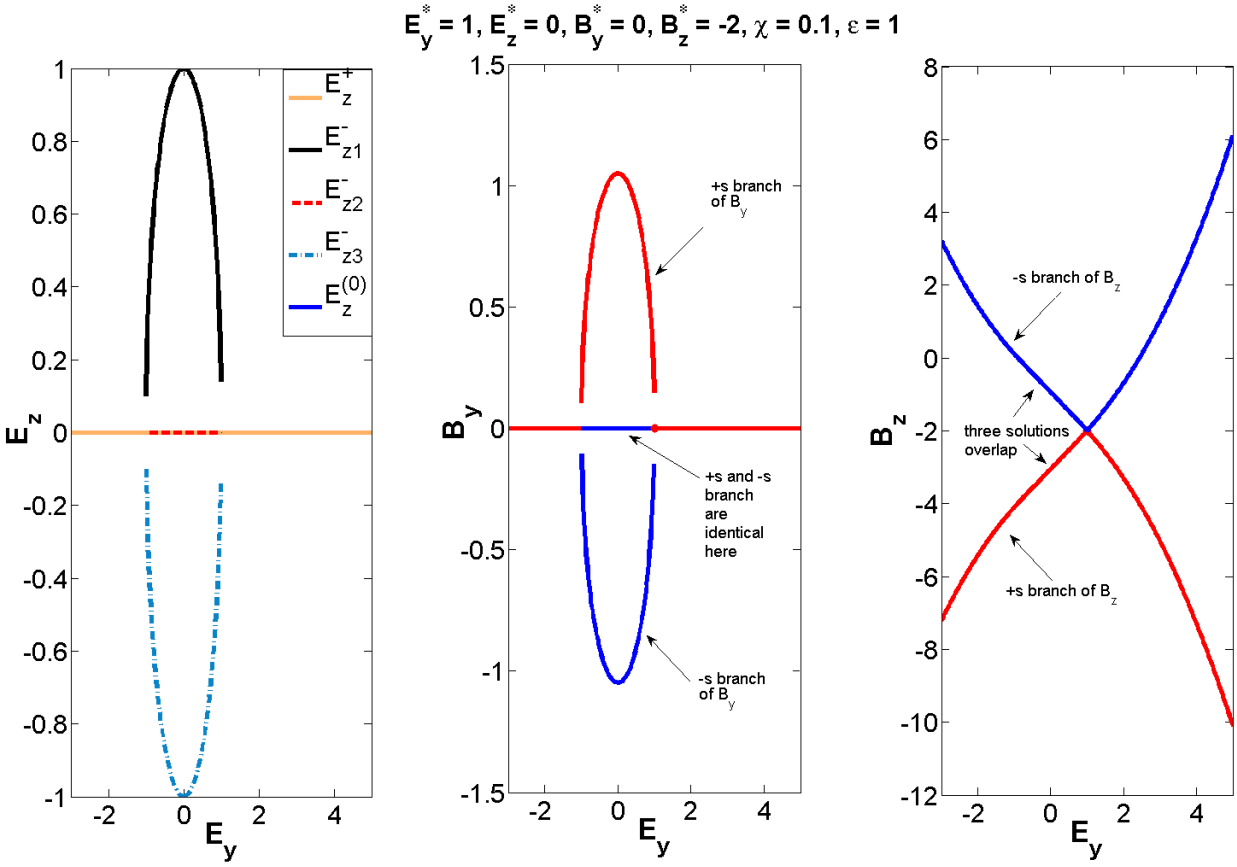



Figure 5.4: Left: All solutions $E_z$ for $E_y^* = 1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = 2$, $\chi = 0.1$, $\epsilon_0 = 1$ and for all cases $\mathcal{D} < 0$, $\mathcal{D} > 0$ and $\mathcal{D} = 0$ which are used to compute $B_y^{\pm}$ and $B_z^{\pm}$. Again, it is $E_y \in [-|E_y^*| - 2, E_y^* + 4]$. Middle and right: $+s$ branches of $B_y$ and $B_z$, respectively, are marked red, $-s$ branches are marked blue.

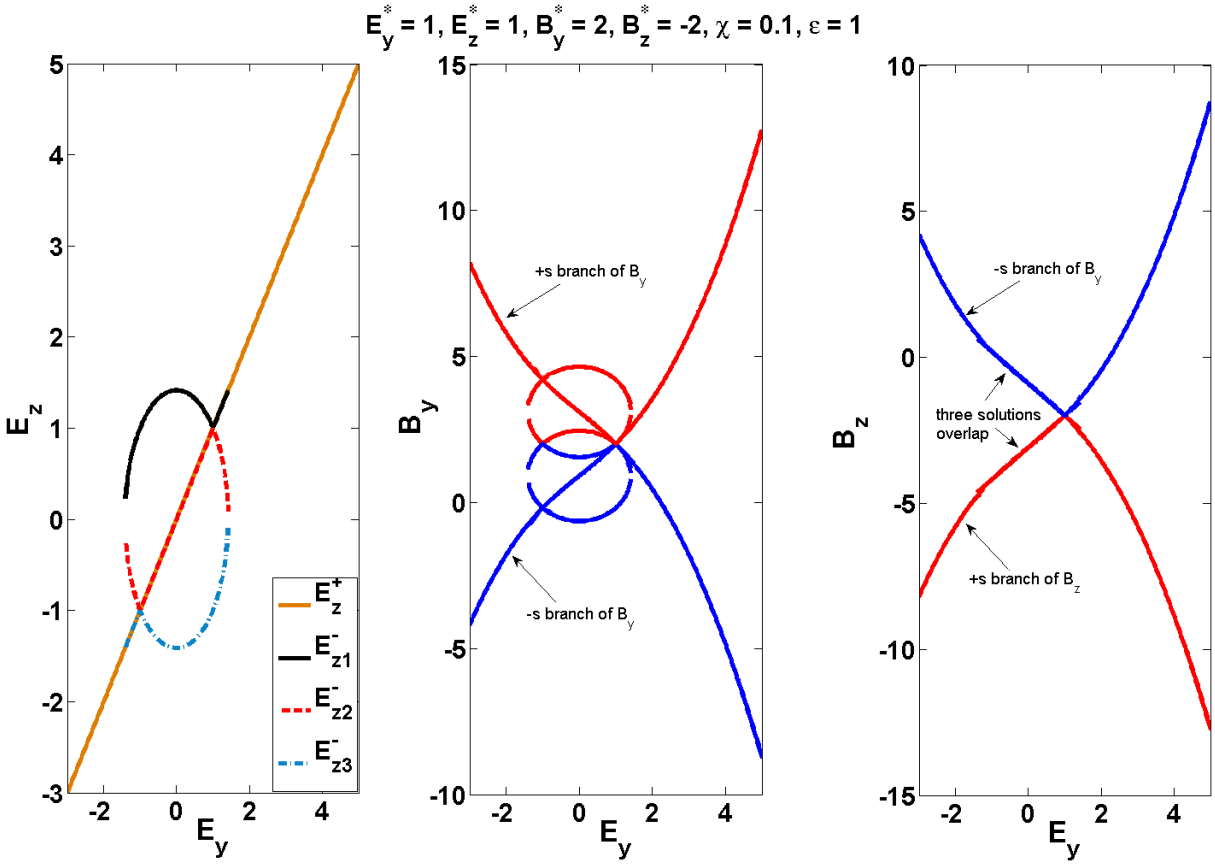



Figure 5.5: The same figure as figure 5.4, but for $E_y^* = 1$, $E_z^* = 1$, $B_y^* = 2$, $B_z^* = -2$, $\chi = 0.1$. Left: All solutions $E_z$ for $\mathcal{D} < 0$, $\mathcal{D} > 0$ and $\mathcal{D} = 0$ which are used to compute $B_y^{\pm}$ and $B_z^{\pm}$. Again, it is $E_y \in [-|E_y^*| - 2, E_y^* + 4]$. Middle and right: $+s$ branches of $B_y$ and $B_z$, respectively, are marked red, $-s$ branches are marked blue.

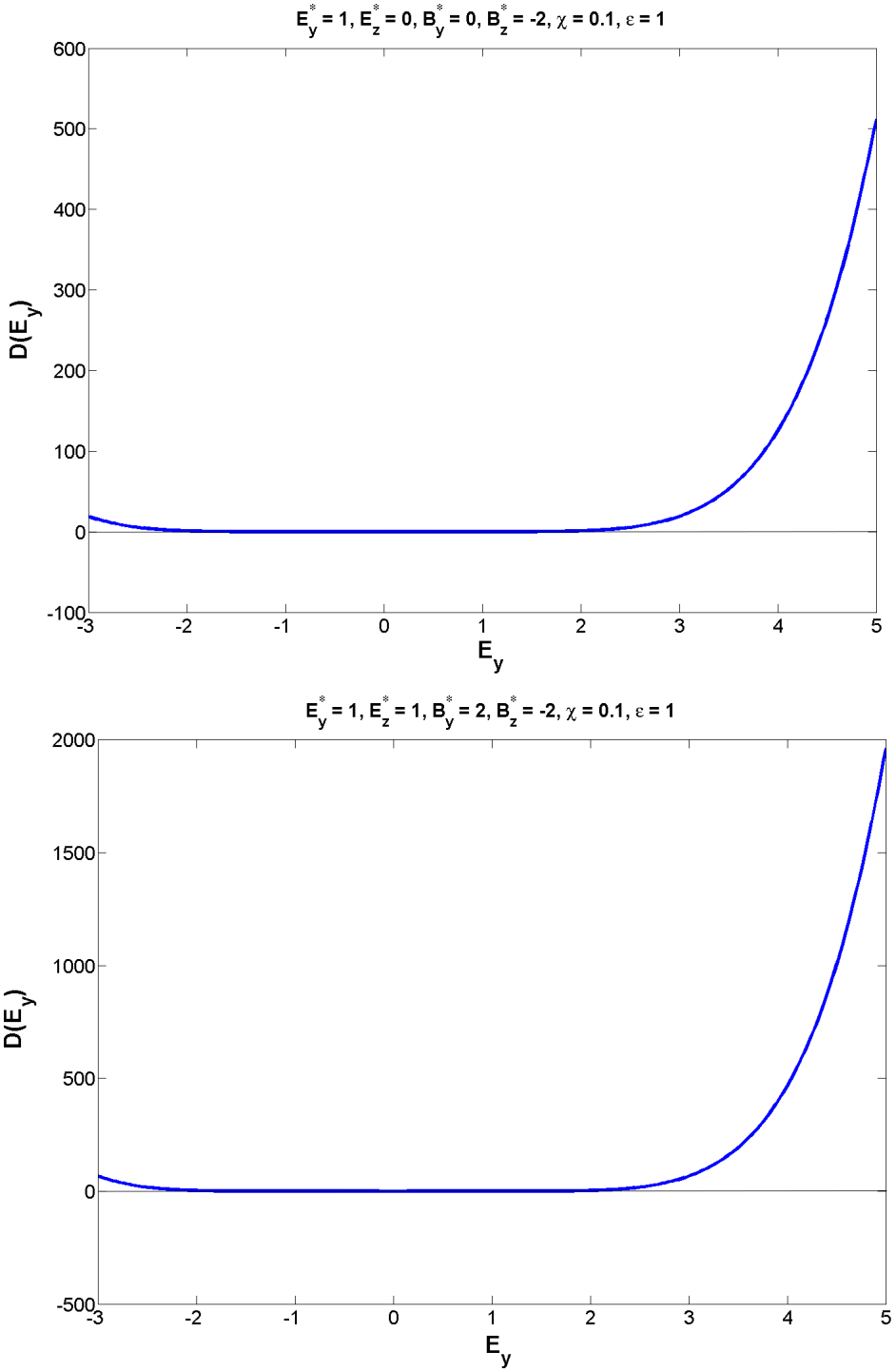



Figure 5.6: Plots of the discriminant $\mathcal{D}$ in dependency of $E_y$ with $E_y \in [-|E_y^*| - 2, E_y^* + 4]$ with (Top) $E_y^* = 1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = 2$, $\chi = 0.1$, $\epsilon_0 = 1$. (Bottom) $E_y^* = 1$, $E_z^* = 1$, $B_y^* = 2$, $B_z^* = -2$, $\chi = 0.1$, $\epsilon_0 = 1$. All the cases $\mathcal{D} > 0$, $\mathcal{D} < 0$ and $\mathcal{D} = 0$ occur.

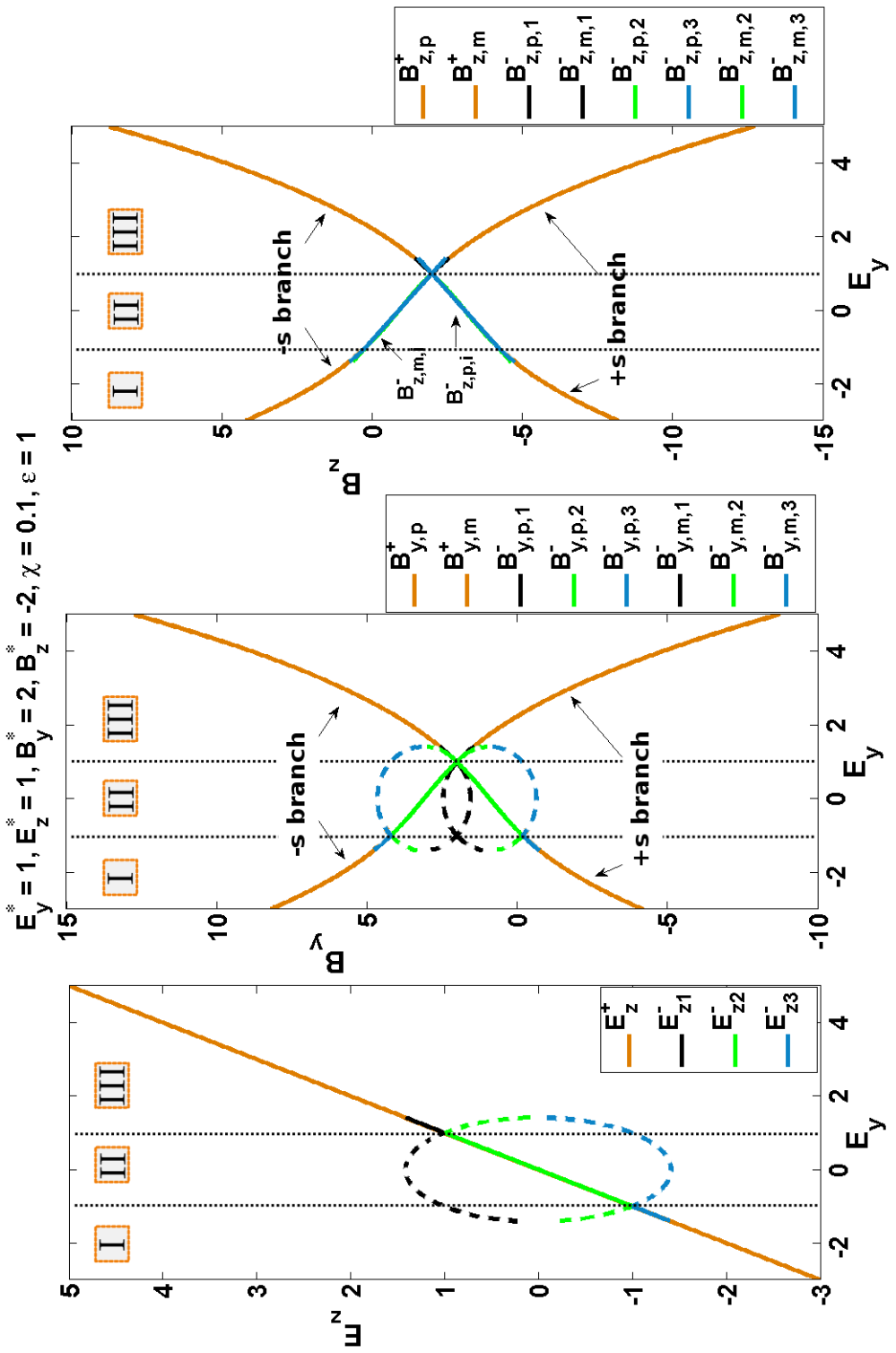



Figure 5.7: Left: $E_z$ after applying the selection rules. The dotted lines show parts of the complete solution that do not fulfill the selection rules. Again, it is $E_y \in [-|E_y^*| - 2, E_y^* + 4]$. Middle and right: The resulting $+s$ branches of $B_y$ and $B_z$, respectively, are marked red, $-s$ branches are marked blue for (Top) $E_y^* = 1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = 2$, $\chi = 0.1$, $\epsilon_0 = 1$. (Bottom) $E_y^* = 1$, $E_z^* = 1$, $B_y^* = 2$, $B_z^* = -2$, $\chi = 0.1$. For this example, where $E_y^* = 1 > 0$ and $E_z^* = 1 > 0$, we have the selection rules as displayed in figure 5.1 (b). Therefore, if $\mathcal{D} < 0$, for $E_y \leq -|E_y^*|$ (region I) $E_{z,3}^-$ is the solution, for $E_y \in [-|E_y^*|, |E_y^*|]$ we choose $E_{z,2}^-$, and for $E_y \geq |E_y^*|$ it is $E_{z,1}^-$. We note that in the right picture for $B_z$ the dotted regions could not be displayed properly.

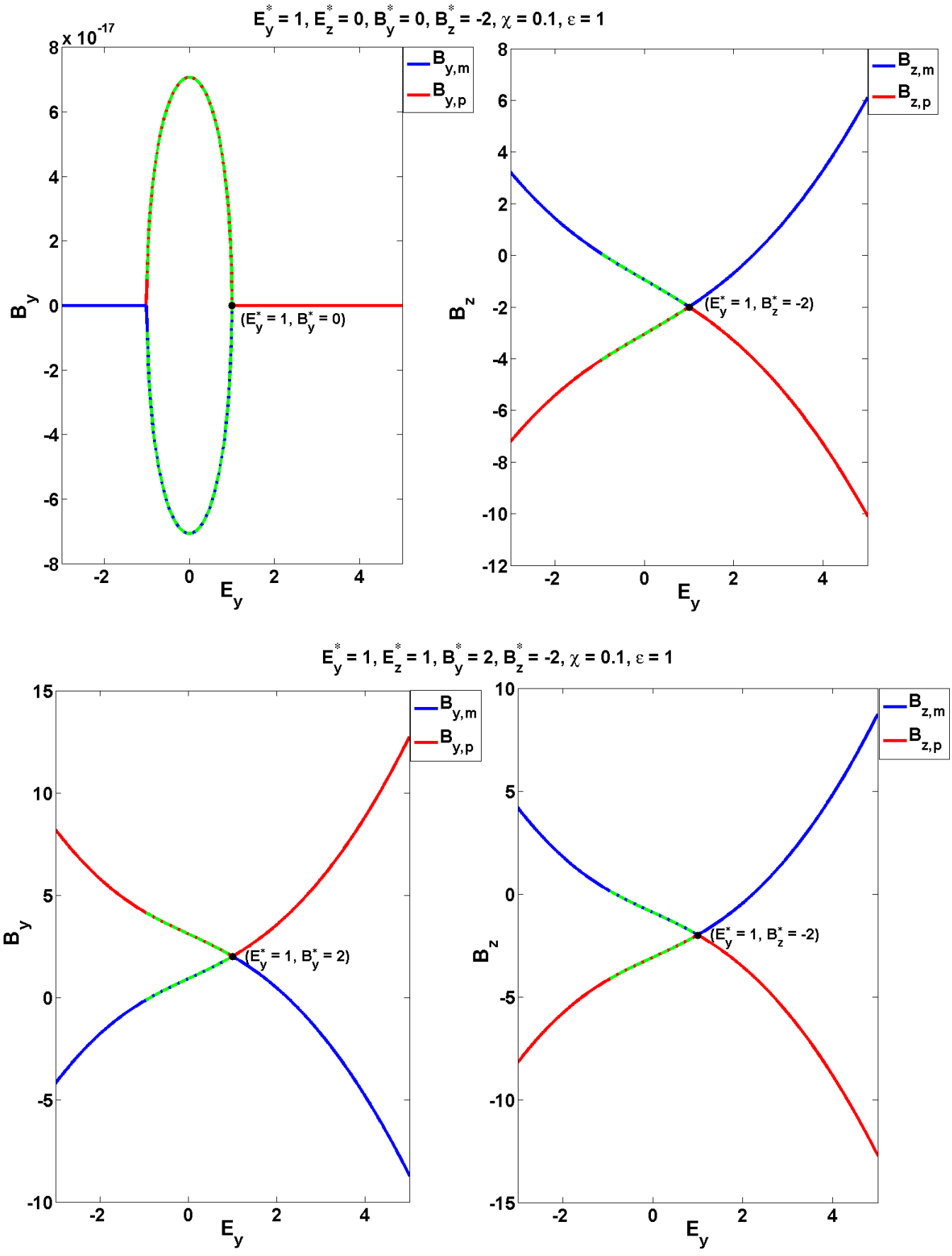



Figure 5.8: One branch of the Hugoniot Locus for the $B_z$ and $B_y$ components after selecting the solutions according to our selection rules with (Top) $E_y^* = 1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = -2$ and (Bottom) $E_y^* = -1$, $E_z^* = 1$, $B_y^* = 1$, $B_z^* = -2$. The dashed parts mark admissible shocks with $\lambda_2^+(\mathbf{E}_L) < s < \lambda_2^+(\mathbf{E}_R)$.

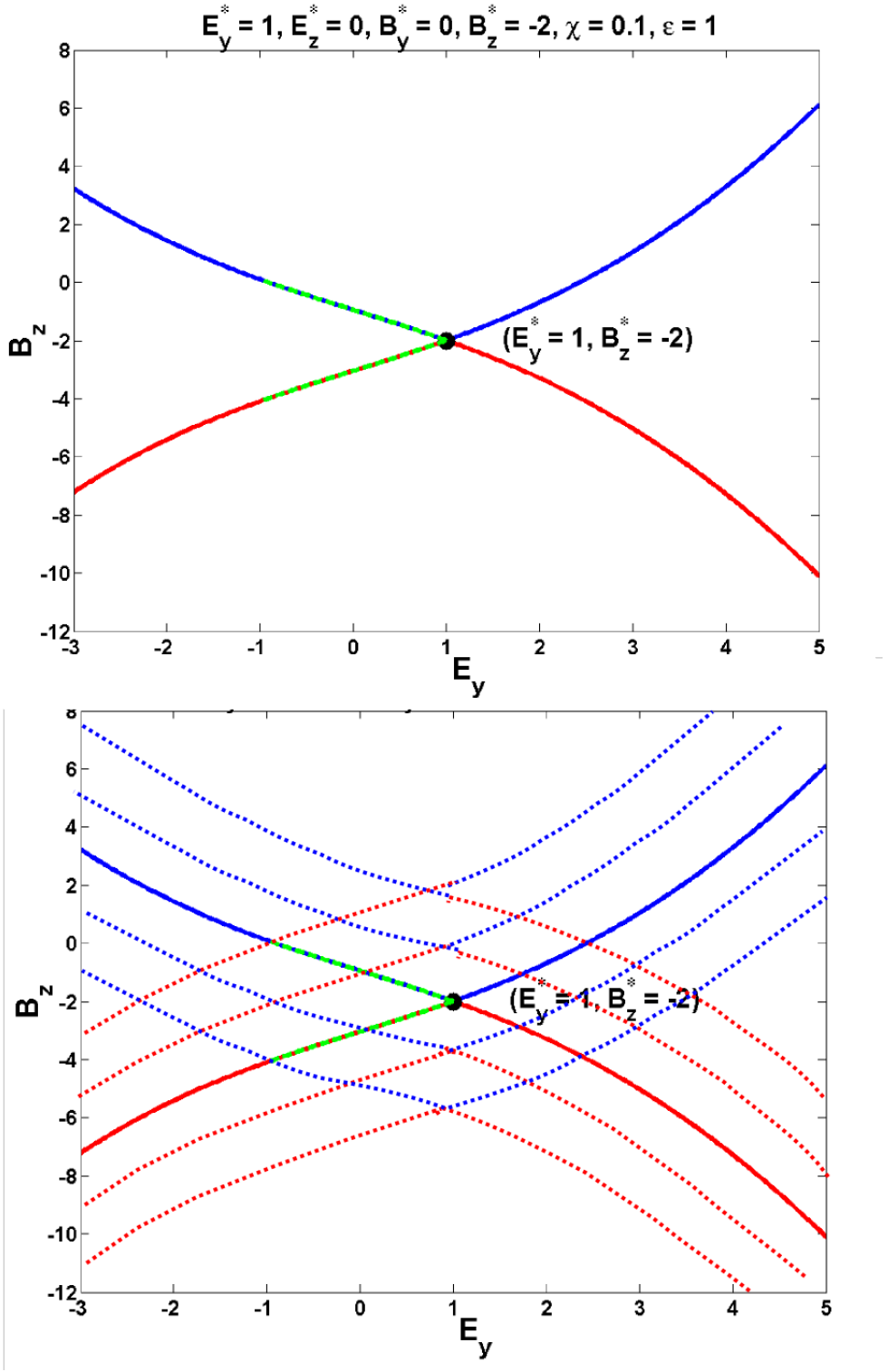



Figure 5.9: Top: One branch of the Hugoniot Locus for the $B_z$ component with $E_y^* = 1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = -2$ after selecting the correct solutions. Bottom: Several corresponding Hugoniot Loci. Blue: $B_z^-$, red: $B_z^+$

### 5.2.2 Admissible Shocks

Not all parts on the Hugoniot Locus give rise to an admissible shock. In section 3.3.2 we gave conditions in order to determine admissible shocks: positivity of entropy and a second condition on the speed of the shock, e.g. the Lax entropy condition, the Liu entropy condition, or the condition of Smoller-Johnson. The author of [71] chooses the condition of Smoller-Johnson in addition to the entropy condition to ensure uniqueness of the solution of the Kerr-nonlinear Riemann problem. Seccia found similar results in [70], yet he uses a reflection and transmission criterion besides the entropy condition. We denote by $\eta$ the entropy of the system, and we check for admissibility by requiring positivity of entropy, that is, $\eta > 0$, and by demanding $\lambda_2(\mathbf{E}_L) < s < \lambda_2(\mathbf{E}_R)$ (condition of Smoller-Johnson); see [71, Section 2.3.6].

#### The Entropy

We want an explicit expression for the entropy. The following demonstration does not comply with mathematical demands and shall give an idea of how to find the entropy function given in [71] or [70]. For doing so, first recall Maxwell's curl-equations

$$\partial_t \mathbf{D} - \nabla \times \mathbf{H} = \mathbf{0}, \tag{5.2.16a}$$
$$\partial_t \mathbf{B} + \nabla \times \mathbf{E} = \mathbf{0}, \tag{5.2.16b}$$

with $\mathbf{B} = \mu_0 \mathbf{H}$ and $\mathbf{D} = \mathcal{E}'\mathbf{E}$. By multiplying equation (5.2.16b) with $\mathbf{B}$ and equation (5.2.16a) by $\mathbf{E}$ we obtain

$$\mathbf{B} \cdot \partial_t \mathbf{B} + \mathbf{B} \cdot (\nabla \times \mathbf{E}) = 0,$$
$$\mathbf{E} \cdot \partial_t \mathbf{D} - \mathbf{E} \cdot (\nabla \times \mathbf{H}) = 0.$$

Adding both equations, using the vector identity $\nabla \cdot (\mathbf{E} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{H})$ and defining $\mathbf{S} := \mathbf{E} \times \mathbf{B}$ we get

$$(\mathbf{B} \cdot \partial_t \mathbf{B} + \mathbf{E} \cdot \partial_t \mathbf{D}) + \nabla \cdot \mathbf{S} = 0, \tag{5.2.17}$$

$\mathbf{S}$ is called the *Poynting vector*. Poynting's theorem (see e.g. [38]) gives (5.2.17) in conservation form as

$$\partial_t w + \nabla \cdot \mathbf{S} = 0, \tag{5.2.18}$$

where $w = w_E + w_M$ is the total energy, $w_E$ is the electric energy and $w_M$ the magnetic energy, which are respectively given as

$$w_E = \int \mathbf{E} \, \mathrm{d}\mathbf{D},$$
$$w_M = \int \frac{1}{\mu_0} \mathbf{B} \, \mathrm{d}\mathbf{B}.$$

For the Kerr-system (5.1.11) it is

$$\begin{aligned} w_E &= \int \mathbf{E} \, \mathrm{d}\mathbf{D} = \int \mathbf{E} \, \mathrm{d}(\mathcal{E}'\mathbf{E}) = \int \mathbf{E} \, \mathrm{d}((\epsilon_0 + \chi|\mathbf{E}|^2)\mathbf{E}) \\ &= \epsilon_0 \int \mathbf{E} \, \mathrm{d}\mathbf{E} + \chi \int \mathbf{E} \, \mathrm{d}(|\mathbf{E}|^2 \mathbf{E}) \\ &= \epsilon_0 \frac{|\mathbf{E}|^2}{2} + \frac{3}{4}\chi |\mathbf{E}|^4, \end{aligned} \tag{5.2.19}$$

$$w_M = \int \frac{1}{\mu_0} \mathbf{B} \, \mathrm{d}\mathbf{B} = \frac{|\mathbf{B}|^2}{2\mu_0}.$$

We note that an alternative way of obtaining the energy $w$ is via the Hamiltonian of the system, which is given as [71]

$$\mathcal{H}(\mathbf{E},\mathbf{B}) := \mathbf{D}\cdot\mathbf{E} - \mathcal{L}(\mathbf{E},\mathbf{B}), \tag{5.2.20}$$

where $\mathcal{L}$ is the Lagrangian,

$$\mathcal{L}(\mathbf{E},\mathbf{B}) := \mathcal{E}\left(\frac{|\mathbf{E}|^2}{2}\right)\mathbf{E} - \frac{|\mathbf{B}|^2}{2\mu_0}, \tag{5.2.21}$$

so that

$$\begin{aligned}\mathcal{H}(\mathbf{E},\mathbf{B}) &= \mathcal{E}'\mathbf{E}\cdot\mathbf{E} - \mathcal{L}(\mathbf{E},\mathbf{B}) = (\epsilon_0+\chi|\mathbf{E}|^2)|\mathbf{E}|^2 - (\epsilon_0\frac{|\mathbf{E}|^2}{2} + \chi\frac{|\mathbf{E}|^4}{4}) + \frac{|\mathbf{B}|^2}{2\mu_0}\\ &= \epsilon_0|\mathbf{E}|^2 + \chi|\mathbf{E}|^4 - \epsilon_0\frac{|\mathbf{E}|^2}{2} - \chi\frac{|\mathbf{E}|^4}{4} + \frac{|\mathbf{B}|^2}{2\mu_0}\\ &= \epsilon_0\frac{|\mathbf{E}|^2}{2} + \frac{3}{4}\chi|\mathbf{E}|^4 + \frac{|\mathbf{B}|^2}{2\mu_0}\\ &= w.\end{aligned} \tag{5.2.22}$$

The Rankine-Hugoniot jump conditions for Maxwell's curl-equations (5.2.16) and (5.2.16a) read (recall (5.2.2))

$$s\,[\![\mathbf{D}]\!] + \mathbf{n}\times[\![\mathbf{H}]\!] = \mathbf{0}, \tag{5.2.23a}$$

$$-s\,[\![\mathbf{B}]\!] + \mathbf{n}\times[\![\mathbf{E}]\!] = \mathbf{0}, \tag{5.2.23b}$$

where the jump was defined as $[\![\mathbf{u}]\!] = \mathbf{u}_L - \mathbf{u}_R$. The divergence conditions $\nabla\cdot\mathbf{B} = 0$ and $\nabla\cdot\mathbf{D} = 0$, together with the fact that the normal components of the **E**- and **H**-field must be zero (assuming no sources), render the conditions

$$s\,[\![\mathbf{D_n}]\!] = [\![\mathcal{E}'\mathbf{E_n}]\!] = \mathbf{0} \quad \text{and} \quad [\![\mathbf{E_n}]\!] = \mathbf{0}, \tag{5.2.24a}$$

$$s\,[\![\mu_0\mathbf{H_n}]\!] = \mathbf{0}, \tag{5.2.24b}$$

where $\mathbf{D_n} := \mathbf{D}\cdot\mathbf{n}$, $\mathbf{H_n} := \mathbf{H}\cdot\mathbf{n}$, and $\mathbf{n}$ as the outer unit normal of $\partial\Omega$. Applying the Rankine-Hugoniot jump condition to Poynting's theorem (5.2.18) gives the so-called generalized entropy [68].

**Definition 5.4.**
*Let $\mathbf{u}(x,t)$ be the entropy solution of the conservation law $\partial_t\mathbf{u} + \partial_x F(\mathbf{u}) = \mathbf{0}$ having a discontinuity in $x(t)$ and moving with shock speed $s$. In the discontinuity $x(t)$ we define the generalized entropy*

$$\eta := -s[\![w]\!] + [\![\mathbf{n}\cdot\mathbf{S}]\!], \tag{5.2.25}$$

*where $S = \mathbf{E}\times\mathbf{H}$ is the Poynting vector.*

**Lemma 5.5.**
*For $s \neq 0$ the entropy can be expressed as*

$$\eta = \frac{1}{\mu_0 s}[\![(3 - s^2\mu_0(\epsilon_0 + \frac{3}{2}\chi|\mathbf{E}|^2)|\mathbf{E}|^2]\!]. \tag{5.2.26}$$

*Note that for $s = 0$ we have the null shock. The entropy condition requires $\eta \geq 0$.*

*Proof.* Equation (5.2.23b) is equivalent to $[\![\mathbf{H}]\!] = \frac{1}{\mu_0 s}\mathbf{n} \times [\![\mathbf{E}]\!]$. Plugging this into (5.2.23a) gives

$$s[\![\mathbf{D}]\!] + \mathbf{n} \times [\![\frac{1}{\mu_0 s}\mathbf{n} \times \mathbf{E}]\!] = 0.$$

The vector identity $\mathbf{n} \times (\mathbf{n} \times \mathbf{E}) = (\mathbf{E} \cdot \mathbf{n})\mathbf{n}$ and the fact that $[\![\mathbf{E_n}]\!] = 0$ (see (5.2.24a)) leads to

$$\mathbf{n} \times [\![\mathbf{n} \times \mathbf{E}]\!] = [\![\mathbf{E_n}]\!]\mathbf{n} - [\![\mathbf{E}]\!] = -[\![\mathbf{E}]\!]. \tag{5.2.27}$$

Altogether we obtain

$$s[\![\mathbf{D}]\!] + \mathbf{n} \times [\![\frac{1}{\mu_0 s}\mathbf{n} \times \mathbf{E}]\!] = s[\![\mathcal{E}'\mathbf{E}]\!] - \frac{1}{\mu_0 s}[\![\mathbf{E}]\!] = 0$$

or, equivalently,

$$[\![\mathcal{E}'\mathbf{E}]\!] = \frac{1}{\mu_0 s^2}[\![\mathbf{E}]\!]. \tag{5.2.28}$$

For the subsequent statements we note that for two vectors $\mathbf{a}$ and $\mathbf{b}$ it holds

$$[\![\mathbf{a} \cdot [\![\mathbf{b}]\!]]\!] = [\![\mathbf{a}]\!] \cdot [\![\mathbf{b}]\!].$$

By using this relation and also observing that

$$\mathbf{n} \cdot \mathbf{S} = \mathbf{n} \cdot (\mathbf{E} \times \mathbf{H}) = \mathbf{H} \cdot (\mathbf{n} \times \mathbf{E}) - \mathbf{E} \cdot (\mathbf{n} \times \mathbf{H})$$

and by applying relations (5.2.23b) and (5.2.23a) we obtain

$$[\![\mathbf{n} \times \mathbf{E}]\!] = [\![\mathbf{H} \cdot s[\![\mathbf{B}]\!] - \mathbf{E} \cdot (-s[\![\mathbf{D}]\!])]\!] = \mu_0 s[\![\mathbf{H}]\!]^2 + s[\![\mathbf{E}]\!] \cdot [\![\mathcal{E}'\mathbf{E}]\!]. \tag{5.2.29}$$

Into this equation we plug the relation found in (5.2.28), which gives

$$\mu_0 s[\![|\mathbf{H}|^2]\!] + s[\![E]\!] \cdot [\![\mathcal{E}'\mathbf{E}]\!] = \mu_0 s[\![|\mathbf{H}|^2]\!] + \frac{1}{\mu_0 s}[\![\mathbf{E}]\!]^2. \tag{5.2.30}$$

Altogether, and by also recalling (5.2.19), the entropy from definition 5.4 becomes

$$\eta = -s[\![w]\!] + [\![\mathbf{n} \cdot (\mathbf{E} \times \mathbf{H})]\!] = \frac{s\mu_0[\![|\mathbf{H}|^2]\!]}{2} - \frac{s\epsilon_0[\![|\mathbf{E}|^2]\!]}{2} - \frac{3s\chi[\![|\mathbf{E}|^4]\!]}{4} + \frac{1}{\mu_0 s}[\![|\mathbf{E}|^2]\!]. \tag{5.2.31}$$

From this equation we eliminate $[\![|\mathbf{H}|^2]\!]$. For this we need the following relations (see (5.2.23b), (5.2.23a))

$$[\![\mathbf{H}]\!] = \frac{1}{\mu_0 s}\mathbf{n} \times [\![\mathbf{E}]\!], \tag{5.2.32a}$$

$$[\![\mathcal{E}'\mathbf{E}]\!] = -\frac{1}{s}\mathbf{n} \times [\![\mathbf{H}]\!], \tag{5.2.32b}$$

the vector identity (5.2.27) and

$$\begin{aligned}
(\mathbf{n} \times [\![|\mathbf{H}|^2]\!]) &= (\mathbf{n} \times [\![\mathbf{H}]\!]) \cdot (\mathbf{n} \times [\![\mathbf{H}]\!]) && (5.2.33)\\
&= (\mathbf{n} \cdot \mathbf{n})([\![\mathbf{H}]\!] \cdot [\![\mathbf{H}]\!]) - (\mathbf{n} \cdot [\![\mathbf{H}]\!])(\mathbf{n} \cdot [\![\mathbf{H}]\!]) && (5.2.34)\\
&= [\![|\mathbf{H}|^2]\!] - [\![|\mathbf{H_n}|^2]\!] && (5.2.35)\\
&= [\![|\mathbf{H}|^2]\!], && (5.2.36)
\end{aligned}$$

where we used $[\![\mathbf{H_n}]\!] = 0$ from (5.2.23b) and (5.2.23a). Taking $\mathbf{n}\times$ (5.2.32a) and using (5.2.27) we encounter

$$\mathbf{n} \times [\![\mathbf{H}]\!] = \frac{1}{s\mu_0}\mathbf{n} \times (\mathbf{n} \times [\![\mathbf{E}]\!]) = -\frac{1}{s\mu_0}[\![\mathbf{E}]\!]. \tag{5.2.37}$$

From equation (5.2.23a) we get

$$\mathbf{n} \times [\![\mathbf{H}]\!] = -s[\![\mathcal{E}'\mathbf{E}]\!]. \tag{5.2.38}$$

Multiplying (5.2.37) with (5.2.38) and using the statement in equation (5.2.33) leads to

$$[\![|\mathbf{H}|^2]\!] = (\mathbf{n} \times [\![|\mathbf{H}|^2]\!]) = \frac{1}{\mu_0}[\![\mathbf{E}]\!][\![\mathcal{E}'\mathbf{E}]\!].$$

Now we use equation (5.2.28), i.e. $[\![\mathcal{E}'\mathbf{E}]\!] = \frac{1}{\mu_0 s^2}[\![\mathbf{E}]\!]$, and plug this into the last relation, which renders

$$[\![|\mathbf{H}|^2]\!] = \frac{1}{s^2\mu_0^2}[\![\mathbf{E}]\!]^2.$$

Finally the entropy in (5.2.31) becomes after some rearrangement in terms

$$\begin{aligned}\eta &= \frac{s\mu_0[\![|\mathbf{H}|^2]\!]}{2} - \frac{s\epsilon_0[\![|\mathbf{E}|^2]\!]}{2} - \frac{3s\chi[\![|\mathbf{E}|]\!]^4}{4} + \frac{1}{\mu_0 s}[\![|\mathbf{E}|^2]\!] \\ &= \frac{3}{2s^2\mu_0}[\![|\mathbf{E}|^2]\!] - \frac{s\epsilon_0[\![|\mathbf{E}|^2]\!]}{2} - \frac{3s\chi[\![|\mathbf{E}|^4]\!]}{4} \\ &= \frac{1}{2s\mu_0}[\![|\mathbf{E}|^2(3 - s^2\mu_0(\epsilon_0 + \frac{3}{2}\chi|\mathbf{E}|^2)]\!].\end{aligned}$$

□

Figure 5.10 shows the shock speed $s^+ = \left(\frac{E_y^* - E_y}{D_y^* - D_y}\right)^{\frac{1}{2}}$ and the eigenvalue $\lambda_2^+ = \frac{1}{\sqrt{\mathcal{E}' + \mathcal{E}''|\mathbf{E}|^2}}$ in dependency of $E_y$ for the exemplary choices from the last subsection, i.e. $E_y^* = 1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = -2$ and for $E_y^* = -1$, $E_z^* = 1$, $B_y^* = 1$, $B_z^* = -2$, respectively. The green dashed parts mark the corresponding admissible shocks. The green dots mark the end points $\lambda_2^+(|\mathbf{E}_L|)$ and $\lambda_2^+(|\mathbf{E}_R|)$ of the condition of Smoller-Johnson $\lambda_2(\mathbf{E}_L) < s < \lambda_2(\mathbf{E}_R)$. Figure 5.11 shows plots of the entropy $\eta$ in (5.2.26) in dependency of $E_y$.

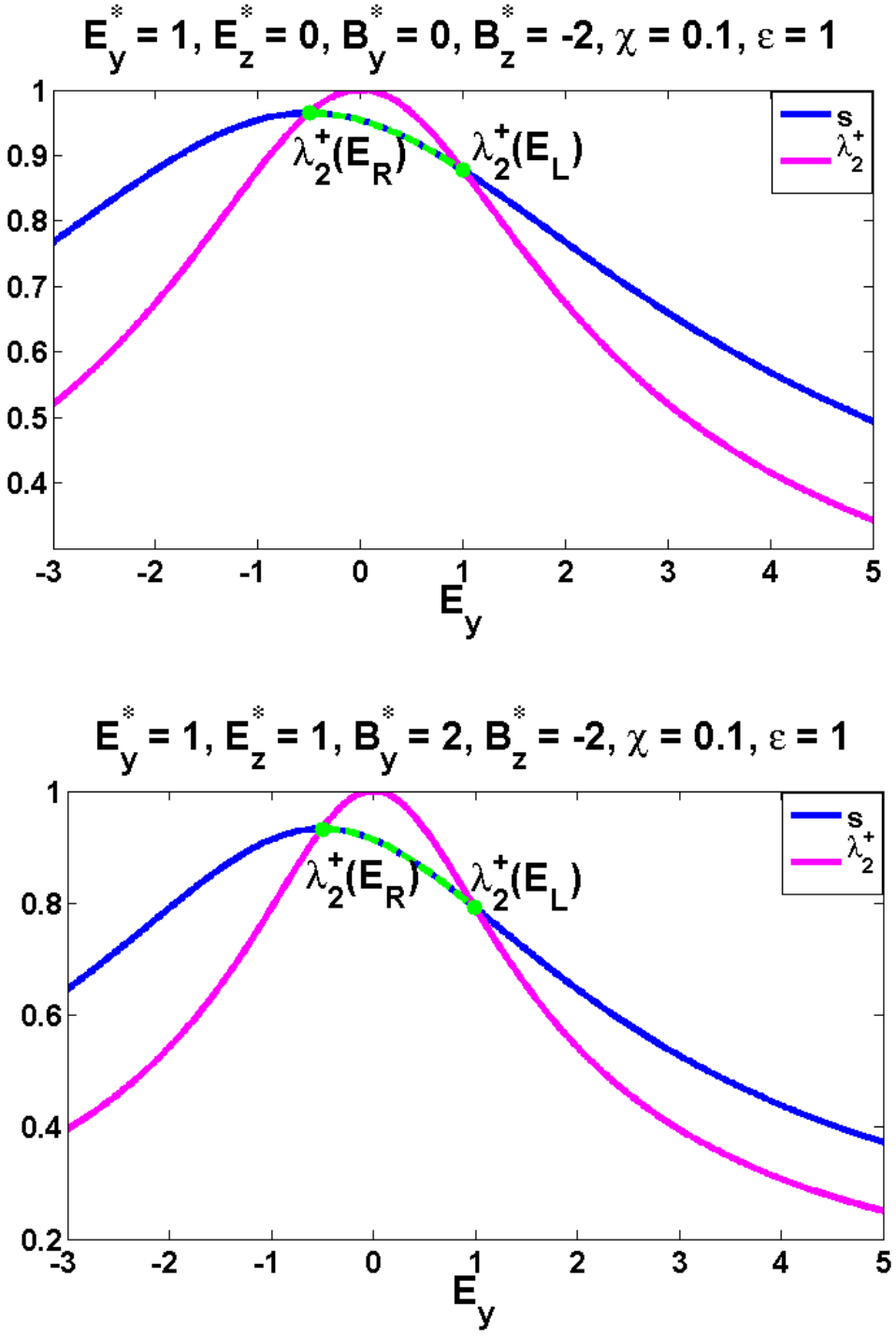



Figure 5.10: Shock speed $s^+ = \left(\frac{E_y^* - E_y}{D_y^* - D_y}\right)^{\frac{1}{2}}$ from (5.2.15) (blue) and the eigenvalue $\lambda_2^+ = \frac{1}{\sqrt{\mathcal{E}' + \mathcal{E}''|\mathbf{E}|^2}}$ (magenta) in dependency of $E_y$ for (Top) $E_y^* = -1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = -2$ and (Bottom) $E_y^* = 1$, $E_z^* = 1$, $B_y^* = 2$, $B_z^* = -2$. The dashed lines mark admissible shocks. $\mathbf{E}_R$ denotes the value right of the line $x = 0$, $\mathbf{E}_L$ denotes the value left of the line $x = 0$.

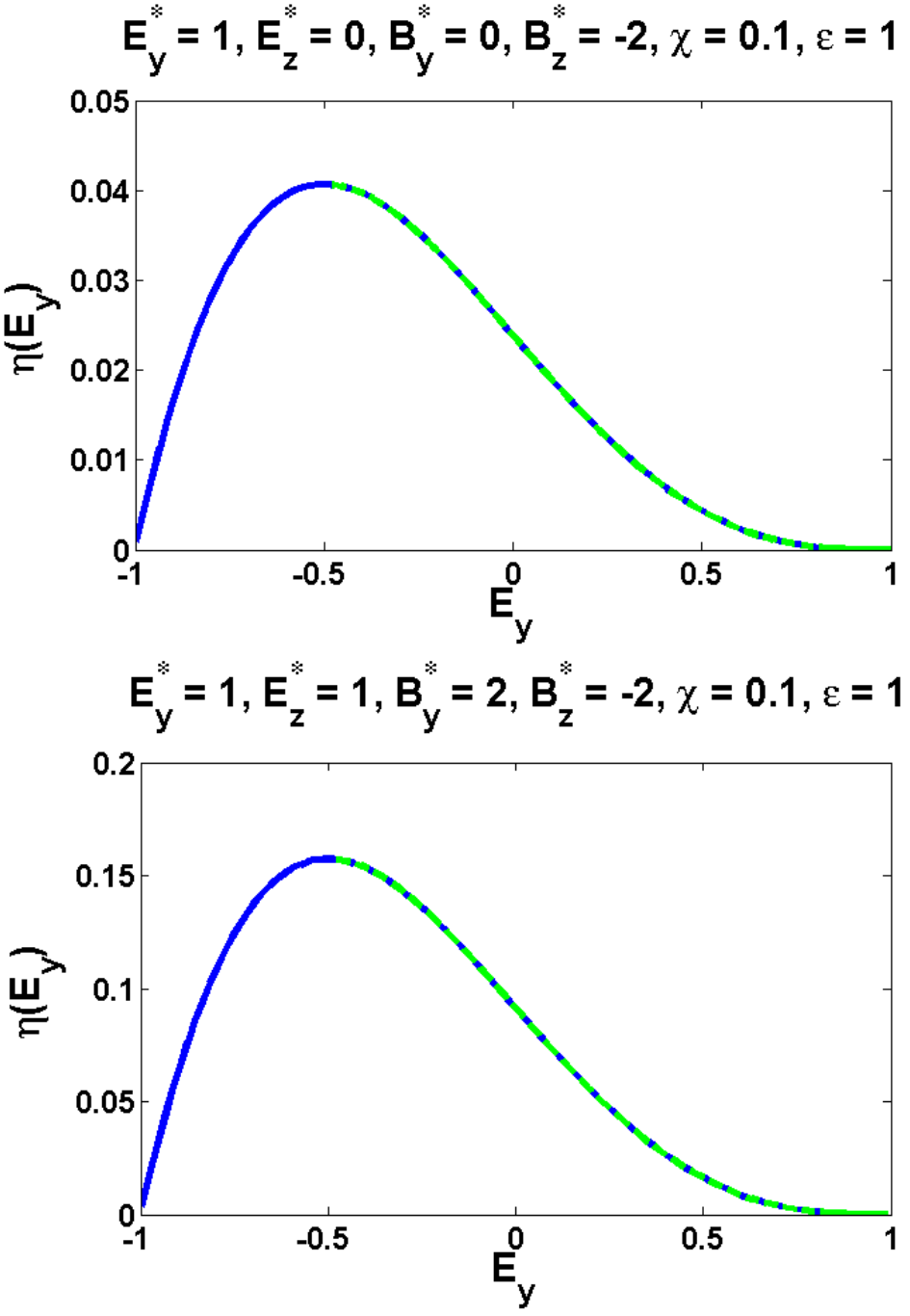



Figure 5.11: Entropy (5.2.26) in dependency of $E_y$ for (Top) $E_y^* = -1$, $E_z^* = 0$, $B_y^* = 0$, $B_z^* = -2$ and (Bottom) $E_y^* = 1$, $E_z^* = 1$, $B_y^* = 2$, $B_z^* = -2$. The dashed lines mark admissible shocks.

### 5.2.3 Riemann Invariants

In section 3.3.2 we presented the definition and some properties of Riemann invariants. A Riemann invariant remains constant across a characteristic wave. It can thus be used to formulate the solution of the Riemann problem.
Recall relations (3.3.19) and (3.3.20) with which the $i$th Riemann invariant $\mathbf{R}_i$ can be determined:

$$\mathbf{R}_i(u) = \int du - \int v_i(u(\xi))d\xi \equiv \text{const} \tag{5.2.39}$$

or, equivalently, one can use

$$\nabla_u \mathbf{R}_i \cdot \mathbf{v}_i = 0. \tag{5.2.40}$$

For the Kerr-nonlinear Riemann problem there are 4 Riemann invariants, which are presented in the subsequent lemma.

**Lemma 5.6** (Riemann invariants for the Kerr-nonlinear Maxwell's equations)**.**
*The Riemann invariants of the Kerr-nonlinear Maxwell's equations* (5.1.11) *in conservative form, i.e.*

$$\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}$$

*with* $\mathbf{u} = (D_y, D_z, \mu_0 H_y, \mu_0 H_z)^T$ *and* $\mathbf{F}(\mathbf{u}) = (H_z, -H_y, -E_z, E_y)^T$*, are given as follows:*

$$\begin{aligned}
&\textit{For } \lambda_1^{\pm}: \\
&\qquad \mathbf{R}_1^{\pm} := \begin{pmatrix} R_{1,1}^{\pm} \\ R_{1,2}^{\pm} \end{pmatrix} = \begin{pmatrix} |\mathbf{E}|^2 \\ |\mathbf{E}|^2 \end{pmatrix}, \\
&\qquad \mathbf{R}_2^{\pm} := \begin{pmatrix} R_{2,1}^{\pm} \\ R_{2,2}^{\pm} \end{pmatrix} = \sqrt{\mathcal{E}'}\mathbf{E} \mp \mathbf{B}, \\
&\textit{For } \lambda_2^{\pm}: \\
&\qquad \mathbf{R}_3^{\pm} := \begin{pmatrix} R_{3,1}^{\pm} \\ R_{3,2}^{\pm} \end{pmatrix} = \frac{\mathbf{E}}{|\mathbf{E}|} \mp \mathbf{B}, \\
&\qquad \mathbf{R}_4^{\pm} := \begin{pmatrix} R_{4,1}^{\pm} \\ R_{4,2}^{\pm} \end{pmatrix} = G(|\mathbf{E}|)\frac{\mathbf{E}}{|\mathbf{E}|} \mp B,
\end{aligned} \tag{5.2.41}$$

*where* $\mathbf{E} = (E_y, E_z)$*,* $\mathbf{B} = B_z, -B_y$*,* $\mathcal{E}' = \mathcal{E}'\left(\frac{|\mathbf{E}|^2}{2}\right) = \epsilon + \chi|\mathbf{E}|^2$ *and* $G(|\mathbf{E}|)$ *is defined as the primitive of*

$$\frac{\mathrm{d}G}{\mathrm{d}|\mathbf{E}|} = \sqrt{\mathcal{E}' + \mathcal{E}''|\mathbf{E}|^2}.$$

*In our case, where* $\mathcal{E}' = \epsilon + \chi|\mathbf{E}|^2$*,* $G(|\mathbf{E}|)$ *can be given explicitly as*

$$G(|\mathbf{E}|) = \frac{1}{2}|\mathbf{E}|\frac{\mathrm{d}G(|\mathbf{E}|)}{\mathrm{d}|\mathbf{E}|} + \frac{\sqrt{3}}{6\sqrt{\chi}}\epsilon_0 \ln\left(\sqrt{3\chi}|\mathbf{E}|\right) + \frac{\mathrm{d}G(|\mathbf{E}|)}{\mathrm{d}|\mathbf{E}|}. \tag{5.2.42}$$

$\mathbf{R}_1^{\pm}$ *and* $\mathbf{R}_2^{\pm}$ *are 1-Riemann invariants corresponding to* $\lambda_1^{\pm}$*;* $\mathbf{R}_2^{\pm}$ *and* $\mathbf{R}_3^{\pm}$ *are 2-Riemann invariants corresponding to* $\lambda_2^{\pm}$*.*

Recall that whenever we write $\mathcal{E}'$ we mean $\mathcal{E}' = \mathcal{E}'\left(\frac{|\mathbf{E}|^2}{2}\right)$.

*Proof.* A proof can be found in [71], using the relation (5.2.40). Here, for demonstration purposes, we will additionally present how to compute the Riemann invariants $\mathbf{R}_1, \mathbf{R}_2$ and $\mathbf{R}_3$ by solving (5.2.39), and exemplary show that the relation (5.2.40) holds for $\mathbf{R}_4$.
The eigenvalues and eigenvectors of the Kerr system are given in (5.2) as

$$\lambda_1^{\pm} = \pm\frac{1}{\sqrt{\mathcal{E}'}}, \qquad \lambda_2^{\pm} = \pm\frac{1}{\sqrt{\mathcal{E}' + \mathcal{E}''|\mathbf{E}|^2}},$$

and recall $\lambda_1^- \leq \lambda_2^- \leq \lambda_2^+ \leq \lambda_1^+$. The corresponding eigenvectors are

$$\mathbf{v}_1^{\pm} := \begin{pmatrix} \mathbf{E}_\perp \\ \lambda_1^{\pm}\mathbf{E}_\perp \end{pmatrix}, \qquad \mathbf{v}_2^{\pm} := \begin{pmatrix} \mathbf{E} \\ \lambda_2^{\pm}\mathbf{E} \end{pmatrix}, \tag{5.2.43}$$

where $\mathbf{E}_\perp = (-E_z, E_y)^T$ is orthogonal to $\mathbf{E} = (E_y, E_z)^T$. We also note that $\nabla_{\mathbf{u}} = (J^{-1}\partial_{\mathbf{E}}, \partial_{\mathbf{B}})$. We want to solve the system of ordinary differential equations

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\xi} = \mathbf{v}_i^{\pm} \qquad (i = 1, 2).$$

For $\lambda_1^{\pm}$ we have the following set of equations:

$$\frac{\mathrm{d}D_y}{\mathrm{d}\xi} = -E_z, \tag{5.2.44a}$$

$$\frac{\mathrm{d}D_z}{\mathrm{d}\xi} = E_y, \tag{5.2.44b}$$

$$\frac{\mathrm{d}B_z}{\mathrm{d}\xi} = -\lambda_1^{\pm}E_z, \tag{5.2.44c}$$

$$\frac{\mathrm{d}B_y}{\mathrm{d}\xi} = -\lambda_1^{\pm}E_y. \tag{5.2.44d}$$

Combining (5.2.44a) and (5.2.44c) gives

$$\frac{\mathrm{d}B_z}{\mathrm{d}\xi} = \lambda_1^{\pm}\frac{\mathrm{d}D_y}{\mathrm{d}\xi}.$$

After integration we obtain

$$B_z = \int \lambda_1^{\pm}(|E(\xi)|)D_y'(\xi)\,\mathrm{d}\xi.$$

It is $D_y' = (\mathcal{E}'E_y)' = (\mathcal{E}')'E_y + \mathcal{E}'E_y'$, and furthermore it holds

$$(\sqrt{\mathcal{E}'})' = \frac{1}{2}\frac{1}{\sqrt{\mathcal{E}'}}(\mathcal{E}')'.$$

Recalling $\lambda_1^{\pm} = \frac{1}{\sqrt{\mathcal{E}'}}$ we thus get

$$\pm B_z = \int \frac{1}{\sqrt{\mathcal{E}'}}((\mathcal{E}')'E_y + \mathcal{E}'E_y')\,\mathrm{d}\xi = \int \frac{1}{\sqrt{\mathcal{E}'}}(\mathcal{E}')'E_y\,\mathrm{d}\xi + \int \mathcal{E}'E_y'\,\mathrm{d}\xi.$$

Partial integration of the first term (and letting $y' := \frac{1}{\sqrt{\mathcal{E}'}}(\mathcal{E}')'$ and $x := E_y$) gives

$$\sqrt{\mathcal{E}'}E_y - \int \sqrt{\mathcal{E}'}E_y'\,\mathrm{d}\xi + \int \mathcal{E}'E_y'\,\mathrm{d}\xi = \sqrt{\mathcal{E}'}E_y.$$

Thus, we get the first component of the Riemann invariant $\mathbf{R}_2^{\pm}$ as

$$\sqrt{\mathcal{E}'}E_y \mp B_z = \text{const} =: R_{2,1}^{\pm}(\mathbf{E}, \mathbf{B}).$$

In an analogous manner we get by combining equations (5.2.44b) with (5.2.44d) the second component of $\mathbf{R}_2^{\pm}$:

$$\sqrt{\mathcal{E}'}E_z \pm B_y = \text{const} =: R_{2,2}^{\pm}(\mathbf{E}, \mathbf{B}).$$

The constants can be determined by choosing an arbitrary point $(\mathbf{E}^*, \mathbf{B}^*)$.
Next, combination of (5.2.44a) with (5.2.44b) gives

$$\begin{pmatrix} D_y' \\ D_z' \end{pmatrix} = \begin{pmatrix} -E_z \\ E_y \end{pmatrix},$$

that is, $\mathbf{D}' = \tilde{\mathbf{E}}$, where we defined $\tilde{\mathbf{E}}$ as $(-E_z, E_y)^T$. Recalling the relation $\partial_{\mathbf{E}}\mathbf{D} = J^{-1}\partial_{\mathbf{D}}\mathbf{E}$ we obtain

$$J^{-1}\mathbf{E}' = \mathbf{E}_{\perp}.$$

Multiplying with $\mathbf{E}$ renders

$$J^{-1}(\mathbf{E}' \cdot \mathbf{E}) = \mathbf{E}_{\perp} \cdot \mathbf{E} = 0 \quad \Longleftrightarrow \quad (|\mathbf{E}|^2)' = 0.$$

And thus we obtain the Riemann invariant $\mathbf{R}_1^{\pm}$.
Now we come to the Riemann invariants corresponding to $\lambda_2^{\pm}$. We need to solve

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\xi} = \mathbf{v}_2^{\pm}.$$

Written out, this reads:

$$\frac{\mathrm{d}D_y}{\mathrm{d}\xi} = E_y, \tag{5.2.45a}$$

$$\frac{\mathrm{d}D_z}{\mathrm{d}\xi} = E_z, \tag{5.2.45b}$$

$$\frac{\mathrm{d}B_z}{\mathrm{d}\xi} = \lambda_2^{\pm}E_y, \tag{5.2.45c}$$

$$\frac{\mathrm{d}B_y}{\mathrm{d}\xi} = -\lambda_2^{\pm}E_z. \tag{5.2.45d}$$

We combine (5.2.45a) and (5.2.45b) and get $\mathbf{D}' = \mathbf{E}'$, or, equivalently,

$$J^{-1}\mathbf{E}' = \mathbf{E} \quad | \cdot \mathbf{E}_{\perp} \quad \Longleftrightarrow \quad \mathbf{E}' \cdot \mathbf{E}_{\perp} = 0. \tag{5.2.46}$$

From this follows that $\mathbf{E}$ has to be such that

$$\mathbf{E}' = a(\mathbf{E})\mathbf{E}' + b(\mathbf{E})\mathbf{E}, \tag{5.2.47}$$

i.e. $\mathbf{E}$ is a product of two functions $f, g$ with $\mathbf{E} = fg$. $f$ and $g$ remain to be determined. Applying the product rule we get

$$\mathbf{E}' = f'g + fg'. \tag{5.2.48}$$

Comparing both expressions (5.2.47) and (5.2.48) with each other, we see that $f' = \mathbf{E}'$ and $g = a(\mathbf{E})$ and $g' = b(\mathbf{E})$. Furthermore, it has to hold:

$$\nabla_{\mathbf{u}} \mathbf{R}_3^{\pm} \cdot \mathbf{v}_2^{\pm} = 0.$$

Recalling $\nabla_{\mathbf{u}} = (J^{-1}\nabla_{\mathbf{E}}, \nabla_{\mathbf{B}})$ and $(\mathbf{v}_{2,1}^{\pm}, \mathbf{v}_{2,2}^{\pm})^T = \mathbf{E}$ we obtain

$$J^{-1}\nabla_{\mathbf{E}} R_{3,1}^{\pm} \cdot \mathbf{E} = 0. \tag{5.2.49}$$

We get $\mathbf{R}_3^{\pm}$ by integrating (5.2.46). Also we assume that $\mathbf{R}_3^{\pm}$ has the form $\mathbf{E} = (fg) = \mathbf{E}\, g(\mathbf{E})$, i.e.

$$\mathbf{R}_3^{\pm} = \mathbf{E}\, g(\mathbf{E}) \qquad \Longrightarrow \qquad \nabla_{\mathbf{E}} R_3^{\pm} = g(\mathbf{E})(1,1)^T + \mathbf{E}\, \nabla_{\mathbf{E}} \cdot g(\mathbf{E}).$$

Plugging this into (5.2.49) gives

$$\begin{aligned} & \Big(g(\mathbf{E})(1,1)^T + \mathbf{E}\nabla_{\mathbf{E}} \cdot g(\mathbf{E})\Big) \cdot \mathbf{E} = 0 \\ \Leftrightarrow \quad & g(\mathbf{E})(\mathbf{E} \cdot (1,1)^T) + \nabla_{\mathbf{E}} \cdot g(\mathbf{E})|\mathbf{E}|^2 = 0. \end{aligned} \tag{5.2.50}$$

Here, by $\nabla_{\mathbf{E}}\mathbf{E}$ the divergence is meant, i.e.

$$\nabla_{\mathbf{E}} \cdot \mathbf{E} = \partial_{E_y} E_y + \partial_{E_z} E_z.$$

For a function $h = h(\mathbf{E})$ it is

$$\nabla_{\mathbf{E}} \cdot h(\mathbf{E}) = \partial_{E_y} h(\mathbf{E}) + \partial_{E_z} h(\mathbf{E}).$$

Continuing, from (5.2.49) we see

$$\nabla_{\mathbf{E}} \cdot g(\mathbf{E}) = -\frac{\mathbf{E} \cdot (1,1)^T}{|\mathbf{E}|^2} g(\mathbf{E}),$$

and after integration this becomes

$$g(\mathbf{E}) = -\int \frac{\mathbf{E} \cdot (1,1)^T}{|\mathbf{E}|^2}\, \mathrm{d}\mathbf{E} = \frac{1}{|\mathbf{E}|},$$

due to the fact that

$$\nabla_{\mathbf{E}} \cdot \frac{1}{|\mathbf{E}|} = \partial_{E_y} \frac{1}{|\mathbf{E}|} + \partial_{E_z} \frac{1}{|\mathbf{E}|} = -\frac{\mathbf{E} \cdot (1,1)^T}{|\mathbf{E}|^2}.$$

Thus, $\mathbf{R}_3^{\pm} = \mathbf{E}\, g(\mathbf{E}) = \frac{\mathbf{E}}{|\mathbf{E}|}$.
For the Riemann invariant $\mathbf{R}_4^{\pm}$ we prove that definition (5.2.40) holds. Recall again $\nabla_{\mathbf{u}} = (J^{-1}\nabla_{\mathbf{E}}, \nabla_{\mathbf{B}})$. We need to show

$$\nabla_{\mathbf{u}} \mathbf{R}_4^{\pm} \cdot \mathbf{v}_2^{\pm} = \mathbf{0} \in \mathbb{R}^4.$$

It holds

$$\nabla_{\mathbf{u}} \mathbf{R}_4^{\pm} = \nabla_{\mathbf{u}} \begin{pmatrix} R_{4,1}^{\pm} \\ R_{4,2}^{\pm} \end{pmatrix} = \left( J^{-1} \begin{pmatrix} \partial_{\mathbf{E}} R_{4,1}^{\pm} \\ \partial_{\mathbf{E}} R_{4,2}^{\pm} \end{pmatrix} \begin{pmatrix} \partial_{\mathbf{B}} R_{4,1}^{\pm} \\ \partial_{\mathbf{B}} R_{4,2}^{\pm} \end{pmatrix} \right).$$

The meaning of this may be confusing, so we write out the components:

$$\left( \begin{pmatrix} \partial_E R_{4,1}^{\pm} \\ \partial_E R_{4,2}^{\pm} \end{pmatrix} \begin{pmatrix} \partial_B R_{4,1}^{\pm} \\ \partial_B R_{4,2}^{\pm} \end{pmatrix} \right) = \begin{pmatrix} \partial_E R_{4,1}^{\pm} & \partial_B R_{4,1}^{\pm} \\ \partial_E R_{4,2}^{\pm} & \partial_B R_{4,2}^{\pm} \end{pmatrix} = \begin{pmatrix} \partial_{E_y} R_{4,1}^{\pm} & \partial_{E_z} R_{4,1}^{\pm} & \partial_{B_z} R_{4,1}^{\pm} & \partial_{(-B_y)} R_{4,1}^{\pm} \\ \partial_{E_y} R_{4,2}^{\pm} & \partial_{E_z} R_{4,2}^{\pm} & \partial_{B_z} R_{4,2}^{\pm} & \partial_{(-B_y)} R_{4,2}^{\pm} \end{pmatrix}.$$

It is

$$\partial_{B_z} R_{4,1}^{\pm} = 0,$$

$$\partial_{(-B_y)} R_{4,1}^{\pm} = \mp 1,$$

$$\partial_{B_z} R_{4,2}^{\pm} = \mp 1,$$

$$\partial_{(-B_y)} R_{4,2}^{\pm} = 0,$$

$$\partial_{E_y} R_{4,1}^{\pm} = G' \frac{E_y^2}{|\mathbf{E}|^2} + \frac{G}{|\mathbf{E}|}(1 - \frac{E_y^2}{|\mathbf{E}|^2}),$$

$$\partial_{E_z} R_{4,1}^{\pm} = G' \frac{E_y E_z}{|\mathbf{E}|^2} - \frac{G}{|\mathbf{E}|^3} E_y E_z,$$

$$\partial_{E_y} R_{4,2}^{\pm} = G' \frac{E_y E_z}{|\mathbf{E}|^2} - \frac{G}{|\mathbf{E}|^3} E_y E_z,$$

$$\partial_{E_z} R_{4,2}^{\pm} = G' \frac{E_z^2}{|\mathbf{E}|^2} + \frac{G}{|\mathbf{E}|}(1 - \frac{E_z^2}{|\mathbf{E}|^2}).$$

So it follows

$$\begin{aligned}
\begin{pmatrix} \partial_{\mathbf{E}} R_{4,1}^{\pm} \\ \partial_{\mathbf{E}} R_{4,2}^{\pm} \end{pmatrix} &= \begin{pmatrix} G' \frac{E_y^2}{|\mathbf{E}|^2} + \frac{G}{|\mathbf{E}|}(1 - \frac{E_y^2}{|\mathbf{E}|^2}) & G' \frac{E_y E_z}{|\mathbf{E}|^2} - \frac{G}{|\mathbf{E}|^3} E_y E_z \\ G' \frac{E_y E_z}{|\mathbf{E}|^2} - \frac{G}{|\mathbf{E}|^3} E_y E_z & G' \frac{E_z^2}{|\mathbf{E}|^2} + \frac{G}{|\mathbf{E}|}(1 - \frac{E_z^2}{|\mathbf{E}|^2}) \end{pmatrix} \\
&= \frac{G'}{|\mathbf{E}|^2} \begin{pmatrix} E_y^2 & E_y E_z \\ E_y E_z & E_z^2 \end{pmatrix} - \frac{G}{|\mathbf{E}|^3} \begin{pmatrix} E_y^2 & E_y E_z \\ E_y E_z & E_z^2 \end{pmatrix} + \frac{G}{|\mathbf{E}|} \operatorname{Id}.
\end{aligned}$$

Here, Id is the identity matrix. Recall that

$$\mathbf{E} \otimes \mathbf{E}^T = \begin{pmatrix} E_y^2 & E_y E_z \\ E_y E_z & E_z^2 \end{pmatrix}$$

was the Kronecker product as defined in 5.1. Thus we obtain

$$\begin{pmatrix} \partial_{\mathbf{E}} R_{4,1}^{\pm} \\ \partial_{\mathbf{E}} R_{4,2}^{\pm} \end{pmatrix} = \frac{G'}{|\mathbf{E}|^2} \mathbf{E} \otimes \mathbf{E}^T - \frac{G}{|\mathbf{E}|^3} \mathbf{E} \otimes \mathbf{E}^T + \frac{G}{|\mathbf{E}|} \operatorname{Id}.$$

At last, it is

$$\begin{aligned}
\nabla_{\mathbf{u}} \begin{pmatrix} R_{4,1}^{\pm} \\ R_{4,2}^{\pm} \end{pmatrix} \cdot \mathbf{v}_2^{\pm} &= \left( J^{-1} \begin{pmatrix} \partial_{\mathbf{E}} R_{4,1}^{\pm} \\ \partial_{\mathbf{E}} R_{4,2}^{\pm} \end{pmatrix} \begin{pmatrix} \partial_{\mathbf{B}} R_{4,1}^{\pm} \\ \partial_{\mathbf{B}} R_{4,2}^{\pm} \end{pmatrix} \right) \cdot \mathbf{v}_2^{\pm} \\
&= \left( J^{-1} \left( \tfrac{G'}{|\mathbf{E}|^2} \mathbf{E} \otimes \mathbf{E}^T - \tfrac{G}{|\mathbf{E}|^3} \mathbf{E} \otimes \mathbf{E}^T + \tfrac{G}{|\mathbf{E}|} \operatorname{Id} \right) \;\middle|\; \mp \operatorname{Id} \right) \cdot \begin{pmatrix} \mathbf{E} \\ \lambda_2^{\pm} \mathbf{E} \end{pmatrix} \\
&= J^{-1} \left[ \left( \frac{G'}{|\mathbf{E}|^2} \mathbf{E} \otimes \mathbf{E}^T - \frac{G}{|\mathbf{E}|^3} \mathbf{E} \otimes \mathbf{E}^T + \frac{G}{|\mathbf{E}|} \operatorname{Id} \right) \mathbf{E} \right] - \lambda_2^{\pm} \mathbf{E} \\
&= J^{-1} \left( \frac{G'}{|\mathbf{E}|^2} (\mathbf{E} \otimes \mathbf{E}^T) \mathbf{E} - \frac{G}{|\mathbf{E}|^3} (\mathbf{E} \otimes \mathbf{E}^T) \mathbf{E} + \frac{G}{|\mathbf{E}|} \operatorname{Id} \mathbf{E} \right) - \lambda_2^{\pm} \mathbf{E}.
\end{aligned}$$

Observing that

$$(\mathbf{E}\otimes\mathbf{E}^T)\mathbf{E} = \begin{pmatrix} E_y^3 + E_y E_z^2 \\ E_z^3 + E_y^2 E_z \end{pmatrix} = \begin{pmatrix} E_y(E_y^2+E_z^2) \\ E_z(E_y^2+E_z^2) \end{pmatrix} = |E|^2\mathbf{E},$$

it follows

$$\begin{aligned} &J^{-1}\left(\frac{G'}{|\mathbf{E}|^2}(\mathbf{E}\otimes\mathbf{E}^T)\mathbf{E} - \frac{G}{|\mathbf{E}|^3}(\mathbf{E}\otimes\mathbf{E}^T)\mathbf{E} + \frac{G}{|\mathbf{E}|}\mathbf{E}\right) - \lambda_2^{\pm}\mathbf{E} \\ &= J^{-1}\left(\frac{G'}{|\mathbf{E}|^2}|E|^2\mathbf{E} - \frac{G}{|\mathbf{E}|^3}|E|^2\mathbf{E} + \frac{G}{|\mathbf{E}|}\mathbf{E}\right) - \lambda_2^{\pm}\mathbf{E} \\ &= G'J^{-1}\mathbf{E} - \lambda_2^{\pm}\mathbf{E} \\ &= G'(\lambda_1^{\pm})^2\mathbf{E} - \lambda_2^{\pm}\mathbf{E} \\ &= \lambda_2^{\pm}\mathbf{E} - \lambda_2^{\pm}\mathbf{E} \\ &= \mathbf{0}. \end{aligned}$$

□

### Geometrical Illustration

Recall that the Riemann invariants given in Lemma 5.6 are constant, i.e. $\mathbf{R}_i^{\pm}(\mathbf{u}) = \mathbf{R}_i^{\pm}(\mathbf{u}^*) =: (\mathbf{R}_i^{\pm})^*$ $(i = 1,\dots,4)$, where $\mathbf{u}^*$ is an arbitrary point. So we choose $\mathbf{u}^* = (\mathbf{E}^*, \mathbf{B}^*)$ and obtain:

$$\begin{aligned} &(R_{1,1}^{\pm})^* = (R_{1,2}^{\pm})^* := R_{1,1}^{\pm}(\mathbf{E}^*,\mathbf{B}^*) = |\mathbf{E}^*|^2 \equiv |\mathbf{E}|^2, \\ &(\mathbf{R}_2^{\pm})^* := \mathbf{R}_2^{\pm}(\mathbf{E}^*,\mathbf{B}^*) = \frac{\mathbf{E}^*}{|\mathbf{E}^*|^2} \equiv \frac{\mathbf{E}}{|\mathbf{E}|^2}, \\ &(\mathbf{R}_3^{\pm})^* := \mathbf{R}_3^{\pm}(\mathbf{E}^*,\mathbf{B}^*) = \sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}^*|^2}{2}\right)}\mathbf{E}^* \mp \mathbf{B}^* \equiv \sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}|^2}{2}\right)}\mathbf{E} \mp \mathbf{B}, \\ &(\mathbf{R}_4^{\pm})^* := \mathbf{R}_4^{\pm}(\mathbf{E}^*,\mathbf{B}^*) = G(|\mathbf{E}^*|)\frac{\mathbf{E}^*}{|\mathbf{E}^*|^2} \mp \mathbf{B}^* \equiv G(|\mathbf{E}|)\frac{\mathbf{E}}{|\mathbf{E}|^2} \mp \mathbf{B}. \end{aligned} \tag{5.2.51}$$

$(\mathbf{E}^*,\mathbf{B}^*)$ is a given point, as well as $(\mathbf{R}_i^{\pm})^*$ $(i = 1,\dots,4)$ are known constants. $\mathbf{E}$ and $\mathbf{B}$ are unknown. To visualize the Riemann invariants geometrically, one chooses either $\mathbf{E}$ or $\mathbf{B}$ as a parameter and then solves the equations (5.2.51) for the remaining unknown variables. Since it is easy to solve for $\mathbf{B}$, we treat $E_y$ as an "inner" parameter and $E_z$ as an "outer" one, i.e. we let $E_y$ vary and fix one $E_z$, then we choose another $E_z$ and vary $E_y$ again, and so forth. Of course, other choices would be possible.
We only consider $\mathbf{R}_2^{\pm}$ and $\mathbf{R}_4^{\pm}$ for visualization. The reason is that later these are used to give explicit formulas of the waves of the Riemann problem. Thus we have the following equations:

For $\mathbf{R}_2^{\pm}$ :

$$\begin{aligned} B_{y,2}^{\pm} &= \pm((R_{2,1}^{\pm})^* - \sqrt{\mathcal{E}'}E_z), \\ B_{z,2}^{\pm} &= \mp((R_{2,2}^{\pm})^* - \sqrt{\mathcal{E}'}E_y). \end{aligned}$$

For $\mathbf{R}_4^{\pm}$ :

$$\begin{aligned} B_{y,4}^{\pm} &= \pm\left((R_{4,1}^{\pm})^* - G(|\mathbf{E}|)\frac{E_z}{|\mathbf{E}|}\right), \\ B_{z,4}^{\pm} &= \mp\left((R_{4,2}^{\pm})^* - G(|\mathbf{E}|)\frac{E_y}{|\mathbf{E}|}\right). \end{aligned}$$

We observe that $B^+_{y,3}$ and $B^-_{y,3}$ intersect in $\mathbf{E}^*$ with $B^+_{y,3}(\mathbf{E}^*) = B^-_{y,3}(\mathbf{E}^*) = B^*_y$. The same holds for $B^+_{z,3}$ and $B^-_{z,3}$, $B^+_{y,4}$ and $B^-_{y,4}$, $B^+_{z,4}$ and $B^-_{z,4}$. The plots in figures 5.12 to 5.15 show that these are not the only intersection points. The Riemann invariants in the $E_y - B_z$-plane (left plots) have an intersection point for all $E_z \in [-3, 3]$, which moves from the right to the left in $E_y$-direction for changing values of $E_z$. On the other hand, the Riemann invariants in the $E_y - B_y$-plane (right plots) wander up and down in $B_y$-direction for changing values of $E_z$ with $E_z \in [-3, 3]$ and do only intersect for certain values of $E_z$. In those cases, whenever $E_z$ is such that the plus and minus branches of $B_y$ intersect, as in figure 5.14, $(\mathbf{E}^*, \mathbf{B}^*)^T$ can be connected to another point $(\mathbf{E}, \mathbf{B})^T$ on the Riemann invariant by a rarefaction wave.

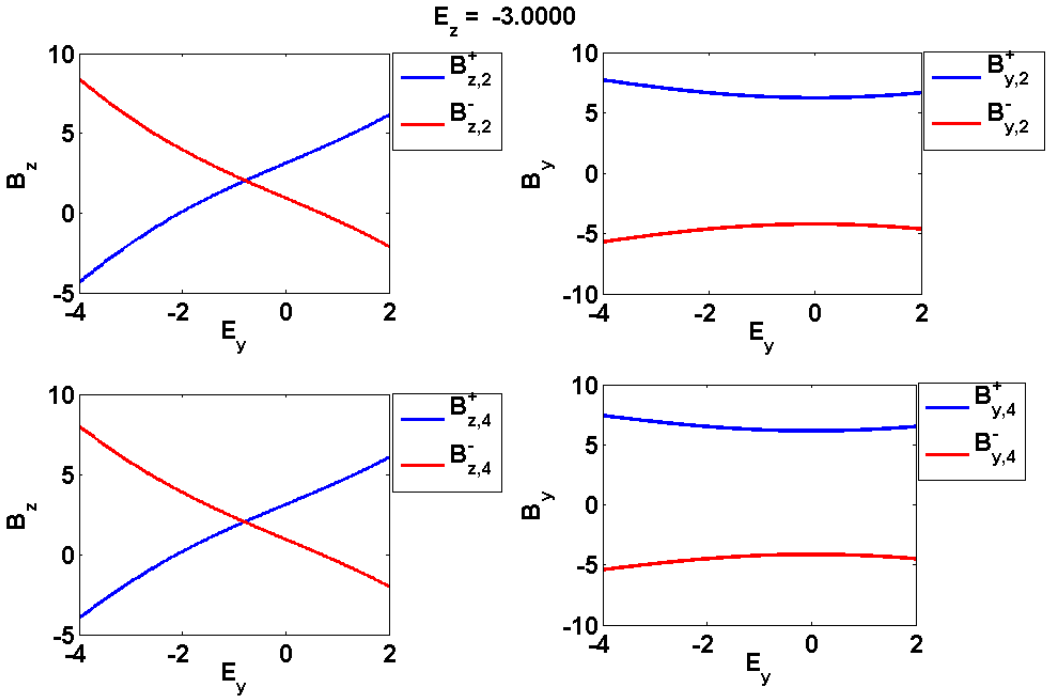



Figure 5.12: Plots of the magnetic field $\mathbf{B}$ for the Riemann invariants $\mathbf{R}_2^{\pm}$ and $\mathbf{R}_4^{\pm}$ at $E_z = -3$, where $E_y \in [-4, 2]$ and $E_z \in [-3, 3]$. The fixed chosen point is $(\mathbf{E}^*, \mathbf{B}^*)^T = (-1, 1, 2, 1)^T$.

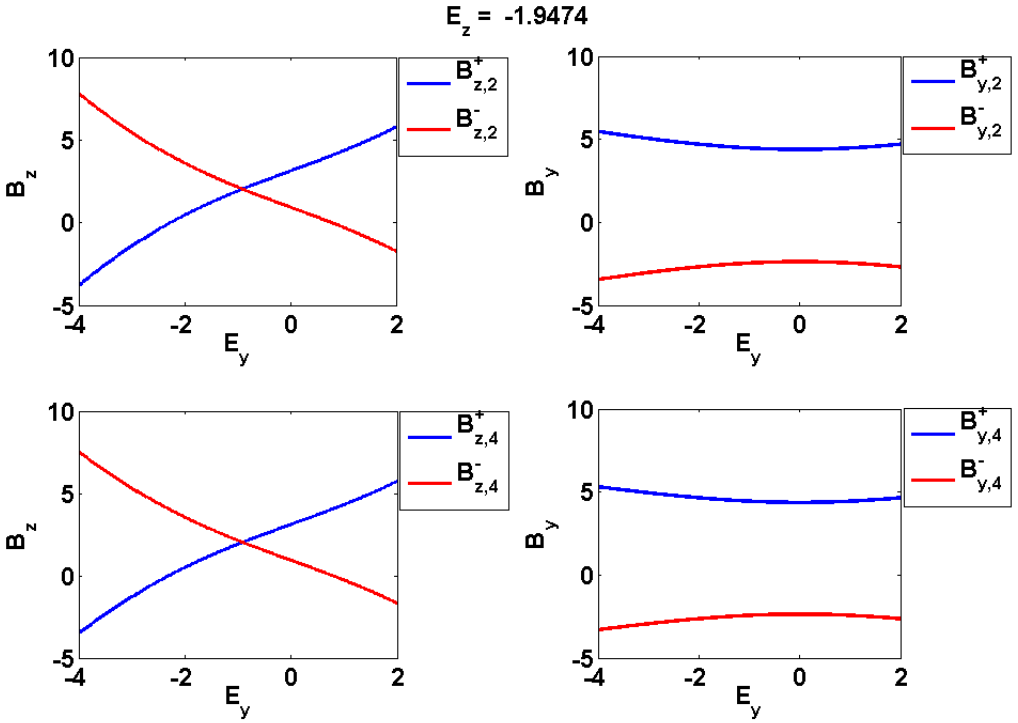



Figure 5.13: Plots of $\mathbf{B}$ at $E_z = -3$.

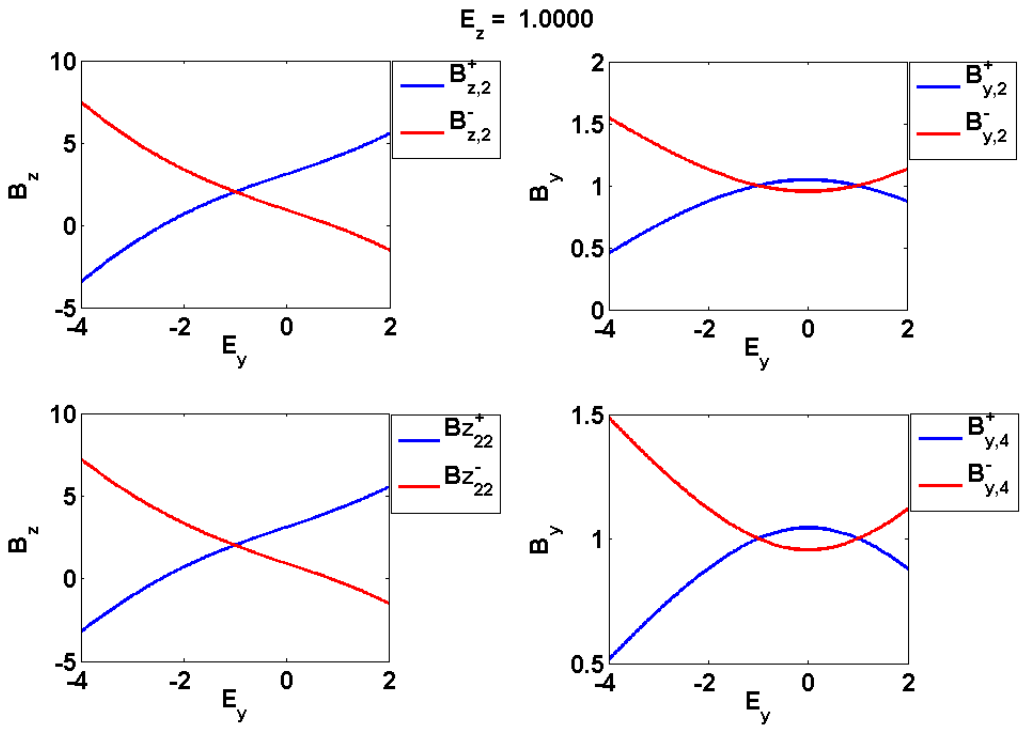



Figure 5.14: Plots of $\mathbf{B}$ at $E_z = 1 = E_z^*$.

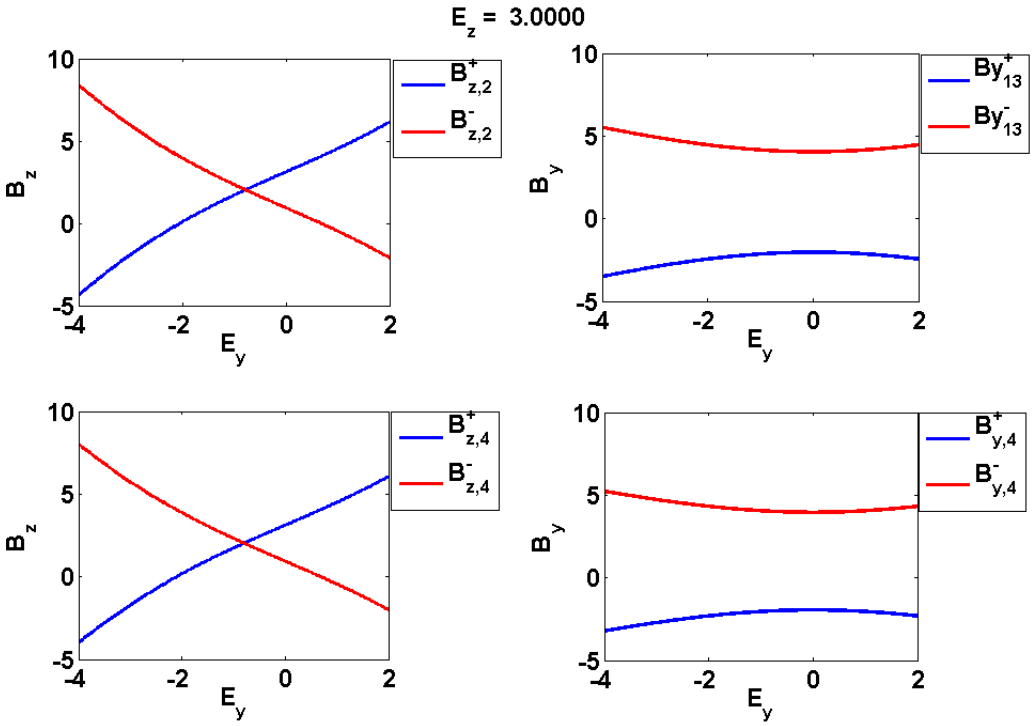



Figure 5.15: Plots of $\mathbf{B}$ at $E_z = 3$.

### 5.2.4 Analytical Solution of the Riemann Problem

As we already pointed out at the beginning of this section, the solution of the Kerr-Riemann problem consists of the following simple waves (see [71]):

(1) For the 2-field, which is genuinely nonlinear: We either have a 2-shock, which can be determined via the Rankine-Hugoniot condition, or a 2-rarefaction which can be obtained by using the 2-Riemann invariants.

(2) For the 1-field, which is linearly degenerate: We have a contact discontinuity which can be determined via the Rankine-Hugoniot condition or via the 1-Riemann invariants.

Due to the fact that $\lambda_1^- \leq \lambda_2^- \leq \lambda_2^+ \leq \lambda_1^+$, the solution of the Riemann problem looks as illustrated in figure 5.16. The different regions between the single waves are denoted, from left to right, by "L", "1", "2", "3" and "R". Table 5.2.4 shows an overview of the wave composition of the Kerr-nonlinear Riemann problem with the corresponding wave speeds. We will go into details in the following subsections.
The author in [71] proves uniqueness of this solution by requiring positivity of entropy $\eta > 0$ and the condition of Smoller-Johnson, see sections 3.3.2 and 5.2.2. Seccia [70] chooses another condition to select admissible shocks, namely a reflection and transmission condition alongside the entropy condition.

#### Contact Discontinuities

The simple wave corresponding to the 1-field with eigenvalues $\lambda_1^{\pm}$ is a contact discontinuity which can be determined by applying the Rankine-Hugoniot jump condition or the 1-Riemann invariants $\mathbf{R}_1^{\pm}, \mathbf{R}_2^{\pm}$ from section 5.6, which are constant across the discontinuity.

| | eigen-value | determined via | wave speed | Riemann invariants |
|---|---|---|---|---|
| left 1-discontinuity | $\lambda_1^-$ | Rankine-Hugoniot Riemann invariants | $s_L = \lambda_1^-(\mathbf{E}_L)$ | $\mathbf{R}_1^-, \mathbf{R}_2^-$ |
| left 2-shock | $\lambda_2^-$ | Rankine-Hugoniot | $\lambda_2^-(\mathbf{E}_1) \leq s_1 \leq \lambda_2^-(\mathbf{E}_2)$ $s_1 = \pm\left[\frac{a_1-a_2}{\mathcal{E}'(a_1^2/2)a_1-\mathcal{E}'(a_2^2/2)a_2}\right]$ | $\mathbf{R}_3^-, \mathbf{R}_4^-$ |
| left 2-rarefaction | $\lambda_2^-$ | Riemann invariants | $\lvert\mathbf{E}_1\rvert > \lvert\mathbf{E}_2\rvert$ head: $x_L^{\text{head}} = x - \lambda_2^-(\lvert\mathbf{E}_1\rvert)t$ tail: $x_L^{\text{tail}} = x - \lambda_2^-(\lvert\mathbf{E}_2\rvert)t$ | $\mathbf{R}_3^-, \mathbf{R}_4^-$ |
| right 2-rarefaction | $\lambda_2^+$ | Riemann invariants | $\lvert\mathbf{E}_2\rvert > \lvert\mathbf{E}_3\rvert$ head: $x_R^{\text{head}} = x - \lambda_2^+(\lvert\mathbf{E}_3\rvert)t$ tail: $x_R^{\text{tail}} = x - \lambda_2^+(\lvert\mathbf{E}_2\rvert)t$ | $\mathbf{R}_3^-, \mathbf{R}_4^-$ |
| right 2-shock | $\lambda_2^+$ | Rankine-Hugoniot | $\lambda_2^+(\mathbf{E}_2) \leq s_2 \leq \lambda_2^+(\mathbf{E}_3)$ $s_2 = \pm\left[\frac{a_2-a_3}{\mathcal{E}'(a_2^2/2)a_2-\mathcal{E}'(a_3^2/2)a_3}\right]$ | $\mathbf{R}_3^-, \mathbf{R}_4^-$ |
| right 1-discontinuity | $\lambda_1^+$ | Rankine-Hugoniot Riemann invariants | $s_R = \lambda_1^+(\mathbf{E}_R) = -s_L$ | $\mathbf{R}_1^-, \mathbf{R}_2^-$ |

Table 5.2: Overview of the wave composition of the Kerr-nonlinear Riemann problem.

The shock speed of the left contact discontinuity is $s_L = \lambda_1^-(|\mathbf{E}_L|)$ and of the right contact is $s_R = \lambda_1^+(|\mathbf{E}_R|)$; it holds $s_L = -s_R$. From $\mathbf{R}_1^\pm$ we see that the modulus of the electric field is constant across the discontinuity, i.e. $|\mathbf{E}| = |\mathbf{E}_L| = |\mathbf{E}_1|$ and $|\mathbf{E}| = |\mathbf{E}_3| = |\mathbf{E}_R|$. For the following we denote the direction of $\mathbf{E}$ as $\mathbf{e} := \mathbf{E}/|\mathbf{E}|$. Thus altogether, it is

$$\mathbf{e}_L = \frac{\mathbf{E}_L}{|\mathbf{E}_L|} = \frac{\mathbf{E}_L}{|\mathbf{E}_R|}, \tag{5.2.52}$$

$$\mathbf{e}_R = \frac{\mathbf{E}_R}{|\mathbf{E}_R|} = \frac{\mathbf{E}_R}{|\mathbf{E}_L|}, \tag{5.2.53}$$

$$\mathbf{e}_1 = \frac{\mathbf{E}_1}{|\mathbf{E}_1|} = \frac{\mathbf{E}_1}{|\mathbf{E}_L|} = \frac{\mathbf{E}_1}{|\mathbf{E}_R|} = \frac{\mathbf{E}_1}{|\mathbf{E}_3|}, \tag{5.2.54}$$

$$\mathbf{e}_2 = \frac{\mathbf{E}_2}{|\mathbf{E}_2|}, \tag{5.2.55}$$

$$\mathbf{e}_3 = \frac{\mathbf{E}_3}{|\mathbf{E}_3|} = \frac{\mathbf{E}_3}{|\mathbf{E}_R|} = \frac{\mathbf{E}_3}{|\mathbf{E}_L|} = \frac{\mathbf{E}_3}{|\mathbf{E}_1|}. \tag{5.2.56}$$

From $\mathbf{R}_2^\pm$ we get

$$\begin{aligned}
&\text{for } \lambda_1^- : \quad \mathbf{B}_1 - \mathbf{B}_L = -\sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_L|^2}{2}\right)}(\mathbf{E}_1 - \mathbf{E}_L) = -\sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_L|^2}{2}\right)}|\mathbf{E}_L|(\mathbf{e}_1 - \mathbf{e}_L), \\
&\text{for } \lambda_1^+ : \quad \mathbf{B}_R - \mathbf{B}_3 = \sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_R|^2}{2}\right)}|\mathbf{E}_R|(\mathbf{e}_R - \mathbf{e}_3).
\end{aligned} \tag{5.2.57}$$

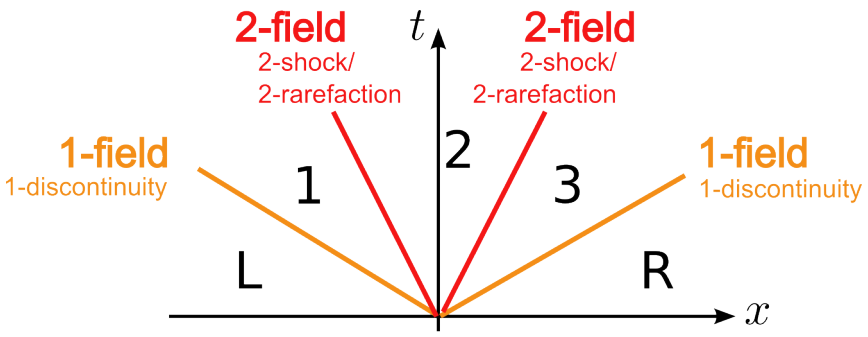



Figure 5.16: A sketch of the wave composition of the Kerr-nonlinear Riemann problem. Locally, the characteristics are straight lines, globally they are winded curves due to the nonlinear nature.

One could have used the Rankine-Hugoniot condition, too, with

$$\begin{aligned} s_L &= \lambda_1^-(|\mathbf{E}_L|) = \lambda_1^-(|\mathbf{E}_1|) = \lambda_1^-(|\mathbf{E}_3|) = \lambda_1^-(|\mathbf{E}_R|) \quad \text{for the left 1-contact,} \\ s_R &= \lambda_1^+(|\mathbf{E}_R|) = \lambda_1^+(|\mathbf{E}_3|) = \lambda_1^+(|\mathbf{E}_1|) = \lambda_1^+(|\mathbf{E}_L|) \quad \text{for the right 1-contact,} \end{aligned}$$

i.e. $s_L = -s_R$.

### Rarefaction Waves

The simple waves corresponding to the 2-field with eigenvalues $\lambda_2^\pm$ can be rarefaction waves. From $\mathbf{R}_3^\pm$ we see that

$$\begin{aligned} \mathbf{e}_2 &= \frac{\mathbf{E}_2}{|\mathbf{E}_2|} = \frac{\mathbf{E}_1}{|\mathbf{E}_1|} = \mathbf{e}_1, \\ \mathbf{e}_2 &= \frac{\mathbf{E}_2}{|\mathbf{E}_2|} = \frac{\mathbf{E}_3}{|\mathbf{E}_3|} = \mathbf{e}_3. \end{aligned} \tag{5.2.58}$$

Using $\mathbf{R}_4^\pm$ we obtain

$$\begin{aligned} \text{for } \lambda_2^- : \quad & \mathbf{B}_2 - \mathbf{B}_1 = \mathbf{e}_2 \left( G(|\mathbf{E}_2|) - G(|\mathbf{E}_L|) \right), \\ \text{for } \lambda_2^+ : \quad & \mathbf{B}_3 - \mathbf{B}_2 = \mathbf{e}_2 \left( G(|\mathbf{E}_2|) - G(|\mathbf{E}_R|) \right), \end{aligned} \tag{5.2.59}$$

where we recall from the previous subsection that $|\mathbf{E}_L| = |\mathbf{E}_1|$ and $|\mathbf{E}_R| = |\mathbf{E}_3|$.

### Admissible Shocks

Besides rarefaction waves, the simple waves corresponding to the 2-field can also be shocks. The author in [71] shows that by applying the Rankine-Hugoniot conditions (5.2.23) and some rearrangement in terms we can differentiate the following cases:

Case 1: If $|\mathbf{E}_L| = |\mathbf{E}_R|$ and $\mathbf{e}_L = \mathbf{e}_R$, then there is no shock.

Case 2: If $|\mathbf{E}_L| = |\mathbf{E}_R|$ and $\mathbf{e}_L \neq \mathbf{e}_R$, the speed of the shock is given as $s = \lambda_1^\pm$, so we have a contact discontinuity.

Case 3: If $|\mathbf{E}_L| \neq |\mathbf{E}_R|$, then $\mathbf{E}_R$ and $\mathbf{E}_L$ are collinear, i.e. its directions are equal, $\mathbf{e}_L = \mathbf{e}_R =: \mathbf{e}$, and there exist constants $a_L, a_R$ with $\mathbf{E}_L = a_L\mathbf{e}$ and $\mathbf{E}_R = a_R\mathbf{e}$. The condition of Smoller-Johnson gives that $a_L$ and $a_R$ must have the same sign.

Furthermore, if we denote the shock speed of the left shock by $s_1$ and the one of the right shock by $s_2$, the shock speeds have to fulfill

$$\begin{aligned}\lambda_2^-(\mathbf{E}_1) \le s_1 \le \lambda_2^-(\mathbf{E}_2),\\ \lambda_2^+(\mathbf{E}_2) \le s_2 \le \lambda_2^+(\mathbf{E}_3).\end{aligned}$$

Thus the shock speeds are given as

$$\begin{aligned} s_1 &= \begin{cases} +\left[\frac{a_1-a_2}{\mathcal{E}'(a_1^2/2)a_1-\mathcal{E}'(a_2^2/2)a_2}\right]^{\frac{1}{2}}, & \text{if } |\mathbf{E}_1| \le |\mathbf{E}_2|,\\ -\left[\frac{a_1-a_2}{\mathcal{E}'(a_1^2/2)a_1-\mathcal{E}'(a_2^2/2)a_2}\right]^{\frac{1}{2}}, & \text{if } |\mathbf{E}_1| \ge |\mathbf{E}_2|, \end{cases}\\ s_2 &= \begin{cases} +\left[\frac{a_2-a_3}{\mathcal{E}'(a_2^2/2)a_2-\mathcal{E}'(a_3^2/2)a_3}\right]^{\frac{1}{2}}, & \text{if } |\mathbf{E}_2| \le |\mathbf{E}_3|,\\ -\left[\frac{a_2-a_3}{\mathcal{E}'(a_3^2/2)a_1-\mathcal{E}'(a_3^2/2)a_3}\right]^{\frac{1}{2}}, & \text{if } |\mathbf{E}_2| \ge |\mathbf{E}_3|, \end{cases} \end{aligned} \tag{5.2.60}$$

where

$$\begin{aligned}&\mathbf{e}_1 = \mathbf{e}_2 = \mathbf{e}_3 =: \mathbf{e},\\ &\mathbf{E}_1 = a_1\mathbf{e}, \quad \mathbf{E}_2 = a_2\mathbf{e}, \quad \mathbf{E}_3 = a_3\mathbf{e}.\end{aligned} \tag{5.2.61}$$

Furthermore, $a_1$ and $a_2$, $a_2$ and $a_3$ must have the same sign, respectively. Thus, an admissible shock has the following form:

$$\begin{aligned}\text{for } \lambda_2^- : \quad \mathbf{B}_2 - \mathbf{B}_1 &= \mathbf{e}_2 \left[(\mathcal{E}'\left(\frac{|\mathbf{E}_L|^2}{2}\right)|\mathbf{E}_L| - \mathcal{E}'\left(\frac{|\mathbf{E}_L|^2}{2}\right)|\mathbf{E}_2|)(|\mathbf{E}_L| - |\mathbf{E}_2|)\right]^{\frac{1}{2}},\\ \text{for } \lambda_2^+ : \quad \mathbf{B}_3 - \mathbf{B}_2 &= \mathbf{e}_2 \left[(\mathcal{E}'\left(\frac{|\mathbf{E}_R|^2}{2}\right)|\mathbf{E}_R| - \mathcal{E}'\left(\frac{|\mathbf{E}_R|^2}{2}\right)|\mathbf{E}_2|)(|\mathbf{E}_R| - |\mathbf{E}_2|)\right]^{\frac{1}{2}}.\end{aligned} \tag{5.2.62}$$

### The unique Solution

The condition on the shock speed, that is, $\lambda_2^{\pm}(\mathbf{E}_L) \le s \le \lambda_2^{\pm}(\mathbf{E}_R)$, decides whether the 2-field is a 2-shock or a 2-rarefaction. Since the eigenvalues are decreasing functions in $|\mathbf{E}|$, it follows that, if $|\mathbf{u}_L| < |\mathbf{u}_R|$ we have a shock, and if $|\mathbf{u}_L| > |\mathbf{u}_R|$ we have a rarefaction. We collect this and all the informations from the last two subsections in defining a function $f : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ by

$$f(x,y) := \begin{cases} G(y) - G(x), & x > y\\ \left[(\mathcal{E}'(y^2/2)y - \mathcal{E}'(x^2/2)x)(y-x)\right]^{\frac{1}{2}}, & x \le y. \end{cases} \tag{5.2.63}$$

Thus, instead of (5.2.59) and (5.2.62) we can write:

$$\text{for } \lambda_2^- : \quad \mathbf{B}_2 - \mathbf{B}_1 = \mathbf{e}_2\, f(|\mathbf{E}_2|, |\mathbf{E}_L|), \tag{5.2.64}$$

$$\text{for } \lambda_2^+ : \quad \mathbf{B}_3 - \mathbf{B}_2 = \mathbf{e}_2\, f(|\mathbf{E}_2|, |\mathbf{E}_R|). \tag{5.2.65}$$

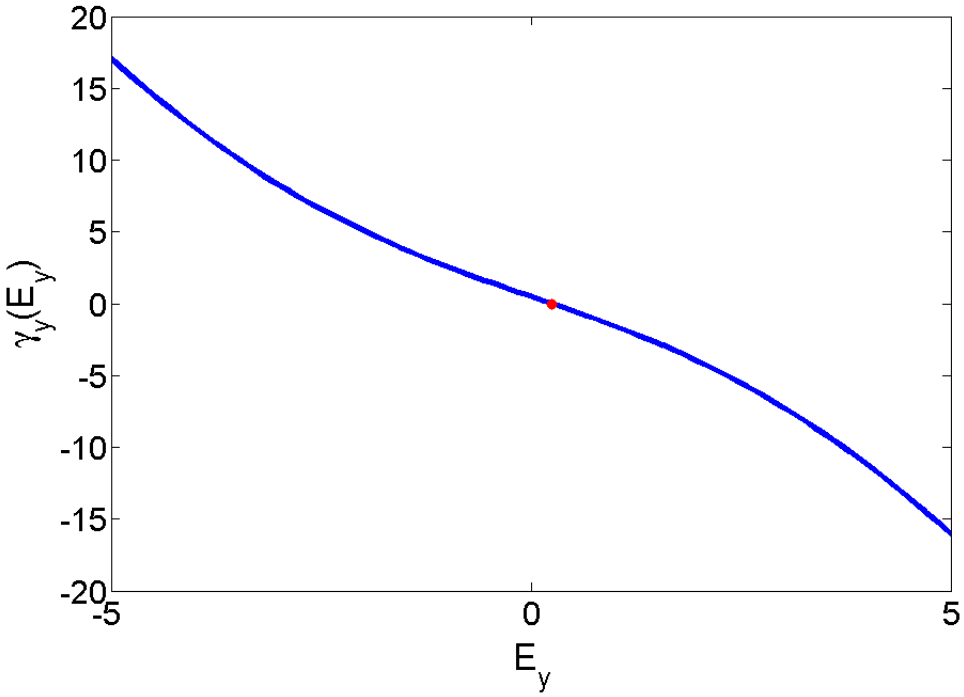



Figure 5.17: The $y$- component of $\gamma$ as defined in (5.2.67), $\gamma_y$, in dependency of $E_y$, where $E_y \in [-5,5]$, $E_z = 0$ and $B_y = 0$. In this case, $\gamma_z \equiv 0$. The – arbitrarily chosen – settings to generate this plot are $E_{L,y} = 0.2$, $E_{R,y} = 0.1$, $E_{L,z} = 0$, $E_{R,z} = 0$, $B_{L,y} = 0$, $B_{R,y} = 0$, $B_{L,z} = -0.1$, $B_{R,z} = 0.1$, $s_R = 1$, $s_L = -1$, $\chi = 0.08$, $\epsilon = 1$. The red dot marks the zero $E_{2,y} \approx 2.496 \cdot 10^{-1}$, where $\gamma_y(E_{2,y}) = 0$.

Combining statements (5.2.57), (5.2.59) and (5.2.62) with each other, we obtain the following relations for the left and right sides of the line $t = 0$:

$$\begin{aligned}
\gamma_L(\mathbf{E}_2; \mathbf{u}_L) &:= \mathbf{B}_L + \sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_L|^2}{2}\right)}\mathbf{E}_L - \mathbf{e}_2 \left(\sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_L|^2}{2}\right)}\,|\mathbf{E}_L| - f(|\mathbf{E}_2|, |\mathbf{E}_L|)\right) = \mathbf{B}_2, \\
\gamma_R(\mathbf{E}_2; \mathbf{u}_R) &:= \mathbf{B}_R - \sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_R|^2}{2}\right)}\mathbf{E}_R + \mathbf{e}_2 \left(\sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_R|^2}{2}\right)}\,|\mathbf{E}_R| - f(|\mathbf{E}_2|, |\mathbf{E}_R|)\right) = \mathbf{B}_2.
\end{aligned} \tag{5.2.66}$$

Thus, $\mathbf{E}_2$ is the zero of the function

$$\gamma(\mathbf{E}_2) := \gamma_L(\mathbf{E}_2; \mathbf{u}_L) - \gamma_R(\mathbf{E}_2; \mathbf{u}_R). \tag{5.2.67}$$

The function $\gamma$ is smooth. As soon as $\mathbf{E}_2$ and $\mathbf{B}_2$ are known, the missing fields $\mathbf{E}_1, \mathbf{B}_1$ and $\mathbf{E}_3, \mathbf{B}_3$ can be determined from equations (5.2.58), (5.2.59), (5.2.61), (5.2.62) and (5.2.66). In [71, Th. 1] it is shown that the above Riemann solution exists and is unique.
As an example, figure 5.17 illustrates the $y$-component of $\gamma$, $\gamma_y$, in dependency of $E_y$, where we have set $E_z = 0$, $B_y = 0$ and $E_y \in [-5,5]$, i.e. $\gamma_z(E_y, E_z) = 0$ for all values of $E_y$ and $E_z = 0$. The – arbitrarily chosen – settings to generate this plot are $E_{L,y} = 0.2$, $E_{R,y} = 0.1$, $E_{L,z} = 0$, $E_{R,z} = 0$, $B_{L,y} = 0$, $B_{R,y} = 0$, $B_{L,z} = -0.1$, $B_{R,z} = 0.1$, $s_R = 1$, $s_L = -1$, $\chi = 0.08$, $\epsilon = 1$. The red dot marks the zero $E_{2,y}$, where $\gamma_y(E_{2,y}) = 0$. In section 5.4.2 we present how to compute the zero $\mathbf{E}_2$ numerically.

## 5.3 DG Space and RK Time Discretization

As for BOR Maxwell's equations in sections 4.2 and 4.4, we apply the RKDG method to Kerr-nonlinear Maxwell's equations

$$\begin{aligned} \partial_t \mathbf{E} - J^{-1} \nabla \times \mathbf{H} &= 0, \\ \partial_t \mathbf{H} + \frac{1}{\mu_0} \nabla \times \mathbf{E} &= 0, \end{aligned} \tag{5.3.1}$$

In this section we will present the semi-discrete form of (5.3.1). We introduce the weak form of Kerr-nonlinear Maxwell's equations and we plug in the DG ansatz. The resulting local matrices do not change as in the BOR case and are computed as presented in, e.g., the book by Hesthaven and Warburton [11]. For this we use their Matlab codes.
The main topic of this section is the numerical flux. Since we have a nonlinear problem the corresponding Riemann problem must be solved in *each* grid point to obtain a numerical flux, which is computationally extremely expensive. Already for very few elements it takes minutes to run the simulation. We therefore choose a Lax-Friedrichs flux and an HLL-like Riemann solver which we modified for the Kerr system. For basic tests we also consider two linear numerical fluxes. We test the performance of our scheme on the analytically given solution for the one-dimensional Kerr-nonlinear Maxwell's equations, as found in [45].

### 5.3.1 Semi-Discrete Scheme

We proceed similarly to BOR Maxwell's equations, that is, we divide the physical domain of interest, $\Omega$, into $K$ elements so that $\Omega = \bigcup_{k=1}^{K} \Omega_k$. The finite element space of discontinuous functions is chosen to be

$$V_h := \{\mathbf{u}_h \in (L^\infty(\Omega))^4 : \mathbf{u}_h|_{\Omega_k} \in V(\Omega_k),\ k = 1, \ldots, K\},$$

where $V(\Omega_k) = \mathcal{P}^p(\Omega_k)$ is the space of one-dimensional polynomials of total degree $p \in \mathbb{N}$. We approximate $\mathbf{u}$ by $\mathbf{u}_h \in V_h$, and use Lagrange interpolation to represent $\mathbf{u}_h$ (cf. Ref. [11]) as

$$\mathbf{u}_h(x,t) = \sum_{i=1}^{N_p} \mathbf{u}_h^k(x_i^k, t)\, l_i^k(x), \tag{5.3.2}$$

where $l_i^k(x)$ are one-dimensional Lagrange polynomials on $\Omega_k$, $N_p = p + 1$ is the number of nodes and $x_i^k$ are suitably chosen interpolation points on $\Omega_k$. For the discontinuous Galerkin ansatz the residual $\mathcal{R}_h := \partial_t \mathbf{u}_h + \partial_x \mathbf{F}(\mathbf{u}_h)$ has to fulfill

$$\int_{\Omega_k} \mathcal{R}_h \cdot \boldsymbol{\psi}\, \mathrm{d}x = -\int_{\partial\Omega_k} (\mathbf{F}_\mathbf{n}(\mathbf{u}_h^k) - \mathbf{F}_\mathbf{n}^{\mathrm{num}}(\mathbf{u}_h^k)) \cdot \psi\, \mathrm{d}x,$$

where $\mathbf{n}$ denotes the outer normal vector of the boundary $\partial\Omega_k$, as usual, and $\mathbf{F}^{\mathrm{num}}$ is the numerical flux. Plugging in the the nodal representation of the fields (5.3.2) into Kerr-Maxwell's equations (5.3.1) and approximating the test functions $\psi$ by $\psi_h \in V_h$ as well,

we obtain the semi-discrete scheme

$$\begin{aligned}
\partial_t \mathbf{E}_\mathbf{y}^\mathbf{k} \int_{\Omega_k} l_i^k\, l_j^k \,\mathrm{d}x + J_{11}\mathbf{H}_\mathbf{z}^\mathbf{k} \int_{\Omega_k} l_i^k\, \partial_x l_j^k \,\mathrm{d}x - J_{12}\mathbf{H}_\mathbf{y}^\mathbf{k} \int_{\Omega_k} l_i^k\, \partial_x l_j^k \,\mathrm{d}x \\
= \int_{\partial\Omega_k} (\mathbf{F}_{E_y} - \mathbf{F}_{E_y}^{\text{num}})\mathbf{n}\, l_i^k\, l_j^k \,\mathrm{d}x, \\
\partial_t \mathbf{E}_\mathbf{z}^\mathbf{k} \int_{\Omega_k} l_i^k\, l_j^k \,\mathrm{d}x + J_{12}\mathbf{H}_\mathbf{z}^\mathbf{k} \int_{\Omega_k} l_i^k\, \partial_x l_j^k \,\mathrm{d}x - J_{22}\mathbf{H}_\mathbf{y}^\mathbf{k} \int_{\Omega_k} l_i^k\, \partial_x l_j^k \,\mathrm{d}x \\
= \int_{\partial\Omega_k} (\mathbf{F}_{E_z} - \mathbf{F}_{E_z}^{\text{num}})\mathbf{n}\, l_i^k\, l_j^k \,\mathrm{d}x, \\
\partial_t \mathbf{H}_\mathbf{y}^\mathbf{k} \int_{\Omega_k} l_i^k\, l_j^k \,\mathrm{d}x - \frac{1}{\mu_0}\mathbf{E}_\mathbf{z}^\mathbf{k} \int_{\Omega_k} l_i^k\, \partial_x l_j^k \,\mathrm{d}x = \int_{\partial\Omega_k} (\mathbf{F}_{H_y} - \mathbf{F}_{H_y}^{\text{num}})\mathbf{n}\, l_i^k\, l_j^k \,\mathrm{d}x, \\
\partial_t \mathbf{H}_\mathbf{z}^\mathbf{k} \int_{\Omega_k} l_i^k\, l_j^k \,\mathrm{d}x + \frac{1}{\mu_0}\mathbf{E}_\mathbf{y}^\mathbf{k} \int_{\Omega_k} l_i^k\, \partial_x l_j^k \,\mathrm{d}x = \int_{\partial\Omega_k} (\mathbf{F}_{H_z} - \mathbf{F}_{H_z}^{\text{num}})\mathbf{n}\, l_i^k\, l_j^k \,\mathrm{d}x.
\end{aligned} \tag{5.3.3}$$

By defining the local matrices

$$\begin{aligned}
(M^k)_{ij} &:= \int_{\Omega_k} l_i^k(x)\, l_j^k(x) \,\mathrm{d}x, \\
(S^k)_{ij} &:= \int_{\Omega_k} l_i^k(x)\, \partial_x l_j^k(x) \,\mathrm{d}x, \\
\mathcal{F}_{ij}^k &:= \int_{\partial\Omega_k} l_i^k(x)\, l_j^k(x) \,\mathrm{d}x
\end{aligned} \tag{5.3.4}$$

and introducing

$$\begin{aligned}
\mathbf{G}_\mathbf{E} &:= (\mathbf{F}_\mathbf{H} - \mathbf{F}_H^{\text{num}})\mathbf{n}, \\
\mathbf{G}_\mathbf{H} &:= (\mathbf{F}_\mathbf{E} - \mathbf{F}_E^{\text{num}})\mathbf{n},
\end{aligned}$$

we can rewrite (5.3.3) as

$$\begin{aligned}
M^k \partial_t \mathbf{E}_\mathbf{y}^\mathbf{k} + (J_{11} S^k \mathbf{H}_\mathbf{z}^\mathbf{k} - \mathcal{F}^k \mathbb{G}_{E_y}) - (J_{12} S^k \mathbf{H}_\mathbf{y}^\mathbf{k} - \mathcal{F}^k \mathbb{G}_{E_y}) = 0, \\
M^k \partial_t \mathbf{E}_\mathbf{z}^\mathbf{k} + (J_{12} S^k \mathbf{H}_\mathbf{z}^\mathbf{k} - \mathcal{F}^k \mathbb{G}_{E_z}) - (J_{22} S^k \mathbf{H}_\mathbf{y}^\mathbf{k} - \mathcal{F}^k \mathbb{G}_{E_z}) = 0, \\
M^k \partial_t \mathbf{H}_\mathbf{y}^\mathbf{k} - \frac{1}{\mu_0} S^k \mathbf{E}_\mathbf{z}^\mathbf{k} + \mathcal{F}^k \mathbb{G}_{H_y} = 0, \\
M^k \partial_t \mathbf{H}_\mathbf{z}^\mathbf{k} + \frac{1}{\mu_0} S^k \mathbf{E}_\mathbf{y}^\mathbf{k} - \mathcal{F}^k \mathbb{G}_{H_z} = 0,
\end{aligned}$$

where $\mathbb{G}_\mathbf{E}$ and $\mathbb{G}_\mathbf{H}$ denote the discrete analogues of $\mathbf{G}_\mathbf{E}$ and $\mathbf{G}_\mathbf{H}$, respectively. After multiplying by $(M^k)^{-1}$ we obtain the final semi-discrete scheme. We note that it holds $(M^k)^{-1} S^k = D_r$, where $D_r$ is the differentiation matrix from [11], i.e.

$$(D_r)_{ij} = \left.\frac{\mathrm{d}l_j(x)}{\mathrm{d}x}\right|_{x_i}.$$

The computation of the local matrices can also be found in [11]. To complete the DG space discretization we need a numerical flux.

### 5.3.2 Numerical Fluxes

We work with the set of equations (5.3.1). As a first approximation we assume to have a constant matrix $J_{LR} := J(\mathbf{E}_{LR})$, and thus the inverse $J_{LR}^{-1} = J^{-1}$ is also constant. The value $\mathbf{E}_{LR}$ is chosen depending on the choice of the numerical flux. For the Lax-Friedrichs

flux, we assume that $J = \epsilon_0 \,\mathrm{Id}$, for the HLL-like flux it is given later in (5.3.16). Thus we obtain

$$\begin{aligned} J_{LR}\,\partial_t \mathbf{E} - \nabla \times \mathbf{H} &= 0, \\ \mu_0\,\partial_t \mathbf{H} + \nabla \times \mathbf{E} &= 0. \end{aligned} \tag{5.3.5}$$

By defining $\mathbf{u} := (E_y, E_z, H_y, H_z)^T = (\mathbf{E}, \mathbf{H})^T$, $\mathbf{F}(\mathbf{u}) := (H_z, -H_y, -E_z, E_y)^T = (\tilde{\mathbf{H}}, \tilde{\mathbf{E}})^T$ and the matrix

$$Q := \begin{pmatrix} J_{LR} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\mu_0}\,\mathrm{Id} \end{pmatrix}$$

we can write this system as a conservation law, namely as

$$\mathbf{Q}\partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}) = \mathbf{0}.$$

We remark that the rotation in system (5.3.5) is computed by using the extended state and flux vectors

$$\bar{\mathbf{u}} := \begin{pmatrix} 0 \\ E_y \\ E_z \\ 0 \\ H_y \\ H_z \end{pmatrix}, \quad \mathbf{F}(\bar{\mathbf{u}})\mathbf{n} = \begin{pmatrix} 0 \\ n_x H_z \\ -n_x H_y \\ 0 \\ -n_x E_z \\ n_x E_y \end{pmatrix} = \begin{pmatrix} -\mathbf{n} \times \bar{\mathbf{H}} \\ \mathbf{n} \times \bar{\mathbf{E}} \end{pmatrix},$$

where $\mathbf{n} = (n_x, 0, 0)^T$ is the outer unit normal of an element $\Omega_k$, and $\bar{\mathbf{H}} := (0, H_y, H_z)^T$, $\bar{\mathbf{E}} := (0, E_y, E_z)^T$, $\bar{\mathbf{D}} := (0, D_y, D_z)^T$ denote the extended analogues of $\mathbf{H}, \mathbf{E}, \mathbf{D}$.

### A Lax-Friedrichs Flux

The Lax-Friedrichs flux often is a good choice of a numerical flux, also for nonlinear problems, although it is a numerical flux for linear problems. It is easily implemented and renders good approximation results comparable to those obtained with a Godunov flux (Riemann solver), although it produces more artificial viscosity; see [12, p. 187]. The authors also remark that, by their numerical experience, the numerical flux plays an increasingly minor role on the quality of the approximation with increasing polynomial order.
The Lax-Friedrichs flux was introduced in section 3.2.2, and it was given as

$$\mathbf{F}_{\mathbf{n}}^{\mathrm{LF}}(\mathbf{u}_L, \mathbf{u}_R) = \frac{1}{2}(\mathbf{F}_{\mathbf{n}}(\mathbf{u}_L) + \mathbf{F}_{\mathbf{n}}(\mathbf{u}_R)) + \frac{1}{2}C(\mathbf{u}_L - \mathbf{u}_R) = \{\{\mathbf{F}_{\mathbf{n}}\}\} + \frac{1}{2}C[\![\mathbf{u}]\!]. \tag{5.3.6}$$

In our case we need an expression for $(\mathbf{F}_{\mathbf{H}} - \mathbf{F}_{\mathbf{H}}^{\mathrm{LF}})\mathbf{n}$ and $(\mathbf{F}_{\mathbf{E}} - \mathbf{F}_{\mathbf{E}}^{\mathrm{LF}})\mathbf{n}$. We abbreviate $\mathbf{F}_L := \mathbf{F}_{\mathbf{n}}(\mathbf{u}_L)$ and $\mathbf{F}_R := \mathbf{F}_{\mathbf{n}}(\mathbf{u}_R)$ in the forthcoming and recall that $\mathbf{G}_{\mathbf{E}}^{\mathrm{LF}} = (\mathbf{F}_{\mathbf{E}} - \mathbf{F}_{\mathbf{E}}^{\mathrm{LF}})\mathbf{n}$; the same holds for $\mathbf{G}_{\mathbf{H}}^{\mathrm{LF}}$.

**Lemma 5.7.**
*The components of the Lax-Friedrichs flux for the Kerr-system* (5.3.1) *are given as*

$$\begin{aligned}
\mathbf{G}_{E_y}^{\mathrm{LF}} &= \frac{1}{2\{\{Z\}\}}\left(Z_R n_x [\![H_z]\!] - \alpha\, C\, [\![E_y]\!]\right),\\
\mathbf{G}_{E_z}^{\mathrm{LF}} &= \frac{1}{2\{\{Z\}\}}\left(-Z_R n_x [\![H_y]\!] - \alpha\, C\, [\![E_z]\!]\right),\\
\mathbf{G}_{H_y}^{\mathrm{LF}} &= \frac{1}{2\{\{Y\}\}}\left(-Z_L n_x [\![E_z]\!] - \alpha\, C\, [\![H_y]\!]\right),\\
\mathbf{G}_{H_z}^{\mathrm{LF}} &= \frac{1}{2\{\{Y\}\}}\left(Z_R n_x [\![E_y]\!] - \alpha\, C\, [\![H_z]\!]\right).
\end{aligned}$$

*Again,* $Z_{\mathrm{L,R}} = \sqrt{\mu_{\mathrm{L,R}}/\epsilon_{\mathrm{L,R}}}$ *is the local impedance,* $Y_{\mathrm{L,R}} = 1/Z_{\mathrm{L,R}}$ *is the local conductance, and* $\alpha \in [0,1]$*. For* $\alpha = 1$*, we have an upwind flux, for* $\alpha = 0$ *it is a central flux. The constant* $C$ *is given as* $C = \max_i |\lambda_i|$*, where* $\lambda_i$ *(*$i = 1, \cdots, 4$*) are the eigenvalues given in* (5.1.16)*. It is assumed to hold* $C \leq \frac{\Delta x}{\Delta t}\,\mathrm{CFL}$*, where* CFL *is the CFL number.*

### A Richtmyer Flux

As another choice of a numerical flux we take the Richtmyer flux which is a two-step version of the Lax-Wendroff flux for nonlinear systems of conservation laws (see e.g. [30, Ch. 5.3.4], [57, Ch. 6.1]). The Lax-Wendroff flux produces finite volume schemes of order 2 (see e.g. [113, Ch. 19.1, p. 844]). It is given as

$$\begin{aligned}
\mathbf{u}^{\mathrm{Ri}} &:= \frac{1}{2}(\mathbf{u}_L + \mathbf{u}_R) + \frac{1}{2}C(\mathbf{F}_{\mathbf{n}}(\mathbf{u}_L) - \mathbf{F}_{\mathbf{n}}(\mathbf{u}_R)),\\
\mathbf{F}_{\mathbf{n}}^{\mathrm{Ri}}(\mathbf{u}_L, \mathbf{u}_R) &:= \mathbf{F}_{\mathbf{n}}(\mathbf{u}^{\mathrm{Ri}}).
\end{aligned}$$

Note that for our purposes we need an expression for $(\mathbf{F}_{\mathbf{H}} - \mathbf{F}_{\mathbf{H}}^{\mathrm{Ri}})\mathbf{n}$ and $(\mathbf{F}_{\mathbf{E}} - \mathbf{F}_{\mathbf{E}}^{\mathrm{Ri}})\mathbf{n}$ so that $\mathbf{G}_{\mathbf{E}}^{\mathrm{Ri}} = (\mathbf{F}_{\mathbf{E}} - \mathbf{F}_{\mathbf{E}}^{\mathrm{Ri}})\mathbf{n}$; analogously for $\mathbf{G}_{\mathbf{H}}^{\mathrm{Ri}}$.
We note that in our case, the Richtmyer flux becomes the Lax-Friedrichs after some rearrangement in terms.

### Another Linear Numerical Flux

Let us first look at the linear case, i.e. $s_1 = s_L$, $s_2 = s_R$. Then it holds $\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{u}_3$ and $\mathbf{F}_1 = \mathbf{F}_2 = \mathbf{F}_3$, and the Rankine-Hugoniot conditions reduce to

$$\mathbf{F}_2 - \mathbf{F}_L = s_L(\mathbf{u}_2 - \mathbf{u}_L), \tag{5.3.7a}$$
$$\mathbf{F}_R - \mathbf{F}_2 = s_R(\mathbf{u}_R - \mathbf{u}_2). \tag{5.3.7b}$$

We solve (5.3.7a) for $\mathbf{u}_2$, which leads to

$$\mathbf{u}_2 = \frac{1}{s_L}(\mathbf{F}_2 - \mathbf{F}_L) + \mathbf{u}_L,$$

and plugging this into (5.3.7a) we obtain

$$\mathbf{F}_L - \mathbf{F}_2 = \frac{s_L}{s_L - s_R}(-s_R(\mathbf{u}_L - \mathbf{u}_R) + (\mathbf{F}_L - \mathbf{F}_R)). \tag{5.3.8}$$

Alternatively one could substract (5.3.7b) from (5.3.7a) and solve for $\mathbf{F}_L - \mathbf{F}_2$, which gives

$$\mathbf{F}_L - \mathbf{F}_2 = \frac{1}{2}[(\mathbf{F}_L - \mathbf{F}_R) + s_L(\mathbf{u}_L - \mathbf{u}_2) + s_R(\mathbf{u}_R - \mathbf{u}_2)]. \tag{5.3.9}$$

Both expressions are equivalent, with the difference that equation (5.3.9) uses $\mathbf{u}_2$ and equation (5.3.8) does not need $\mathbf{u}_2$ at all. All flux expressions are equivalent to the Lax-Friedrichs flux in the previous section 5.3.2. Equations (5.3.8) and (5.3.9) can be used to test the quality of the approximation of the zero $\mathbf{u}_2$ of the function $\gamma$ defined in (5.2.67).

### A Nearly Exact Numerical Flux

We now turn back to the full problem (5.3.10) and solve for $\mathbf{F}_L - \mathbf{F}_2$. At this point we make a first approximation to the exact Riemann solution. We suppose to have only shocks, thus ignoring rarefaction waves, but computing the middle state $\mathbf{u}_2$ still exactly (in the sense that it is determined as the zero of the function $\gamma$ in (5.2.67) via some zero finding method). By applying the Rankine-Hugoniot conditions we obtain

$$\mathbf{F}_1 - \mathbf{F}_L = s_L(\mathbf{u}_1 - \mathbf{u}_L), \tag{5.3.10a}$$
$$\mathbf{F}_2 - \mathbf{F}_1 = s_1(\mathbf{u}_2 - \mathbf{u}_1), \tag{5.3.10b}$$
$$\mathbf{F}_3 - \mathbf{F}_2 = s_2(\mathbf{u}_3 - \mathbf{u}_2), \tag{5.3.10c}$$
$$\mathbf{F}_R - \mathbf{F}_3 = s_R(\mathbf{u}_R - \mathbf{u}_3). \tag{5.3.10d}$$

By $\mathbf{F}_i$ we denote the flux in region $i$ ($i = L, 1, 2, 3, R$) and by $\mathbf{u}_i$ the solution in region $i$. Recall that $s_L$ and $s_R$ are the wave speeds of the left and right contact discontinuity (see definition 3.21), respectively, and $s_1$, $s_2$ are the shock speeds of the left and right shock, respectively (see definition 3.20 and (5.2.60)).
By adding equations (5.3.10a) and (5.3.10b), and equations (5.3.10c) and (5.3.10d) we eliminate $\mathbf{F}_1$ and $\mathbf{F}_3$ and obtain

$$\mathbf{F}_2 - \mathbf{F}_L = s_L(\mathbf{u}_1 - \mathbf{u}_L) + s_1(\mathbf{u}_2 - \mathbf{u}_1), \tag{5.3.11}$$
$$\mathbf{F}_R - \mathbf{F}_2 = s_2(\mathbf{u}_3 - \mathbf{u}_2) + s_R(\mathbf{u}_R - \mathbf{u}_3). \tag{5.3.12}$$

Substracting equation (5.3.11) from (5.3.12) and adding $\mathbf{F}_L$ on both sides gives after some rearrangement in terms

$$\begin{aligned}\mathbf{F}_L - \mathbf{F}_2 =& \frac{1}{2}\left[(\mathbf{F}_L - \mathbf{F}_R) + s_L(\mathbf{u}_L - \mathbf{u}_1) + s_1(\mathbf{u}_1 - \mathbf{u}_2)\right.\\ &\left.+ s_2(\mathbf{u}_3 - \mathbf{u}_2) + s_R(\mathbf{u}_R - \mathbf{u}_3)\right].\end{aligned} \tag{5.3.13}$$

We implement this numerical flux by considering the following. The wave speeds $s_L$, $s_R$ correspond to the left and right contact discontinuities, respectively, with $s_L = \lambda_1^-(|\mathbf{E}_L|)$ and $s_R = \lambda_1^+(|\mathbf{E}_R|)$, if $|\mathbf{E}_L|$ is the upwind value, i.e. $|\mathbf{E}_L| \leq |\mathbf{E}_R|$. In the other case, it is the other way around. For a shock, the wave speeds $s_1$ and $s_2$ are given by equation (5.2.60). In case of a rarefaction wave, we choose $s_1$ and $s_2$ to be the speeds of the head of the left and right rarefaction fan, respectively, that is, $s_1 = \lambda_2^-(|\mathbf{E}_L|)$ and $s_2 = \lambda_2^+(|\mathbf{E}_R|)$, taking into account that $|\mathbf{E}_3| = |\mathbf{E}_R|$ and $|\mathbf{E}_1| = |\mathbf{E}_L|$.
This numerical flux still needs $\mathbf{u}_2$, which has to be computed for every grid point anew. Although we use a fast zero finding routine, see section 5.4.2, the simulation is still very slow. We therefore want a *global* approximation to $\mathbf{u}_2$.

### An HLL-like Flux

In section 5.2.4 the exact Riemann solution of the Kerr system was given and in the last subsection we presented an implementation of a nearly exact numerical flux resulting from this Kerr-Riemann solution. We pointed out that using this as an analytical numerical flux is computationally too expensive. In this section we present an HLL-like Riemann solver, which is a modified HLL solver to suit our needs. The original HLL Riemann solver was introduced by Harten, Lax and van Leer [29], and it is an approximation to the exact solution of the Riemann problem. The authors also showed that if the scheme with an HLL flux converges, it converges to the weak solution. Furthermore, several works demonstrate that the HLL flux (and its modified versions, like the HLLC or HLLE solvers [30]) give very good numerical performances; see e.g. [31], where several versions of the

HLL flux are compared for the magneto-hydrodynamic equations. For some background and properties of the HLL flux, see e.g. [30, Ch. 10].
The HLL flux assumes to have two shock waves and thus ignores rarefaction waves or contact discontinuities. Applying the Rankine-Hugoniot jump conditions, one obtains the numerical flux $\mathbf{F}^{\text{num}} = \mathbf{F}^{\text{HLL}}$ with (see e.g. [30, Ch. 10], [114])

$$\mathbf{F}^{\text{HLL}}(\mathbf{u}_l, \mathbf{u}_r) = \begin{cases} \mathbf{F}(\mathbf{u}_l) & \text{if } 0 \leq s_1^*, \\ \frac{s_2^*\mathbf{F}(\mathbf{u}_l) - s_1^*\mathbf{F}(\mathbf{u}_r) + s_1^* s_2^* (\mathbf{u}_r - \mathbf{u}_l)}{s_2^* - s_1^*} & \text{if } s_1^* \leq 0 \leq s_2^*, \\ \mathbf{F}(\mathbf{u}_r) & \text{if } s_2^* \leq 0. \end{cases} \tag{5.3.14}$$

Here, "$r$" denotes the value right to the shock wave with wave speed $s_1^*$, and "$l$" denotes the value left to the shock wave with wave speed $s_2^*$; see figure 5.18 (c) for a sketch. Note that equation (5.3.14) is to be understood as a general formula for the HLL flux, so $\mathbf{u}_l$ and $\mathbf{u}_r$ must not be confused with $\mathbf{u}_L$ and $\mathbf{u}_R$. The wave speeds $s_1^*$ and $s_2^*$ are determined via so-called *wave speed estimates*. Several wave speed estimates were proposed, where we refer to the ones in Toro [30] and in Batten et al. [114].
Recall that the solution of the Riemann problem was given as displayed in figure 5.18(a).

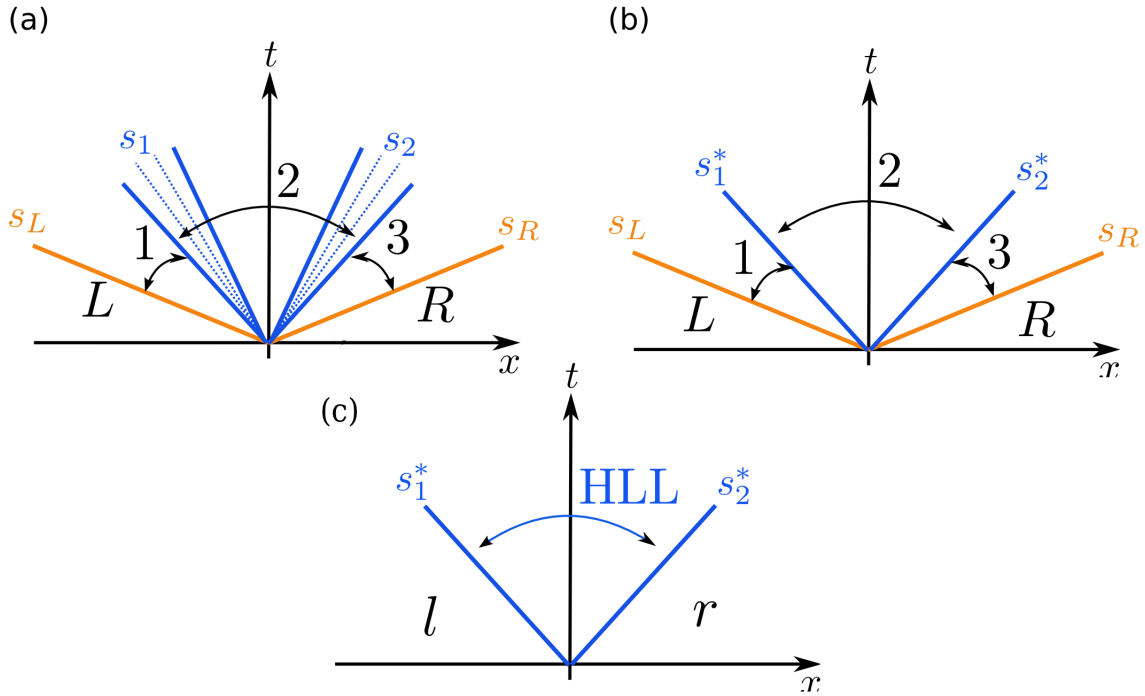



Figure 5.18: (a) Exact Riemann solution for Kerr-nonlinearity (from left to right): left contact discontinuity with speed $s_L$, left rarefaction wave/left shock with speed $s_1$, right rarefaction wave/right shock with speed $s_2$, right contact discontinuity with speed $s_R$; (b) Riemann solution for Kerr-nonlinearity when using an HLL-like approximation (from left to right): left contact, left shock with estimated speed $s_1^*$, right shock with estimated speed $s_2^*$, right contact; (c) HLL solver: left and right shock.

$s_1^*$ and $s_2^*$ are global approximations to the shock speeds $s_1$ and $s_2$. The following formulas are collective information from section 5.2.4 and are essential for the subsequent discussion.

$$|\mathbf{E}_1| = |\mathbf{E}_L|, \quad |\mathbf{E}_3| = |\mathbf{E}_R|, \tag{5.3.15a}$$

$$\mathbf{e}_1 = \frac{\mathbf{E}_1}{|\mathbf{E}_L|}, \quad \mathbf{e}_2 = \frac{\mathbf{E}_2}{|\mathbf{E}_2|}, \quad \mathbf{e}_3 = \frac{\mathbf{E}_3}{|\mathbf{E}_R|} \tag{5.3.15b}$$

$$\mathbf{e}_1 = \mathbf{e}_2 = \mathbf{e}_3, \tag{5.3.15c}$$

$$\tilde{\mathbf{H}}_1 = \tilde{\mathbf{H}}_L - \sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_L|^2}{2}\right)}|\mathbf{E}_L|(\mathbf{e}_1 - \mathbf{e}_L), \tag{5.3.15d}$$

$$\tilde{\mathbf{H}}_3 = \tilde{\mathbf{H}}_R - \sqrt{\mathcal{E}'\left(\frac{|\mathbf{E}_R|^2}{2}\right)}|\mathbf{E}_R|(\mathbf{e}_R - \mathbf{e}_2). \tag{5.3.15e}$$

Note that $\mathbf{E}_i = [E_{i,y}, E_{i,z}]^T, \tilde{\mathbf{H}}_i = [H_{i,z}, -H_{i,y}]^T$ $(i = 1, 2, 3)$ and so on. Also, it is $\mathbf{E}_1 \neq \mathbf{E}_3 \neq \mathbf{E}_2$ in general.
The HLL Riemann solver assumes to have two shocks in between a left region "$l$" and a right region "$r$"; see figure 5.18 (c) for a visualization. Across the shocks the Rankine-Hugoniot condition holds. By combining the Rankine-Hugoniot condition for the left and right shock, one can show [30, Ch. 10] that $\mathbf{u}_2$ is approximated by a state $\mathbf{u}^{\text{HLL}}$ in the following manner (recall (5.3.14)):

$$\mathbf{u}^{\text{HLL}} = \frac{s_2^*\mathbf{u}_r - s_1^*\mathbf{u}_l + \mathbf{F}_l - \mathbf{F}_r}{s_2^* - s_1^*}. \tag{5.3.6}$$

For our case, we have a left and right contact discontinuity, and we now assume two have only a left and right shock, thus neglecting rarefaction waves completely, analogously to the HLL solver. The wave speeds $s_1$ and $s_2$ are approximated by the wave speeds $s_1^*$ and $s_2^*$, respectively. Across all four waves the Rankine-Hugoniot jump conditions (5.3.10) hold. We now substitute $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{H}_3$ by approximations $\mathbf{E}_1^{\text{HLL}}$, $\mathbf{E}_2^{\text{HLL}}$, $\mathbf{E}_3^{\text{HLL}}$, $\mathbf{H}_1^{\text{HLL}}$, $\mathbf{H}_2^{\text{HLL}}$ and $\mathbf{H}_3^{\text{HLL}}$ given by (5.3.6). Thus we obtain

$$\mathbf{E}_2 \approx \mathbf{E}_2^{\text{HLL}} = \frac{s_2^*\mathbf{E}_3^{\text{HLL}} - s_1^*\mathbf{E}_1^{\text{HLL}} + \tilde{\mathbf{H}}_1^{\text{HLL}} - \tilde{\mathbf{H}}_3^{\text{HLL}}}{s_2^* - s_1^*}, \tag{5.3.7a}$$

$$\mathbf{H}_2 \approx \mathbf{H}_2^{\text{HLL}} = \frac{s_2^*\mathbf{H}_3^{\text{HLL}} - s_1^*\mathbf{H}_1^{\text{HLL}} + \tilde{\mathbf{E}}_1^{\text{HLL}} - \tilde{\mathbf{E}}_3^{\text{HLL}}}{s_2^* - s_1^*}, \tag{5.3.7b}$$

$$\mathbf{E}_3 \approx \mathbf{E}_3^{\text{HLL}} = \frac{s_R\mathbf{E}_R - s_2^*\mathbf{E}_2^{\text{HLL}} + \tilde{\mathbf{H}}_2^{\text{HLL}} - \tilde{\mathbf{H}}_R}{s_R - s_2^*}. \tag{5.3.7c}$$

First we observe that equation (5.3.7b) can also be written as

$$\tilde{\mathbf{H}}_2^{\text{HLL}} = \frac{s_2^*\tilde{\mathbf{H}}_3^{\text{HLL}} - s_1^*\tilde{\mathbf{H}}_1^{\text{HLL}} + \mathbf{E}_1^{\text{HLL}} - \mathbf{E}_3^{\text{HLL}}}{s_2^* - s_1^*}, \tag{5.3.8}$$

and from equations (5.3.15a), (5.3.15b) and (5.3.15c) one obtains

$$\mathbf{E}_1^{\text{HLL}} = \frac{|\mathbf{E}_L|}{|\mathbf{E}_R|}\mathbf{E}_3^{\text{HLL}}, \tag{5.3.9}$$

so we can express $\mathbf{E}_1^{\mathrm{HLL}}$ in terms of $\mathbf{E}_3^{\mathrm{HLL}}$. The same holds for $\tilde{\mathbf{H}}_1^{\mathrm{HLL}}$ and $\tilde{\mathbf{H}}_3^{\mathrm{HLL}}$, as we can see from equations (5.3.15d) and (5.3.15e) if we insert (5.3.9):

$$\tilde{\mathbf{H}}_1^{\mathrm{HLL}} = \tilde{\mathbf{H}}_L - \sqrt{\mathcal{E}'_L}\left(\frac{|\mathbf{E}_L|}{|\mathbf{E}_R|}\mathbf{E}_3^{\mathrm{HLL}} - \mathbf{E}_L\right), \tag{5.3.10a}$$

$$\tilde{\mathbf{H}}_3^{\mathrm{HLL}} = \tilde{\mathbf{H}}_R - \sqrt{\mathcal{E}'_R}(\mathbf{E}_R - \mathbf{E}_3^{\mathrm{HLL}}). \tag{5.3.10b}$$

Here we abbreviated $\mathcal{E}'(|\mathbf{E}_i|^2/2) =: \mathcal{E}'_i$ for $i = L, R$. By plugging relations (5.3.9), (5.3.10a) and (5.3.10b) into (5.3.8), we encounter

$$\tilde{\mathbf{H}}_2^{\mathrm{HLL}} = \mathbf{a}_{LR} + b_{LR}\mathbf{E}_3^{\mathrm{HLL}}, \tag{5.3.11}$$

where we introduced

$$\mathbf{a}_{LR} := \frac{1}{s_2^* - s_1^*}\left[s_2^*(\tilde{\mathbf{H}}_R - \sqrt{\mathcal{E}'_R}\mathbf{E}_R) - s_1^*(\tilde{\mathbf{H}}_L + \sqrt{\mathcal{E}'_L}\mathbf{E}_L)\right],$$

$$b_{LR} := \frac{1}{s_2^* - s_1^*}\left[s_2^*\sqrt{\mathcal{E}'_R} + s_1^*\sqrt{\mathcal{E}'_L}\frac{|\mathbf{E}_L|}{|\mathbf{E}_R|} + \frac{|\mathbf{E}_L|}{|\mathbf{E}_R|} - 1\right].$$

Note that $\mathbf{a}_{LR}$ is a two-dimensional vector and $b_{LR}$ is a scalar. Analogously, we insert equations (5.3.9), (5.3.10a) and (5.3.10b) into (5.3.7a) and obtain

$$\mathbf{E}_2^{\mathrm{HLL}} = \mathbf{c}_{LR} + d_{LR}\mathbf{E}_3^{\mathrm{HLL}} \tag{5.3.12}$$

with

$$\mathbf{c}_{LR} := \frac{1}{s_2^* - s_1^*}\left[\tilde{\mathbf{H}}_L - \tilde{\mathbf{H}}_R + \sqrt{\mathcal{E}'_L}\mathbf{E}_L + \sqrt{\mathcal{E}'_R}\mathbf{E}_R)\right],$$

$$d_{LR} := \frac{1}{s_2^* - s_1^*}\left[s_2^* - s_1^*\frac{|\mathbf{E}_L|}{|\mathbf{E}_R|} - \sqrt{\mathcal{E}'_L}\frac{|\mathbf{E}_L|}{|\mathbf{E}_R|} - \sqrt{\mathcal{E}'_R}\right].$$

Next, we solve equation (5.3.7c) for $\mathbf{E}_2^{\mathrm{HLL}}$, that is,

$$\mathbf{E}_2^{\mathrm{HLL}} = \frac{1}{s_2^*}(s_R\mathbf{E}_R - \tilde{\mathbf{H}}_R + \tilde{\mathbf{H}}_2^{\mathrm{HLL}} - (s_R - s_2^*)\mathbf{E}_3^{\mathrm{HLL}}), \tag{5.3.13}$$

and plug in equation (5.3.11), which gives

$$\mathbf{E}_2^{\mathrm{HLL}} = \mathbf{A}_{LR} + B_{LR}\mathbf{E}_3^{\mathrm{HLL}}, \tag{5.3.14}$$

where $\mathbf{A}_{LR}$ and $B_{LR}$ are defined as

$$\mathbf{A}_{LR} := \frac{1}{s_2^*}\left[s_R\mathbf{E}_R + \mathbf{a}_{LR} - \tilde{\mathbf{H}}_R\right],$$

$$B_{LR} := \frac{1}{s_2^*}\left(b_{LR} - s_R + s_2^*\right).$$

We equalize equations (5.3.12) and (5.3.14) to obtain

$$\mathbf{E}_3^{\mathrm{HLL}} = \frac{\mathbf{c}_{LR} - \mathbf{A}_{LR}}{B_{LR} - d_{LR}}.$$

Inserting this into (5.3.12) finally gives

$$\mathbf{E}_2^{\mathrm{HLL}} = \mathbf{c}_{LR} + d_{LR}\left(\frac{\mathbf{c}_{LR} - \mathbf{A}_{LR}}{B_{LR} - d_{LR}}\right) = \frac{\mathbf{c}_{LR}B_{LR} - d_{LR}\mathbf{A}_{LR}}{B_{LR} - d_{LR}}.$$

If $B_{LR} = d_{LR}$, we choose the linear numerical flux in equation (5.3.8). This is justified by the following considerations. Plugging alternatively equation (5.3.11) into (5.3.7c) gives

$$\mathbf{E}_3^{\mathrm{HLL}} = \frac{1}{s_R - s_2^* - b_{LR}} \left( s_R \mathbf{E}_R - s_2^* \mathbf{E}_2^{\mathrm{HLL}} + \mathbf{a}_{LR} - \tilde{\mathbf{H}}_R \right).$$

If we insert this into (5.3.12) we obtain

$$\mathbf{E}_2^{\mathrm{HLL}} = \frac{s_R - s_2^* - b_{LR}}{1 + s_2^* d_{LR}} \left[ \mathbf{c}_{LR} + \frac{d_{LR}}{s_R - s_2^* - b_{LR}} (s_R \mathbf{E}_R + \mathbf{a}_{LR} - \tilde{\mathbf{H}}_R) \right].$$

The cases $s_R - s_2^* - b_{LR} = 0$ or $B_{LR} = d_{LR}$ are of particular interest. The case $s_R - s_2^* - b_{LR} = 0$ can only occur if $s_R = s_2^*$ and if $b_{LR} = 0$. This happens if $|\mathbf{E}_L| = |\mathbf{E}_R|$, and in this case, $\mathbf{E}_1 = \mathbf{E}_3$, so that $s_1^* = -s_2^*$. Then it is also $B_{LR} = d_{LR}$, i.e. we only have the 2-region and we obtain the solution of the linear Riemann problem corresponding to linear Maxwell's equations. Therefore we can choose the linear numerical flux in (5.3.8) or the Lax-Friedrichs flux as given in (5.3.6). With view to numerics, we therefore make the following important observation as a consequence of equation (5.3.9): Whenever the values $|\mathbf{E}_L|$ and $|\mathbf{E}_R|$ are close to each other, that is, $||\mathbf{E}_L| - |\mathbf{E}_R|| \leq \mathrm{tol}$ for some given tolerance, we are – inside this tolerance – close to the linear case. Then the linear numerical flux can be looked upon as a good approximation, meaning, inside this tolerance. For instance, for problems with smooth solutions, as e.g. a Gaussian pulse which we consider in section 5.4.1 or a (linear) standing wave as a basic test, our numerical results suggest a tolerance of around $10^{-14}$ up to $10^{-4}$, depending on $\chi$. For instance, if $\chi = 10^{-14}$, $\mathrm{tol} \sim 10^{-14}$, for $\chi = 0.08$, it is $\mathrm{tol} \sim 10^{-4}$.

### Wave Speed Estimates

What remains are explicit expressions for the wave speeds $s_1^*, s_2^*$. These can be obtained via so-called *wave speed estimates*. We choose the wave speed estimates given by Batten et. al. [114], which were first introduced by Einfeldt et. al. in [115], due to the fact that these are less diffusive compared to other estimates, they give rise to a robust algorithm and shocks are resolved exactly.
Following Batten et. al. [114] the wave speed approximations $s_1^*$ and $s_2^*$ are given for an $m$-dimensional system as

$$\begin{aligned} s_1^* &= \min\{\lambda_1(\mathbf{u}_l), \lambda_1(\mathbf{u}^{\mathrm{Roe}})\}, \\ s_2^* &= \max\{\lambda_m(\mathbf{u}_r), \lambda_m(\mathbf{u}^{\mathrm{Roe}})\}. \end{aligned} \tag{5.3.15}$$

Here, $\mathbf{u}^{\mathrm{Roe}}$ is the so-called Roe average of the Roe solver [28] which has to be determined. A practical introduction to the Roe solver can be found in e.g. [30, Ch. 11]. There, also the Roe-Pike method is explained which gives a way of avoiding the computation of the Roe matrix explicitly, which is needed to determine the Roe average $\mathbf{u}^{\mathrm{Roe}}$. Yet this ansatz leads to extremely large and complicated expressions for our case, in parts only implicitly given. A note in the book by Laney [116, Ch. 5.3.2, p. 86] gives us another approach. The author remarks that the Roe average $\mathbf{u}^{\mathrm{Roe}}$ can be written as

$$\mathbf{u}^{\mathrm{Roe}} =: \mathbf{u}_{RL} = \theta \mathbf{u}_L + (1 - \theta)\mathbf{u}_R. \tag{5.3.16}$$

If $\theta \in [0, 1]$ such Roe averages are called *linear averages*, *convex interpolations* or *convex linear combinations*. For example, for Euler's equations, the Roe average indeed fulfills (5.3.16), where $\theta$ is a fixed fraction consisting of the Euler variables so that $0 < \theta < 1$. Our approach therefore is to assume that $\mathbf{u}^{\mathrm{Roe}}$ can be expressed as in (5.3.16) with $\theta \in [0, 1]$,

where $\theta$ is treated as an unknown.
In order to finalize the wave speed estimates, we need the eigenvalues of the Kerr system which were given in (5.1.16). For our purposes we rename the eigenvalues in order of their magnitude as

$$\lambda_1(|\mathbf{E}|) = -\frac{1}{\sqrt{\mathcal{E}'(\frac{|\mathbf{E}|^2}{2})}}, \tag{5.3.17}$$

$$\lambda_2(|\mathbf{E}|) = -\frac{1}{\sqrt{\mathcal{E}'(\frac{|\mathbf{E}|^2}{2}) + \mathcal{E}''|\mathbf{E}|^2}}, \tag{5.3.18}$$

$$\lambda_3(|\mathbf{E}|) = \frac{1}{\sqrt{\mathcal{E}'(\frac{|\mathbf{E}|^2}{2}) + \mathcal{E}''|\mathbf{E}|^2}}, \tag{5.3.19}$$

$$\lambda_4(|\mathbf{E}|) = \frac{1}{\sqrt{\mathcal{E}'(\frac{|\mathbf{E}|^2}{2})}}. \tag{5.3.20}$$

$\lambda_1, \lambda_2$ are strictly monotonic increasing with $|\mathbf{E}|$, and $\lambda_3, \lambda_4$ are strictly monotonic decreasing with $|\mathbf{E}|$. Thus, using (5.3.15) and the ansatz (5.3.16) with $\theta \in [0, 1]$ we obtain the wave speeds

$$s_1^* = \min_\theta\{\lambda_1(|\mathbf{E}_L|), \lambda_1(|\mathbf{E}_{RL}(\theta)|)\}, \tag{5.3.21}$$

$$s_2^* = \max_\theta\{\lambda_4(|\mathbf{E}_R|), \lambda_4(|\mathbf{E}_{RL}(\theta)|)\}, \tag{5.3.22}$$

where $\mathbf{E}_{RL}(\theta) = \theta\mathbf{E}_L + (1-\theta)\mathbf{E}_R$. Also note, since the eigenvalues are increasing or decreasing functions in $|\mathbf{E}|$, the maximum and minimum, respectively, given in (5.3.21) always exists, thus $s_1^*, s_2^*$ are well-defined.
We could not find explicit expressions for $s_1^*, s_2^*$, only implicit ones which involved the computation of zeros of implicitly defined functions, which would contradict the intention of the HLL solver, namely to be computationally efficient and affordable. In practice we therefore chose the following procedure: We define an array theta = linspace$(0, 1, \text{num})$, where num is an arbitrary number which should not be too big with view to efficiency; besides it is unnecessary to choose a big value of num. We then compute $s_1^*, s_2^*$ as given in (5.3.21) and take the maximum and the minimum, respectively, over all chosen values of $\theta$.

The HLL-like flux should be applied whenever $s_1 \neq s_L$ and $s_2 \neq s_R$. If $s_1$ and $s_L$, $s_2$ and $s_R$ are close to each other within a tolerance tol, we are close to the linear case. Furthermore, throughout our simulations we observed $s_L = -s_R$ and $s_1 = -s_2$. Whenever $|s_1 - s_L| \leq \text{tol}$ (in this case it is also $|s_2 - s_R| \leq \text{tol}$) one can use a linear flux, like e.g. the Lax-Friedrichs flux (5.3.6) or the linear flux (5.3.8).

## 5.4 Numerical Tests

### 5.4.1 Gaussian Pulse for the One-Dimensional Kerr System

If we set $E_y = 0$ and $H_z = 0$ we obtain the one-dimensional Kerr-nonlinear Maxwell's equations as

$$\partial_t E_z - J_{22}\partial_x H_y = 0, \tag{5.4.1}$$

$$\partial_t H_y - \frac{1}{\mu_0}\partial_x E_z = 0 \tag{5.4.2}$$

where $J_{22}$ is an entry of the matrix $J^{-1}$ from (5.1.13), which for this case is given as

$$J^{-1} = \begin{pmatrix} \frac{1}{\epsilon_0+\chi E_z^2} & 0 \\ 0 & \frac{1}{\epsilon_0+3\chi E_z^2} \end{pmatrix} =: \begin{pmatrix} J_{11} & J_{12} \\ J_{12} & J_{22} \end{pmatrix}.$$

Equivalently, one could set $E_z = 0$ and $H_y = 0$, which gives

$$J^{-1} = \begin{pmatrix} \frac{1}{\epsilon_0+3\chi E_y^2} & 0 \\ 0 & \frac{1}{\epsilon_0+3\chi E_y^2} \end{pmatrix}$$

and

$$\partial_t E_y + J_{11}\partial_x H_z = 0, \tag{5.4.3}$$

$$\partial_t H_z + \frac{1}{\mu_0}\partial_x E_y = 0 \tag{5.4.4}$$

In [45] the analytical solution of (5.4.1) is given implicitly as

$$E_z = E_z(x,t) = G\left( x \pm \frac{1}{\sqrt{\sqrt{\epsilon_0\mu_0} + \sqrt{\frac{\mu_0}{\epsilon_0}}3\chi E_z(x,t)^2}}\, t \right), \tag{5.4.5}$$

$$H_y(x,t) = \frac{\mu_0}{\epsilon_0} F\left( \frac{3\chi E_z(x,t)^2}{\epsilon_0} \right). \tag{5.4.6}$$

The initial data for $E_z$ is assumed to be continuously differentiable, that is,
$E_z(x, t = 0) \in C^1(\mathbb{R})$. Furthermore, the function $F$ is for the right traveling wave, i.e. for the minus sign, defined as the power series

$$F(x) := \sum_{n=0}^{\infty} a_n x^n \tag{5.4.7}$$

with the coefficients

$$a_n = (-1)^n \frac{(2n-3)!!}{2^n(2n+1)n!} \qquad \text{for } n > 0,$$

$$a_0 = -1.$$

Here, $m!!$ is the double factorial which is defined as

$$m!! := (2m-1)!, \quad m \in \mathbb{N},\ m \le n,$$

$$m!! = 1 \text{ for } m \le 0.$$

For the left traveling wave, one substitutes $F \to -F$.
We note that with the substitution $E_z \to E_y$ and $H_y \to -H_z$ (5.4.5) is also a solution of (5.4.3).

In our simulations we choose a Gaussian pulse as initial condition, i.e.

$$G(E_z) := E_0 \exp\left[ -\frac{1}{2\sigma}x - \left( \frac{1}{\sqrt{\sqrt{\epsilon_0\mu_0} + \sqrt{\frac{\mu_0}{\epsilon_0}}3\chi E_z^2}}\, t \right)^2 \right], \tag{5.4.8}$$

where $E_0$ is the amplitude and $\sigma$ is the width of the Gaussian pulse. $E_z$ is determined as the zero of the function

$$f(E_z) := E_z - G(E_z).$$

Analogously, one cant determine $E_y$ as the zero of $f(E_y) = E_y - G(E_y)$. We use a modified regula falsi method to compute its zero, which is as fast as a Newton's method (which is of order 2) or much faster than a bisection scheme, avoiding the computation of the derivative of $f$. For computing the analytical solution, speed is essential, since the computation of $E_z$ takes most of the simulation time. Figure 5.19 shows snapshots of a Gaussian pulse with the settings $\chi = 0.08$, $\sigma = 0.1$, $E_0 = 1$, $x_0 = 0$, the simulation interval was $[-1, 5]$ and the simulation time $t_{\text{final}} = 3$. Please note that theoretically, the Gaussian pulse is for $t > t_{\max} = 0.947$ not analytical anymore; indeed, we can see for $t = 3$ in figure 5.19, that the pulse steepens more and more, thus resulting in a shock.
We compute the double factorial via the formula

$$(2m-1)!! = \frac{(2m-1)!}{2^{m-1}(m-1)!}, \quad \text{for } m > 0.$$

Furthermore, we determine the coefficients $a_n$ via the Horner scheme (see e.g. [80] [81]); in order to do so the power series $F$ is rewritten as

$$F(x) = \sum_{n=0}^{\infty} a_n x^n = (\dots(a_n x + a_{n-1})x + \dots)x + a_0.$$

In order to ensure existence and uniqueness of the analytical solution the assumptions of the implicit function theorem must be fulfilled. This leads to the condition $t < t_{\max}$ for the time variable $t$, where $t_{\max}$ is given as follows [117, Lemma 2.15]

$$t_{\max} := \left(1 + \frac{3\chi E_z^2}{\epsilon_0}\right)^{\frac{3}{2}} \left|\frac{\epsilon_0\sqrt{\epsilon_0\mu_0}}{3\chi E_0^2}\right| \sigma e^{\frac{1}{2}}.$$

If $\chi = 0$, the Gaussian pulse is given as

$$\begin{aligned} E_y &= e^{-\frac{(x-x_0+t)^2}{2\sigma^2}}, \\ H_z &= e^{-\frac{(x-x_0+t)^2}{2\sigma^2}}. \end{aligned} \tag{5.4.9}$$

In this case, $t_{\max}$ would be infinity, and the total simulation time can be chosen arbitrarily. Furthermore, as we mentioned in section 2.1, $\chi$ has to be small enough so that the power series converges.

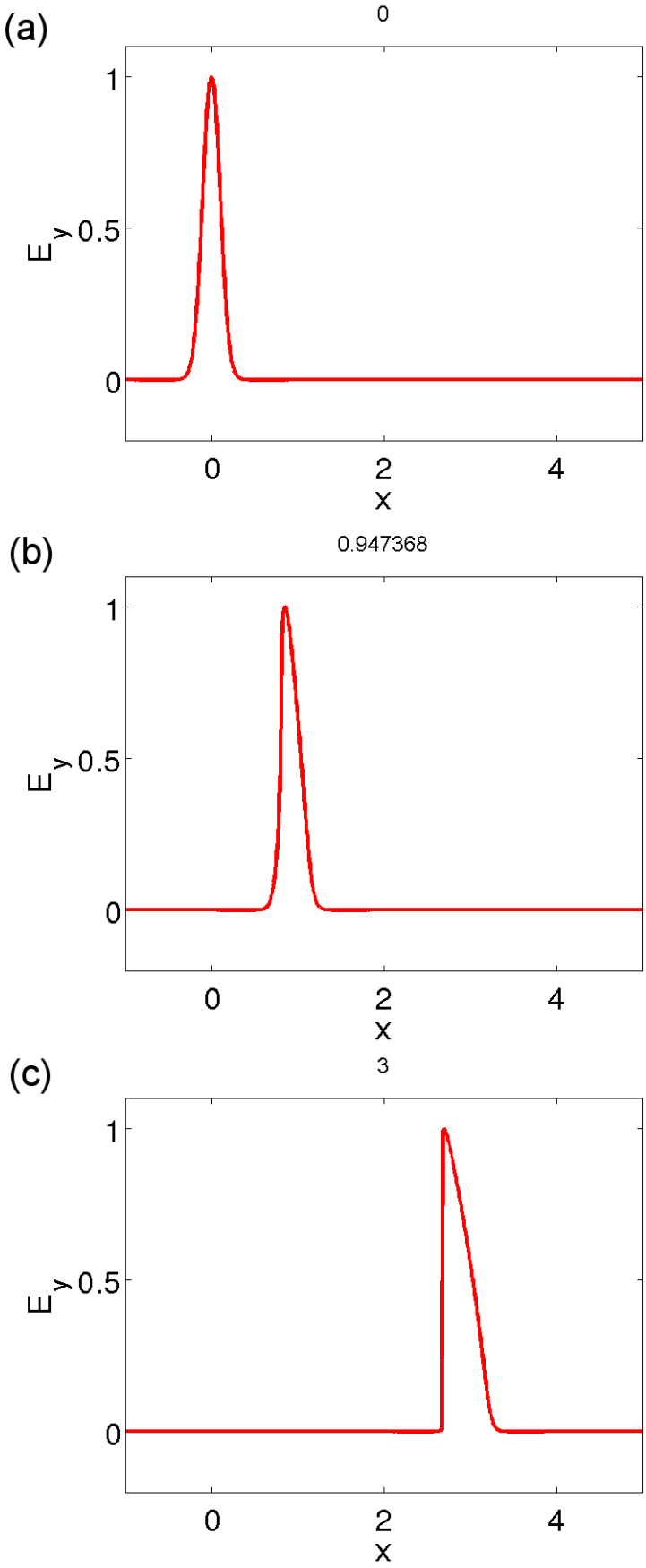



Figure 5.19: Snapshots of the Gaussian pulse (5.4.8) with $\chi = 0.08$, $\sigma = 0.1$, $E_0 = 1$, $x_0 = 0$, $x \in [-1, 5]$ and the final simulation time $t_{\text{final}} = 3$, at time (a) $t = 0$, (b) $t = 0.947$, (c) $t = 3$.

#### Zero Finding Routine for the Numerical Computation of $\mathbf{E}_z$

In our implementations we use an efficient modified regula falsi method to determine the zero of the function $f(E_z) := E_z - G(E_z)$, which is a modified version of the regula falsi scheme (see e.g. [113, Ch. 9.2]). Let $f \in C^2(D)$, $D \subset \mathbb{R}$ open, and $[a_0, b_0]$ be an interval in which a zero exists with $f(a_0) \cdot f(b_0) < 0$. Then we perform the following steps:

1. Bracketing in case the starting guess is far away from the actual zero. Define $z := \frac{1}{2}(a_m + b_m)$ with $m > 0$ and let the new interval $[a_{m+1}, b_{m+1}] = [a_{m+1}, z]$ or $[z, b_{m+1}]$, depending on $\operatorname{sign}(f(z) \cdot f(a_m))$.

2. Regula falsi step. Define

$$x := \frac{a_m f(b_m) - b_m f(a_m)}{f(b_m) - f(a_m)}.$$

3. Newton step with overshoot. Let

$$y := x - 2\frac{b_m - a_m}{f(b_m) - f(a_m)}.$$

4. We distinguish the following cases:

Case 1: If $y \notin [a_m, b_m]$: No change.

Case 2: If $f(x) \cdot f(y) > 0$: Choose either $[a_m, x]$ or $[x, b_m]$ as a new interval, depending on $\operatorname{sign}(f(x) \cdot f(a_m))$.

Case 3: If $f(x) \cdot f(y) < 0$: Either $[x, y]$ (if $y > x$) or $[y, x]$ (if $y < x$).

In cases 2 and 3 we add a bisection step to increase the convergence speed.
The $y$ in step 2 can be looked upon as a "trial balloon". For smooth functions we are very soon, i.e. for some $m > m_0$, where $m_0$ is a small integer, in case 1 of step 3 so that the computational effort remains small.

Step 1 ensures stability and convergence of the method and step 2 increases its speed with a total convergence order of 2, in contrast to the simple regula falsi scheme which can be very slow. It is thus as fast as Newton's method but avoids the computation of the derivative $f'$, and it is faster than the secant or regula falsi methods. The idea of the modified regula falsi scheme can be compared to the idea of the Illinois variant of the regula falsi method, see e.g. [118]. Our modified regula falsi method is adjusted such that it uses `fzero` if the initial guess is a scalar and not an interval. All simulations were performed on an Intel Core 2 processor.

### 5.4.2 Zero Finding Routine for the Numerical Computation of $\mathbf{E}_2$

In order to compute the exact numerical flux and the wave speeds $s_1$ and $s_2$ we need the zero $\mathbf{E}_2$ of the function $\gamma$ in (5.2.67) which we determine via a zero finding routine. $\gamma$ is continuously differentiable, as already mentioned in section 5.2.4. We therefore expect that a zero finding routine does converge.
The modified regula falsi method which we used in section 5.4.1 to compute a Gaussian pulse solution is designed for real scalars, like e.g. fzero by Matlab. Yet in our case, $\mathbf{E}_2$ is a two-dimensional vector. We have to run the routine for $E_{2,y}$ and $E_{2,z}$, respectively. Besides the question of time, we know that solving the function $\gamma$ in (5.2.67) for $\mathbf{E}_2$ also requires its norm $|\mathbf{E}_2|^2 = E_{2,y}^2 + E_{2,z}^2$. Therefore we cannot apply the regula falsi method componentwise, since we have a coupled problem in $E_{2,y}$ and $E_{2,z}$. We decided to use Muller's method (see e.g. [113, Ch. 9.5.2]) which is able to determine complex zeros as

| | convergence order | total time | no. of iterations | no. of function evaluations | $\lvert f(\mathbf{E}_2^{\text{appr}})\rvert$ |
|---|---|---|---|---|---|
| modified regula falsi | 2 | 0.005973 | 5 | 11 | `4.26e-14` |
| regula falsi | 1 | 0.006474 | 5 | 11 | `4.26e-14` |
| secant method | golden ratio $\approx 1.6$ | 0.006705 | 7 | 9 | `3.73e-14` |
| `fzero` | at most 2 | 0.008126 | 6 | 31 | 0 |

Table 5.3: Comparison of the performance of different zero finding routines, where we have chosen the following (arbitrary) settings: $\mathbf{E}_L = (2,1)^T$, $\mathbf{E}_R = (1,1)^T$, $\mathbf{B}_L = (-1,1)^T$, $\mathbf{B}_R = (1,0)^T$, $\epsilon = 1$, $\mu = 1$, $\chi = 1/3$. The initial interval is $[-5,8]$ and the initial guess for fzero is 0. These guesses are not close to the zero. $\lvert f(\mathbf{E}_2^{\text{appr}})\rvert$ is the function value at the obtained approximative zero $\mathbf{E}_2^{\text{appr}}$.

well. We thus write the vector $\mathbf{E}_2 = (E_{2,y}, E_{2,z})^T$ as a complex number $E_2 := E_{2,y} + \mathrm{i}E_{2,z}$; the function $\gamma$ becomes a complex scalar function $\tilde{\gamma}$ in dependency of $E_2$. We then apply Muller's method to find the zero $E_2$ of $\tilde{\gamma}$ so that $E_{2,y} = \ oRe(E_2)$ and $E_{2,z} = \mathrm{Im}(E_2)$. Whenever one of the components $E_y$ or $E_z$ is zero, we can use fzero or regula falsi to compute $E_{2,y}$ or $E_{2,z}$, since then the coupling nature is absent. In that case, we have compared fzero with Muller's method and obtain a maximal difference of around $10^{-16}$. We have also run a simulation with the exact numerical flux from section 5.3.2, once using fzero and once using Muller's method. We chose $\chi = 10^{-16}$, $p = 3$, $K = 50$, $t_{\text{final}} = 4$ and the Gaussian pulse with the settings from above. The total simulation time was about 2 hours for fzero, and about 17 minutes for Muller's method (on a single CPU Inter Core 2). The maximal $L^1$-error was in both cases about $8 \cdot 10^{-3}$.
Muller's method is a modified secant method. In contrast to the secant method it needs three initial guesses and its corresponding three function values, and then interpolates quadratically. For completeness we give Muller's method here.

1. Find three initial guesses $x_{i-2}$, $x_{i-1}$, $x_i$.
2. Evaluate $\tilde{\gamma}(x_{i-2})$, $\tilde{\gamma}(x_{i-1})$ and $\tilde{\gamma}(x_i)$.
3. Define
$$q := \frac{x_i - x_{i-1}}{x_{i-1} - x_{i-2}}.$$
4. Define the quantities
$$A := q\tilde{\gamma}(x_i) - q(1+q)\tilde{\gamma}(x_{i-1}) + q^2\tilde{\gamma}(x_{i-2}),$$
$$B := (2q+1)\tilde{\gamma}(x_i) - (1+q)^2\tilde{\gamma}(x_{i-1}) + q^2\tilde{\gamma}(x_{i-2}),$$
$$C := (1+q)\tilde{\gamma}(x_i).$$
5. The new value is
$$x_{i+1}^{(1,2)} = x_i - (x_i - x_{i-1})\frac{2C}{B \pm \sqrt{B^2 - 4AC}}.$$

In our simulations we use information from the linear numerical flux (5.3.8) to find the three initial guesses for Muller's method. We let

$$\mathbf{E}_0 := \frac{s_L}{s_L - s_R}(-s_R(\mathbf{E}_L - \mathbf{E}_R) + (\mathbf{B}_L - \mathbf{B}_R))$$

and define the three initial guesses

$$x_1 = E_{0,y} + \mathrm{i}E_{0,z},$$

$$x_0 = x_1 - \varepsilon,$$

$$x_2 = x_1 + \varepsilon,$$

where $\varepsilon > 0$. Recall that $s_L = \lambda_1^-(\mathbf{E}_L)$ and $s_R = \lambda_1^+(\mathbf{E}_R)$. We also need a stopping criterion for Muller's method, which we define as follows.

- Choose a tolerance `tol`, a maximal iteration number `max_iter` and a maximal counter number `max_count`. Initially define the iteration step number `iter = 1` and a counter `count = 1`.
- Let $f_{\text{new}} = \min(\tilde{\gamma}(x_0), \tilde{\gamma}(x_1), \tilde{\gamma}(x_2))$ and denote the corresponding value $x_0$, $x_1$ or $x_2$ by $x_{\text{old}}$.
- if $|f_{\text{new}}| \leq$ tol

  $E_2^{\text{new}} = x_{\text{old}}$

  else

  while $|f_{\text{new}}| >$ tol or `count == max_count`

  Do a Muller's step and obtain the two new values $x_{i+1}^{(1,2)}$.

  Evaluate $\tilde{\gamma}(x_{i+1}^{(1,2)})$ and choose $f_{\text{new}} = \min(\tilde{\gamma}(x_{i+1}^{(1)}), \tilde{\gamma}(x_{i+1}^{(2)}))$ with the corresponding value $x_{i+1}^{(1)}$ or $x_{i+1}^{(2)}$, which gives the new starting guess $E_2^{\text{new}}$ for the next Muller step, from which we determine three new initial guesses $x_{i-2}$, $x_{i-1}$, $x_i$ and their three corresponding function values.

  Increase `iter = iter`+1.

  if `iter == max_iter`

  reduce the tolerance by `number` so that `tol = tol*number`. Set `count = count`+1. Repeat if necessary until `count = max_count`.

  end

  end

  end.

The last if-loop enforces the ending of the while-loop in case Muller's method cannot find a function of the given tolerance. Note that until then the tolerance might be too high in order to talk of convergence of the method. It is meant as a sure stopping criterion for Muller's method.
In our simulations we have chosen $\varepsilon = 0.01$, `tol = 1e-15`, `max_iter = 10`, `max_count = 3` and `number = 5/count`.

The wave speeds $s_1$ and $s_2$ are given as follows (see section 5.2.4):

- We need to identify the upwind and downwind value.

- If $|\mathbf{E}_L| < |\mathbf{E}_R|$, then $|\mathbf{E}_L|$ is the upwind value and we have a 2-shock. The speed $s_1$ of the left 2-shock and the speed $s_2$ of the right 2-shock are given according to (5.2.60). In this case, we also have $\mathbf{E}_1 = \frac{|\mathbf{E}_L|}{|\mathbf{E}_2|}\mathbf{E}_2$ and $\mathbf{E}_3 = \frac{|\mathbf{E}_R|}{|\mathbf{E}_2|}\mathbf{E}_2$.

- If $|\mathbf{E}_L| > |\mathbf{E}_R|$, $|\mathbf{E}_R|$ is the upwind value. Then $\mathbf{E}_1 = \frac{|\mathbf{E}_R|}{|\mathbf{E}_2|}\mathbf{E}_2$ and $\mathbf{E}_3 = \frac{|\mathbf{E}_L|}{|\mathbf{E}_2|}\mathbf{E}_2$. In this case we have a 2-rarefaction wave; we choose $s_1$ to be the speed of the head of the left rarefaction fan, that is, $s_1 = \lambda_2^-(|\mathbf{E}_L|)$, and $s_2 = \lambda_2^+(|\mathbf{E}_R|)$ is the speed of the head of the right rarefaction fan.

### 5.4.3 Comparison of the Numerical Fluxes and the Exact Numerical Flux

In this section we compare the following numerical fluxes with each other: three linear numerical fluxes (the Lax-Friedrichs flux from section 5.3.2, the Richtmyer flux from section 5.3.2 and the linear flux from section 5.3.2), the exact numerical flux from section 5.3.2 and the HLL-like flux from section 5.3.2. For all fluxes we compute the $L^1$-error between the exact solution $E_y$, i.e. the Gaussian pulse (5.4.5), and its approximation $E_{y,h}$, produced via the DG method, over the entire time, for increasing polynomial order $p$ and decreasing maximal edge length $h$. For implementation, we approximate the $L^1$-error by

$$e(E_{y,h} - E_y) := h \sum_i \sum_j |(E_{y,h} - E_y)_{ij}|,$$

where $(E_{y,h} - E_y)_{ij}$ denotes an element of the $N_p \times K$-array $E_{y,h} - E_y$. The simulation time in case of the implicitly given Gaussian pulse (5.4.5) was $t_{\text{final}} = t_{\max} - 0.1 = 0.84$, the domain was chosen to be $\Omega = [-1, 5]$, and we set $\chi = 10^{-16}$ or $\chi = 0.08$ with $\epsilon = \mu = 1$ in order to compare the behavior of the scheme for the linear and nonlinear case, respectively. The time step size was $\Delta_t = 0.4 \min(\Delta_x)$, the Gaussian pulse width was $\sigma = 0.1$, its center $x_0 = 0$ and its amplitude $E_0 = \sqrt{\epsilon/\mu}$. In the power series (5.4.7) we chose $n = 15$. For the regula falsi algorithm we have set the tolerance on the function value $f$ at the approximate zero $x$ and on $x$ itself to $10^{-18}$, respectively, with a maximal iteration number of 1000.
Table 5.4 gives an overview of the numerical results with the different numerical fluxes for $\chi = 10^{-16}$ (once with the implicitly defined Gaussian pulse (5.4.5), where zero approximation is needed, and once with the explicitly given Gaussian pulse (5.4.9), which does not need the computation of a zero) and $\chi = 0.08$. In figures 5.20 to 5.30 the convergence results are plotted in logarithmic scale. We observe an error behavior of approximately $O(h^3)$ to $O(h^4)$ for all numerical fluxes if $\chi = 0.08$ (for varying polynomial order $p = 3, \ldots, 9$), and $p$-convergence $O(h^{p+1})$ for $\chi = 10^{-16}$, which we would expect from theory. In the following we present details.

Numerical tests suggest that the convergence order 3 to 4 for $\chi = 0.08$, which is not according to theory, may arise due to two reasons. First, it may be due to the usage of a linear flux, since, for $\chi = 0$, the order of convergence is between $O(h^4)$ to $O(h^5)$, as figure 5.23 illustrates. We first observe $p$-convergence, and then the order of convergence stagnates, which is due to the Runge-Kutta time integration, which is of order 4. Figures 5.23, 5.24, 5.26, 5.27 and 5.29 show $p$-convergence in case of $\chi = 10^{-16}$ as well, whereas for $\chi = 0.08$, this behavior is lost, see figures 5.20, 5.21, 5.22, 5.28 and 5.30. Second, this convergence behavior may also be due to an error coming from the approximation to the Gaussian pulse for the nonlinear Kerr-system as given in (5.4.5). The following two numerical tests indicate to this. First, if we make a very simple linear test with $\chi = 0$,

using initially a standing wave in an empty cavity, i.e.

$$E_y = -\sin(k(x-a))\cos(\omega t),$$

$$H_z = \cos(k(x-a))\sin(\omega t),$$

where $L := b - a$ is the length of the simulation interval $[a, b]$, $k = \frac{m\pi}{L}$ is the wave number, $m$ the number of modes and $\omega = \sqrt{\frac{\mu}{\epsilon}}k$ the frequency, we approximately observe the well-known convergence behavior $O(h^{p+1})$, again taking into account that we use a $(4, 5)$-Runge-Kutta method. We have chosen the Lax-Friedrichs flux. See figure 5.25 for a plot. We have set $a = -1$, $b = 1$ and $m = 1$. Secondly, we repeat this test with $\chi = 0$ and use initially a Gaussian pulse for the linear problem as given in (5.4.9), which does not need a zero finding routine to be established. We have again used a Lax-Friedrichs flux. We observe a convergence behavior that is similar to the one for the standing wave, as figure 5.26 illustrates.

Convergence plots for the exact numerical flux from section 5.3.2 are shown in figures 5.27, where we have set $\chi = 10^{-16}$ and $\chi = 0.08$. We have used fzero and Muller's method to compute the zero $\mathbf{E}_2$ of the function $\gamma$ in (5.2.67). Furthermore, now we have $\Delta_t = 0.2\min(\Delta_x)$, but a bigger time step size as e.g. $\Delta_t = 0.4\min(\Delta_x)$ would work as well. Yet, with a smaller time step size we expect a smaller error coming from the Runge-Kutta time integration, and thus a better convergence behavior, which is what we indeed obtain; see figures 5.20 and 5.24 for example. Throughout, the maximal error between a zero found with fzero and a zero found with Muller's method was around $10^{-16}$. The total simulation time needed to produce the results with fzero was about 5.5 days to more than 7 days on our computer (a single CPU Inter Core 2). With Muller's method, the same test system needs about 2 days. For $\chi = 10^{-16}$, we observe $p$-convergence; the stagnation starting with polynomial order $p = 6$ is due to the Runge-Kutta scheme. We tested the error coming from the approximation to the implicitly defined Gaussian pulse (5.4.5), for $\chi = 10^{-16}$, with the different numerical fluxes. Thus, we have run simulations with the exact numerical flux and computed the implicitly defined Gaussian pulse as given in (5.4.5), i.e. via regula falsi, and once we have run the same simulation with the explicitly given Gaussian pulse (5.4.9); see figures 5.26 and 5.23 as examples for $\chi = 10^{-16}$; in both cases, a Lax-Friedrichs flux was used. Again we find indications that the computation of the implicitly given Gaussian pulse is responsible for a slightly lesser accuracy and convergence order. Yet, the results with $\chi = 0.08$, where we find a convergence order of about 2.5 to 4, suggest there are also other reasons, possibly due to the nonlinear behavior which manifests itself stronger if $\chi$ $O(1)$, instead of $\chi$ 0.
For the HLL-like flux from section 5.3.2, figures 5.29 and 5.30 show the $L^1$-error for the same test system from above with $\chi = 10^{-16}$ and $\chi = 0.08$. Here, the total simulation time was about 5 hours (for $\chi = 10^{-16}$ and $\chi = 0.08$), which is an immense improvement in time to the exact flux. Furthermore, we have approximately the same convergence order. Additionally, we observe that the wave speeds $s_1^*$ and $s_2^*$ are good approximations to the exact wave speeds $s_1$ and $s_2$. We will say more about this in the following subsection.

Throughout our simulations and tests we observed that we are often in the situation $|s_1 - s_L| \leq \text{tol}$ and $|s_2 - s_R| \leq \text{tol}$, where $s_L = -s_R$ in our case. Numerical results suggest that the maximum of the differences $|s_1 - s_L|$ and $|s_2 - s_R|$, respectively, is of the magnitude of $|\chi|$; for example, if $\chi = 0.08$, $\max|s_1 - s_L| \approx 0.07$. Simulations also suggest that this occurs in grid points around the peak of the Gaussian pulse, whereas in grid points away from the peak, we have $|s_1 - s_L| \sim 10^{-16}$. An intuitive explanation comes from the fact that the self-focusing of the Gaussian pulse is exactly observed around the peak. In this

$\chi = 10^{-16}$, implicitly defined Gaussian pulse (5.4.5):

| numerical flux | order of convergence | error behavior | running time |
|---|---|---|---|
| exact flux (fzero) | $p$-convergence | $10^{-9}$ | $> 7$ days |
| linear fluxes | $p$-convergence | $10^{-11}$ | 1 h |
| HLL flux | $p$-convergence | $10^{-11}$ | 2 h |

$\chi = 10^{-16}$, explicitly given Gaussian pulse (5.4.9):

| numerical flux | order of convergence | error behavior | running time |
|---|---|---|---|
| exact flux (fzero) | $p$-convergence | $10^{-10}$ | 5.5 days |
| linear fluxes | $p$-convergence | $10^{-11}$ | 1 h |
| HLL flux | no tests | – | – |

$\chi = 0.08$:

| numerical flux | order of convergence | error behavior | running time |
|---|---|---|---|
| exact flux (muller) | 2.5 to 4 | $10^{-5}$ | 2 days |
| linear fluxes | 2.6 to 4 | $10^{-5}$ | 2.2 h |
| HLL flux | 2.5 to 4 | $10^{-5}$ | 2.7 h |

Table 5.4: Overview of the most important numerical results for $\chi = 10^{-16}$ and $\chi = 0.08$ with the different numerical fluxes: exact flux, linear fluxes (Lax-Friedrichs, linear flux from (5.3.7) or (5.3.9), Richtmyer flux) and HLL-like flux. In case of the exact numerical flux, we have written in brackets if fzero (fzero) or Muller's method (muller) was used to compute $\mathbf{E}_2$ of $\gamma$ of (5.2.67). With $p$-convergence we mean that the convergence behavior of the $L^1$-error looks like $p$-convergence at the beginning and then stagnates for certain polynomial degrees, resulting in an order of convergence of 4, coming from the Runge-Kutta time integration, which is of order 4.

case, whenever $|s_1 - s_L| \leq \text{tol}$ (then $|s_2 - s_R| \leq \text{tol}$ holds also), we are close to the linear case, and the corresponding Riemann problem is linear, thus giving a linear numerical flux.

Summarized we find the following: All numerical fluxes produce convergence orders that are approximately similar to each other. For $\chi = 10^{-16}$, we find $p$-convergence. For $\chi = 10^{-5}$, the order of convergence is between 2.6 and 4. After a thorough comparison of the five different numerical fluxes presented here, we think that using a linear flux is a suitable choice with respect to efficiency. It is the fastest among the numerical fluxes, while giving the same convergence results. Furthermore we have observed in our simulations that we are often in the linear case, that is, we have $s_1 = s_L$ and $s_2 = s_R$. Several numerical tests suggest that the not optimal convergence order between 2.6 and 4 is a consequence of the approximation to the implicitly defined Gaussian pulse (5.4.5), other numerical error sources (like e.g. Muller's method or fzero), or simply the nonlinear nature of the problem at hand.

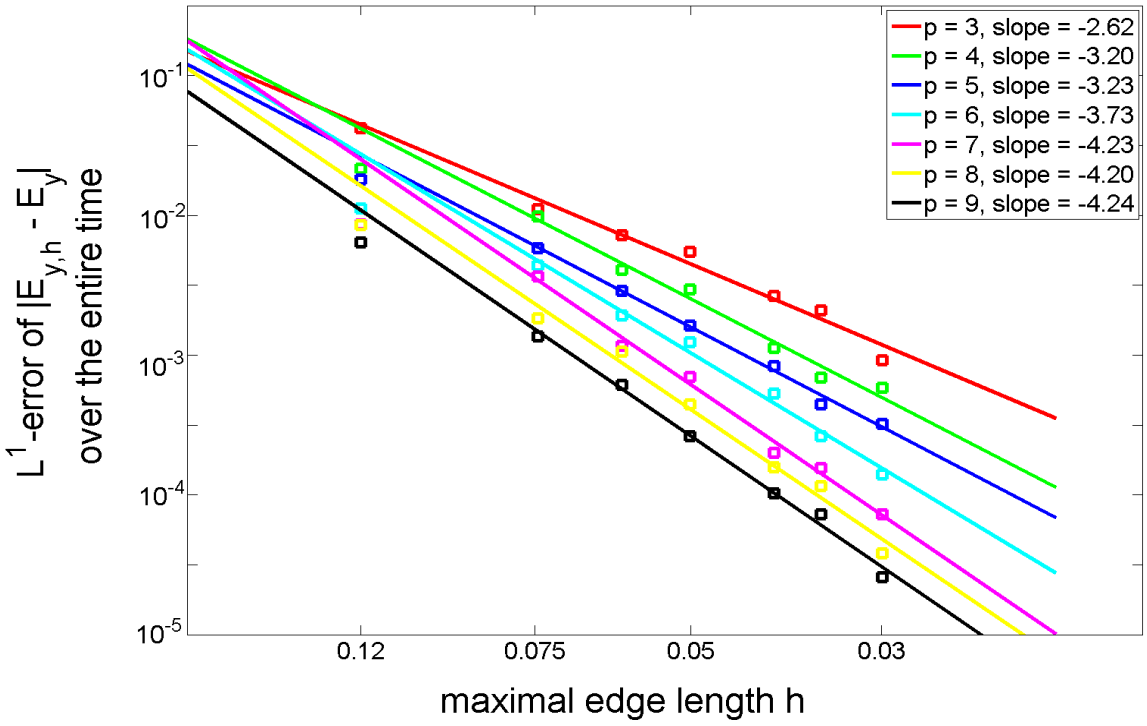



Figure 5.20: $L^1$-error plot for $E_y$ for polynomial order $p \in \{3, \ldots, 9\}$ and number of elements $K = 50, 80, 100, 120, 150, 170, 200$ and with the linear flux from equation (5.3.8); initially the implicitly defined Gaussian pulse (5.4.5) is chosen with width $\sigma = 0.1$, center $x_0 = 0$, amplitude $E_0 = 1$ and number of coefficients $n = 15$, where the Gaussian pulse was computed according to the exact solution for the one-dimensional Kerr-problem. The time step size was $\Delta_t = 0.4 \min(\Delta_x)$. The simulation domain is $[-1, 5]$, and the material parameters are $\chi = 0.08$, $\mu = 1$, $\epsilon = 1$ and the final simulation time was $t_{\text{final}} = 0.84$.

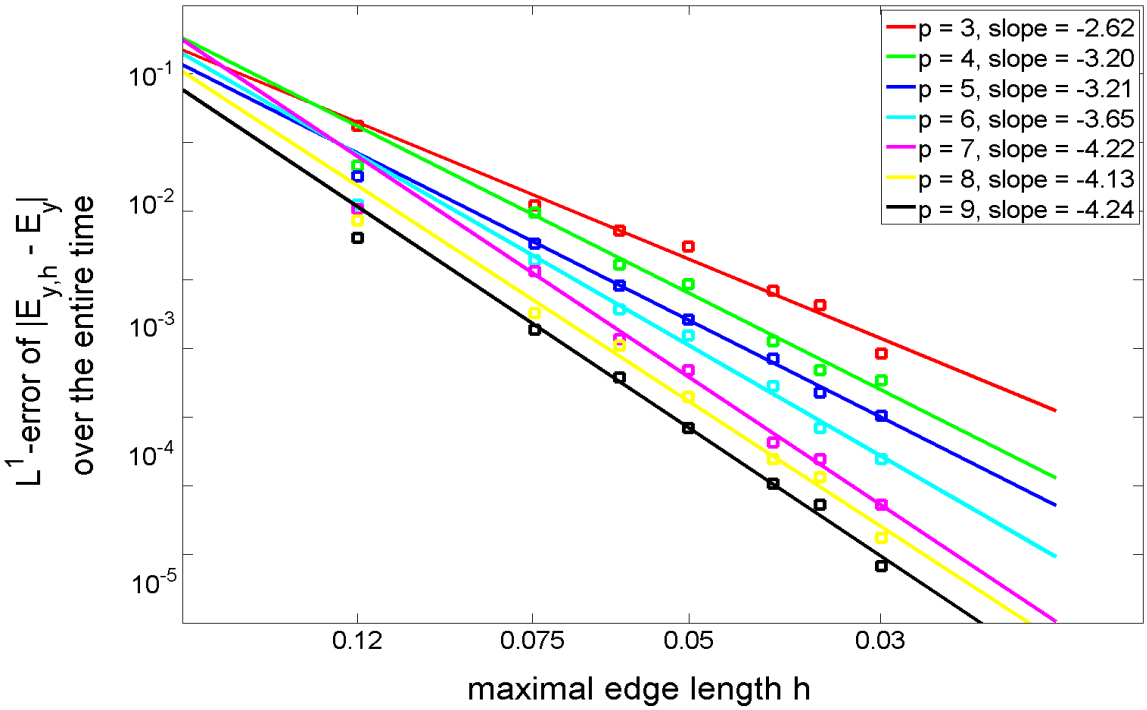



Figure 5.21: $L^1$-error plot for the implicitly defined Gaussian pulse as in figure 5.20 with a Lax-Friedrichs flux and with $\chi = 0.08$; the total simulation time was $t_{\text{final}} = 0.84$ over the domain $[-1, 5]$. The time step size was $\Delta_t = 0.4 \min(\Delta_x)$.

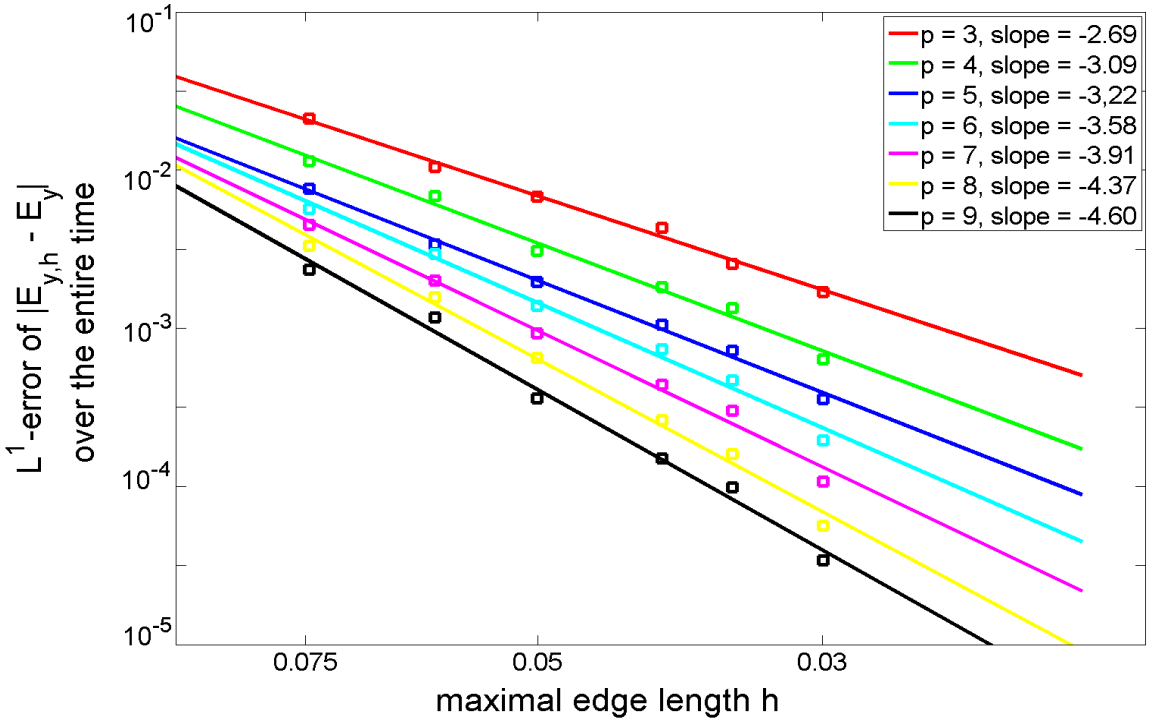



Figure 5.22: $L^1$-error plot for the implicitly defined Gaussian pulse as in figure 5.21 with a Richtmyer flux as given in section 5.3.2.

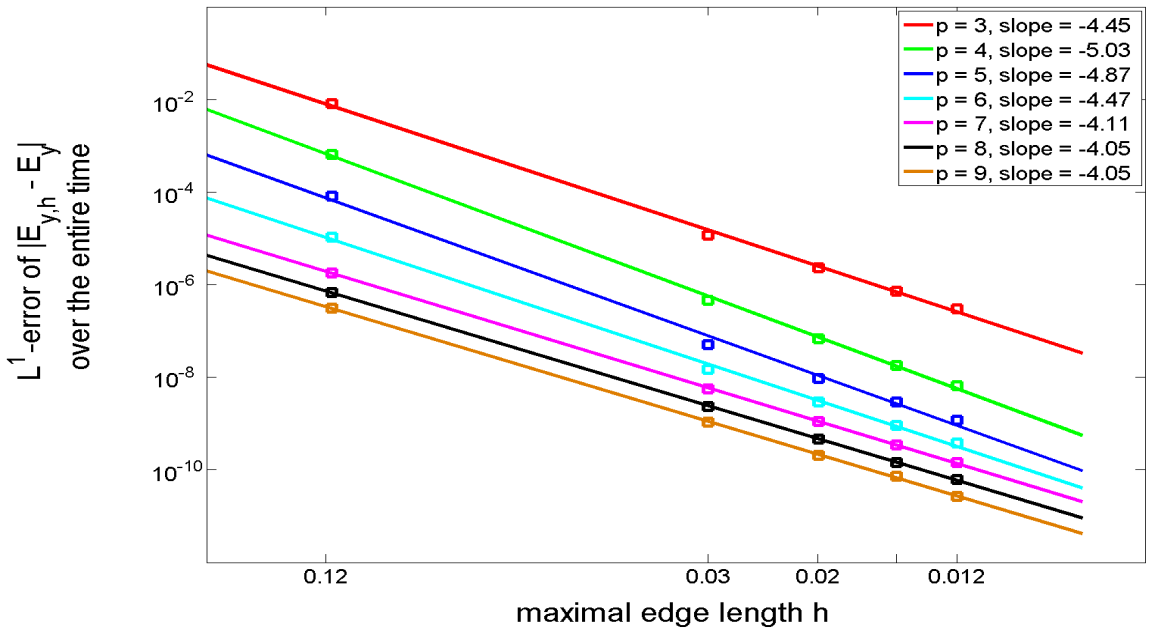



Figure 5.23: $L^1$-error plot for the implicitly given Gaussian pulse (5.4.5) for $p \in \{3, \dots, 9\}$ and $K = 50, 200, 300, 400, 500$ with a Lax-Friedrichs flux and with $\chi = 0$; the total simulation time was $t_{\text{final}} = 4$ over a domain $[-1, 5]$; for $\chi = 0$ the maximal allowed time $t_{\max}$ is infinity. The time step size was $\Delta_t = 0.4 \min(\Delta_x)$.

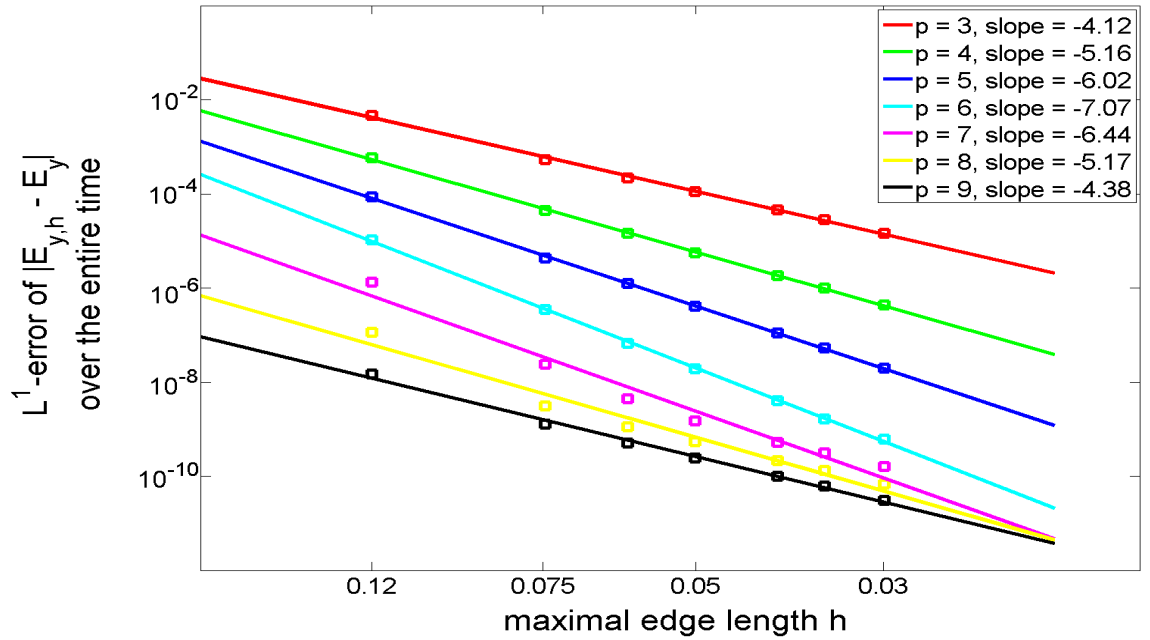



Figure 5.24: $L^1$-error plot for the explicitly given Gaussian pulse (5.4.9) with the linear flux (5.3.8) and with $\chi = 10^{-16}$, $K = 50, 80, 100, 120, 150, 170, 200$. The total simulation time was $t_{\text{final}} = 4$ over a domain $[-1, 5]$; here, the time step size was $\Delta_t = 0.2 \min(\Delta_x)$. With a smaller time step size we expect a better convergence behavior and smaller error, which is what we indeed obtain.

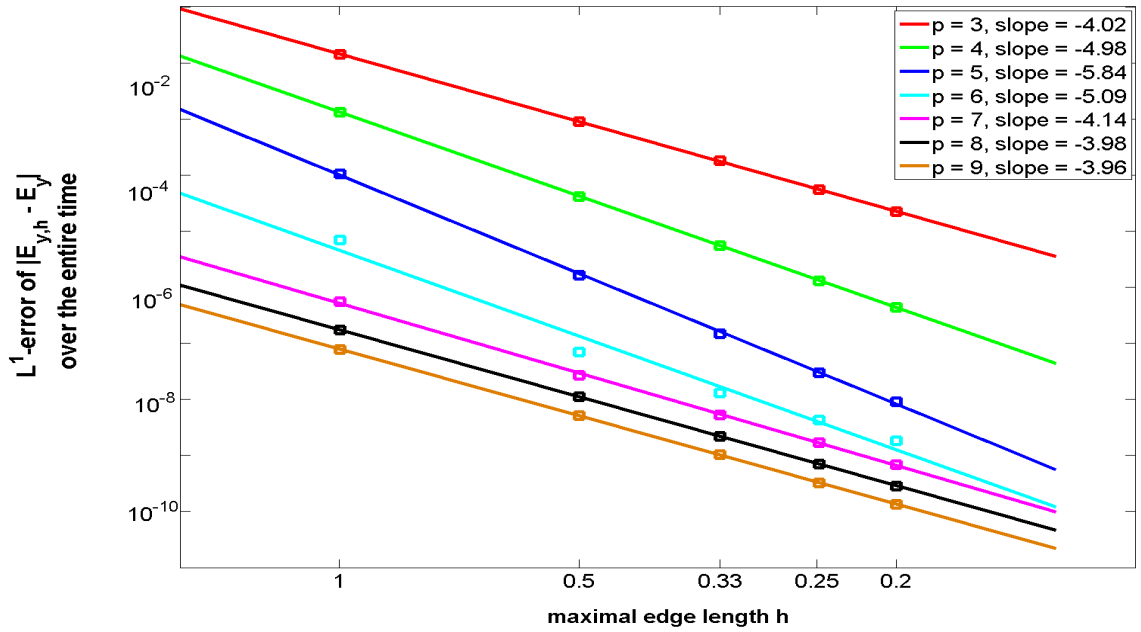



Figure 5.25: $L^1$-error for a standing wave in an empty cavity for $p \in \{3, \dots, 9\}$ and $K = 2, 4, 6, 8, 10$ with a Lax-Friedrichs flux and $\chi = 0$. The simulation domain was $[-1, 1]$ and the simulation time was $t_{\text{final}} = 5$. The time step size was $\Delta_t = 0.4 \min(\Delta_x)$.

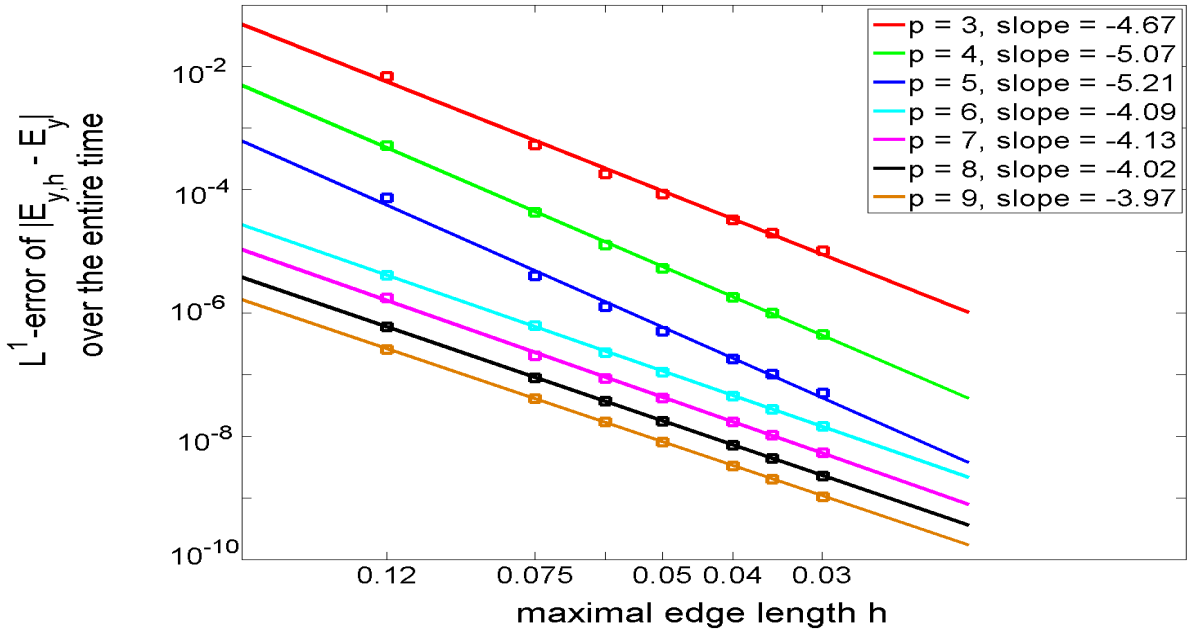



Figure 5.26: $L^1$-error for the explicitly defined Gaussian pulse according to (5.4.9) for the linear problem with $\chi = 0$, $p \in \{3, \dots, 9\}$ and $K = 50, 80, 100, 120, 150, 170, 200$, where no zero approximation is needed to compute the Gaussian pulse. A Lax-Friedrichs flux was used. The simulation domain was $[-1, 5]$ and the simulation time was $t_{\text{final}} = 4$.

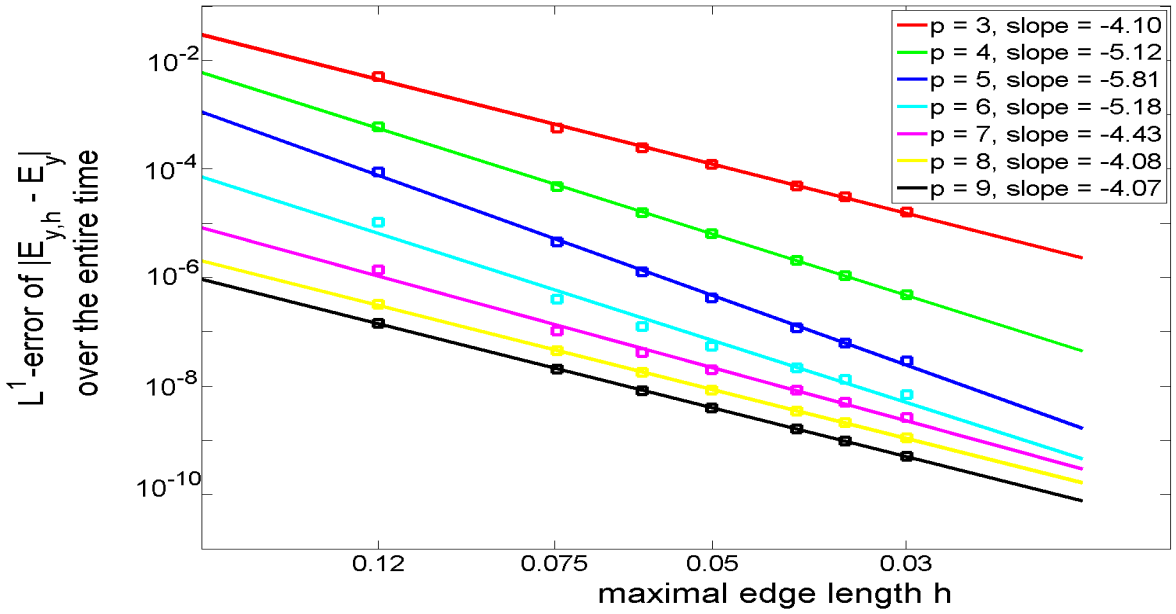



Figure 5.27: $L^1$-error for the explicitly given Gaussian pulse (5.4.9) with the exact numerical flux from section 5.3.2 and with $\chi = 10^{-16}$, $p \in \{3, \dots, 9\}$ and $K = 50, 80, 100, 120, 150, 170, 200$, where fzero was used to compute the zero $\mathbf{E}_2$ of the function $\gamma$ in (5.2.67). The simulation domain was $[-1, 5]$ and the simulation time was $t_{\text{final}} = 0.84$. The total running time of the simulation was about 5.5 days.

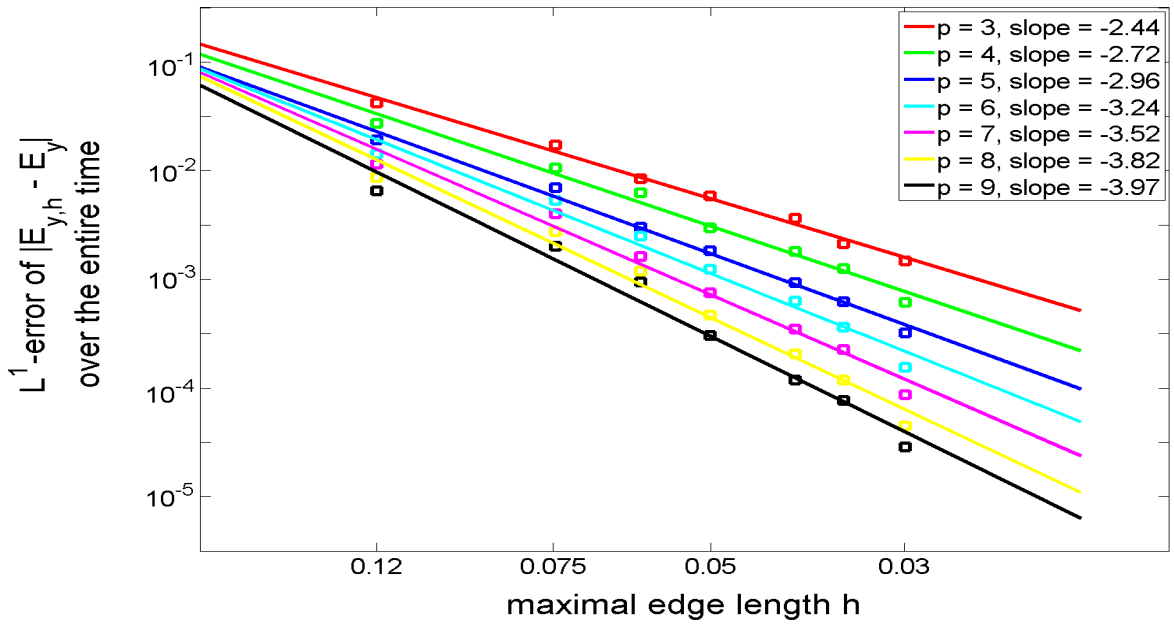



Figure 5.28: $L^1$-error plot as in figure 5.27, but now with $\chi = 0.08$. The exact numerical flux with Muller's method was used to compute the zero $\mathbf{E}_2$ of the function $\gamma$ in (5.2.67). The total running time was about 2 days.

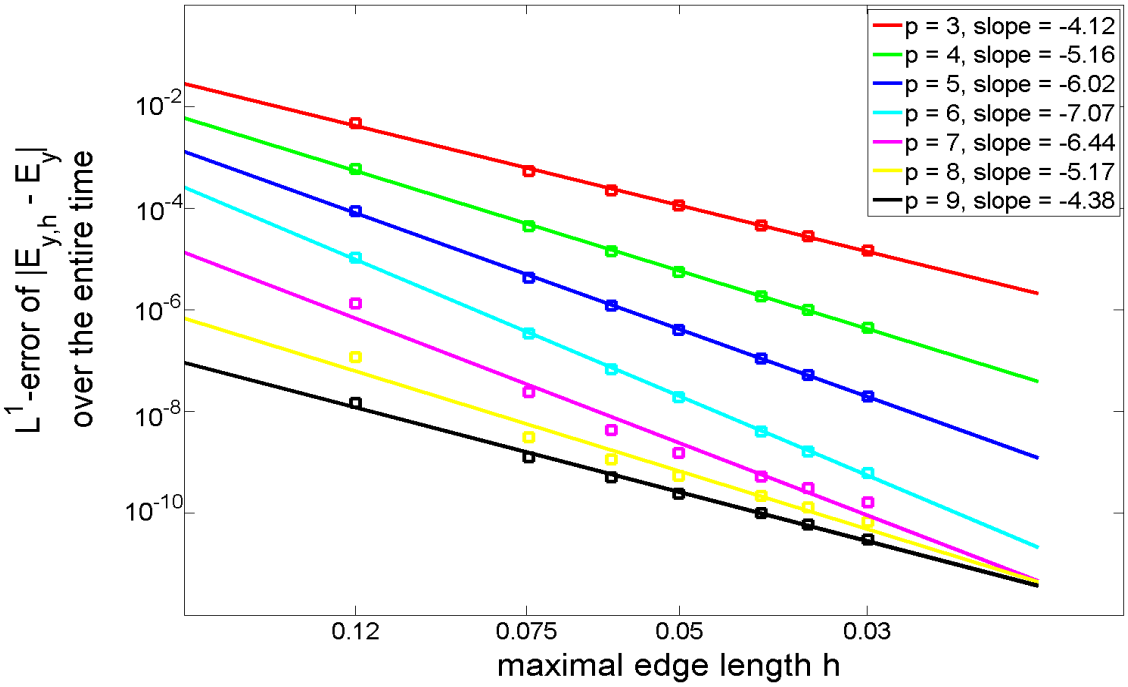



Figure 5.29: $L^1$-error for the implicitly defined Gaussian pulse (5.4.5) with $\chi = 10^{-16}$ and with the HLL-like flux of section 5.3.2. The simulation domain was $[-1, 5]$ and the running time was $t_{\text{final}} = 0.84$.

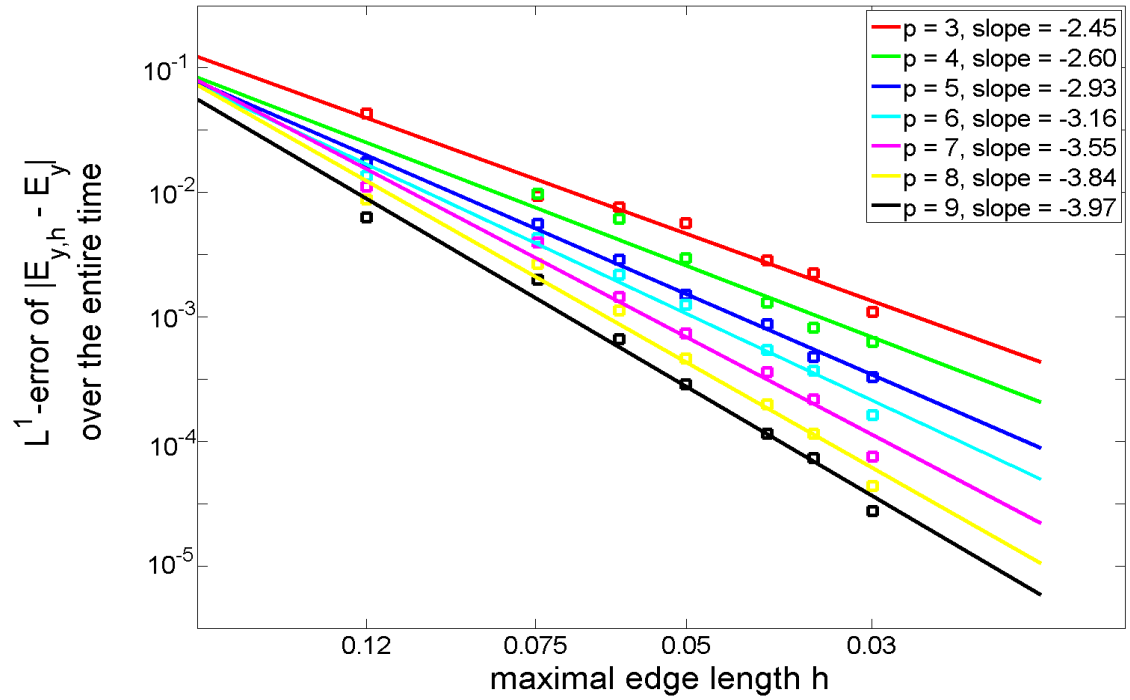



Figure 5.30: $L^1$-error for a Gaussian pulse as in figure 5.29, but with $\chi = 0.08$.

### 5.4.4 Comparison Between the Exact Wave Speeds and the Wave Speed Estimates

We compare the exact wave speeds $s_1$ and $s_2$ with the wave speed estimates $s_1^*$ and $s_2^*$ by computing the $l^1$-error

$$e(s_i, s_i^*) := h \sum_j (|(s_i - s_i^*)_j|) \qquad i = 1, 2,$$

where $(s_i - s_i^*)_j$ denotes the $j$-th component of the vector $s_i - s_i^*$ $(i = 1, 2)$. $s_1^*$ and $s_2^*$ were given in (5.3.21), and we compute $s_1$ and $s_2$ as follows:

In the left region:
If $|\mathbf{E}_2| > |\mathbf{E}_L|$, we have a left rarefaction, and $s_1$ is chosen to be the speed of the head of the rarefaction fan, that is, $s_1 = \lambda_2^-(|\mathbf{E}_L|)$.
On the other hand, if $|\mathbf{E}_2| \leq |\mathbf{E}_L|$, we have a left shock; $s_1$ is given in (5.2.60).

In the right region, we proceed analogously. If $|\mathbf{E}_2| > |\mathbf{E}_R|$, we have a rarefaction wave, and $s_2 = \lambda_2^+(|\mathbf{E}_R|)$. If $|\mathbf{E}_2| \leq |\mathbf{E}_R|$, $s_2$ is given via (5.2.60).

Table 5.5 shows the numerical results for a traveling Gaussian pulse, for $\chi = 10^{-18}$ and for $\chi = 0.08$. We see that the error is of a magnitude around $10^{-4}$ for $\chi = 0.08$, hinting again to a reason for the convergence order of 3 to 4 of the RKDG method with a linear flux, as we observed in the previous section. Furthermore, it is $e(s_1, s_1^*) = e(s_2, s_2^*)$ and $e(s_1, s_L) = e(s_2, s_R)$. We also observe that for $\chi = 0.08$ we indeed have $s_1 \neq s_L$ and $s_2 \neq s_R$ with $e(s_1, s_L) \sim 10^{-2}$, that is, we have truly a nonlinear problem. For $\chi = 10^{-18}$ we obtain $e(s_1, s_1^*) = 0$ and $s_1 = s_L$, as in the linear case. Figure 5.31 shows the maximum of the error $e(s_1, s_1^*)$ over the entire time, and table 5.6 shows $e(s_1, s_L)$ over the entire time for $p \in \{3, \cdots, 9\}$ and $K = 50, 80, 100, 120, 150, 170, 200$. We conclude that $s_1^*$ and $s_2^*$ are relatively good approximations to $s_1$ and $s_2$, and above all, $s_1^*$ and $s_2^*$ are given globally and their computation is fast.

$\chi = 0.08$, $K = 50$, $N = 3$:

| time | 0.1326 | 0.2652 | 0.3978 | 0.5304 | 0.6629 | 0.7955 |
|---|---|---|---|---|---|---|
| $e(s_1, s_1^*)$ | 7.8965e-05 | 1.2676e-04 | 2.2906e-04 | 2.8590e-04 | 2.6828e-04 | 1.6404e-04 |
| $e(s_1, s_L)$ | 0.0174 | 0.0177 | 0.0177 | 0.0173 | 0.0164 | 0.0151 |

$\chi = 10^{-18}$, $K = 50$, $N = 3$:

| time | 0.1326 | 0.2652 | 0.3978 | 0.5304 | 0.6629 | 0.7955 |
|---|---|---|---|---|---|---|
| $e(s_1, s_1^*)$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $e(s_1, s_L)$ | 0 | 0 | 0 | 0 | 0 | 0 |

Table 5.5: Error $e(s_1, s_1^*)$ and $e(s_1, s_L)$ for a Gaussian pulse (settings as in section 5.4.3) for $K = 50$, $N = 3$ and (top) $\chi = 0.08$, (bottom) $\chi = 10^{-18}$ at time $t = 0.1326, 0.2652, 0.3978, 0.5304, 0.6629, 0.7955$. For $\chi = 0.08$, it is $e(s_1, s_L) \neq 0$, which means we are indeed in the nonlinear regime. For $\chi = 10^{-18}$, $e(s_1, s_1^*) = e(s_1, s_L) = 0$ for all times, i.e. we are in the linear case, as expected.

| $e(s_1, s_1^*)$ | $K = 50$ | $K = 80$ | $K = 100$ | $K = 120$ | $K = 150$ | $K = 170$ | $K = 200$ |
|---|---|---|---|---|---|---|---|
| $p = 3$ | 0.0177 | 0.0189 | 0.0174 | 0.0172 | 0.0173 | 0.0172 | 0.0172 |
| $p = 4$ | 0.0195 | 0.0192 | 0.0181 | 0.0176 | 0.0174 | 0.0173 | 0.0173 |
| $p = 5$ | 0.0198 | 0.0183 | 0.0182 | 0.0175 | 0.0174 | 0.0174 | 0.0172 |
| $p = 6$ | 0.0215 | 0.0187 | 0.0178 | 0.0177 | 0.0174 | 0.0173 | 0.0173 |
| $p = 7$ | 0.0209 | 0.0185 | 0.0180 | 0.0177 | 0.0174 | 0.0173 | 0.0173 |
| $p = 8$ | 0.0207 | 0.0185 | 0.0179 | 0.0177 | 0.0175 | 0.0174 | 0.0189 |
| $p = 9$ | 0.0207 | 0.0187 | 0.0179 | 0.0177 | 0.0175 | 0.0176 | 0.0187 |

Table 5.6: Error behavior of $e(s_1, s_1^*)$ over the entire time for polynomial order $p \in \{3, \dots, 9\}$ and number of elements $K = 50, 80, 100, 120, 150, 170, 200$. The final simulation time was $t_{\text{final}} = 0.84$.

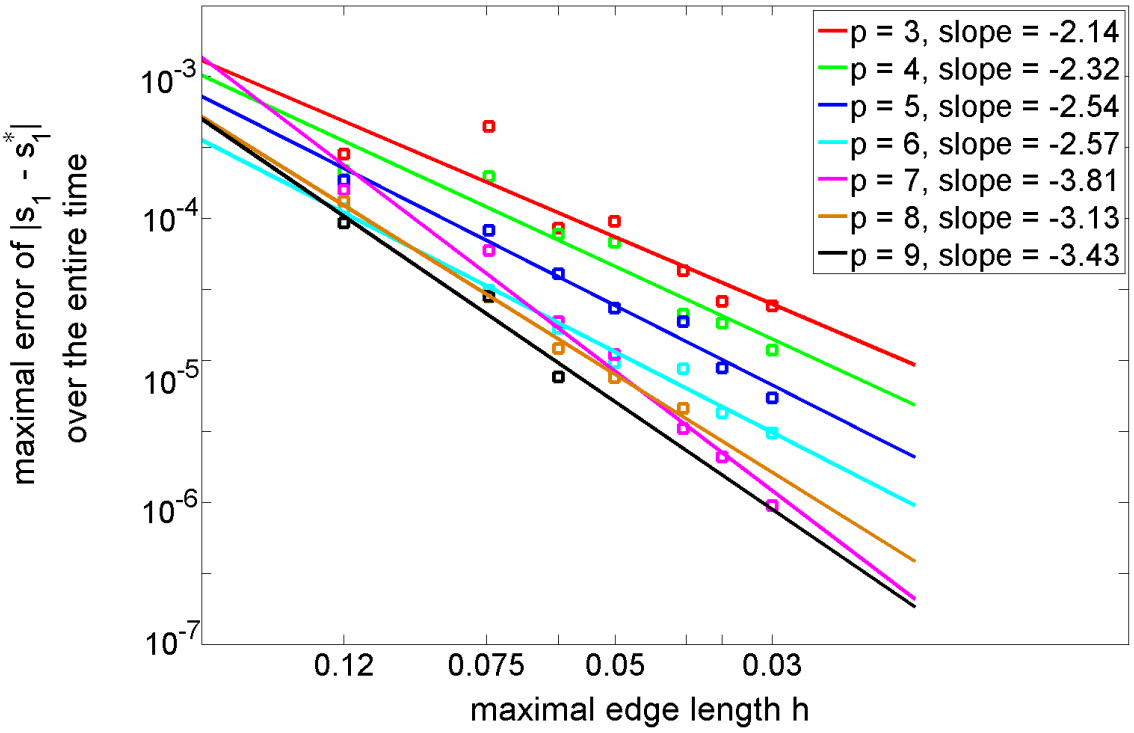



Figure 5.31: Plot of the error $\max(|s_1 - s_1^*|)$ for a Gaussian pulse according to (5.4.5) with $\chi = 0.08$, $p \in \{3, \dots, 9\}$ and $K = 50, 200, 300, 400, 500$. A Lax-Friedrichs flux was used. The simulation domain was $[-1, 5]$ and the simulation time was $t_{\text{final}} = 0.84$.

## 5.5 Summary

In this chapter we applied the Discontinuous Galerkin method to Kerr-nonlinear Maxwell's equations. The numerical flux was chosen as the solution of the corresponding Riemann problem. We presented the exact solution structure of the Riemann problem as proven in [71]. It consists of four simple waves, of which two are a contact discontinuity and the other two are either a shock or a rarefaction wave. We investigated the simple waves by exploring the Hugoniot Locus and the Riemann invariants, containing valuable information about the exact form of the corresponding simple wave.
Due to the nonlinear behavior of the Kerr system the numerical flux is not given globally, but locally, which leads to the necessity of solving the Riemann problem in each grid point to establish the numerical flux. This is extremely expensive in view of simulations, leading to the demand of more efficient alternatives. We presented here several global approximate numerical fluxes, the well-known and widely used Lax-Friedrichs flux, a Richtmyer flux, a linear flux which assumes to have only two contact discontinuities in spite of two contacts and two shocks or rarefaction waves, respectively, and an HLL-like flux, which assumes to have two contact discontinuities and two shocks. The wave speeds are given by wave speed estimates. In order to compare the performance of these fluxes we implemented a nearly exact numerical flux which is computed locally in every grid point, as the Kerr-Riemann problem requires, yet with the restriction of not taking into account rarefaction waves, but considering only shocks, as in case of the HLL-like flux. Yet we compute the exact shock speeds, and in case of a rarefaction wave we take the speed of the head of the rarefaction fan as the shock speed. Furthermore, the middle states are computed locally by determining the zero $\mathbf{E}_2$ via Muller's method, in every grid point. The results show that linear numerical fluxes require the least resources in time and memory, whereby producing a DG scheme of order 2.6 to 4, which is comparable to the order of convergence with the exact numerical flux and HLL-like flux. Therefore, a linear flux might be a good choice with respect to efficiency.

Chapter 6

# 6 Summary and Outlook

The first part of this dissertation was devoted to background theory which is important for the problems of our interest: the linear BOR Maxwell's equations and the Kerr-nonlinear Maxwell's equations. We first presented Maxwell's equations and some of their basic properties, like the constitutive relations between the electromagnetic fields, their behavior at interfaces and boundary conditions. We introduced the weak form of Maxwell's equations, which is indispensable for our purposes. Furthermore we rewrote Maxwell's equations as a conservation law and saw that they are a hyperbolic system of equations.
We then introduced the Runge-Kutta Discontinuous Galerkin method which was intended to be applied to the linear BOR Maxwell's equations and the Kerr-nonlinear Maxwell's equations. We have seen that the numerical flux is an essential ingredient of the DG method. We chose a numerical flux that is the solution to the corresponding Riemann problem. In this context we studied the general Riemann problem and its solution, and we introduced the notion of admissible, that is, physically relevant shock waves and rarefaction waves.

The second part of this work dealt with the application of the theory given in the first part to linear BOR Maxwell's equations and the Kerr-nonlinear Maxwell's equations.
For BOR Maxwell's equations, we presented an efficient implementation of the Runge-Kutta Discontinuous Galerkin method in two and three space dimensions. We showed that all elementary matrices can be constructed by exploiting the set of orthogonal Jacobi polynomials. Above all, we demonstrated how the stiffness matrices can be directly constructed from two global template matrices.
Our approach still requires to pre-compute and invert the BOR mass matrix for each element. Still it reduces the required memory by at least a factor of two when compared to a quadrature-based approach. Since BOR systems are effectively two-dimensional due to a dimension reduction of 1, owing to the rotational symmetry, in most cases this memory requirement does not lead to significant limitations in terms of applicability. For cases where memory is very scarce, our method leads to reductions by roughly a factor of four at the price of some additional matrix-vector products.
Finally, in a set of numerical experiments, we demonstrated that our implementation yields optimal $p$-convergence and is a promising method for solving the time-dependent Maxwell equations in BOR systems. First, we observed higher order convergence of our scheme for a homogeneous cavity. We successively added PML and incoming waves to our systems and showed simulations of electromagnetic waves traveling along a half fiber and scattering by a sphere, which was placed beneath the fiber in a distance of $1\mu$m. Our plots show the radiation is absorbed by the surrounding PML.

In the last chapter we applied the Discontinuous Galerkin method to Kerr-nonlinear Maxwell's equations in one space dimension for the $y$- and $z$-components of the electromagnetic fields. Again, the numerical flux was chosen to be the solution of the corresponding Riemann problem. We presented the exact solution structure of the Riemann problem as proven in [71]. It consists of four simple waves, of which two are a contact discontinuity and the other two are either a shock or a rarefaction wave. We investigated the simple waves by exploring the Hugoniot Locus and the Riemann invariants.

Due to the nonlinear behavior of the Kerr system the numerical flux is not given globally, but locally, which leads to the necessity of solving the Riemann problem in each grid point to establish the numerical flux. This is extremely expensive in view of simulations, leading to the demand of more efficient alternatives. We presented here three global approximate numerical fluxes, the well-known and widely used Lax-Friedrichs flux, a linear flux and an HLL-like flux. We also implemented a nearly exact numerical flux which assumes to have only shocks but which uses the exact wave speeds. This flux is computed locally by solving the Kerr-Riemann problem in every grid point anew. The middle state, which is needed to construct the solution of the Riemann problem, is the zero of a smooth function. We determined this zero by applying Muller's method (in every grid point). Muller's method is much faster than, e.g., fzero by Matlab. The results show that the Lax-Friedrichs flux requires the least resources in time and memory, whereby producing a DG scheme of order 3 to 4. Numerical observations suggest that a linear flux is a good choice with respect to efficiency. The HHL-like flux seems to be a promising alternative to a linear flux, but still needs improvement and corrections.

For time integration we used the low-storage Runge-Kutta scheme of order 4 with 5 stages [7]. The error coming from a RK scheme depends on the choice of the Runge-Kutta coefficients. Thus, in order to increase the accuracy of the RKDG method and to decrease the $L^1$-error between the approximative and exact solution, one could choose another time integration procedure, like e.g. a Runge-Kutta method of higher order which can be easily integrated into our RKDG scheme. We mention the RK 8(7) method by Dormand and Prince (see e.g. [119, Ch. 5.4, Ch. 4] and the references theirein), with which the error coming from the RK time integration might be decreased from around $10^{-4}$ to $10^{-6}$. The paper by Verner [120] suggests that higher accuracy can be achieved without increasing the computational effort by choosing an efficient set of Runge-Kutta coefficients. If high accuracy is needed, one could even use a Runge-Kutta method of order 10, see [121], which includes a table of the RK coefficients.
With view to increasing optimality one could also think about replacing Muller's method with another zero finding routine, like e.g. a nonlinear SOR method; see e.g. [113, Ch. 19.5].

# Acknowledgments

This thesis would not have been possible without the support and assistance of many people to whom I owe my earnest thankfulness.
I wish to express my deep thanks to Prof. Dr. Willy Dörfler for his many suggestions and corrections during my PhD time.
I am thankful to Prof. Dr. Christian Wieners who kindly agreed on surveying my PhD thesis as a co-referent on short notice.
I would like to give my sincere and hearty acknowledgments to Prof. Dr. Kurt Busch who gave me the possibility to work on an interesting and challenging research field. I am deeply grateful to him for his continued motivation, encouragement and assistance.
My very special thanks go also to Prof. Dr. Marlies Hochbruck for her kind support and guidance. I am truly indebted to her.
I am also deeply grateful for the help and support by Dr. Jens Niegemann, who even assisted me from Zurich. Without his knowledge and experience many things would not have been achieved. Additionally, I owe him my overall thankfulness for proofreading my thesis and giving valuable hints and corrections.
I owe numerous thanks to Dr. Michael König for our fruitful discussions and his constructive suggestions. His advice was always much appreciated.
I am also obliged to Dr. Richard Diehl for his hints and inspirations. I owe many thanks to him.
I was immensely fortunate to work in the Photonics Group and to enjoy many pleasurable events such as barbecues, cakes and pizza seminars, or the joy of listening to the Funky Photons at Christmas and PhD parties. I also wish to express my deep thankfulness for all the corrections and constructive hints for improving my presentations, which would have been by far less informative and enjoyable without their contribution. I am indeed full of thankfulness for this very special, kind and professional support.
I am sincerely thankful for the financial support by the DFG Research Training Group 1294 “Analysis, Simulation and Design of Nanotechnological Processes” at the KIT, which also gave me the possibility to travel abroad to attend many conferences and workshops.
I could not conclude my thesis without expressing my deep and hearty gratefulness to my boyfriend Steve, who supported me in the kindest and most patient way with his encouragement and backup; and to my sister Angela for her never-ending backing throughout all these years. Of course, I am truly and with all my heart thankful to my parents. Thank you all!

# Publications and Presentations

## Publications

- E. Blank, T. Dohnal. Families of Surface Gap Solitons and Their Stability via the Numerical Evans Function Method. *SIAM Journal on Applied Dynamical Systems*, Vol. 10 (2): pp. 667-706, 2011.
- E. Blank, K. Busch, W. Dörfler, M. König, J. Niegemann. The Discontinuous Galerkin Method Applied to BOR Maxwell's Equations, 2012. Submitted.

## Presentations

- E. Blank, T. Dohnal. Stability of Surface Gap Solitons Using the Evans Function Method. GAMM-Tagung, Karlsruhe, 22. - 26.03.2010.
- E. Blank, T. Dohnal. Surface Gap Solitons in the Periodic Nonlinear Schroedinger Equation at a Nonlinearity Interface. OSA conference on Nonlinear Photonics, Karlsruhe, 21. - 24.06.2010 (poster).
- E. Blank, T. Dohnal. Surface Gap Solitons in the Gross Pitaevskii Equation with a Nonlinear Interface. SIAM conference on Nonlinear Waves, Philadelphia, 16-19.08.2010.
- E. Blank, K. Busch, W. Dörfler, M. König, J. Niegemann. Discontinuous Galerkin Method for BOR Maxwell's Equations. 7th Workshop on Numerical Methods for Optical Nano Structures, ETH Zürich, 04. - 06.07.2011.
- E. Blank, K. Busch, W. Dörfler, M. König, J. Niegemann. Discontinuous Galerkin Method for BOR Maxwell's Equations. ICIAM conference, Vancouver, 18. - 22.07.2011.

# Bibliography


[1] W. H. Reed and T. R. Hill. Triangular Mesh methods for the Neutron Transport Equation. *Los Alamos Scientific Laboratory Report LA-UR-73-479*, 1973.

[2] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. The Development of Discontinuous Galerkin Methods. *CiteSeerX*, 1999.

[3] V. Thomée. The Discontinuous Galerkin Time Stepping Method in: Galerkin Finite Element Methods for Parabolic Problems. *Springer Series in Computational Mathematics*, 25:203–230, 2006.

[4] D. Schötzau and C. Schwab. Time Discretization of Parabolic Problems by the hp-Version of the Discontinuous Galerkin Finite Element Method. *SIAM Journal on Numerical Analysis*, 38(3):837–875, 2001.

[5] J. Cesenek and M. Feistauer. Theory of the Space-Time Discontinuous Galerkin Method for Nonstationary Parabolic Problems with Nonlinear Convection and Diffusion. *SIAM Journal on Numerical Analysis*, 50(3):1181–1206, 2012.

[6] J. H. Williamson. Low-storage Runge-Kutta schemes. *Journal of Computational Physics*, 35:48–56, 1980.

[7] M. H. Carpenter and C. A. Kennedy. Fourth-Order 2N-Storage Runge-Kutta Schemes. *NASA Technical Memorandum 109112*, 1994.

[8] B. Cockburn and Chi-Wang Shu. The Runge–Kutta Discontinuous Galerkin Method for Conservation Laws V: Multidimensional Systems. *Journal of Computational Physics*, 141:199–224, 1998.

[9] B. Cockburn. An Introduction to the Discontinuous Galerkin Method for Convection-Dominated Problems. *C.I.M.E. lecture notes*, 1997.

[10] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. Discontinuous Galerkin Methods. Theory, Computation and Applications. *Lecture Notes in Computational Science and Engineering, 11. Springer-Verlag*, 2000.

[11] J. Hesthaven and T. Warburton. Nodal Discontinuous Galerkin Methods – Algorithms, Analysis and Applications. *Springer*, 2008.

[12] B. Cockburn and Chi-Wang Shu. Runge-Kutta Discontinuous Galerkin Methods for Convection-Dominated Problems. *Journal of Scientific Computing*, 16(3):173–261, 2001.

[13] K. Busch, M. König, and J. Niegemann. Discontinuous Galerkin Methods in Nanophotonics. *Laser Photonics Review*, 2011.

[14] G. R. Richter. An Optimal-Order Error Estimate for the Discontinuous Galerkin Method. *Mathematics of Computation*, 50(181):75–88, 1988.

[15] C. Johnson and J. Pitkäranta. An Analysis of the Discontinuous Galerkin Method for a Scalar Hyperbolic Equation. *Mathematics of Computation*, 46:1–26, 1986.

[16] M. Grote, A. Schneebeli, and D. Schoetzau. Interior Penalty Discontinuous Galerkin Method for Maxwell's Equations: Optimal $L^2$-Norm Error Estimates. *IMA Journal of Numerical Analysis*, 28:440–468, 2008.

[17] B. Cockburn, S. Hou, and C.-W. Shu. The Runge-Kutta Local Projection Discontinuous Galerkin Finite Element Method for Conservation Laws IV: The Multidimensional Case. *AMS Mathematical Computation*, 54:545–581, 1990.

[18] J. Jaffre, C. Johnson, and A. Szepessy. Convergence of the Discontinuous Galerkin Finite Element Method for Hyperbolic Conservation Laws. *Math. Models Methods Appl. Sci.*, 5, 1995.

[19] C. E. Baumann and J. T. Oden. A Discontinuous hp Finite Element Method for the Euler and Navier–Stokes Equations. *Wiley International Journal for Numerical Methods in Fluids*, 31:79–95, 1999.

[20] J. R. Mautz and R. F. Harrington. Radiation and Scattering from Bodies of Revolution. *Appl. Sci. Res.*, 20:405–434, 1969.

[21] Jianming Jin. Computational Electrodynamics: The Finite Element Method in Electromagnetics, Second Edition. John Wiley & Sons, 2002.

[22] A. Taflove and S. C. Hagness. Computational Electrodynamics – The Finite-Difference Time-Domain Method. *Artech House Antennas and Propagation Library*, 3rd Ed., 2005. Ch. 12.

[23] D. B. Davidson and R. W. Ziolkowski. Body-of-Revolution Finite-Difference Time-Domain Modeling of Space-Time Focusing by a Three-Dimensional Lens. *J. Opt. Soc. Am. A*, 11(4):1471–1490, 1994.

[24] D. W. Prather and S. Shi. Formulation and Application of the Finite-Difference Time-Domain Method for the Analysis of Axially Symmetric Diffractive Optical Elements. *J. Opt. Soc. Am. A*, 16(5):1131–1142, May 1999.

[25] A. Mohammadi, V. Sandoghdar, and M. Agio. Gold Nanorods and Nanospheroids for Enhancing Spontaneous Emission. *New Journal of Physics*, 10(10):105–115.

[26] M. von Ardenne, G. Musiol, and U. Klemradt. Elektrooptische Effekte in: Effekte der Physik und ihre Anwendungen. *Verlag Harri Deutsch*, 3rd Ed.:721, 2005.

[27] A. E. Siegman. Lasers. *Univ Science Books*, 1986.

[28] P. L. Roe. Approximate Riemann Solvers, Parameter Vectors and Difference Schemes. *Journal of Computational Physics*, 43:357–372, 1981.

[29] A. Harten, P. D. Lax, and B. Van Leer. On Upstream Differencing and Godunov-type Schemes for Hyperbolic Conservation Laws. *SIAM Review*, 25:35–61, 1983.

[30] E. F. Toro. Riemann Solvers and Numerical Methods for Fluid Dynamics – A Practical Introduction. *Springer*, 2009.

[31] V. Wheatley, P. Huguenot, and H. Kumar. On the Role of Riemann Solvers in Discontinuous Galerkin Methods for Magnetohydrodynamics. *Research Report No. 2009-39*, 2009.

[32] A. Karlsson and G. Kristensson. Constitutive Relations, Dissipation, and Reciprocity for the Maxwell Equations in the Time Domain. *Journal of Electromagnetic Waves and Applications*, 6, 1992.

[33] M. Gustafsson. Time Domain Theory of the Macroscopic Maxwell Equations. *Technical report LUTEDX/(TEAT-7062)/1-24, Department of Electromagnetic Theory, Lund Institute of Technology*, 1997.

[34] R. W. Boyd. Nonlinear Optics. *Elsevier*, 3rd Ed., 2008.

[35] J. V. Moloney and A. C. Newell. Nonlinear Optics. *Westview Press*, 2004.

[36] B.-N. Jiang, J. Wu, and L.A. Povinelli. The Origin of Spurious Solutions in Computational Electromagnetics. *Journal of Computational Physics*, 125:104–123, 1996.

[37] B. Cockburn, F. Li, and C.-W. Shu. Locally Divergence-Free Discontinuous Galerkin Methods for the Maxwell Equations. *Journal of Computational Physics*, 194:588–610, 2003.

[38] D. Jackson. Classical Electrodynamics. *Wiley*, 1999.

[39] P. Monk. Finite Element Methods for Maxwell's Equations. *Oxford University Press*, 2003.

[40] R. Adams and J. Fournier. Sobolev Spaces. *Academic Press*, 2003. 2nd Ed.

[41] H.W. Alt. Linear Functional Analysis. An Application Oriented Introduction (Lineare Funktionalanalysis. Eine Anwendungsorientierte Einführung. *Springer*, 5th Ed., 2006.

[42] A. Buffa and P. Ciarlet Jr. On Traces for Functional Spaces Related to Maxwell's Equations Part I: An Integration by Parts Formula in Lipschitz Polyhedra. *Mathematical Methods in the Applied Sciences*, 24:9–30, 2001.

[43] E. Stein (Ed.), R. De Borst (Ed.), and T. J. R. Hughes (Ed.). Encyclopedia of Computational Mechanics. *John Wiley & Sons*, 1st Ed., 2004.

[44] R. Leis. Initial boundary value problems in mathematical physics. *B.G. Teubner*, 1986.

[45] M. Pototschnig, J. Niegemann, L. Tkeshelashvili, and K. Busch. Time-Domain Simulations of the Nonlinear Maxwell Equations Using Operator-Exponential Methods. *IEEE Transactions on Antennas and Propagation*, 57:475, 2009.

[46] J. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computational Physics*, 114(2):185–200, 1994.

[47] S.D. Gedney. An Anisotropic Perfectly Matched Layer Absorbing Medium for the Truncation of FDTD Lattices. *Antennas and Propagation, IEEE Transactions*, 44(12):1630–1639, 1996.

[48] W. C. Chew and W. H. Weedon. A 3D Perfectly Matched Medium from Modified Maxwell's Equations with Stretched Coordinates. *Microwave and Optical Technology Letters*, 7:599–604, 1994.

[49] F. L. Teixeira and W. C. Chew. General Closed-form PML Constitutive Tensors to Match Arbitrary Bianisotropic and Dispersive Linear Media. *IEEE Microwave and Guided Wave Letters*, 8(6):223–225, 1998.

[50] J. Niegemann. Higher-Order Methods for Solving Maxwell's Equations in the Time-Domain. 2009. Karlsruhe Institute of Technology.

[51] M. König. Discontinuous Galerkin Methods in Nanophotonics. 2011. Karlsruhe Institute of Technology.

[52] S. Abarbanela and D. Gottlieb. A Mathematical Analysis of the PML Method. *Journal of Computational Physics*, 134(2):357–363, 1997.

[53] S. Abarbanela and D. Gottlieb. On the Construction and Analysis of Absorbing Layers in CEM. *Applied Numerical Mathematics*, 27:331–340, 1998.

[54] R. Strichartz. Fourier Transforms and Distribution Theory. *Boca Raton, FL: CRC Press*, 1993.

[55] J. Wloka. Partielle Differentialgleichungen: Sobolevräume und Randwertaufgaben. *Teubner Verlag Stuttgart*, 1982.

[56] F. L. Teixeira and W. C. Chew. Systematic Derivation of Anisotropic PML Absorbing Media in Cylindrical and Spherical Coordinates. *IEEE Microwave and Guided Wave Letters*, 7(11), 1997.

[57] R. J. LeVeque. Finite Volume Methods for Hyperbolic Problems. *Cambridge Texts in Applied Mathematics*, 2002.

[58] D. Kroener, M. Ohlberger, and C. Rohde. An Introduction to Recent Developments in Theory and Numerics for Conservation Laws. *Proceedings of the International School, Freiburg/Littenweiler, Germany, October 20-24, 1997. Lecture Notes in Computational Science and Engineering, 5. Berlin, Heidelberg (u.a.): Springer*, 1999.

[59] E. Godlewski and P.-A. Raviart. Numerical Approximation of Hyperbolic Systems of Conservation Laws. *Springer*, 1996.

[60] S. J. Ruuth and R. J. Spiteri. High-Order Strong-Stability-Preserving Runge-Kutta Methods With Downwind-Biased Spatial Discretizations. *SIAM Journal on Numerical Analysis*, 42(2):974–996, 2004.

[61] L. Tobón, J. Chen, and Qing Huo Liu. Spurious Solutions in Mixed Finite Element Method for Maxwell's Equations: Dispersion Analysis and New Basis Functions. *Journal of Computational Physics*, 230:7300–7310, 2011.

[62] J. C. Nedelec. Mixed Finite Element in 3D in H(div) and H(curl). *Lecture Notes in Mathematics*, 1192/1986:321–325, 1986.

[63] J. S. Hesthaven and T. Warburton. High-Order Nodal Discontinuous Galerkin Methods for the Maxwell Eigenvalue Problem. *Philos. Transact. A. Math. Phys. Eng. Sci.*, 362:493–524, 2004.

[64] A. Harten, J. M. Hyman, and P. D. Lax. On Finite-Difference Approximations and Entropy Conditions for Shocks. *Communications on Pure and Applied Mathematics*, 29:297–322, 1976.

[65] N. N. Kuznetsov. Accuracy of Some Approximate Methods for Computing the Weak Solutions of a First-Order Quasi-Linear Equation. *USSR Computational Mathematics and Mathematical Physics*, 16:105–119, 1976.

[66] M. G. Crandall and A. Majda. Monotone Difference Approximations for Scalar Conservation Laws. *AMS Mathematics of Computation*, 34(149):1–21, 1980.

[67] B. Cockburn. Discontinuous Galerkin Methods for Convection-Dominated Problems. *NATO/von Karman Institute for Fluid Dynamics/ NASA (LNCSE Volume 9) Lecture Notes*, 1999.

[68] T. Ruggeri and A. Strumia. Main field and Convex Covariant Density for Quasi-Linear Hyperbolic Systems: Relativistic Fluid Dynamics. *Annales de l'I.H.P., section A*, 34(1):65–84, 1981.

[69] J. A. Smoller and J. L. Johnson. Global Solutions for an Extended Class of Hyperbolic Systems of Conservation Laws. *Archive for Rational Mechanics and Analysis*, 32:169–189, 1969.

[70] L. Seccia. Shock Wave Propagation and Admissibility Criteria in a Nonlinear Dielectric Medium. *Continuum Mechanics and Thermodynamics*, 7:277–296, 1995.

[71] A. de La Bourdonnaye. High-Order Scheme for a Nonlinear Maxwell System Modelling Kerr Effect. *Journal of Computational Physics*, 160(2):500–521, 2000.

[72] T.-P. Liu. Admissible Solutions of Hyperbolic Conservation Laws. *Memoirs of the American Mathematical Society*, 30(240), 1981.

[73] T.-P. Liu. The Riemann Problem for General Systems of Hyperbolic Conservation Laws. *Journal of Differential Equations*, 18:218–234, 1975.

[74] T.-P. Liu and T. Yang. Weak Solutions of General Systems of Hyperbolic Conservation Laws. *Communications in Mathematical Physics*, 230(2):289–327, 2003.

[75] H. Ohwa and K. Kishi. On the Existence of Shock Curves in 2x2 Hyperbolic Systems of Conservation Laws. *Nonlinear Evolution Equations and Mathematical Modeling*, 1640:23–46, 2009.

[76] B. L. Keyfitz and H. C. Kranzer. Existence and Uniqueness of Entropy Solutions to the Riemann Problem for Hyperbolic Systems of Two Nonlinear Conservation Laws. *Journal of Differential Equations*, 27:444–476, 1978.

[77] B. Cockburn and C.-W. Shu. TVB Runge-Kutta Local Projection Discontinuous Galerkin Finite Element Method for Conservation Laws. II: General Framework. *Mathematics of Computation*, 52(186):411–435, 1989.

[78] C.-W. Shu. Total-Variation-Diminishing Time Discretizations. *SIAM Journal on Scientific and Statistical Computing*, 9(6):1073–1084, 1988.

[79] J. C. Butcher. Numerical Methods for Ordinary Differential Equations. *Wiley*, 2nd Ed., 2008.

[80] E. Süli and D. F. Meyers. An Introduction to Numerical Analysis. *Cambridge University Press*, 2003.

[81] F. B. Hildebrand. Introduction to Numerical Analysis. *Dover Pubn Inc.*, 2nd Ed., 1987.

[82] S. J. Ruuth and R. J. Spiteri. A New Class of Optimal High-Order Strong-Stability-Preserving Time Discretization Methods. *SIAM Journal on Numerical Analysis*, 4(2):496–491, 2002.

[83] A. Harten. High Resolution Schemes for Hyperbolic Conservation Laws. *Journal of Computational Physics*, 49:357–393, 1983.

[84] X.-D. Liu, S. Osher, and T. Chan. Weighted Essentially Non-Oscillatory Schemes. *Journal of Computational Physics*, 115(1):200–212, 1994.

[85] T. J. Barth and H. Deconinck. High-Order Methods for Computational Physics. *Lecture Notes in Computational Science and Engineering, Springer*, 9, 1999.

[86] C.-W. Shu. Essentially and Weighted Essentially Non-Oscillatory Methods for Hyperbolic Conservation Laws. *ICASE Report No. 97-65*, 1997.

[87] J. Qiu. Runge-Kutta Discontinuous Galerkin Method Using WENO Limiters. *SIAM Journal on Scientific Computing*, 26(3):907–929, 2005.

[88] X. Zhong and C.-W. Shu. A Simple Weighted Essentially Non-Oscillatory Limiter for Runge-Kutta Discontinuous Galerkin Methods. *Journal of Computational Physics*, 2012. to appear.

[89] M. Wierse. A New Teoretically Motivated Higher Order Upwind Scheme on Unstructured Grids of Simplices. *Advances in Computational Mathematics*, 7(3):303–335, 1997.

[90] G. Jiang and C.-W. Shu. On Cell Entropy Inequality for Discontinuous Galerkin Methods. *AMS Mathematics of Computation*, 62:531–538, 1994.

[91] E. Blank, K. Busch, W. Dörfler, M. König, and J. Niegemann. The Discontinuous Galerkin Method Applied to BOR Maxwell's Equations. *Submitted*, 2013.

[92] Yinchao Chen, R. Mittra, and P. Harms. Finite-Difference Time-Domain Algorithm for Solving Maxwell's Equations in Rotationally Symmetric Geometries. *IEEE Transactions on Microwave Theory and Techniques*, 44 (6), 1996.

[93] P. Ciarlet Jr. and S. Labrunie. Numerical Solution of Maxwell's Equations in Axisymmentric Domains with the Fourier Singular Complement Method. *Differential Equations and Applications*, 3(1):113–155, 2001.

[94] J. G. van Bladel. Electromagnetic Fields. *John Wiley & Sons*, 2nd Ed., 2007. Ch.16.

[95] M.F. Wong, M. Prak, and V. F. Hanna. Axisymmetric Edge-Based Finite Element Formulation For Bodies Of Revolution: Application To Dielectric Resonators. *Microwave Symposium Digest, IEEE MTT-S International*, 1:285–288, 1995.

[96] T. Lu, P. Zhang, and W. Cai. Discontinuous Galerkin Methods for Dispersive and Lossy Maxwell's Equations and PML Boundary Conditions. *Journal of Computational Physics*, 200:549–580, 2004.

[97] T. Lu, P. Zhang, and W. Cai. Discontinuous Galerkin Methods for Dispersive and Lossy Maxwell's Equations and PML Boundary Conditions. *Elsevier, Journal of Computational Physics*, 200:549–580, 2004.

[98] A. Stegun and M. Abramowitz. Handbook of Mathematical Functions. *Dover Pubn Inc.*, 1965.

[99] C. Balanis. Advanced Engineering Electromagnetics. *Mathematical Methods in the Applied Sciences*, 1989.

[100] A. Wachter and H. Hoeber. Repetitorium Theoretische Physik. *Springer*, 2005. 219-224.

[101] T. Warburton. An Explicit Construction for Interpolation Nodes on the Simplex. *Journal of Engineering Mathematics*, 56(3):247–262, 2005.

[102] T. Koornwinder. Two-Variable Analogues of the Classical Orthogonal Polynomials – Theory and Application of Special Functions. *Academic Press*, pages 435–495, 1975.

[103] J. Proriol. Sur une Famille de Polynomes á Deux Variables Orthogonaux Dans un Triangle. *Comptes Rendus de l'Académie des Sciences Paris*, 257:2459–2461, 1957.

[104] M. Dubiner. Spectral Methods on Triangles and Other Domains. *Journal of Scientific Computing*, 6(4):345–390, 1991.

[105] T. Lu, P. Zhang, and W. Cai. Discontinuous Galerkin Methods for Dispersive and Lossy Maxwell's Equations and PML Boundary Conditions. *Journal of Computational Physics*, 200(2):549–580, 2004.

[106] M. König, Chr. Prohm, K. Busch, and J. Niegemann. Stretched-Coordinate PMLs for Maxwell's Equations in the Discontinuous Galerkin Time-Domain Method. *Optics Express*, 19(5):4618–4631, 2011.

[107] G.P. Agrawal. Fiber-Optic Communication Systems. *Wiley Series in Microwave and Optical Engineering*, 2010.

[108] J. Schöberl. NETGEN: An Advancing Front 2D/3D-Mesh Generator Based on Abstract Rules. *Computing and Visualization in Science*, 1(1):41–52, 1997.

[109] F. B. Hildebrand. Advanced Calculus for Engineers. *Prentice Hall Inc.*, 6th Ed., 1956. 163-164.

[110] J. Gieseler. Discontinuous Galerkin Finite Element Time Domain Methods for the Numerical Treatment of the Nonlinear Maxwell's Equations. 2008. Karlsruhe Institute of Technology.

[111] J. Bewersdorff. Algebra für Einsteiger: Von der Gleichungsauflösung zur Galois-Theorie. *Vieweg+Teubner Verlag*, 2004.

[112] P. Pesic. Abels Beweis: Die Geschichte rund um die Lösungsformeln vom Grad 2 bis 4 und der komplette Beweis von Abel. *Springer*, 2005.

[113] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical Recipes in C – The Art of Scientific Computing. *Cambridge University Press*, 2nd Ed., 2002.

[114] P. Batten, N. Clarke, C. Lambert, and D. M. Causon. On the Choice of Wavespeeds for the HLLC Riemann Solver. *SIAM Journal on Scientific Computing*, 18(6):1553–1570, 1997.

[115] B. Einfeldt, C. D. Munz, P. L. Roe, and B. Sjorgreen. On Godunov-type Methods Near Low Densities. *Journal of Computational Physics*, 92:273–295, 1991.

[116] C. B. Laney. Computational Gasdynamics. *Cambridge University Press*, 1998.

[117] M. Pototschnig. Analysis and Simulation of the Emission from Two-Level Systems in Photonic Crystals. *Diploma Thesis, KIT*, 2006.

[118] J.A. Ford. Improved Algorithms of Illinois-type for the Numerical Solution of Nonlinear Equations. *Technical Report CSM-257, University of Essex Press*, 1995.

[119] P. Deuflhard and F. Bornemann. Gewöhnliche DifferentialgleichungenNumerical Mathematics 2. *Walter de Gruyter*, 3rd Ed., 2008.

[120] J. H. Verner. Numerically Optimal Runge–Kutta Pairs with Interpolants. *Numerical Algorithms*, 53(2-3):383–396, 2010.

[121] T. Feagin. A Tenth-Order Runge-Kutta Method with Error Estimate. *Proceedings of the IAENG Conference on Scientific Computing*, 2007.