Error analysis of implicit and exponential time integration of linear Maxwell's equations

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

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

DISSERTATION VON

Mag. math. TOMISLAV PAŽUR

aus Novi Marof, Kroatien

Tag der mündlichen Prüfung: 11.12.2013
Referentin: Prof. Dr. Marlis Hochbruck
Korreferent: Prof. Dr. Christian Wieners

Mojoj mami
"Sve što radim ja, ja radim zbog nje"

Acknowledgements

I would like to thank most sincerely my supervisor, Prof. Dr. Marlis Hochbruck, first for giving me the opportunity to work on this research project, and secondly for great supervising and constant support during the last 3 years in which this work has been done. She has dedicated plenty of hours to discussing my problems and to our joint work and thus made this dissertation possible. Nothing less grateful I am for her support and understanding on a personal level, especially through some bad times.

I also thank to my second supervisor Prof. Dr. Christian Wieners for his helpful remarks on my work, his interesting ideas and also for providing me a possibility to use his M++ software for parallel computing.

As a member of the Research Training Group 1294: "Analysis, Simulation and Design of Nanotechnological Processes" of the German Research Foundation (DFG) I am also grateful to them for financing my work. I thank my colleagues and friends from Research Training Group for a pleasant working atmosphere and lots of interesting discussions. In particular, I would like to mention Abdullah Demirel, Anton Verbitsky, Bilal Haddou-Temsamani, Hans-Jürgen Freisinger and Hannes Gerner. I also thank to Andreas Sturm, HiWi of the Research Training Group, for our joint work on numerical experiments which can be found in the last chapter of this thesis.

A big thank goes to my friends and colleagues Branimir Anić and Jelena Patarčić, for helping me a lot when I moved to Karlsruhe. I am also very grateful to other friends who have made my days in Karlsruhe less rainy: Oliver Tavić, Emanuel Lacić, Denis Štogl, Zrinka Bočkaj, Tomislav Đuričić, Pavo Brković, Tomislav Kos-Grabar, Ivana Kajić, Matija Ilijaš and Mihael Plut. A special thank goes to my best friend Miran Turkalj for positive energy that he shares. I am very happy to have him in my life. I also thank my dear friend Nina Čupić for proofreading the dissertation.

At the end, I would like to express my greatest gratitude to my parents Zorica and Stjepan Pažur for unconditional love, support and understanding they have been giving me and for teaching me the importance of good education from my early age. At the very end, I would like to thank my sisters Katarina, Jelena and Ana for being my happiness and brightening my days.

Abstract

This thesis is concerned with the numerical analysis of some well-known time integration methods, such as implicit collocation methods and exponential integrators, for linear Maxwell's equations in time-domain. The error analysis of time integrators is done both for:

- continuous Maxwell's equations in a semigroup theory framework
- space discrete problem obtained by discretizing Maxwell's equations in space by using discontinuous Galerkin (dG) finite element method.

In both cases error bounds for Gauss and Radau collocation methods can be obtained by applying some already known results which are based on Hille-Phillips operational calculus. In the continuous case we prove an equivalent error bound by using the other approach - energy technique. By using this technique error bounds can be improved in the space discrete case. For algebraically stable and coercive methods, such as Gauss and Radau collocation methods, it is proved that the full discretization error is of order $\mathcal{O}(\tau^{s+1} + h^{k+1/2})$, where s is the number of stages of the collocation method, and h and k are mesh size and polynomial degree of the dG method, respectively. As expected in the case of partial differential equations, we do not obtain the classical order of convergence for these methods. This is called the order reduction phenomena and it is also demonstrated by means of numerical experiments.

We believe that the energy technique is also applicable to the more complicated case of quasilinear Maxwell's equations in time domain and our next goal will be to provide error bounds for Gauss and Radau collocation methods in this case.

Contents

CHAPTER 1

Introduction

Numerical methods for Maxwell's equations

In the last couple of decades there has been a great interest in solving Maxwell's equations because of their great importance and diversity of applications. The first numerical method for solving time-dependent Maxwell's equations was the finite-difference time domain (FDTD) scheme proposed by Yee in 1966 in [65], which uses a staggered grid both in space and time. This is an efficient fully discrete method which is explicit in time and simple to implement. Therefore, it is not surprising that it has become very popular and a large number of papers on FDTD schemes followed up. As a good textbook for FDTD schemes and its application, we refer reader to [62]. However, this method, like all finite difference methods, is difficult to generalize to unstructured grids and can handle only regular domains. Other disadvantages are: no adaptivity, the numerical analysis requires high regularity and it is only conditionally stable (CFL condition). A very efficient, unconditionally stable method based on a finite-difference scheme was proposed in 2000 [66].

Finite element based methods can handle irregular domains, achieve higher order and allow adaptivity and error control. They also use a variational approach which inherits many properties of the continuous problem. This makes a rigorous error analysis possible. In recent years, there has been a great interest in solving Maxwell's equations numerically by using discontinuous Galerkin (dG) finite element methods for the spatial discretization, see the recent textbooks [55, 33]. Some of the main advantages of dGFEMs compared to a standard continuous FEM are: non-conforming meshes are handled much easier, they are highly parallelizable and the mass matrix is block diagonal. This approach is particularly attractive if one is interested in the simulation of wave propagation in composite materials, where the electric permittivity and the magnetic permeability are discontinuous. For the full discretization, discontinuous Galerkin

methods have to be supplemented with suitable time integration schemes. Explicit time integrators can exploit the block diagonal structure of the mass matrix of discontinuous Galerkin schemes and thus lead to fully explicit schemes. The RKDG methods of [10] achieve high-order convergence both in space and time by using strong stability preserving Runge-Kutta schemes in time. On the other hand, explicit methods suffer from step size restrictions due to stability requirements (CFL condition). Implicit methods can be used with larger time steps at the cost of solving linear or even nonlinear systems.

The aim of the thesis

Within this PhD project we want to analyze efficient numerical methods for solving linear Maxwell's equations in time-domain which are given as

$$\begin{aligned} \mu\partial_t \mathbf{H} + \nabla \times \mathbf{E} &= 0, \\ \epsilon\partial_t \mathbf{E} - \nabla \times \mathbf{H} &= -\mathbf{J}, \end{aligned} \tag{1.1}$$

with the corresponding initial and boundary conditions. Here, the electric field $\mathbf{E}$ and the magnetic field intensity $\mathbf{H}$ are unknowns and the electric current density $\mathbf{J}$ is given. Both temporal and spatial derivatives appear in (1.1), but our focus is on time discretization. In [39] we showed that implicit and exponential time integration of linear Maxwell's equations constitute an efficient alternative even for very large problems. It is the aim of this thesis to analyze the discretization error of these two classes of time integrators applied to linear Maxwell's equations. The error analysis of time integrators is done both for:

1) continuous Maxwell's equations (1.1)

2) space discrete problems obtained by discretizing Maxwell's equations in space by using the discontinuous Galerkin (dG) finite element method.

Results

1) To study the time integration of the continuous Maxwell's equations, we first formulate problem (1.1) as an abstract Cauchy problem

$$u'(t) + Au(t) = f(t), \quad u(0) = u_0, \tag{1.2}$$

where A is a generator of a unitary C_0-group. Implicit collocation methods for a more general case, where A is a generator of a bounded C_0-semigroup, have been studied in [4], generalizing earlier work in [5, 6] for the homogeneous problem $f \equiv 0$. The results shown in these papers can be applied to our situation and yield convergence of order $s+1$ in time for s-stage Gauß and Radau collocation methods. Elegant proofs are based on a Hille-Phillips operational calculus by using Laplace transformations. We follow a different approach by using the energy technique which was motivated by the analysis for quasilinear parabolic problems and L-stable Runge–Kutta methods presented in [46]. Our analysis applies to implicit Runge–Kutta methods which are algebraically

stable and coercive and the convergence results obtained are equivalent to those in [4]. Nevertheless, this technique enable us to get better estimates in the fully discrete case and it also has a potential to be generalized to nonlinear problems, which is to best of our knowledge not the case with the technique used in [4].

2) Our main goal is to study fully discrete schemes for Maxwell's equations. We discretize (1.1) in space by using discontinuous Galerkin (dG) methods, see [32]. The main ingredient of dG methods are numerical fluxes, which originate from the finite volume method [45]. Two main choices of fluxes are upwind fluxes and central fluxes. If we denote the mesh size with h and the order of polynomial approximation with k, then dG schemes for Maxwell's equations with central and upwind fluxes are convergent with errors of order $\mathcal{O}(h^k)$ and $\mathcal{O}(h^{k+1/2})$, respectively.

After discretizing in space by using dG methods we end up with an abstract Cauchy problem on a finite dimensional space given by

$$u_h'(t) + A_h u_h(t) = \pi_h f(t), \quad u_h(0) = \pi_h u_0. \tag{1.3}$$

Here, A_h is a discrete operator that approximates A and π_h the L^2-projection on a finite dimensional space. Equation (1.3) can be seen as a system of ordinary differential equations. To obtain a fully discrete scheme we discretize this problem in time.

Some results for fully discrete schemes for Maxwell's equations have been shown recently. In a different framework and for more general first-order systems, error bounds for explicit Runge–Kutta methods of order two and three have been proved in [9]. For constant permittivity and permeability, an optimal convergence rate of $k + 1/2$ in space and s in time, where s denotes the number of stages of the explicit Runge-Kutta method, has been shown. For dG methods with central fluxes and the leap-frog method, error bounds of order k in space and order two in time have been shown in [23]. These methods are explicit and therefore suffer from step size restrictions. Also in the context of dG methods with central fluxes, a locally-implicit time integration method was suggested in [56] and analyzed in [18] and [49].

For algebraically stable and coercive methods, such as Gauss and Radau collocation methods, we prove by using the energy technique, that the full discretization error is of order:

- $\mathcal{O}(\tau^{s+1} + h^{k+1/2})$ when upwind fluxes are used,
- $\mathcal{O}(\tau^{s+1} + h^{k})$ when central fluxes are used.

Here s is the number of stages of the collocation method. This is our main result which can also be found in [38]. Here, piecewise constant coefficients ϵ and μ are considered. As expected in the case of partial differential equations, we do not obtain the classical order of convergence for these methods. This phenomenon is called order reduction and we demonstrate it also by numerical experiments. The full discretization error estimates for Gauss and Radau methods can also be obtained by using results and techniques from [4], but they are suboptimal compared to our results.

1.1 Outline of thesis

This thesis is organized as follows. In Chapter 2 we present some mathematical tools and introduce some notation that is used throughout the thesis.

Chapter 3 is dedicated to the Maxwell's equations; physical modelling and mathematical analysis, including the analytical framework for linear isotropic materials with perfectly conducting boundary conditions, are provided. Moreover, we formulate linear Maxwell's equations as an abstract Cauchy problem (1.2).

In Chapter 4 we study the discontinuous Galerkin finite element method. After motivating the method by a simple one-dimensional example, we introduce some basic tools needed for the analysis of the method. In the last section therein we apply the dG method with both central and upwind flux to Maxwell's equations providing the convergence analysis.

In Chapter 5, we shortly introduce implicit collocation methods and exponential integrators. We present error estimates for the time discretization of the abstract initial value problem (1.2) by s-stage implicit Runge–Kutta methods by using an energy technique. We show full temporal order of convergence for the implicit Euler method and for the implicit midpoint rule while in general, the temporal order will suffer from order reduction to order $s + 1$. It is well known that full temporal order will not be achieved for stiff problems without additional regularity assumptions which are often unnatural in the context of time-dependent partial differential equations, cf. [4]. Our error bounds depend in an explicit form on the regularity of the solution. We also discuss the relation to the earlier work [4] and show the same result with a different technique.

Chapter 6 contains convergence results for the full discretization for the explicit and implicit Euler scheme and for general higher order implicit Runge–Kutta methods. These results are proven by generalizing the results from Chapter 5 to the discrete Maxwell operator A_h. We also show that for the homogeneous Maxwell equations, the divergence is conserved in a weak sense. In the last section therein we provide the error bounds for exponential integrators applied on discrete Maxwell problem (1.3).

Finally, in Chapter 7 we discuss some implementation issues and present numerical experiments which confirm the order reduction phenomena.

CHAPTER 2

Preliminaries

This chapter collects some mathematical concepts and introduces the notation used in this thesis.

2.1 Sobolev spaces

Here we state some fundamental definitions and results in the context of Sobolev spaces. Proofs are omitted and for more details we refer the reader to [8] and [22]. Let $U \subset \mathbb{R}^d$ be a measurable open set with boundary ∂U. We start with the definition of Lebesgue spaces. For $p \in [1, \infty\rangle$ we define the space

$$L^p(U) := \left\{ f : U \to \mathbb{R} : f \text{ measurable and } \int_U |f|^p \leq \infty \right\}$$

with norm $\|f\|_{L^p(U)} := \left(\int_U |f|^p\right)^{1/p}$. For $p = \infty$ we have

$$L^\infty(U) := \left\{ f : U \to \mathbb{R} : f \text{ measurable and } \operatorname*{ess\,sup}_U |f(x)| \leq \infty \right\}$$

with norm $\|f\|_{L^\infty(U)} := \operatorname{ess\,sup}_U |f(x)|$. All integrals here are understood in the Lebesgue sense. For $p \in [1, \infty]$, $(L^p(U), \|\cdot\|_{L^p(U)})$ is a Banach space. With p' we denote its conjugate exponent defined by $\frac{1}{p} + \frac{1}{p'} = 1$. Then $f \in L^p(U)$ and $g \in L^{p'}(U)$ imply $fg \in L^1(U)$ and the **Hölder inequality** holds

$$\|fg\|_{L^1(U)} \leq \|f\|_{L^p(U)} \|g\|_{L^{p'}(U)} . \tag{2.1}$$

The index $p = 2$ is of special interest, since $L^2(U)$ is a Hilbert space with inner product

$$(f, g)_{L^2(U)} = \int_U fg. \tag{2.2}$$

We denote by $(\cdot, \cdot)_{0,U}$ the inner product in $L^2(U)$ and the corresponding norm by $\|\cdot\|_{0,U}$. If it is clear, from the context, what U is, we may just write $(\cdot, \cdot)_0$ and $\|\cdot\|_0$. In $L^2(U)$, the Hölder inequality becomes the **Cauchy-Schwarz inequality**

$$(f, g)_{0,U} \leq \|f\|_{0,U} \|g\|_{0,U}. \tag{2.3}$$

For $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ we define $|\alpha| := \sum_{i=1}^{d} \alpha_i$ and $\partial^\alpha := \frac{\partial^{|\alpha|}}{\partial_1^{\alpha_1} \ldots \partial_d^{\alpha_d}}$. For two integers k and p with $k \geq 0$ and $p \in [1, \infty]$, we define the Sobolev space $W^{k,p}(U)$ as

$$W^{k,p}(U) := \{v \in L^p(U) : \partial^\alpha v \in L^p(U),\ |\alpha| \leq k\},$$

where the derivatives are understood in the weak sense. $W^{k,p}(U)$ is a Banach space with norm $\|v\|^2_{W^{k,p}(U)} := \sum_{|\alpha| \leq k} \|\partial^\alpha v\|^2_{L^p(U)}$. We need only the case $p = 2$ and denote $H^k(U) = W^{k,2}(U)$. With the inner product

$$(u, v)_{k,U} := \sum_{|\alpha| \leq k} \int_U \partial^\alpha u \, \partial^\alpha v,$$

the space $H^k(U)$ is a Hilbert space. We denote

$$\|v\|_{k,U} := \|v\|_{H^s(U)}, \qquad |v|^2_{k,U} := \sum_{|\alpha|=k} \|\partial^\alpha v\|^2_{L^2(U)}.$$

For $s \in \langle 0, 1 \rangle$, the fractional Sobolev spaces $H^s(U)$ are defined as

$$H^s(U) := \left\{ v \in L^2(U) : \frac{v(x) - v(y)}{\|x - y\|^{s+d/2}} \in L^2(U \times U) \right\}.$$

The dual space of $H^s(U)$ is denoted by $H^{-s}(U)$ and the duality product between $f \in H^{-s}(U)$ and $u \in H^s(U)$ we denote by

$$\langle f, u \rangle_{-s \times s, U} := \langle f, u \rangle_{H^{-s}(U) \times H^s(U), U}.$$

If U is a bounded domain with Lipschitz boundary, then there exists a continuous **trace operator**

$$\gamma^{\partial U} : H^1(U) \to H^{1/2}(\partial U)$$

such that $\gamma^{\partial U} v = v|_{\partial U}$ for all $v \in C^\infty(\bar{U})$. The kernel of $\gamma^{\partial U}$ is denoted by $H^1_0(U)$.

2.2 Operators and semigroups

Throughout the thesis, $(X, \|\cdot\|_X)$ will denote a generic Banach space. By $\mathcal{B}(X)$ we denote the set of all bounded linear operators on X. For $A \in \mathcal{B}(X)$, the operator norm is defined as

$$\|A\|_{X\leftarrow X} := \max_{x\neq 0} \frac{\|Ax\|_X}{\|x\|_X}.$$

More interesting for us will be unbounded operators which generate strongly continuous semigroups. Some fundamentals of the semigroup theory needed in this thesis is presented in what follows. For more details, we refer the reader to textbooks [21, 53] but also to Internet Seminar (ISEM) lectures from 2012 and 2013 [3, 40].

Definition 2.1. *A map $S(\cdot) : \mathbb{R}_+ \to \mathcal{B}(X)$ is called **strongly continuous semigroup** or just C_0-**semigroup** if the following conditions are fulfilled:*

1. *$S(0) = I$ and $S(t+s) = S(t)S(s)$ for all $t, s \geq 0$.*
2. *For each $x \in X$ the map $S(\cdot)x : \mathbb{R}_+ \to R$, $t \mapsto S(t)x$ is continuous.*

The operator defined by

$$D(A) := \left\{ x \in X \mid \exists \lim_{t\to 0^+} \frac{1}{t}(S(t)x - x) \text{ in } X \right\}$$

and

$$Ax := \lim_{t\to 0^+} \frac{1}{t}(S(t)x - x) \quad \text{for} \quad x \in D(A),$$

*is called the **generator** of $S(\cdot)$.*

Let $S(\cdot)$ be a C_0-semigroup generated by A. We write $S(t) = e^{tA}$. Then the domain of A is dense in X and A is a closed operator. There exist constants $M \geq 1$ and $\omega \geq 0$ such that

$$\|S(t)\|_{X\leftarrow X} \leq Me^{\omega t} \quad \text{for all } t \geq 0. \tag{2.4}$$

If a semigroup satisfies (2.4), we say that it is of **type** (M, ω). For a C_0-semigroup of type (M, ω), we have that for all $n \in \mathbb{N}$ and all $\lambda \in \mathbb{C}$ with $\operatorname{Re}\lambda > \omega$ it holds

$$\|(\lambda I - A)^{-n}\|_{X\leftarrow X} \leq \frac{M}{(\operatorname{Re}\lambda - \omega)^n}. \tag{2.5}$$

Further on, the homogeneous Cauchy problem

$$u'(t) = Au(t), \quad t \geq 0, \quad u(0) = u_0, \tag{2.6}$$

has a unique solution $u = S(\cdot)u_0 \in C^1(\mathbb{R}_+, X) \cap C(\mathbb{R}_+, D(A))$ for each given initial value $u_0 \in D(A)$.

C_0-semigroups of type $(M, 0)$ are called **uniformly bounded semigroups**. If moreover $M = 1$, then they are called **semigroups of contractions**.

Semigroups of contractions are closely related to dissipative operators. An operator A on a Hilbert space $(X, (\cdot,\cdot)_X)$ is **dissipative** if for every $x \in D(A)$

$$\mathrm{Re}(Ax, x) \leq 0 \tag{2.7}$$

holds. The Lumer-Phillips theorem establish the connection between contraction semigroups and dissipative operators.

Theorem 2.2 (Lumer-Phillips, 1961)**.** *For a densely defined, dissipative linear operator A on a Hilbert space X the following statements are equivalent:*

(i) The closure $\overline{A}$ of A generates a contraction semigroup.

(ii) The range of $(\lambda - A)$ is dense in X for some (hence all) $\lambda > 0$.

If in Definition 2.1 $\mathbb{R}_+$ is replaced by $\mathbb{R}$, "$t \to 0^+$" by "$t \to 0$" and "$t, s \geq 0$" by "$t, s \in \mathbb{R}$", then we have a C_0-group. The Stone theorem characterizes unitary C_0-groups.

Definition 2.3. *The **adjoint operator** A^* of A is defined by*

$$D(A^*) := \{x \in X : \exists y \in X \text{ s. t. } \forall u \in D(A) \text{ holds } (Au, x)_X = (u, y)_X\}.$$

*and $A^*x := y$ for $x \in D(A^*)$. The operator A is called **self-adjoint** if $A^* = A$, and **skew-adjoint** if $A^* = -A$.*

Theorem 2.4 (Stone, 1930)**.** *Let A be a densely defined linear operator on a Hilbert space X. Then A generates a C_0-group of unitary operators if and only if A is skew-adjoint.*

We end this section with the well-posedness of the inhomogeneous Cauchy problem

$$u'(t) = Au(t) + f(t), \quad t \in [0, T], \quad u(0) = u_0. \tag{2.8}$$

A function $u : [0, T] \to X$ is a solution of (2.8) if u belongs to $C^1([0, T], X)$, $u(t) \in D(A)$ for all $t \in [0, T]$ and (2.8) holds.

Theorem 2.5. *Let A generate a C_0-semigroup $S(\cdot)$ and $u_0 \in D(A)$. Assume either that $f \in C^1([0, T], X)$ or that $f \in C^0([0, T], D(A))$. Then the unique solution of (2.8) is given by*

$$u(t) = S(t)u_0 + \int_0^t S(t - s)f(s)ds, \quad t \in [0, T]. \tag{2.9}$$

2.3 Gronwall's lemma

We use the following versions of continuous and discrete Gronwall's lemmas in our analysis. The first one is the continuous Gronwall lemma in integral form, cf. [20, Proposition 2.1].

Lemma 2.6. *Let $T \in \mathbb{R}_+ \cup \{\infty\}$, a, $b \in L^\infty(0,T)$ and $\lambda \in L^1(0,T)$, $\lambda(t) \geq 0$ for almost all $t \in [0,T]$. If, further, on b is monotonically increasing, continuous function, then,*

$$a(t) \leq b(t) + \int_0^t \lambda(s)a(s)ds \quad a.e \text{ in } [0,T]$$

implies for almost all $t \in [0,T]$

$$a(t) \leq e^{\Lambda(t)}b(t),$$

where $\Lambda(t) := \int_0^t \lambda(\tau)d\tau$.

For the analysis of time integration methods we will use two versions of the discrete Gronwall lemma in sum form, one for explicit and one for implicit methods. For both versions see [20, Proposition 4.1].

Lemma 2.7. *Let $\{a_n\}, \{b_n\} \subset \mathbb{R}$, $\lambda \in \mathbb{R}_0^+$ and $\tau \in \mathbb{R}^+$ satisfy*

$$a_{n+1} \leq b_{n+1} + \tau\lambda \sum_{j=0}^{n} a_j, \quad n = 0, 1, \ldots$$

with initial condition $a_0 \leq b_0$. Then if $\{b_n\}$ is monotonically increasing, it follows

$$a_n \leq b_n \left(1 + \lambda\tau\right)^n \leq b_n \mathrm{e}^{\lambda n\tau}, \quad n = 0, 1, \ldots.$$

Lemma 2.8. *Let $\{a_n\}, \{b_n\} \subset \mathbb{R}$, $\lambda \in \mathbb{R}_0^+$, $\tau \in \mathbb{R}^+$ and $1 - \lambda\tau > 0$. If $\{b_n\}$ is monotonically increasing, then, inequality*

$$a_{n+1} \leq b_{n+1} + \tau\lambda \sum_{j=0}^{n} a_{j+1}, \quad n = 0, 1, \ldots$$

with initial condition $a_0 \leq b_0$ implies

$$a_n \leq b_n \left(\frac{1}{1 - \lambda\tau}\right)^n, \quad n = 0, 1, \ldots$$

The expression $(1 - \lambda\tau)^n$ can also be bounded with the exponential function. We will actually use the following corollary of the previous Lemma.

Corollary 2.9. *Let the assumptions of Lemma 2.8 be satisfied and let in addition $\lambda\tau \leq \frac{3}{4}$. Then a_n is bounded by*

$$a_n \leq b_n \mathrm{e}^{2\lambda n\tau}, \quad n = 0, 1, \ldots$$

Proof. $1 + x \geq e^{2x}$ holds on $[-\frac{3}{4}, 0]$. This gives $(1 - \lambda\tau)^{-1} \leq \mathrm{e}^{2\lambda\tau}$. □

2.4 Some notation and other useful facts

Throughout this thesis, $\Omega \subset \mathbb{R}^3$ is a bounded domain with Lipschitz-continuous boundary $\partial\Omega$ and outward unit normal $\mathbf{n}$. T is the final time of the simulation.

k denotes the polynomial degree used in the discontinuous Galerkin method and s is number of inner stages of Runge–Kutta methods.

We also recall that differential operators gradient ∇, divergence $\nabla\cdot$ and curl $\nabla\times$ are, for $u : \mathbb{R}^3 \to \mathbb{R}$ and $\mathbf{v} : \mathbb{R}^3 \to \mathbb{R}^3$, defined as

$$\begin{aligned} \nabla u &= \begin{pmatrix} \partial_x u \\ \partial_y u \\ \partial_z u \end{pmatrix}, \\ \nabla \cdot \mathbf{v} &= \partial_x v_x + \partial_y v_y + \partial_z v_z, \\ \nabla \times \mathbf{v} &= \begin{pmatrix} \partial_y v_z - \partial_z v_y \\ \partial_z v_x - \partial_x v_z \\ \partial_x v_y - \partial_y v_x \end{pmatrix}. \end{aligned}$$

It holds

$$\nabla \cdot (\nabla \times \mathbf{v}) = 0 \tag{2.10}$$

and

$$\nabla \times \nabla u = 0. \tag{2.11}$$

The Young's inequality will be used very often.

Theorem 2.10 (Young's inequality)**.** *Let $a, b \geq 0$ be real numbers. For any $\gamma > 0$ we have*

$$ab \leq \frac{\gamma}{2} a^2 + \frac{1}{2\gamma} b^2. \tag{2.12}$$

CHAPTER 3

Maxwell's equations

In this chapter first we give a short physical introduction to Maxwell's equations following [19, 50, 43]. We start by stating Maxwell's equations in its general form. By using the appropriate constitutive relations we simplify them and get linear Maxwell's equations which we want to solve numerically. Further on, we discuss boundary and interface conditions and provide reductions of Maxwell's equations in two space dimensions.

In addition to that, we try to give a mathematical insight into deriving linear Maxwell's equations. We define relevant function spaces and provide a setting in which the system of linear Maxwell's equations that we are solving, is a well-posed problem, i. e., it posses a unique solution.

3.1 Physical modelling of Maxwell's equation

Maxwell's equations form the fundamentals of classical electrodynamics and classical optics. They were completely formulated by James Clerk Maxwell in the period from 1861 to 1865. They consists of a set of partial differential equations which describe the propagation of electromagnetic waves trough media. We begin with the general Maxwell's equations in differential form. For the given electric current density $\mathbf{J} : \Omega \times \mathbb{R}_+ \to R^3$ and the electric charge density $\rho : \Omega \times \mathbb{R}_+ \to \mathbb{R}$, we seek for vector fields $\mathbf{D}, \mathbf{E}, \mathbf{B}, \mathbf{H} : \Omega \times \mathbb{R}_+ \to \mathbb{R}^3$ such that

$$\partial_t \mathbf{B}(t, \mathbf{x}) + \nabla \times \mathbf{E}(t, \mathbf{x}) = 0, \tag{3.1a}$$

$$\partial_t \mathbf{D}(t, \mathbf{x}) - \nabla \times \mathbf{H}(t, \mathbf{x}) = -\mathbf{J}(t, \mathbf{x}), \tag{3.1b}$$

$$\nabla \cdot \mathbf{D}(t, \mathbf{x}) = \rho, \tag{3.1c}$$

$$\nabla \cdot \mathbf{B}(t, \mathbf{x}) = 0, \tag{3.1d}$$

holds. Here $\mathbf{D}$ is the electric displacement, $\mathbf{E}$ the electric field, $\mathbf{B}$ the magnetic induction and $\mathbf{H}$ the magnetic field intensity.

The first equation is Faraday's law of induction and it describes how the time-varying magnetic field effects the electric field. The second equation is the generalization of Ampere's law and it describes the effect of the electric current (external and induced) on the magnetic field. The last two equations are Gauß's electric law and Gauß's magnetic law, respectively. Gauß's electric law is the generalization of Coulomb's law and it describes the source of electromagnetic displacement. Gauß's magnetic law says that there are no free magnetic poles. For a more physical insight in these equations see [41].

After differentiating (3.1c) with respect to time and using (3.1b), we derive the continuity equation

$$\partial_t \rho + \nabla \cdot \mathbf{J} = 0 \tag{3.2}$$

which expresses the conservation of the charge of the system. Equations (3.1a)-(3.1b) are also called curl-equations, while (3.1c)-(3.1d) are div-equations or divergence equations. For time-evolution only the curl-equations are important and the div-equations can be seen as constraints that have to be fulfilled for all times t. However, it is not hard to see that if the continuity equation (3.2) holds, then from the curl-equations it follows that $\nabla \cdot \mathbf{D}$ and $\nabla \cdot \mathbf{B}$ are constant in time. Indeed, taking the divergence of the curl equations and using (2.10) yields

$$\begin{aligned}
&\partial_t(\nabla \cdot \mathbf{B}(t,\mathbf{x})) = 0,\\
&\partial_t(\nabla \cdot \mathbf{D}(t,\mathbf{x}) - \rho(t,\mathbf{x})) = \nabla \cdot (\nabla \times \mathbf{H} - \mathbf{J}(t,\mathbf{x})) + \nabla \cdot \mathbf{J}(t,\mathbf{x}) = 0.
\end{aligned}$$

Hence, the divergence equations are not independent relations, and if they hold at some initial time t_0, they hold for all times $t > t_0$.

3.1.1 Constitutive relations

As we have seen, the set of equations (3.1) contains 6 independent scalar equations and 12 scalar unknowns and therefore is not consistent. Constitutive relations are of the form

$$\mathbf{D} = \mathbf{D}(\mathbf{E},\mathbf{H}) \quad \text{and} \quad \mathbf{B} = \mathbf{B}(\mathbf{E},\mathbf{H}),$$

and therefore they couple the unknown fields. In free space (vacuum), for example, we have

$$\mathbf{D}(t,\mathbf{x}) = \epsilon_0 \mathbf{E}(t,\mathbf{x}), \quad \mathbf{B}(t,\mathbf{x}) = \mu_0 \mathbf{H}(t,\mathbf{x}),$$

where ϵ_0 is the permittivity of free space and μ_0 the permeability of free space. Those constants are related to the speed of light in free spaces via

$$c := \frac{1}{\sqrt{\epsilon_0 \mu_0}}.$$

In matter, however, the situation can be much more complicated. Nevertheless, in what follows we suppose **linear** constitutive relations

$$\begin{aligned}
\mathbf{D}(t,\mathbf{x}) &= \epsilon \mathbf{E}(t,\mathbf{x}) = \epsilon_0 \epsilon_r \mathbf{E}(t,\mathbf{x})\\
\mathbf{B}(t,\mathbf{x}) &= \mu \mathbf{H}(t,\mathbf{x}) = \mu_0 \mu_r \mathbf{H}(t,\mathbf{x}).
\end{aligned} \tag{3.3}$$

Here, ϵ (ϵ_r) is (relative) permittivity and μ (μ_r) (relative) permeability. Further on, we suppose that the medium is **isotropic** or directionally independent but possibly **inhomogeneous**, which implies that ϵ and μ are real-valued functions, i.e. $\epsilon, \mu : \mathbb{R}^3 \to \mathbb{R}_+$.

In conducting media one additional constitutive relation is given by the fact that electromagnetic field induces an electric current. This is approximated by Ohm's law

$$\mathbf{J}(t, \mathbf{x}) = \sigma(\mathbf{x})\mathbf{E}(t, \mathbf{x}) + \mathbf{J}_e(t, \mathbf{x}). \tag{3.4}$$

Here, $\sigma : \mathbb{R}^3 \to \mathbb{R}$ is called conductivity and $\mathbf{J}_e$ is the external (applied) current density. Good conductors, like metals, have a large value of σ.

3.1.2 Interface and boundary conditions

In most physical problems, we have to deal with more than one material. In our modeling, we assume that ϵ and μ vary smoothly along one material, but may have jumps at material interfaces. From high school physics classes we know that light changes direction when it propagates from one material to another (the refraction phenomenon). In this section we shortly discuss what happens with electromagnetic fields at material interfaces.

Let S be a surface that separates two materials, and $\mathbf{n}$ be a normal vector pointing from material 1 to material 2. From the curl-equations, with the help of Stokes' theorem, we can deduce

$$\begin{aligned} \mathbf{n} \times (\mathbf{E}_1 - \mathbf{E}_2) &= 0 \\ \mathbf{n} \times (\mathbf{H}_1 - \mathbf{H}_2) &= \mathbf{J}_S. \end{aligned}$$

where $\mathbf{E}_1(\mathbf{H}_1)$ and $\mathbf{E}_2(\mathbf{H}_2)$ are electric (magnetic) fields in material 1 and 2, respectively, and $\mathbf{J}_S$ is the surface current density. However, if the current distribution remains finite, then $\mathbf{J}_S = 0$ and we have that both magnetic and electric field are continuous in tangential direction. From the div-equations, with the help of Gauß's divergence theorem, one gets

$$\begin{aligned} \mathbf{n} \cdot (\mathbf{D}_1 - \mathbf{D}_2) &= \rho_S, \\ \mathbf{n} \cdot (\mathbf{B}_1 - \mathbf{B}_2) &= 0, \end{aligned}$$

where ρ_S is the surface charge density. Again, assuming that the charge distribution is finite, i. e. $\rho_S = 0$, we get the continuity in normal directions of the fields $\mathbf{D}$ and $\mathbf{B}$. However, this implies that the normal component of $\mathbf{E}$ is discontinuous for $\epsilon_1 \neq \epsilon_2$ and the normal component of $\mathbf{H}$ is discontinuous for $\mu_1 \neq \mu_2$. More precisely,

$$\begin{aligned} \mathbf{n} \cdot (\epsilon_1 \mathbf{E}_1 - \epsilon_2 \mathbf{E}_2) &= 0, \\ \mathbf{n} \cdot (\mu_1 \mathbf{H}_1 - \mu_2 \mathbf{H}_2) &= 0. \end{aligned}$$

Boundary conditions

We consider boundaries where our domain is surrounded by a perfect conductor. Perfect conductors are characterized by the formal limit $\sigma \to \infty$. This implies that electric field and magnetic induction both vanish in the perfect conductor, see [19, Section 1.1.6]. Therefore, if one side of S is occupied by a perfect conductor, the interface conditions imply the boundary conditions

$$\mathbf{n} \times \mathbf{E} = 0 \quad \text{and} \quad \mathbf{n} \cdot \mathbf{B} = 0. \tag{3.5}$$

If we are dealing with unbounded domains then we need to simulate absorbing boundary conditions (when electromagnetic wave hits the boundary it should not be reflected but absorbed; this corresponds to the situation in which it leaves the domain and does not come back). In the literature this is called *Silver-Müller* boundary condition and it is modeled as

$$\mathbf{n} \times \mathbf{E} = c\mu(\mathbf{n} \times \mathbf{H}) \times \mathbf{n}. \tag{3.6}$$

In this thesis we consider perfectly conducting boundary conditions (3.5) only.

3.1.3 Linear Maxwell's equations

Finally, we are interested in solving linear Maxwell's equations with perfectly conducting boundary conditions. This yields the following set of equations

$$\begin{aligned}
&\mu\partial_t \mathbf{H} + \nabla \times \mathbf{E} = 0 && \text{on } \Omega \times (0,T),\\
&\epsilon\partial_t \mathbf{E} - \nabla \times \mathbf{H} = -\mathbf{J} && \text{on } \Omega \times (0,T),\\
&\nabla \cdot (\epsilon\mathbf{E}) = \rho, \quad \nabla \cdot (\mu\mathbf{H}) = 0 && \text{on } \Omega \times (0,T),\\
&\mathbf{n} \times \mathbf{E} = 0, \quad \mathbf{n} \cdot (\mu\mathbf{H}) = 0 && \text{on } \partial\Omega \times (0,T),\\
&\mathbf{E}(0) = \mathbf{E}_0, \quad \mathbf{H}(0) = \mathbf{H}_0 && \text{on } \Omega.
\end{aligned} \tag{3.7}$$

As we have already discussed, the div-equations are satisfied automatically if the continuity equation (3.2) holds. We will also see that it is enough to consider only the boundary condition for the **E**-field; the one for the **H**-field is satisfied then too. Therefore, throughout this thesis we will consider the question of solving numerically

$$\begin{aligned}
&\mu\partial_t \mathbf{H} + \nabla \times \mathbf{E} = 0 && \text{on } \Omega \times (0,T),\\
&\epsilon\partial_t \mathbf{E} - \nabla \times \mathbf{H} = -\mathbf{J} && \text{on } \Omega \times (0,T),\\
&\mathbf{n} \times \mathbf{E} = 0 && \text{on } \partial\Omega \times (0,T),\\
&\mathbf{E}(0) = \mathbf{E}_0, \quad \mathbf{H}(0) = \mathbf{H}_0 && \text{on } \Omega.
\end{aligned} \tag{3.8}$$

In the following section we show that this is a mathematically well-posed problem, i. e. it possesses a unique solution. But first we will mention the reductions of the problem to two dimensions. In order to do this, it is helpful to write the curl-equations componentwise. We get the following 6 equations

$$
\begin{aligned}
&\mu\partial_t H_x + \partial_y E_z - \partial_z E_y = 0,\\
&\mu\partial_t H_y + \partial_z E_x - \partial_x E_z = 0,\\
&\mu\partial_t H_z + \partial_x E_y - \partial_y E_x = 0,\\
&\epsilon\partial_t E_x - \partial_y H_z + \partial_z H_y = -J_x,\\
&\epsilon\partial_t E_y - \partial_z H_x + \partial_x H_z = -J_y,\\
&\epsilon\partial_t E_z - \partial_x H_y + \partial_y H_x = -J_z.
\end{aligned}
$$

3.1.4 Reductions to two dimensions

If we suppose that the underlying physical system has some symmetries, it is possible to reduce the dimensionality of the system. In applications we often have that the system is homogeneous in one direction. Without loss of generality, let us say that it is the z-direction in which the system is homogeneous. This means that all z-derivatives in the aforementioned six equations will vanish. This gives us two decoupled sets of three equations.

The first one contains H_x, H_y and E_z components and it is called **transverse-magnetic (TM) polarization**. It describes the propagation where the electric field is perpendicular to the plane of propagation.

$$\mu\partial_t H_x + \partial_y E_z = 0 \tag{3.9a}$$
$$\mu\partial_t H_y - \partial_x E_z = 0 \tag{3.9b}$$
$$\epsilon\partial_t E_z - \partial_x H_y + \partial_y H_x = -J_z. \tag{3.9c}$$

The set of the other three equations contains E_x, E_y and H_z components only, and it is called **transverse-electric (TE) polarization**. It describes the propagation where the electric field lies in the plane of propagation.

$$\epsilon\partial_t E_x - \partial_y H_z = -J_x \tag{3.10a}$$
$$\epsilon\partial_t E_y + \partial_x H_z = -J_y \tag{3.10b}$$
$$\mu\partial_t H_z + \partial_x E_y - \partial_y E_x = 0. \tag{3.10c}$$

Our numerical experiments will mainly deal with TM polarization. After this short introduction from the physical point of view we pursue further to the mathematics of the problem.

3.2 Analysis of linear Maxwell's equation

In this section we show that the linear Maxwell's equations (3.8) can be written as an abstract Cauchy problem. Using the semigroup theory we provide the well-posedness result. For this purpose first we have to choose adequate function spaces.

3.2.1 Function spaces

For the coefficients we suppose that

$$\mu, \epsilon \in L^\infty(\Omega), \quad \epsilon, \mu \geq \delta > 0. \tag{3.11}$$

The basic Hilbert space we work with is $V := L^2(\Omega)^3 \times L^2(\Omega)^3$ with the inner product defined by

$$(\begin{pmatrix} \mathbf{H}_1 \\ \mathbf{E}_1 \end{pmatrix}, \begin{pmatrix} \mathbf{H}_2 \\ \mathbf{E}_2 \end{pmatrix})_V := \int_\Omega \mu \mathbf{H}_1 \cdot \mathbf{H}_2 + \epsilon \mathbf{E}_1 \cdot \mathbf{E}_2.$$

By our assumption on μ and ϵ, this inner product is obviously equivalent to the standard inner product on $L^2(\Omega)^6$. The norm induced by this inner product corresponds to the electromagnetic energy of the physical system

$$\left\| \begin{pmatrix} \mathbf{H} \\ \mathbf{E} \end{pmatrix} \right\|_V = \int_\Omega \mu \left|\mathbf{H}\right|^2 + \epsilon \left|\mathbf{E}\right|^2. \tag{3.12}$$

Two spaces play an important role in the analysis of Maxwell's equations, namely the **div**-space

$$H(\mathrm{div}, \Omega) := \left\{ v \in L^2(\Omega)^3 : \nabla \cdot v \in L^2(\Omega) \right\}, \tag{3.13}$$

and the **curl**-space

$$H(\mathrm{curl}, \Omega) := \left\{ v \in L^2(\Omega)^3 : \nabla \times v \in L^2(\Omega)^3 \right\}. \tag{3.14}$$

Following [19, 48, 55], we give some properties of these spaces.

Properties of $H(\mathrm{div}, \Omega)$

With the inner product

$$(u, v)_{H(\mathrm{div},\Omega)} := (u, v)_{L^2(\Omega)^3} + (\nabla \cdot u, \nabla \cdot v)_{L^2(\Omega)},$$

$H(\mathrm{div}, \Omega)$ is a Hilbert space. The associated graph norm is

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

$H(\mathrm{div}, \Omega)$ is the closure of $C^\infty(\overline{\Omega})^3$ with respect to $\|\cdot\|_{H(\mathrm{div},\Omega)}$. The normal trace operator is well defined on $C^\infty(\overline{\Omega})^3$

$$\gamma_n v = \mathbf{n} \cdot v|_{\partial\Omega}, \quad v \in C^\infty(\overline{\Omega})^3. \tag{3.15}$$

[48, Theorem 3.24] shows that functions in $H(\mathrm{div}, \Omega)$ also have a well-defined normal trace, i. e. the mapping γ_n can be extended by continuity to a continuous linear map from $H(\mathrm{div}, \Omega)$ onto $H^{-1/2}(\partial\Omega)$. Moreover, the following integration by parts formula holds: for all $v \in H(\mathrm{div}, \Omega)$ and $w \in H^1(\Omega)$ we have

$$\int_\Omega v \cdot \nabla w + \nabla \cdot v \, w = \langle \gamma_n v, w \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial\Omega}. \tag{3.16}$$

We are now allowed to define the subspace $H_0(\mathrm{div}, \Omega)$ of $H(\mathrm{div}, \Omega)$ as

$$H_0(\mathrm{div}, \Omega) := \{v \in H(\mathrm{div}, \Omega) : \mathbf{n} \cdot v|_{\partial\Omega} = 0\}.$$

It can be shown [48, Theorem 3.25] that this is the closure of $C_0^\infty(\overline{\Omega})^3$ with respect to $\|\cdot\|_{H(\mathrm{div},\Omega)}$.

Properties of $H(\mathrm{curl}, \Omega)$

Again, with the inner product

$$(u, v)_{H(\mathrm{curl},\Omega)} := (u, v)_{L^2(\Omega)^3} + (\nabla \times u, \nabla \times v)_{L^2(\Omega)^3},$$

$H(\mathrm{curl}, \Omega)$ is a Hilbert space which implies the closedness of the $\nabla\times$ operator. The graph norm is defined as

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

This space is the closure of $C^\infty(\overline{\Omega})^3$ with respect to $\|\cdot\|_{H(\mathrm{curl},\Omega)}$. The tangential trace operator is well defined on $C^\infty(\overline{\Omega})^3$

$$\gamma_t v = \mathbf{n} \times v|_{\partial\Omega}, \quad v \in C^\infty(\overline{\Omega})^3. \tag{3.17}$$

[48, Theorem 3.29] shows that functions in $H(\mathrm{curl}, \Omega)$ have also a well-defined tangential trace, i. e., the mapping γ_t can be extended by continuity to a continuous linear map from $H(\mathrm{curl}, \Omega)$ into $H^{-1/2}(\partial\Omega)^3$. Moreover, the following integration by parts formula holds: for all $v \in H(\mathrm{curl}, \Omega)$ and $w \in H^1(\Omega)^3$ we have

$$\int_\Omega (\nabla \times v) \cdot w - v \cdot (\nabla \times w) = \langle \gamma_t v, w \rangle_{-\frac{1}{2}\times\frac{1}{2},\partial\Omega}. \tag{3.18}$$

We can now define the subspace of $H(\mathrm{curl}, \Omega)$ with zero tangential trace

$$H_0(\mathrm{curl}, \Omega) := \{v \in H(\mathrm{curl}, \Omega) : \mathbf{n} \times v|_{\partial\Omega} = 0\}.$$

It can be shown [48, Theorem 3.33] that this is the closure of $C_0^\infty(\overline{\Omega})^3$ with respect to $\|\cdot\|_{H(\mathrm{curl},\Omega)}$.

3.2.2 Well-posedness

We define the Maxwell's operator by

$$D(A_M) = H(\mathrm{curl}, \Omega) \times H_0(\mathrm{curl}, \Omega)$$

and

$$A_M \begin{pmatrix} \mathbf{H} \\ \mathbf{E} \end{pmatrix} := \begin{pmatrix} \mu^{-1} \nabla \times \mathbf{E} \\ -\epsilon^{-1} \nabla \times \mathbf{H} \end{pmatrix}. \tag{3.19}$$

The corresponding graph norm is $\|v\|_{D(A_M)} := (\|v\|_V^2 + \|A_M v\|_V^2)^{1/2}$. Closedness of the $\nabla\times$ operator implies closedness of the operator A_M. Equivalently, $D(A_M)$ equipped with the graph norm is a Banach space. With this notation we can rewrite the problem (3.8) in a compact form as an abstract ordinary differential equation.

For a given $f = \big(0, -\mathbf{J}\big)^T$ we seek for $u = \big(\mathbf{H}, \mathbf{E}\big)^T \in C^1([0,T], V)$, $u(t) \in D(A_M)$ for all $t \in [0,T]$ such that

$$\partial_t u(t) + A_M u(t) = f(t), \quad t \geq 0 \qquad u(0) = u_0. \tag{3.20}$$

We show that this is a well-posed problem. Let us consider the homogeneous problem first.

Homogeneous case

This corresponds to the physical situation without electric current and electric charge ($\mathbf{J} = 0$ and $\rho = 0$). The first step to prove that there exists a unique solution of the homogeneous problem

$$\partial_t u(t) + A_M u(t) = 0, \quad t \geq 0 \qquad u(0) = u_0 \tag{3.21}$$

is to show that A_M is a skew-adjoint operator.

Proposition 3.1. *Under the assumption (3.11) on the coefficients μ and ϵ, the Maxwell operator A_M defined in (3.19) is a skew-adjoint operator on V.*

Proof. To prove $A_M^* = -A_M$ it is enough to show that A_M is a skew-symmetric operator and that $D(A_M^*) = D(-A_M)$. We first show skew-symmetry. For $u = \big(u_1, u_2\big)^T, v = \big(v_1, v_2\big)^T \in D(A_M)$ using the integration by parts formula (3.18) we have

$$\begin{aligned}
(A_M u, v)_V &= (\begin{pmatrix} \mu^{-1}\nabla \times u_2 \\ -\epsilon^{-1}\nabla \times u_1 \end{pmatrix}, \begin{pmatrix} v_1 \\ v_2 \end{pmatrix})_V \\
&= (\nabla \times u_2, v_1)_{0,\Omega} - (\nabla \times u_1, v_2)_{0,\Omega} \\
&= (u_2, \nabla \times v_1)_{0,\Omega} + \langle \mathbf{n} \times u_2, v_1 \rangle_{-\frac{1}{2}\times\frac{1}{2},\partial\Omega} \\
&\quad - (u_1, \nabla \times v_2)_{0,\Omega} - \langle \mathbf{n} \times u_1, v_2 \rangle_{-\frac{1}{2}\times\frac{1}{2},\partial\Omega} \\
&= (u_2, \nabla \times v_1)_{0,\Omega} - (u_1, \nabla \times v_2)_{0,\Omega} \\
&= -(\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \begin{pmatrix} \mu^{-1}\nabla \times v_2 \\ -\epsilon^{-1}\nabla \times v_1 \end{pmatrix})_V = -(u, A_M v)_V
\end{aligned}$$

since the boundary terms are equal to zero by the definition of $D(A_M)$. Hence

$$(A_M u, v)_V = -(u, A_M v)_V, \text{ for } u, v \in D(A_M), \tag{3.22}$$

i. e., A_M is skew-symmetric. Since $D(-A_M) = D(A_M)$, it remains to show $D(A_M^*) = D(A_M)$. The inclusion $D(A_M) \subset D(A_M^*)$ follows trivially from Definition 2.3. To show

the converse we take $v = \big(v_1, v_2\big)^T \in D(A_M^*)$. Then, there is $w = \big(w_1, w_2\big)^T \in V$ such that for every $u = \big(u_1, u_2\big)^T \in D(A_M)$ holds

$$(A_M u, v)_V = (u, w)_V$$

or

$$(\nabla \times u_2, v_1)_{0,\Omega} - (\nabla \times u_1, v_2)_{0,\Omega} = (\mu u_1, w_1)_{0,\Omega} + (\epsilon u_2, w_2)_{0,\Omega}.$$

After setting $u_1 = 0$ we get

$$\int_\Omega (\nabla \times u_2) \cdot v_1 = \int_\Omega u_2 \cdot (\epsilon w_2), \quad \forall u_2 \in H_0(\mathrm{curl}, \Omega),$$

and in particular for all $u_2 \in C_c^\infty(\Omega)^3$. By the definition of weak derivative we have $\nabla \times v_1 = \epsilon w_2 \in L^2(\Omega)^3$, and therefore, $v_1 \in H(\mathrm{curl}, \Omega)$. Similarly, by taking $u_2 = 0$

$$\int_\Omega (\nabla \times u_1) \cdot v_2 = \int_\Omega u_1 \cdot (-\mu w_1), \quad \forall u_1 \in H(\mathrm{curl}, \Omega),$$

and in particular for all $u_1 \in C_c^\infty(\Omega)^3$. Therefore we have $\nabla \times v_2 = -\mu w_1 \in L^2(\Omega)^3$ again. Moreover, by taking $u_1 \in H^1(\Omega)^3 \subset H(\mathrm{curl}, \Omega)$ and using the integral equality from above and integration by parts formula (3.18) we get

$$\langle \mathbf{n} \times v_2, u_1 \rangle_{-\frac{1}{2} \times \frac{1}{2}, \partial\Omega} = 0 \quad \forall u_1 \in H^1(\Omega)^3,$$

which implies $\mathbf{n} \times v_2 = 0$ and therefore $v_2 \in H_0(\mathrm{curl}, \Omega)$. We have proved $v \in D(A_M)$. □

The following result is a direct consequence of the proposition.

Corollary 3.2. *For all $u \in D(A_M)$ we have $(A_M u, u)_V = 0$.*

Now, by applying Stone's theorem (Theorem 2.4) we obtain the following results.

Theorem 3.3. *The operator $-A_M$ generates a C_0-group of unitary operators $S_M(t) = e^{-tA_M}$ on V for $t \in \mathbb{R}$. For $u_0 \in D(A_M)$ the homogeneous problem (3.21) has a unique solution given by $u(t) = S_M(t)u_0$. The solution satisfies $u \in C^1(\mathbb{R}, V) \cap C(\mathbb{R}, D(A_M))$.*

Remark. Since $-A_M$ generates a unitary C_0-group, we have that the **electromagnetic energy is conserved**, i. e., $\|u(t)\|_V = \|u(0)\|_V$ for all $t \in \mathbb{R}$.

In the homogeneous case the continuity equation (3.2) is satisfied and therefore the divergence equations are automatically fulfilled. Then, as we have mentioned in Section 3.1.3, the extended set of linear Maxwell's equations (3.7) is satisfied too (with $\rho = 0$ and $\mathbf{J} = 0$ of course). This can be mathematically formalized in the following way. We define the subspace V_0 of the Hilbert space V as

$$V_0 := \big\{(\mathbf{H}, \mathbf{E}) \in L^2(\Omega)^6 |\ \nabla \cdot (\epsilon \mathbf{E}) = 0,\ \nabla \cdot (\mu \mathbf{H}) = 0,\ \mathbf{n} \cdot (\mu \mathbf{H}) = 0\big\} \tag{3.23}$$

and the operator $A_{M,0}$ as the restriction of A_M on $D(A_M) \cap V_0$, i. e.,

$$D(A_{M,0}) = D(A) \cap V_0, \qquad A_{M,0} = A|_{D(A_{M,0})}.$$

Then, we have the following result [36, Proposition 3.5.].

Theorem 3.4. *The operator $-A_{M,0}$ generates a C_0-group of unitary operators $S_{M,0}(t) = e^{-tA_{M,0}}$ on V_0 for $t \in \mathbb{R}$. For $u_0 \in D(A_{M,0})$ the homogeneous problem*

$$\partial_t u + A_{M,0} u = 0, \qquad u(0) = u_0.$$

has a unique solution given by $u(t) = S_{M,0}(t)u_0$. The solution satisfies $u \in C^1(\mathbb{R}, V_0) \cap C(\mathbb{R}, D(A_{M,0}))$.

Inhomogeneous case

To show the well-posedness of the inhomogeneous problem (3.20) we apply Theorem 2.5.

Theorem 3.5 (well-posedness)**.** *Assume either that $f \in C^1([0,T], V)$ or that $f \in C^0([0,T], D(A_M))$. For $u_0 \in D(A_M)$ the unique solution of (3.20) is given by*

$$u(t) = S_M(t)u_0 + \int_0^t S_M(t-s)f(s)ds, \quad t \in [0,T].$$

The divergence equations will be fulfilled if they hold for the initial data and if the continuity equation (3.2) holds. Therefore, if $\mathbf{H}_0$, $\mathbf{E}_0$ and ρ are chosen such that $\partial_t \rho = -\nabla \cdot \mathbf{J}$, $\rho_0 = \nabla \cdot (\epsilon \mathbf{E}_0)$ and $\nabla \cdot (\mu \mathbf{H}_0) = 0$, then the extended set of linear Maxwell's equations (3.7) is satisfied too.

The well-posedness of the extended set (3.7) can be formally shown by application of Theorem 2.5 again.

Theorem 3.6. *Assume either that $f \in C^1([0,T], V_0)$ or that $f \in C^0([0,T], D(A_{M,0}))$. For $u_0 \in D(A_{M,0})$ the inhomogeneous problem*

$$\partial_t u + A_{M,0} u = f, \qquad u(0) = u_0,$$

has a unique solution given by

$$u(t) = S_{M,0}(t)u_0 + \int_0^t S_{M,0}(t-s)f(s)ds, \quad t \in [0,T].$$

CHAPTER 4

Discontinuous Galerkin method

The discontinuous Galerkin finite element method was first proposed by Reed and Hill in 1973 for the two dimensional steady neutron transport equation [57]. This is a first order hyperbolic partial differential equation of the form

$$\mu u + \beta \cdot \nabla u = f, \tag{4.1}$$

also known as the steady advection-reaction equation. The first analysis of this method was done by Lesaint and Raviart in 1974 in [44], where they proved an L^2-norm error estimate of order $\mathcal{O}(h^k)$ on general triangulations and of order $\mathcal{O}(h^{k+1})$ on a Cartesian grid if the exact solution is smooth enough. As mentioned in Section 2.4, k denotes the polynomial degree used in the discontinuous Galerkin method. This estimate was improved by Johnson and Pitkäranta in 1986. In [42] they showed the L^p-norm estimate of order $\mathcal{O}(h^{k+1/2})$ for $p \in [1, \infty]$ on general grids. More precisely, if u_h is the discrete solution obtained from the numerical method and $u \in W^{k+1,p}(\Omega)$ the exact solution of (4.1), then

$$\|u - u_h\|_{L^p(\Omega)} \leq C h^{k+1/2} \, |u|_{W^{k+1,p}(\Omega)} \, .$$

In 1991, Peterson confirmed the sharpness of this estimate numerically [54]. On specially constructed meshes ("Peterson meshes"), one can really see the order of $k + 1/2$. Therefore, the aforementioned estimate is optimal, in the sense that exponent of h cannot be increased while the regularity of u is kept. On the other hand, the estimate is not optimal in the sense that the polynomial interpolation or projection of degree k has order of $k + 1$. Also, in most numerical examples one really can see order $k + 1$. This raises the following question: which conditions on meshes should be satisfied in order to obtain full order $k + 1$? This problem was considered by Richter in 1988. In [58] he proved the optimal order of convergence on triangular meshes whose faces are uniformly not aligned with the flow direction, i. e. $|\beta \cdot \mathbf{n}|$ is uniformly positive for all normals $\mathbf{n}$

of all elements. The same question was discussed by Cockburn et al. in 2008. In [11] they proved the optimal convergence rate under some other (yet weaker) conditions on meshes.

From our point of view, unsteady problems are more interesting, i. e., problems which involve the partial derivative with respect to time, for example linear advection equation

$$\mu\partial_t u + \beta \cdot \nabla u = f. \tag{4.2}$$

In a series of papers [12, 13, 14, 15, 16], Cockburn, Shu and coworkers developed the so-called Runge–Kutta discontinuous Galerkin (RKDG) methods for the full discretization of a more general problem - nonlinear hyperbolic conservation laws

$$\partial_t u + \nabla \cdot F(u) = f.$$

RKDG methods combine the discontinuous Galerkin method for space discretization with strong stability preserving Runge–Kutta methods (see [24]) for time discretization. We now show that Maxwell's equations can be written in form of a linear multidimensional hyperbolic system, which is a generalization of (4.2) to a vector case. For the $\nabla\times$ operator we have

$$\begin{aligned}\nabla \times \mathbf{E} &= \begin{pmatrix} \partial_y E_z - \partial_z E_y \\ \partial_z E_x - \partial_x E_z \\ \partial_x E_y - \partial_y E_x \end{pmatrix} = \partial_x \begin{pmatrix} 0 \\ -E_z \\ E_y \end{pmatrix} + \partial_y \begin{pmatrix} E_z \\ 0 \\ -E_x \end{pmatrix} + \partial_z \begin{pmatrix} -E_y \\ E_x \\ 0 \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}}_{=:R_x} \partial_x E + \underbrace{\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}}_{=:R_y} \partial_y E + \underbrace{\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{=:R_z} \partial_z E\end{aligned}$$

By denoting $M := \begin{pmatrix} \mu I_3 & 0_{3,3} \\ 0_{3,3} & \epsilon I_3 \end{pmatrix}$ and $A_i := \begin{pmatrix} 0_{3,3} & R_i \\ R_i^T & 0_{3,3} \end{pmatrix}$ for $i \in \{x, y, z\}$ we can write the curl-equations as

$$M\partial_t u + A_x\partial_x u + A_y\partial_y u + A_z\partial_z u = f. \tag{4.3}$$

Here $A_i \in \mathbb{R}^{6\times 6}$ for $i = x, y, z$ are constant symmetric matrices, while $M = M(\mathbf{x})$ is space dependent. Hyperbolicity means that for every $\mathbf{x} \in \Omega$ any linear combination of the matrices $M^{-1}A_x$, $M^{-1}A_y$ and $M^{-1}A_z$ is diagonalizable with real eigenvalues [45].

The chapter is organized as follows: First, we consider a simple motivational example. We derive the discontinuous Galerkin method for the linear advection equation in one space dimension. Explaining the basic concepts and basic ideas of the numerical method for this simple example should facilitate the transition to the more complicated case of Maxwell's equations. In the second section, we introduce the basic concepts needed for the analysis and the understanding of the discontinuous Galerkin finite element method in more than one spatial dimension. Finally, in the last section, we apply the discontinuous Galerkin method to Maxwell's equations and provide the convergence analysis.

4.1 One dimensional linear advection equation

Let $I = [\underline{x}, \overline{x}] \subset \mathbb{R}$ be an interval. The linear advection equation can be written as conservation law

$$\partial_t u(x,t) + \partial_x F(u(x,t)) = f(x,t), \quad \text{in} \quad I \times [0,T], \tag{4.4}$$

where $F(u) = au$, with constant $a > 0$. The function F is called the flux. When supplied with the initial data and periodic boundary conditions (4.4) yields a well-posed problem

$$\begin{aligned} \partial_t u(x,t) + a\partial_x u(x,t) &= f(x,t), \\ u(x,0) &= u_0(x), \\ u(\underline{x},t) &= u(\overline{x},t). \end{aligned} \tag{4.5}$$

The same example is considered in [33, Chapter 2.]. The exact solution can be easily computed by using the method of characteristics. The characteristics for this equation are defined as a solution of a trivial ordinary differential equation

$$X'(t) = a, \quad X(0) = x_0 \in \mathbb{R}.$$

This gives

$$X(t) = at + x_0.$$

If we now define $U(t) := u(X(t), t)$, then

$$\frac{d}{dt}U(t) = f(X(t), t). \tag{4.6}$$

In the homogeneous case this implies that the solution is constant along characteristics. In particular, for $x_0 \in I$ we can compute the solution at a point $(x_0 + at, t)$ by following the characteristic back to the point $(x_0, 0)$ and by reading the initial value at the point x_0, i. e. $u(x_0 + at, t) = u_0(x_0)$, or equivalently

$$u(x,t) = u_0(x - at).$$

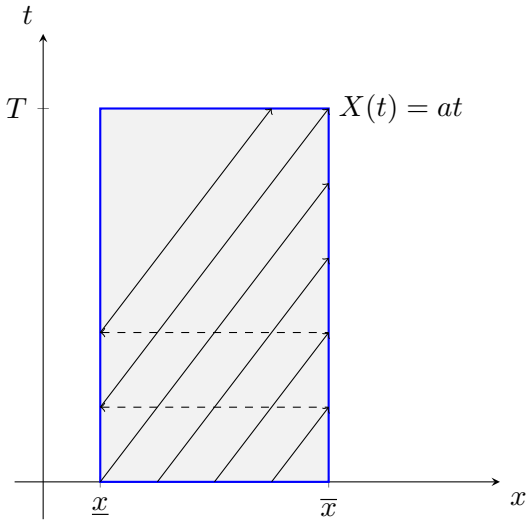
In Figure 4.1 the domain and characteristics are sketched. Since $a > 0$, both characteristics and the solution are propagating from left to right. As we can see in the figure, for higher characteristics $x = at + x_0$, with $x_0 < \underline{x}$, we have to use the periodicity argument. In the inhomogeneous case, we get, for $x_0 \in I$

$$u(x,t) = u_0(x - at) + \int_0^t f(x - a(t-s), s)ds,$$

i. e. the solution is not constant on characteristics anymore, but changes with the influence of the source term. Nevertheless, it is still easily computable.

For $u_0 \in H^1(\Omega)$, $u_0(\underline{x}) = u_0(\overline{x})$, and $f \in L^2(\Omega)$ the problem is well-posed and we have $u(.,t) \in H^1(I)$ for all $t \in [0,T]$. Since in one dimension $H^1(\underline{x}, \overline{x}) \subset C(\underline{x}, \overline{x})$ holds, this implies the continuity of the solution u.

Figure 4.1: Domain and characteristics



4.1.1 Discrete space and notation

Let $N \in \mathbb{N}$ and let

$$\underline{x} = x_0 \leq x_1 < \ldots < x_{n-1} < x_n < \ldots < x_N = \overline{x}$$

be a subdivision of the interval I and let $I_n = [x_{n-1}, x_n]$. Points $\{x_1, \ldots, x_{n-1}\}$ are called **interfaces**. Let h be the length of largest interval I_n, $n = 1, \ldots N$. We define the discrete space

$$V_h = \left\{ v_h \in L^2(\underline{x}, \overline{x}) : v_h|_{I_n} \in \mathbb{P}^k(I_n),\ \forall n = 1, \ldots, N \right\}, \tag{4.7}$$

which we call the discontinuous Galerkin space. This is the function space in which we approximate the solution. The functions $v_h \in V_h$ may not be continuous over interfaces and therefore can take two values on each x_n, $n = 1, \ldots, N-1$. We denote these two values by $v_h(x_n^-)$ and $v_h(x_n^+)$ as values from the left and the right, respectively. Furthermore, we use the following notation for average and jump over the interface x_n

$$\begin{aligned} \{\{v_h\}\}_{x_n} &= \frac{v_h(x_n^-) + v_h(x_n^+)}{2}, \\ [\![v_h]\!]_{x_n} &= v_h(x_n^+) - v_h(x_n^-). \end{aligned}$$

For the boundary points we use

$$\begin{aligned} \{\{v_h\}\}_{x_0} = \{\{v_h\}\}_{x_N} &= \frac{v_h(x_N^-) + v_h(x_0^+)}{2}, \\ [\![v_h]\!]_{x_0} = [\![v_h]\!]_{x_N} &= v_h(x_0^+) - v_h(x_N^-). \end{aligned} \tag{4.8}$$

With π_h we denote the standard L^2-projection on the discrete space V_h.

4.1.2 Heuristical derivation of the method

We multiply (4.4) by a test function $\phi_h \in V_h$ and integrate over I_n to get

$$\int_{I_n} (\partial_t u(x,t) + a\partial_x u(x,t))\, \phi_h(x) dx = \int_{I_n} f(x,t)\phi_h(x) dx, \quad \forall n = 1, \dots, N.$$

In the rest of the section, for the sake of readability, we omit the variables whenever it is possible. Integrating the second term by parts we obtain that the exact solution satisfies

$$\int_{I_n} (\partial_t u)\phi_h - \int_{I_n} (au)\partial_x \phi_h + (au)(x_n)\phi_h(x_n^-) - (au)(x_{n-1})\phi_h(x_{n-1}^+) = \int_{I_n} f\phi_h$$

for all test functions $\phi_h \in V_h$ and for all $n = 1, \dots, N$. We notice here, that the expressions $(au)(x_n)$ and $(au)(x_{n-1})$ are well-defined, since au is a continuous function. For the function ϕ_h we have used values from the interval I_n. When trying to replace the exact solution u by a function u_h from the discrete space V_h, we encounter difficulties since $(au_h)(x_n)$ and $(au_h)(x_{n-1})$ are not well-defined. Therefore we replace these terms by **numerical fluxes** $(au_h)^*(x_n)$ and $(au_h)^*(x_{n-1})$ which are to be defined later. The name comes from the fact that they approximate the **flux** function $F(u_h)$ at interfaces. We obtain a numerical scheme: find a discrete function u_h such that

$$\int_{I_n} (\partial_t u_h)\phi_h - \int_{I_n} (au_h)\partial_x \phi_h + (au_h)^*(x_n)\phi_h(x_n^-) - (au_h)^*(x_{n-1})\phi_h(x_{n-1}^+) = \int_{I_n} f\phi_h, \quad (4.9)$$

for all test function $\phi_h \in V_h$ and for all $n = 1, \dots, N$. This form of the method is also called the **weak form**. If we integrate by parts once again, and use the values from the interval I_n for u_h we get the **strong form**

$$\begin{aligned} \int_{I_n} (\partial_t u_h + \partial_x(au_h))\phi_h &+ \phi_h(x_n^-)\left((au_h)^*(x_n) - (au_h)(x_n^-)\right) \\ &- \phi_h(x_{n-1}^+)\left((au_h)^*(x_{n-1}) - (au_h)(x_{n-1}^+)\right) = \int_{I_n} f\phi_h. \end{aligned} \quad (4.10)$$

It is clear that the weak and the strong form are equivalent and from now on we will use the strong form only.

To complete the numerical scheme, we now deal with the choice of the numerical flux. This is the central point of the discontinuous Galerkin method. In the next subsection we present some of the most popular choices. These choices will have their analogs in the more complicated case of Maxwell's equations.

4.1.3 Choice of numerical flux

The numerical flux $F^*(u_h) = (au_h)^*$ at the point x_n is chosen to be a function of the function value from the left $u_h(x_n^-)$ and the function value from the right $u_h(x_n^+)$

$$F^*(u_h(x_n)) = g(u_h(x_n^-), u_h(x_n^+)).$$

A reasonable request on the numerical flux F^* is to be equal to the exact flux F in the case when u_h is a continuous function, i. e. when $u(x_n^-) = u(x_n^+)$. This is known as the **consistency** condition of g and can be written compactly as

$$g(u,u) = F(u), \quad \text{for every} \quad u \in \mathbb{R}.$$

Central flux

The simplest choice one can think of is the one of the central flux. Here the idea is to take the mean of the left and the right value, i. e.

$$F^*(u_h(x_n)) = (au_h)^*(x_n) = \{\{au_h\}\}_{x_n}. \tag{4.11}$$

Inserting this choice into the strong form above yields

$$\int_{I_n} (\partial_t u_h + \partial_x(au_h))\phi_h + \frac{1}{2}\phi_h(x_n^-)[\![au_h]\!]_{x_n} + \frac{1}{2}\phi_h(x_{n-1}^+)[\![au_h]\!]_{x_{n-1}} = \int_{I_n} f\phi_h,$$

for all $n = 2, \dots N-1$ and all $\phi_h \in V_h$. This is the **local formulation** of the discrete problem. For the boundary points x_0 and x_N, the condition

$$u_h(x_N^-) = u_h(x_0^+)$$

will be satisfied weakly in the dG scheme. Boundary points x_0 and x_N can be seen as one point at which u_h has to be continuous. We use the central flux (4.11) also for the boundary points, as defined in (4.8). Taking these considerations into account and summing over all subintervals I_n, we obtain the following **global formulation**

$$\int_{\underline{x}}^{\overline{x}} (\partial_t u_h)\phi_h + \sum_{n=1}^{N} \int_{I_n} \partial_x(au_h)\phi_h + \sum_{n=0}^{N-1} \{\{\phi_h(x_n)\}\}[\![au_h]\!]_{x_n} = \int_{\underline{x}}^{\overline{x}} f\phi_h, \quad \forall \phi_h \in V_h. \tag{4.12}$$

For the initial condition we take $u_h(\cdot, 0) = \pi_h u_0$. As we will see in the case of Maxwell's equations, the central flux, although very simple, it is not the best choice and leads to suboptimal error estimates. Therefore, the following alternative is a more popular choice.

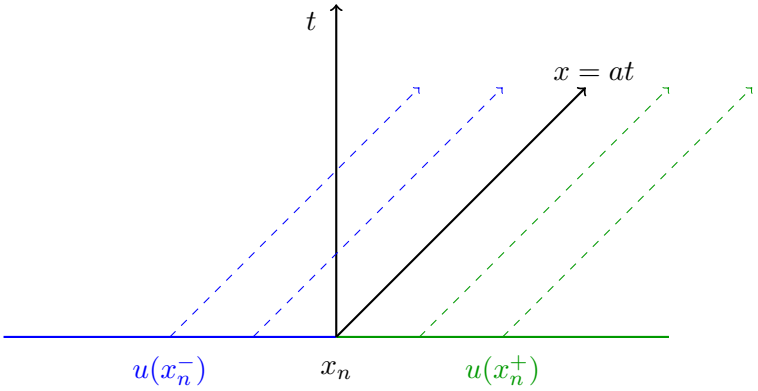
Upwind flux

The construction of the upwind flux can be easily explained by Figure 4.2. We would like to approximate the value of u_h at the point x_n. Values from the left, $u_h(x_n^-)$, and from the right, $u_h(x_n^+)$, are given. Since u_h is smooth (polynomial) on both I_n and I_{n+1}, we further approximate $u_h \approx u_h(x_n^-)$ on some small interval left from x_n and $u_h \approx u_h(x_n^+)$ on some small interval right from x_n, and take this as the initial value for our problem (4.4). We know that the exact solution will travel from the left to the right on characteristics of the form $x = at + const$. This means that for some small $t_* > 0$ the

solution on the line from $(x_n, 0)$ to (x_n, t_*) will be equal to $u_h(x_n^-)$ (or approximately equal if there is a force term $f \neq 0$). Therefore, we take the value from the left to approximate the flux $(au_h)^*(x_n)$, i. e.

$$F^*(u_h(x_n)) = (au_h)^*(x_n) = au_h(x_n^-). \tag{4.13}$$

Figure 4.2: Construction of upwind fluxes



The problem that appeared above, which consists of a partial differential equation and an initial value of the form

$$u_0(x) = \begin{cases} u_L, & x < x_*, \\ u_R, & x > x_*, \end{cases}$$

is known as Riemann problem [45, Chapter 3]. In this case, we had a trivial one to solve, but in the case of linear system of partial differential equations or in a multidimensional case, the situation becomes more complicated and one should study the theory of Riemann solvers in more details.

We can rewrite (4.13) as a correction of central fluxes with the additional jump term

$$F^*(u_h(x_n)) = \{\{au_h\}\}_{x_n} - \frac{1}{2}[\![au_h]\!]_{x_n}. \tag{4.14}$$

Inserting this into the strong form (4.10) yields

$$\int_{I_n} (\partial_t u_h + \partial_x(au_h))\phi_h + \frac{1}{2}\phi_h(x_n^-)\Big([\![au_h]\!]_{x_n} - [\![au_h]\!]_{x_n}\Big) + \frac{1}{2}\phi_h(x_{n-1}^+)\Big([\![au_h]\!]_{x_{n-1}} + [\![au_h]\!]_{x_{n-1}}\Big) = \int_{I_n} f\phi_h,$$

for all $n = 2, \dots N-1$ and all $\phi_h \in V_h$. The flux for the boundary points is determined by the same formula (4.14) as the flux for the inner ones, using (4.8) again. If we just

naively sum and subtract the terms in brackets, after summing over all intervals, we get the following global formulation

$$\int_{\underline{x}}^{\overline{x}} (\partial_t u_h)\phi_h + \sum_{n=1}^{N} \int_{I_n} \partial_x (a u_h)\phi_h + \sum_{n=0}^{N-1} \phi_h(x_n^+) [\![a u_h]\!]_{x_n} = \int_{\underline{x}}^{\overline{x}} f\phi_h, \quad \forall \phi_h \in V_h.$$

We see that the only difference between this equation and the global formulation (4.12) for the central flux method is that in the second sum the mean value of the test function is replaced by the value from the right. Without summing and subtracting terms in brackets we get the global formulation

$$\begin{aligned} \int_{\underline{x}}^{\overline{x}} (\partial_t u_h)\phi_h + \sum_{n=1}^{N} \int_{I_n} \partial_x (a u_h)\phi_h + \sum_{n=0}^{N-1} \{\{\phi_h\}\}_{x_n} [\![a u_h]\!]_{x_n} \\ + \frac{1}{2} \sum_{n=0}^{N-1} [\![\phi_h]\!]_{x_n} [\![a u_h]\!]_{x_n} = \int_{\underline{x}}^{\overline{x}} f\phi_h, \qquad \forall \phi_h \in V_h, \end{aligned} \tag{4.15}$$

which may seem more complicated, but as we will see has some advantages. The term

$$\frac{1}{2} \sum_{n=0}^{N-1} [\![\phi_h]\!]_{x_n} [\![a u_h]\!]_{x_n}$$

is called a "stabilization" term and we see that this scheme differs from the scheme with central flux (4.12) by this term only.

Other numerical fluxes

As we will see in the next subsection, any convex combination of central flux and upwind flux will provide us with a stable method. Stability here means that the L^2-norm of the discrete solution u_h does not grow with time, as long as the norm of the exact solution stays bounded too. So, any flux of the form

$$F^*(u_h(x_n)) = \{\{a u_h\}\}_{x_n} - \frac{1}{2}\alpha [\![a u_h]\!]_{x_n}, \tag{4.16}$$

with $\alpha \in [0, 1]$ is a good choice for a numerical method.

4.1.4 Operators, stability and consistency

Continuous case

Let us consider the continuous problem (4.5) again. If we define

$$A : D(A) \to L^2(\underline{x}, \overline{x}), \qquad Au = a\partial_x u, \tag{4.17}$$

with domain $D(A) = \{v \in H^1(\underline{x}, \overline{x}) \mid v(\underline{x}) = v(\overline{x})\}$, then we can rewrite it as

$$\begin{aligned} \partial_t u(t) + Au(t) &= f(t), \\ u(0) &= u_0. \end{aligned} \tag{4.18}$$

In this notation, $u(t)$ is a space function for every $t > 0$, and we are looking for a solution $u \in C^1([0,T], L^2(\underline{x}, \overline{x}))$, such that $u(t) \in D(A)$ for all $t \in [0,T]$. For the operator A the skew-symmetry property holds

$$(Au, v)_0 = -(u, Av)_0 \quad \text{for every } u, v \in D(A).$$

Indeed $(Au, v)_0 = (a\partial_x u, v)_0 = -(au, \partial_x v)_0 + a((uv)(\overline{x}) - (uv)(\underline{x})) = -(u, Av)_0$. In particular this implies

$$(Au, u)_0 = 0 \quad \text{for every } u \in D(A). \tag{4.19}$$

In the absence of the source ($f = 0$) this gives us **stability**, or more precisely, the norm of the exact solution is constant

$$\frac{d}{dt} \|u(t)\| = 0.$$

We would like to have similar properties in the discrete case too.

Central flux

The dG scheme with central flux (4.12) can be written in the compact form

$$\begin{aligned} \partial_t u_h(t) + A_h^{\mathrm{cf}} u_h(t) &= \pi_h f(t), \\ u_h(0) &= \pi_h u_0, \end{aligned} \tag{4.20}$$

with $u_h \in C^1([0,T], V_h)$ and the discrete operator $A_h^{\mathrm{cf}} : V_h \to V_h$ defined by

$$(A_h^{\mathrm{cf}} u_h, \phi_h)_0 := \sum_{n=1}^{N} \int_{I_n} \partial_x (a u_h) \phi_h + \sum_{n=0}^{N-1} \{\{\phi_h\}\}_{x_n} [\![a u_h]\!]_{x_n} \tag{4.21}$$

for every $u_h, \phi_h \in V_h$. The following discrete analog of (4.19) holds.

Lemma 4.1. *For every $u_h \in V_h$ we have*

$$(A_h^{cf} u_h, u_h)_0 = 0. \tag{4.22}$$

Proof. We have

$$(A_h^{\mathrm{cf}} u_h, u_h)_0 = \sum_{n=1}^{N} \int_{I_n} \partial_x (a u_h) u_h + \sum_{n=0}^{N-1} \{\{u_h\}\}_{x_n} [\![a u_h]\!]_{x_n}.$$

The integration by parts formula gives

$$\int_{I_n} \partial_x(au_h)u_h = \frac{1}{2}a\Big(u_h^2(x_n^-) - u_h^2(x_{n-1}^+)\Big).$$

For the second sum we use

$$\{\{u_h\}\}_{x_n}[\![au_h]\!]_{x_n} = \frac{1}{2}a\Big(u_h^2(x_n^+) - u_h^2(x_n^-)\Big) \quad \text{and}$$
$$\{\{u_h\}\}_{x_0}[\![au_h]\!]_{x_0} = \frac{1}{2}a\Big(u_h^2(x_0^+) - u_h^2(x_N^-)\Big).$$

After summing up, all terms cancel, and we proved the result. □

The dG method with central flux is therefore **stable** and for the discrete solution u_h of (4.20) with $f = 0$ we have

$$\frac{d}{dt}\|u_h(t)\| = 0.$$

Another important property of the central flux method is the **consistency** property. We can extend the operator A_h^{cf} to $V_h + D(A)$ using the same formula (4.21). Then for $u \in D(A)$, because of continuity, holds

$$(A_h^{\mathrm{cf}}u, \phi_h)_0 = \sum_{n=1}^{N}\int_{I_n} a(\partial_x u)\phi_h + \sum_{n=0}^{N-1}\{\{\phi_h\}\}_{x_n}[\![au]\!]_{x_n} = \int_{\underline{x}}^{\overline{x}} a\partial_x u\phi_h = (Au, \phi_h)_0,$$

for all $\phi_h \in V_h$, or equivalently $\pi_h Au = A_h^{\mathrm{cf}}u$. Now, if we apply the L^2-projection on the first equation of (4.18) and use $\pi_h\partial_t = \partial_t\pi_h$, we get that the exact solution satisfies

$$\partial_t\pi_h u(t) + A_h^{\mathrm{cf}}u(t) = \pi_h f(t).$$

Subtracting this equation from the first equation in (4.20) gives the error equation

$$\partial_t e_h(t) + A_h^{\mathrm{cf}}e_h(t) = A_h^{\mathrm{cf}}e_\pi(t), \tag{4.23}$$

with $e_h = u_h - \pi_h u$ and $e_\pi = u - \pi_h u$.

Using Lemma 4.1 and polynomial approximation properties in $\mathbb{R}$, the convergence of order $\mathcal{O}(h^{k+1/2})$ can be shown. For more details we refer the reader to [33, Section 4.5].

Upwind flux

Although the central flux method is stable from a theoretical point of view, numerically it can be unstable. From (4.22) it follows that all eigenvalues of A_h^{cf} are on the imaginary axis. Taking the rounding errors into account we might have eigenvalues with positive real part which can cause instability. As we have already seen, the upwind flux dG

method has the additional stabilization term which prevents this from happening. The dG discrete operator with the upwind flux $A_h^{\mathrm{upw}} : V_h \to V_h$ is defined by

$$(A_h^{\mathrm{upw}} u_h, \phi_h)_0 := \sum_{n=1}^{N} \int_{I_n} \partial_x (a u_h) \phi_h + \sum_{n=0}^{N-1} \{\{\phi_h\}\}_{x_n} [\![a u_h]\!]_{x_n} + \frac{1}{2} \sum_{n=0}^{N-1} [\![\phi_h]\!]_{x_n} [\![a u_h]\!]_{x_n},$$

for every $u_h, \phi_h \in V_h$. If we define the stabilization form $S_h : V_h \to V_h$ by

$$(S_h u_h, \phi_h)_0 := \frac{1}{2} \sum_{n=0}^{N-1} [\![\phi_h]\!]_{x_n} [\![a u_h]\!]_{x_n},$$

then it is clear that $A_h^{\mathrm{upw}} = A_h^{\mathrm{cf}} + S_h$. The dG scheme with upwind flux (4.15) can be rewritten as

$$\begin{aligned} \partial_t u_h(t) + A_h^{\mathrm{upw}} u_h(t) &= \pi_h f(t), \\ u_h(0) &= \pi_h u_0, \end{aligned} \tag{4.24}$$

where $u_h \in C^1([0,T], V_h)$. Since obviously $(S_h u_h, u_h)_0 \geq 0$ for all $u_h \in V_h$, Lemma 4.1 immediately yields

Corollary 4.2. *For every $u_h \in V_h$ we have*

$$(A_h^{\mathrm{upw}} u_h, u_h)_0 = \frac{a}{2} \sum_{n=0}^{N-1} [\![u_h]\!]_{x_n}^2 \geq 0.$$

As before, we recover the **stability** of the dG scheme, i. e.

$$\frac{d}{dt} \|u_h(t)\| \leq 0$$

in the case without source term. As in the case of the central flux method, we extend the operator to $V_h + D(A)$ and see that

$$\pi_h A u = A_h^{\mathrm{upw}} u, \quad \text{for all } u \in D(A). \tag{4.25}$$

By using this, we obtain the error equation analogously to the central flux case

$$\partial_t e_h(t) + A_h^{\mathrm{upw}} e_h(t) = A_h^{\mathrm{upw}} e_\pi(t). \tag{4.26}$$

The convergence of order $\mathcal{O}(h^{k+1})$ can be shown, see [33, Section 4.5].

Other fluxes

As already stated, any convex combination of central flux and upwind flux will provide us with a stable and convergent numerical method. For $\alpha \in [0,1]$ we define the operator $A_h^\alpha : V_h + D(A) \to V_h$ by

$$A_h^\alpha := A_h^{\mathrm{cf}} + \alpha S_h.$$

Obviously, $(A_h^\alpha u_h, u_h)_0 \geq 0$ for all $u_h \in V_h$ and the method is stable. The consistency and the error equation are obtained as before.

4.2 Introduction to the discontinuous Galerkin method

As in other numerical methods, the idea of the discontinuous Galerkin method is to approximate the exact solution in a finite-dimensional function space. To construct the discontinuous Galerkin function space, we follow mostly [55]. We start by an assumption on the domain Ω.

Definition 4.3 (cf. Definition 1.47 of [22])**.** *A* ***polyhedron*** *in* $\mathbb{R}^3$ *is a domain whose boundary is a finite union of polygons.*

Throughout this thesis we suppose the following:

Assumption 4.4. *The domain* Ω *is a polyhedron in* $\mathbb{R}^3$*.*

This assumption enables us to cover the domain with a mesh consisting of polyhedral elements only. Otherwise, if we would allow domains with a curved boundary, then the meshes with isoparametric elements would have to be used to approximate the domain correctly. This would make things more complicated. More about isoparametric finite elements can be found in [33, Chapter 9].

Since Ω is a polyhedron, we know that the outward unit norma $\mathbf{n}$ is defined a. e. on $\partial\Omega$. The first step in constructing a discrete function space is to discretize the domain Ω. Therefore, we introduce the concept of meshes. Secondly, we define some important **broken functional spaces** and state some useful properties. In particular, we define **broken polynomial spaces**, which are known to be discontinuous Galerkin function spaces, i. e. function spaces in which we search for a numerical solution. Finally, we discuss conditions on the meshes such that the **optimal polynomial approximation properties** hold. For this purpose, the concept of **admissible mesh sequences** is introduced.

4.2.1 Meshes

We start with the most familiar and most simple case, that of simplicial meshes.

Definition 4.5. *Let* $S = \{a_0, \dots, a_d\}$ *be a set of* $d+1$ *points in* $\mathbb{R}^d$ *such that the vectors* $\{a_1 - a_0, \dots, a_d - a_0\}$ *are linearly independent. The interior of the convex hull of* S *is called a* ***non-degenerate simplex*** *in* $\mathbb{R}^d$*, and the points in* S *are its* ***vertices****.*

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

Definition 4.6. *A finite set* $\mathcal{T} = \{K\}$ *is called a* ***simplical mesh*** *of the domain* Ω *if:*

i) every $K \in \mathcal{T}$ *is a non-degenerate simplex*

ii) for every $K_i, K_j \in \mathcal{T}$*,* $K_i \neq K_j$ *we have* $K_i \cap K_j = \emptyset$

iii) $\mathcal{T}$ *forms a partition of* Ω*, i. e.* $\overline{\Omega} = \bigcup_{K \in \mathcal{T}} \overline{K}$

Each $K \in \mathcal{T}$ is called a ***mesh element****.*

Although, we will use simplical meshes in our numerical experiments exclusively, the convergence results we present in this thesis are valid for a more general class of meshes, namely **general meshes** that consist of polyhedra.

Definition 4.7. *A* ***general mesh*** *$\mathcal{T}$ of the domain Ω is a finite set of polyhedra $\mathcal{T} = \{K\}$ satisfying ii) and iii) of Definition 4.6. Each $K \in \mathcal{T}$ is called a* ***mesh element****.*

A simplicial mesh is obviously just a special case of a general mesh.

Definition 4.8. *Let $\mathcal{T}$ be a general mesh of the domain Ω. For all $K \in \mathcal{T}$, we denote the* ***diameter*** *of K by h_K. By r_K we denote the* ***radius*** *of the largest ball inscribed in K. The* ***meshsize*** *is then defined as*

$$h := \max_{K \in \mathcal{T}} h_K. \tag{4.27}$$

We use the notation $\mathcal{T}_h$ for a mesh $\mathcal{T}$ with meshsize h.

Definition 4.9. *Let $\mathcal{T}_h$ be a general mesh of the domain Ω and $K \in \mathcal{T}_h$. We define $\mathbf{n}_K$ almost everywhere on ∂K as the* ***outward unit normal to*** *K.*

Mesh faces, averages and jumps

These concepts play an essential role in the design and the analysis of discontinuous Galerkin methods.

Definition 4.10. *Let $\mathcal{T}_h$ be a mesh of the domain Ω. A closed subspace F of $\overline{\Omega}$ is a* ***mesh face*** *if F has a positive $d-1$ dimensional Hausdorff measure and either of the two following conditions is satisfied:*

i) there are distinct mesh elements K_1 and K_2 such that $F = \partial K_1 \cap \partial K_2$; then, F is called ***interface***

ii) there is a mesh element K such that $F = \partial K \cap \partial\Omega$; then, F is called ***boundary face****.*

We denote the set of all interfaces by $\mathcal{F}_h^i$ and the set of all boundary faces by $\mathcal{F}_h^b$. Further on, we set

$$\mathcal{F}_h := \mathcal{F}_h^i \cup \mathcal{F}_h^b.$$

For any mesh element $K \in \mathcal{T}_h$, the set

$$\mathcal{F}_K := \{F \in \mathcal{F}_h \mid F \subset \partial K\}$$

collects mesh faces composing the boundary of K.

We observe that in the case of simplicial meshes, interfaces are always parts of hyperplanes, but this must not be the case for general meshes. We continue with the definition of averages and jumps across the interfaces for piecewise smooth functions. Let in the following $v : \Omega \to \mathbb{R}$ be a function such that for every $K \in \mathcal{T}_h$, $v|_K$ is smooth enough to have traces a. e. on ∂K (on example $v|_K \in H^1(K)$). Then, on all $F \in \mathcal{F}_h^i$, v admits a possibly two-valued trace.

For each $F \in \mathcal{F}_h^i$ the choice of K_1 and K_2 is arbitrary but fixed in what follows. This will be important for the following definitions.

Definition 4.11. *Let $F = \partial K_1 \cap \partial K_2 \in \mathcal{F}_h^i$. For a. e. $\mathbf{x} \in F$ we define*

*i) the **average** of v as $\{\{v\}\}_F(\mathbf{x}) := \frac{1}{2}\left(v_{K_1}(\mathbf{x}) + v_{K_2}(\mathbf{x})\right)$ and*

*ii) the **jump** of v as $[\![v]\!]_F(\mathbf{x}) := v_{K_2}(\mathbf{x}) - v_{K_1}(\mathbf{x})$.*

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

Definition 4.12. *For all $F \in \mathcal{F}_h$ and a. e. $\mathbf{x} \in F$ we define the unit normal $\mathbf{n}_F$ to F at $\mathbf{x}$ as*

i) $\mathbf{n}_{K_1}$, the unit normal to F at $\mathbf{x}$ pointing from K_1 to K_2 if $F = \partial K_1 \cap \partial K_2$,

ii) $\mathbf{n}$, the unit outward normal to Ω if $F \in \mathcal{F}_h^b$.

Remark. Unit normal $\mathbf{n}_F$ is fixed for each face F. Both jump $[\![v]\!]_F$ and unit normal $\mathbf{n}_F$ depend on the choice of K_1 and K_2 but in the numerical scheme we use only the cross-product $\mathbf{n}_F \times [\![v]\!]_F$ which is in dependent of this choice.

4.2.2 Broken functional spaces

A range of broken functional spaces can be constructed with respect to the mesh on Ω. The discontinuous Galerkin finite element space is chosen as a space consisting of piecewise polynomial functions.

Broken polynomial space (dG space)

Let $k \geq 0$ be an integer. We define the set $\mathbb{A}_d^k$ as

$$\mathbb{A}_d^k := \left\{ \alpha \in \mathbb{N}^d : |\alpha|_{l_1} \leq k \right\}. \tag{4.28}$$

The set of all polynomials of d variables of total degree at most k is defined as

$$\mathbb{P}_d^k := \left\{ p : \mathbb{R}^d \to \mathbb{R} \mid \exists \gamma_\alpha \in \mathbb{R} \text{ for } \alpha \in \mathbb{A}_k^d \text{ s. t. } p(\mathbf{x}) = \sum_{\alpha \in \mathbb{A}_d^k} \gamma_\alpha \mathbf{x}^\alpha \right\}$$

where for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $\alpha = (\alpha_1, \ldots, \alpha_d)$ is $\mathbf{x}^\alpha := \prod_{i=1}^d x_i^{\alpha_i}$. The dimension of the vector space $\mathbb{P}_d^k$ is the same as the cardinal number of the set $\mathbb{A}_d^k$ and it equals $\binom{k+d}{k}$. For any polyhedral set $K \subset \mathbb{R}^d$, we define the space of polynomials of d variables, of total degree at most k on K as a set of restrictions to K of all polynomials in $\mathbb{P}_d^k$, i. e.

$$\mathbb{P}_d^k(K) := \left\{ p|_K : K \to \mathbb{R} : p \in \mathbb{P}_d^k \right\}. \tag{4.29}$$

The **broken polynomial space** on the mesh $\mathcal{T}_h$ is now defined as

$$\mathbb{P}_d^k(\mathcal{T}_h) := \left\{ v \in L^2(\Omega) \,|\, \forall K \in \mathcal{T}_h,\ v|_K \in \mathbb{P}_d^k(K) \right\}. \tag{4.30}$$

Obviously

$$\dim(\mathbb{P}_d^k(\mathcal{T}_h)) = \mathrm{card}(\mathcal{T}_h) \times \dim(\mathbb{P}_d^k).$$

$\mathbb{P}_d^k(\mathcal{T}_h)$ is known as the discontinuous Galerkin finite element space and we denote it by V_h.

Definition 4.13. *By π_h we denote the L^2-**orthogonal projection** on a dG space $\mathbb{P}_d^k(\mathcal{T}_h)$.*

We misuse the notation: for a vector-valued functions $v \in L^2(\Omega)^n$, $\pi_h v$ is the L^2-orthogonal projection onto $\mathbb{P}_d^k(\mathcal{T}_h)^n$.

Broken Sobolev spaces

Broken Sobolev spaces and broken gradient operators are of great importance for the analysis of discontinuous Galerkin methods. For $m \in \mathbb{N}$ we define a broken Sobolev space

$$H^m(\mathcal{T}_h) := \left\{ v \in L^2(\Omega) |\, \forall K \in \mathcal{T}_h,\ v|_K \in H^m(K) \right\}, \tag{4.31}$$

which is a Hilbert space with the $\|\cdot\|^2_{H^m(\mathcal{T}_h)}$ defined via

$$|u|^2_{H^m(\mathcal{T}_h)} := \sum_{K \in \mathcal{T}_h} |u|^2_{H^m(K)}, \qquad \|u\|^2_{H^m(\mathcal{T}_h)} := \sum_{j=0}^m |u|^2_{H^j(\mathcal{T}_h)}.$$

Obviously $H^{m_1}(\mathcal{T}_h) \subset H^{m_2}(\mathcal{T}_h)$ for $m_1 \geq m_2$. The broken gradient can now be defined on $H^1(\mathcal{T}_h)$.

Definition 4.14. *The broken gradient $\nabla_h : H^1(\mathcal{T}_h) \to L^2(\Omega)^d$ is defined such that for all $v \in H^1(\mathcal{T}_h)$,*

$$(\nabla_h v)|_K := \nabla(v|_K) \quad \textit{for all } K \in \mathcal{T}_h.$$

It is not hard to see that the usual Sobolev spaces are subspaces of broken Sobolev spaces, i. e. $H^m(\Omega) \subset H^m(\mathcal{T}_h)$ holds for all $m \in \mathbb{N}$. Moreover, $\nabla_h v = \nabla v$ in $L^2(\Omega)^d$ holds for all $v \in H^1(\Omega)$, see [55, Lemma 1.22]. The reverse inclusion does not hold in general. The reason is because functions in the Sobolev space $H^1(\Omega)$ have zero jumps across interfaces, while this does not hold for functions in the broken Sobolev space $H^1(\mathcal{T}_h)$. The following characterization of $H^1(\Omega)$ is given in Lemma 1.23 of [55].

Lemma 4.15. *A function $v \in H^1(\mathcal{T}_h)$ belongs to $H^1(\Omega)$ if and only if*

$$[\![v]\!]_F = 0 \qquad \forall F \in \mathcal{F}_h^i.$$

Similarly, functions in $H(\mathrm{div};\Omega)$ introduced in Section 3.2 can be characterized by using jumps across the interfaces, but this time the **normal** component of jumps only.

Lemma 4.16. *A function $v \in H^1(\mathcal{T}_h)^d$ belongs to $H(\mathrm{div};\Omega)$ if and only if*

$$\mathbf{n}_F \cdot [\![v]\!]_F = 0 \qquad \forall F \in \mathcal{F}_h^i. \tag{4.32}$$

Proof. [55, Lemma 1.24] and the fact that $L^2(K) \subset L^1(K)$ for K bounded. □

The broken version of $H(\mathrm{curl},\Omega)$

We also define the broken version of the $H(\mathrm{curl},\Omega)$ space introduced in Section 3.2 as

$$H(\mathrm{curl},\mathcal{T}_h) := \left\{ v \in L^2(\Omega)^3 \mid \forall K \in \mathcal{T}_h,\ v \in H(\mathrm{curl},K) \right\}.$$

The broken curl operator $\nabla_h \times : H(\mathrm{curl},\mathcal{T}_h) \to L^2(\Omega)$ is defined as

$$(\nabla_h \times v)|_K := \nabla \times (v|_K) \quad \text{for all } K \in \mathcal{T}_h.$$

As before, $H(\mathrm{curl},\Omega) \subset H(\mathrm{curl},\mathcal{T}_h)$ and $\nabla_h \times v = \nabla \times v$ for $v \in H(\mathrm{curl},\Omega)$. Functions in $H(\mathrm{curl},\Omega)$ are characterized by a zero **tangential** jump across the interfaces.

Lemma 4.17. *A function $v \in H^1(\mathcal{T}_h)^3$ belongs to $H(\mathrm{curl},\Omega)$ if and only if*

$$\mathbf{n}_F \times [\![v]\!]_F = 0 \qquad \forall F \in \mathcal{F}_h^i. \tag{4.33}$$

Proof. Let $\phi \in C_0^\infty(\Omega)^3$. Applying the integration by parts formula (3.18) on each mesh element yields

$$\begin{aligned}\int_\Omega (\nabla_h \times v)\cdot\phi = \sum_{K\in\mathcal{T}_h}\int_K \nabla\times(v|_K)\cdot\phi &= \sum_{K\in\mathcal{T}_h}\int_K v\cdot(\nabla\times\phi) + \sum_{K\in\mathcal{T}_h}\int_{\partial K}(\mathbf{n}_K\times v|_K)\cdot\phi \\ &= \int_\Omega v\cdot(\nabla\times\phi) - \sum_{F\in\mathcal{F}_h^i}\int_F (\mathbf{n}_F\times[\![v]\!]_F)\cdot\phi.\end{aligned}$$

Hence, if (4.33) holds, we have

$$\int_\Omega (\nabla_h\times v)\cdot\phi = \int_\Omega v\cdot(\nabla\times\phi), \quad \forall\phi\in C_0^\infty(\Omega)^3$$

implying that $\nabla\times v = \nabla_h\times v \in L^2(\Omega)^3$. This proves sufficiency. To prove the necessity, let us assume $v \in H(\mathrm{curl},\Omega)$. In that case $\nabla_h\times v = \nabla\times v$ holds and the aforementioned identity yields

$$\sum_{F\in\mathcal{F}_h^i}\int_F (\mathbf{n}_F\times[\![v]\!]_F)\cdot\phi = 0, \quad \forall\phi\in C_0^\infty(\Omega)^3.$$

Now by choosing ϕ to have the support that intersects one face only, we prove the claim. □

4.2.3 Admissible mesh sequences

The aim of this subsection is to state some important, technical tools necessary to analyze the convergence of dG method as the meshsize tends to zero. We consider a mesh sequence

$$\mathcal{T}_{\mathcal{H}} := (\mathcal{T}_h)_{h\in\mathcal{H}}$$

where $\mathcal{H}$ is a countable subset of $\mathbb{R}_+$ having 0 as the only accumulation point.

In order to derive these technical tools, the mesh sequence $\mathcal{T}_{\mathcal{H}}$ has to be constructed in an appropriate way. First we present the concept of shape and contact regular mesh sequence, which provides us the inverse and trace inequalities. To have, in addition, the optimal polynomial approximation properties, we need to pose an additional requirement on the mesh sequence, which leads to the concept of admissible mesh sequence.

Shape and contact regularity

Definition 4.18. *$\mathcal{T}_h$ is a **matching simplicial mesh** if it is a simplicial mesh and if for any $K \in \mathcal{T}_h$ with vertices $\{a_0, \dots, a_d\}$, the set $\partial K \cap \partial K'$ for any $K' \in \mathcal{T}_h$, $K' \neq K$, is the convex hull of a (possibly empty) subset of $\{a_0, \dots, a_d\}$.*

In $\mathbb{R}^3$ the set $\partial K \cap \partial K'$ for two distinct elements K and K' of a matching simplicial mesh is either empty, or a common vertex, or a common edge, or a common face of the two elements.

Definition 4.19. *$\mathcal{T}_h'$ is a **matching simplicial submesh** of a general mesh $\mathcal{T}_h$ if*

(i) $\mathcal{T}_h'$ is a matching simplicial mesh,

(ii) for all $K' \in \mathcal{T}_h'$, there is only one $K \in \mathcal{T}_h$ such that $K' \subset K$,

(iii) for all $F' \in \mathcal{F}_h'$, the set collecting mesh faces of $\mathcal{T}_h'$, there is at most one $F \in \mathcal{F}_h$ such that $F' \subset F$.

*The simplices in $\mathcal{T}_h'$ are called **subelements**, and the mesh faces in $\mathcal{F}_h'$ **subfaces**. We set, for all $T \in \mathcal{T}_h$*

$$\mathcal{T}_K' := \left\{K' \in \mathcal{T}_h' : \ K' \subset K\right\} \qquad \mathcal{F}_K' := \left\{F' \in \mathcal{F}_h' : \ F' \subset \partial K\right\}.$$

Definition 4.20. *A mesh sequence $\mathcal{T}_{\mathcal{H}}$ is **shape- and contact-regular** if for all $h \in \mathcal{H}$, $\mathcal{T}_h$ admits a matching simplicial submesh $\mathcal{T}_h'$ such that*

(i) The mesh sequence $\mathcal{T}_{\mathcal{H}}'$ is shape-regular in the usual sense, i. e., there is $\rho_1 > 0$, independent of h, such that, for all $K' \in \mathcal{T}_h'$

$$\rho_1 h_{K'} \leq r_{K'}.$$

(ii) There is a parameter $\rho_2 > 0$, independent of h, such that, for all $K \in \mathcal{T}_h$ and for all $K' \in \mathcal{T}_K'$

$$\rho_2 h_K \leq h_{K'}.$$

The parameters $\rho = (\rho_1, \rho_2)$ are called the ***mesh regularity parameters****.*

It is not difficult to show that the diameters of the neighboring elements in a shape- and contact-regular mesh sequence are comparable in the following sense (which is contact-regularity in the usual sense), see [55, Lemma 1.43].

Lemma 4.21 (Diameter comparison for neighboring elements)**.** *Let $\mathcal{T}_h$ be a shape- and contact-regular mesh sequence with parameters $\rho = (\rho_1, \rho_2)$. Then for all $h \in \mathcal{H}$ and all $K_1, K_2 \in \mathcal{T}_h$ sharing a face F, there holds*

$$h_{K_1} \geq \rho_1 \rho_2 h_{K_2}. \tag{4.34}$$

For some of our results we also need a quasi-uniformity of a mesh sequence, cf. [22, Definition 1.140].

Definition 4.22. *A mesh sequence $\mathcal{T}_\mathcal{H}$ is* ***quasi-uniform*** *if it is shape-regular and there is c such that*

$$\forall h,\ \forall K \in \mathcal{T}_h, \quad h_K \geq ch. \tag{4.35}$$

Inverse and trace inequalities

Inverse and trace inequalities are very important tools in the analysis of dG methods. They hold on the broken polynomial space $V_h = \mathbb{P}_d^k(\mathcal{T}_h)$. This is a finite dimensional function space, and it is well known that all norms on a finite dimensional vector space are equivalent. The inverse inequality provides us the equivalence constant between broken H^1-norm and L^2-norm on V_h, which is of course h-dependent.

Lemma 4.23 (Inverse inequality, cf. Lemma 1.44 of [55])**.** *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence with parameters ρ. Then, for all $h \in \mathcal{H}$, all $v_h \in \mathbb{P}_d^k(\mathcal{T}_h)$, and all $K \in \mathcal{T}_h$*

$$\|\nabla v_h\|_{L^2(K)^d} \leq C_{\mathrm{inv}}\, h_K^{-1}\, \|v_h\|_{L^2(K)}\,, \tag{4.36}$$

where C_{inv} depends only on ρ, d and k.

The discrete trace inequality enables us to control the L^2-norm of traces with the L^2-norm on the elements.

Lemma 4.24 (Discrete trace inequality, cf. Lemma 1.45 of [55])**.** *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence with parameters ρ. Then, for all $h \in \mathcal{H}$, all $v_h \in \mathbb{P}_d^k(\mathcal{T}_h)$, all $K \in \mathcal{T}_h$ and all $F \in \mathcal{F}_K$*

$$\|v_h\|_{L^2(F)} \leq C_{\mathrm{tr}}\, h_K^{-1/2}\, \|v_h\|_{L^2(K)}\,, \tag{4.37}$$

where C_{tr} depends only on ρ, d and k.

Remark. Dependence of constants on the polynomial degree k: C_{inv} scales as k^2 [61], while C_{tr} scales as $\sqrt{k(k+d)}$ on $\mathbb{P}_d^k(\mathcal{T}_h)$ as proven in [64].

We will also need the following form of the continuous trace inequality.

Lemma 4.25 (Continuous trace inequality, cf. Lemma 1.49 of [55]). *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence with parameters $\rho = (\rho_1, \rho_2)$. Then, for all $h \in \mathcal{H}$, all $v \in H^1(\mathcal{T}_h)$, all $K \in \mathcal{T}_h$, and all $F \in \mathcal{F}_K$,*

$$\|v\|^2_{L^2(F)} \le C_{\mathrm{cti}}\Big(2\,\|\nabla v\|_{L^2(K)^d} + dh_K^{-1}\,\|v\|_{L^2(K)}\Big)\,\|v\|_{L^2(K)}\,, \tag{4.38}$$

with $C_{\mathrm{cti}} := \rho_1^{-1}$ if $\mathcal{T}_h$ is matching and simplicial and $C_{\mathrm{cti}} := (1+d)(\rho_1\rho_2)^{-1}$ otherwise.

Optimal polynomial properties

The meshes should be constructed in a way that the following property holds.

Definition 4.26. *The mesh sequence $\mathcal{T}_\mathcal{H}$ has **optimal polynomial approximation properties** if, for all $h \in \mathcal{H}$, all $K \in \mathcal{T}_h$, and all polynomials of degree k, there is a linear interpolation operator $\mathcal{I}_K^k : L^2(K) \to \mathbb{P}_d^k(K)$ such that, for all $s \in \{0, \dots, k+1\}$ and all $v \in H^s(K)$ there holds*

$$\left|v - \mathcal{I}_K^k\right|_{H^m(K)} \le C_{\mathrm{app}} h_K^{s-m}\,|v|_{H^s(K)}\,, \qquad \forall m \in \{0, \dots, s\}\,, \tag{4.39}$$

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

Definition 4.27. *The mesh sequence $\mathcal{T}_\mathcal{H}$ is **admissible** if it is shape- and contact-regular and if it has optimal polynomial approximation properties.*

We give two sufficient conditions on the mesh sequence $\mathcal{T}_\mathcal{H}$ to be admissible.

Definition 4.28. *Polyhedron P is **star-shaped with respect to ball** if there is a ball $B_P \subset P$ such that, for all $\mathbf{x} \in P$, the convex hull of $\mathbf{x} \cup B_P$ is included in $\overline{P}$.*

Lemma 4.29 (cf. Lemma 1.61. of [55]). *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence. Assume that, for all $h \in \mathcal{H}$ and all $K \in \mathcal{T}_h$, the mesh element K is star-shaped with respect to a ball with uniformly comparable diameter with respect to h_K. Then, the mesh sequence $\mathcal{T}_\mathcal{H}$ is admissible.*

Proof. Using averaged Taylor polynomials, see [7, Chapter 4]. □

Definition 4.30. *The mesh sequence $\mathcal{T}_\mathcal{H}$ is **finitely shaped** if there is a finite set $\hat{R} = \{\hat{K}\}$ whose elements are reference polyhedra in $\mathbb{R}^d$ such that, for all $h \in \mathcal{H}$ and all $K \in \mathcal{T}_h$ there is a reference polyhedron $\hat{K} \in \hat{R}$ and an affine bijective map F_K such that $K = F_K(\hat{K})$.*

Lemma 4.31 (cf. Lemma 1.62 of [55]). *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence. If $\mathcal{T}_\mathcal{H}$ is a finitely shaped mesh sequence, then it is admissible.*

Mesh sequence whose elements are simplices satisfy both conditions and therefore are admissible. Throughout this Thesis, the following assumption is made.

Assumption 4.32. *The mesh sequence $\mathcal{T}_\mathcal{H}$ is admissible.*

In the analysis of the method we usually work with the L^2-projection, instead of the interpolation operator. Therefore, the following two results, that hold on admissible mesh sequences, are very important.

Lemma 4.33 (Optimality of L^2-orthogonal projection, cf. Lemma 1.59 of [55]). *Let $\mathcal{T}_\mathcal{H}$ be an admissible mesh sequence. Then, for all $s \in \{0, \dots, k+1\}$ and all $v \in H^s(K)$, there holds*

$$|v - \pi_h v|_{H^m(K)} \leq C'_{\text{app}} h_K^{s-m} |v|_{H^s(K)} \qquad \forall m \in \{0, \dots, s\}, \tag{4.40}$$

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

Lemma 4.34 (Polynomial approximation on mesh faces, cf. Lemma 1.58 of [55]). *Let $\mathcal{T}_\mathcal{H}$ be an admissible mesh sequence, $s \in \{1, \dots, k+1\}$ and $v \in H^s(K)$. Then for all $h \in \mathcal{H}$, all $K \in \mathcal{T}_h$, and all $F \in \mathcal{F}_K$, there holds*

$$\|v - \pi_h v\|_{L^2(F)} \leq C''_{\text{app}} h_K^{s-1/2} |v|_{H^s(K)}, \tag{4.41}$$

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

Proof. This is direct consequence of (4.40) and the continuous trace inequality (4.38). □

4.3 Discontinuous Galerkin method applied to Maxwell's equations

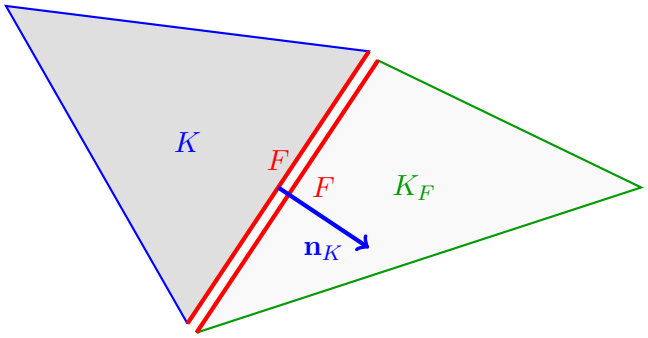
In this section we apply the discontinuous Galerkin method to the Maxwell's equations (3.20) and show that the numerical solution which it yields converges to the exact solution of the problem. As we have seen at the beginning of the chapter, the optimal convergence rate of the dG method applied to a first order hyperbolic equation is $\mathcal{O}(h^{k+1/2})$ when polynomials of degree k are used. The upwind flux dG method for Maxwell's equation is considered in [31] and the error estimate of order $\mathcal{O}(h^k)$) is proved. In [63], the same method is analyzed for Maxwell's equation in dispersive media and the order $\mathcal{O}(h^{k+1/2})$ is shown. For Maxwell's equations with smooth coefficients written as a second order PDE system and interior penalty dG method, order $\mathcal{O}(h^{k+1})$ was proved [25, 26].

We proceed in a similar way as in the one dimensional problem in Section 4.1 and first derive the simpler method by using the central flux. We prove that that method converges with the suboptimal order. The upwind flux, obtained from the theory of Riemann solvers, is a better choice and the optimal convergence order is shown.

Let the domain Ω and the mesh sequence $\mathcal{T}_\mathcal{H}$ satisfy Assumptions 4.4 and 4.32. The discontinuous Galerkin space is defined as a piecewise polynomial space, as discussed in the previous section, in each of 6 components

$$V_h = \left\{ v_h \in L^2(\Omega) \mid v_h|_K \in \mathbb{P}_3^k(K) \right\}^6 = \mathbb{P}_3^k(\mathcal{T}_h)^6. \tag{4.42}$$

Figure 4.3: Basic dG notation



In general, $V_h \not\subset D(A_M)$, so that the method is nonconforming. For the L^2-orthogonal projection $\pi_h : V \to V_h$ we have, by definition,

$$(v_h, u - \pi_h u)_{0,\Omega} = 0 \qquad \text{for all} \qquad u \in V,\ v_h \in V_h. \tag{4.43}$$

As in the previous section, we denote elements by K and faces by F. K_F is the neighboring element to K with respect to the face F. With $v_K := v_h|_K$ we denote the restriction of the discrete function v_h on the element K. In addition to jumps defined with respect to faces (see Definition 4.11), we also need jumps defined with respect to elements.

Definition 4.35. *For $K \in \mathcal{T}_h$, $F \in \mathcal{F}_K$ and a. e. $\mathbf{x} \in F$ we define*

$$[\![v_h]\!]^K(\mathbf{x}) = \begin{cases} [\![v_h]\!]_F(\mathbf{x}), & \text{if} \quad \mathbf{n}_K = \mathbf{n}_F, \\ -[\![v_h]\!]_F(\mathbf{x}), & \text{if} \quad \mathbf{n}_K = -\mathbf{n}_F, \end{cases}$$

or, equivalently, $[\![v_h]\!]^K(\mathbf{x}) = v_{K_F}(\mathbf{x}) - v_K(\mathbf{x})$.

We rewrite Maxwell's equations (3.20) as

$$\begin{aligned} \mu \partial_t \mathbf{H} + \nabla \times \mathbf{E} &= \tilde{f}_{\mathbf{H}}, \\ \epsilon \partial_t \mathbf{E} - \nabla \times \mathbf{H} &= \tilde{f}_{\mathbf{E}}, \end{aligned} \tag{4.44}$$

where $\tilde{f} = \begin{pmatrix} \tilde{f}_{\mathbf{H}} \\ \tilde{f}_{\mathbf{E}} \end{pmatrix} = \begin{pmatrix} \mu f_{\mathbf{H}} \\ \epsilon f_{\mathbf{E}} \end{pmatrix}$.

Assumption 4.36. *We suppose that $\mu_K := \mu|_K$ and $\epsilon_K := \epsilon|_K$ are constant for each $K \in \mathcal{T}_h$.*

Note that for piecewise constant coefficients, we also have

$$(v_h, u - \pi_h u)_V = 0 \qquad \text{for all} \qquad u \in V,\ v_h \in V_h. \tag{4.45}$$

The integration by parts formula (3.18) for the $\nabla\times$ operator will be used frequently. Therefore, we write it in a more familiar form: $\mathbf{E} \in H^1(K)^3$ satisfies

$$(\nabla \times \mathbf{E}, \phi)_{0,K} = (\mathbf{E}, \nabla \times \phi)_{0,K} + \sum_{F \in \mathcal{F}_K} (n_F \times \mathbf{E}, \phi)_{0,F}.$$

To derive the method we proceed similar to Section 4.1: we multiply equations (4.44) by test functions $\phi_h, \psi_h \in \mathbb{P}_3^k(\mathcal{T}_h)^3$, respectively, and integrate over element K. After integrating by parts we get

$$\begin{aligned}
\int_K \mu \partial_t \mathbf{H} \cdot \phi_h + \int_K \mathbf{E} \cdot (\nabla \times \phi_h) + \int_{\partial K} (\mathbf{n}_K \times \mathbf{E}) \cdot \phi_K &= \int_K \tilde{f}_{\mathbf{H}} \cdot \phi_h, \\
\int_K \epsilon \partial_t \mathbf{E} \cdot \psi_h - \int_K \mathbf{H} \cdot (\nabla \times \psi_h) - \int_{\partial K} (\mathbf{n}_K \times \mathbf{H}) \cdot \psi_K &= \int_K \tilde{f}_{\mathbf{E}} \cdot \psi_h.
\end{aligned}$$

Since both $\mathbf{E}$ and $\mathbf{H}$ are in $H(\mathrm{curl}; \Omega)$, according to (4.33), the terms $\mathbf{n}_K \times \mathbf{E}$ and $\mathbf{n}_K \times \mathbf{H}$ are well-defined on the boundary of the elements. We want to replace $\mathbf{H}$ and $\mathbf{E}$ by functions $\mathbf{H}_h$ and $\mathbf{E}_h$ from the discrete space V_h, but since they are not continuous in tangential directions on the boundary of elements, boundary integrals would not be well defined. Therefore, we replace $\mathbf{n}_K \times \mathbf{H}$ and $\mathbf{n}_K \times \mathbf{E}$ on the boundary by **numerical fluxes** $(\mathbf{n}_K \times \mathbf{H}_h)^*$ and $(\mathbf{n}_K \times \mathbf{E}_h)^*$ whose choice will determine the numerical scheme:

$$\begin{aligned}
\int_K \mu \partial_t \mathbf{H}_h \cdot \phi_h + \int_K \mathbf{E}_h \cdot (\nabla \times \phi_h) + \int_{\partial K} (\mathbf{n}_K \times \mathbf{E}_h)^* \cdot \phi_K &= \int_K \tilde{f}_{\mathbf{H}} \cdot \phi_h, \\
\int_K \epsilon \partial_t \mathbf{E}_h \cdot \psi_h - \int_K \mathbf{H}_h \cdot (\nabla \times \psi_h) - \int_{\partial K} (\mathbf{n}_K \times \mathbf{H}_h)^* \cdot \psi_K &= \int_K \tilde{f}_{\mathbf{E}} \cdot \psi_h.
\end{aligned}$$

This is called the **local weak form** of a general dG scheme for Maxwell's equations (the spatial derivatives are applied to the test functions, cf. the weak form (4.9)). Integrating by parts once more and taking function values from the element K we derive the **local strong form** (spatial derivatives are applied to the unknown functions):

$$\begin{aligned}
\int_K \mu \partial_t \mathbf{H}_h \cdot \phi_h + \int_K (\nabla \times \mathbf{E}_h) \cdot \phi_h + \int_{\partial K} ((\mathbf{n}_K \times \mathbf{E}_h)^* - \mathbf{n}_K \times \mathbf{E}_K) \cdot \phi_K &= \int_K \tilde{f}_{\mathbf{H}} \cdot \phi_h, \\
\int_K \epsilon \partial_t \mathbf{E}_h \cdot \psi_h - \int_K (\nabla \times \mathbf{H}_h) \cdot \psi_h - \int_{\partial K} ((\mathbf{n}_K \times \mathbf{H}_h)^* - \mathbf{n}_K \times \mathbf{H}_K) \cdot \psi_K &= \int_K \tilde{f}_{\mathbf{E}} \cdot \psi_h.
\end{aligned}$$

4.3.1 Central flux

Let $F = \partial K \cap \partial K_F$. The simplest choice for the numerical fluxes is the central flux, i.e.

$$(\mathbf{n}_K \times \mathbf{E}_h)^*|_F = \mathbf{n}_K \times \frac{\mathbf{E}_K + \mathbf{E}_{K_F}}{2}, \qquad (\mathbf{n}_K \times \mathbf{H}_h)^*|_F = \mathbf{n}_K \times \frac{\mathbf{H}_K + \mathbf{H}_{K_F}}{2}.$$

Inserting this into the strong form above we obtain

$$\begin{aligned}
\int_K \mu \partial_t \mathbf{H}_h \cdot \phi_h + \int_K (\nabla \times \mathbf{E}_h) \cdot \phi_h + \int_{\partial K} \frac{1}{2} \phi_K \cdot (\mathbf{n}_K \times [\![\mathbf{E}_h]\!]^K) &= \int_K \tilde{f}_{\mathbf{H}} \cdot \phi_h, \\
\int_K \epsilon \partial_t \mathbf{E}_h \cdot \psi_h - \int_K (\nabla \times \mathbf{H}_h) \cdot \psi_h - \int_{\partial K} \frac{1}{2} \psi_K \cdot (\mathbf{n}_K \times [\![\mathbf{H}_h]\!]^K) &= \int_K \tilde{f}_{\mathbf{E}} \cdot \psi_h.
\end{aligned}$$

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

$$\frac{1}{2}\int_F (\mathbf{n}_F \times [\![\mathbf{E}_h]\!]_F) \cdot (\phi_K + \phi_{K_F})$$

and analogously for the second field. If $F = \partial K \cap \partial\Omega$ is a boundary face, we model the boundary conditions in the following way:

$$\begin{aligned} &(\mathbf{n}_F \times \mathbf{E}_h)^*|_F = 0 \text{ since we have } \mathbf{n}_F \times \mathbf{E} = 0 \text{ for the exact solution} \\ &(\mathbf{n}_F \times \mathbf{H}_h)^*|_F = (\mathbf{n}_F \times \mathbf{H}_K)|_F \text{ since we have no b.c. for } \mathbf{H}. \end{aligned} \tag{4.46}$$

Therefore, we get the following **global formulation**

$$\begin{aligned} &\int_\Omega \mu\partial_t \mathbf{H}_h \cdot \phi_h + \int_\Omega (\nabla_h \times \mathbf{E}_h) \cdot \phi_h + \sum_{F\in\mathcal{F}_h^i} \int_F (\mathbf{n}_F \times [\![\mathbf{E}_h]\!]_F) \cdot \{\{\phi_h\}\}_F \\ &\qquad + \sum_{F\in\mathcal{F}_h^b} \int_F (-\mathbf{n}_F \times \mathbf{E}_h) \cdot \phi_h = \int_\Omega \tilde{f}_{\mathbf{H}} \cdot \phi_h, \\ &\int_\Omega \epsilon\partial_t \mathbf{E}_h \cdot \psi_h - \int_\Omega (\nabla_h \times \mathbf{H}_h) \cdot \psi_h - \sum_{F\in\mathcal{F}_h^i} \int_F (\mathbf{n}_F \times [\![\mathbf{H}_h]\!]_F) \cdot \{\{\psi_h\}\}_F = \int_\Omega \tilde{f}_{\mathbf{E}} \cdot \psi_h. \end{aligned} \tag{4.47}$$

We define the discontinuous Galerkin operator $A_h^{\mathrm{cf}} : V_h \to V_h$ by

$$\begin{aligned} (A_h^{\mathrm{cf}} u_h, w_h)_V := &(\nabla_h \times \mathbf{E}_h, \phi_h)_{0,\Omega} - (\nabla_h \times \mathbf{H}_h, \psi_h)_{0,\Omega} + \\ &+ \sum_{F\in\mathcal{F}_h^i} \Big((\mathbf{n}_F \times [\![\mathbf{E}_h]\!]_F, \{\{\phi_h\}\}_F)_{0,F} - (\mathbf{n}_F \times [\![\mathbf{H}_h]\!]_F, \{\{\psi_h\}\}_F)_{0,F} \Big) \\ &+ \sum_{F\in\mathcal{F}_h^b} (-\mathbf{n}_F \times \mathbf{E}_h, \phi_h)_{0,F} \end{aligned} \tag{4.48}$$

for all $u_h = \begin{pmatrix} \mathbf{H}_h \\ \mathbf{E}_h \end{pmatrix}, w_h = \begin{pmatrix} \phi_h \\ \psi_h \end{pmatrix} \in V_h$. The discrete problem (4.47) can now be written in a compact form. We seek for a function $u_h = \begin{pmatrix} \mathbf{H}_h \\ \mathbf{E}_h \end{pmatrix} \in C^1([0,T], V_h)$ which satisfies

$$(\partial_t u_h, w_h)_V + (A_h^{\mathrm{cf}} u_h, w_h)_V = (f, w_h)_V \quad \text{for all } w_h \in V_h. \tag{4.49}$$

Next we discuss the properties of the operator A_h^{cf}.

Properties of the operator A_h^{cf}

The central flux dG method applied to the Maxwell's equations yields the following semidiscrete problem

$$\partial_t u_h(t) + A_h^{\mathrm{cf}} u_h(t) = f_h(t), \quad u_h(0) = \pi_h u_0, \tag{4.50}$$

where $f_h(t) := \pi_h f(t)$. The function u_h approximates the exact solution u of Maxwell's equations (3.20). In order to prove that u_h converges to u as the meshsize h tends to zero, we first prove some auxiliary results for the discrete operator.

We first note that A_h^{cf} can be extended to a larger space $V_h + D(A_M)$. For $u \in D(A_M)$, we define $A_h^{\mathrm{cf}} u$ by the very same formula (4.48). Then, the following consistency property holds.

Lemma 4.37. *For $u \in D(A_M)$ we have*

$$A_h^{\mathrm{cf}} u = \pi_h A_M u. \tag{4.51}$$

Proof. For $u = (\mathbf{H}, \mathbf{E}) \in D(A_M)$ we have $\mathbf{n}_F \times [\![\mathbf{E}]\!]_F = 0$ and $\mathbf{n}_F \times [\![\mathbf{H}]\!]_F = 0$ for $F \in \mathcal{F}_h^i$ and $\mathbf{n} \times \mathbf{E} = 0$ on $\partial\Omega$, so the sum over faces in (4.48) is zero. Therefore, for all $w_h = (\phi_h, \psi_h) \in V_h$ we have

$$\begin{aligned}(A_h^{\mathrm{cf}} u, w_h)_V &= (\nabla_h \times \mathbf{E}, \phi_h)_{0,\Omega} - (\nabla_h \times \mathbf{H}, \psi_h)_{0,K} \\ &= (\nabla \times \mathbf{E}, \phi_h)_{0,\Omega} - (\nabla \times \mathbf{H}, \psi_h)_{0,\Omega} = (A_M u, w_h)_V,\end{aligned}$$

where we have used $\nabla_h \times v = \nabla \times v$ for $v \in H(\mathrm{curl}, \Omega)$. This is equivalent with (4.51). □

The next lemma is the discrete analog to Corollary 3.2.

Lemma 4.38. *For all $u_h = (\mathbf{H}_h, \mathbf{E}_h) \in V_h$ we have $(A_h^{\mathrm{cf}} u_h, u_h)_V = 0$.*

Proof. By (4.48) we have

$$\begin{aligned}(A_h^{\mathrm{cf}} u_h, u_h)_V = &\sum_K \Big((\nabla \times \mathbf{E}_h, \mathbf{H}_h)_{0,K} - (\nabla \times \mathbf{H}_h, \mathbf{E}_h)_{0,K} \Big) \\ &+ \sum_{F \in \mathcal{F}_h^i} \Big((\mathbf{n}_F \times [\![\mathbf{E}_h]\!]_F, \{\{\mathbf{H}_h\}\}_F)_{0,F} - (\mathbf{n}_F \times [\![\mathbf{H}_h]\!]_F, \{\{\mathbf{E}_h\}\}_F)_{0,F} \Big) \\ &- \sum_{F \in \mathcal{F}_h^b} (\mathbf{n}_F \times \mathbf{E}_h, \mathbf{H}_h)_{0,F}.\end{aligned} \tag{4.52}$$

Integration by parts in the first term gives

$$\begin{aligned}&\sum_K \Big((\nabla \times \mathbf{E}_h, \mathbf{H}_h)_{0,K} - (\nabla \times \mathbf{H}_h, \mathbf{E}_h)_{0,K} \Big) = \\ &\sum_{F \in \mathcal{F}_h^i} \Big((\mathbf{n}_F \times \mathbf{E}_K, \mathbf{H}_K)_{0,F} + (\mathbf{n}_{K_F} \times \mathbf{E}_{K_F}, \mathbf{H}_{K_F})_{0,F} \Big) + \sum_{F \in \mathcal{F}_h^b} (\mathbf{n}_F \times \mathbf{E}_h, \mathbf{H}_h)_{0,F}.\end{aligned}$$

For the sum over all internal faces we have

$$\begin{aligned}
&\sum_{F\in\mathcal{F}_h^i}\Big(+(\mathbf{n}_F\times[\![\mathbf{E}_h]\!]_F,\{\{\mathbf{H}_h\}\}_F)_{0,F}-(\mathbf{n}_F\times[\![\mathbf{H}_h]\!]_F,\{\{\mathbf{E}_h\}\}_F)_{0,F}\Big)\\
&\quad=\frac{1}{2}\sum_{F\in\mathcal{F}_h^i}\Big((\mathbf{n}_F\times\mathbf{E}_{K_F},\mathbf{H}_K+\mathbf{H}_{K_F})_{0,F}-(\mathbf{n}_F\times\mathbf{E}_K,\mathbf{H}_K+\mathbf{H}_{K_F})_{0,F}\\
&\qquad\qquad-(\mathbf{n}_F\times\mathbf{H}_{K_F},\mathbf{E}_K+\mathbf{E}_{K_F})_{0,F}+(\mathbf{n}_F\times\mathbf{H}_K,\mathbf{E}_K+\mathbf{E}_{K_F})_{0,F}\Big)\\
&\quad=\sum_{F\in\mathcal{F}_h^i}\Big((\mathbf{n}_F\times\mathbf{E}_{K_F},\mathbf{H}_{K_F})_{0,F}-(\mathbf{n}_F\times\mathbf{E}_K,\mathbf{H}_K)_{0,F}\Big).
\end{aligned}$$

Here we have used $a\times b=-b\times a$ and $(a\times b)\cdot c=(c\times a)\cdot b$ for vectors $a,b,c\in\mathbb{R}^3$. By inserting these considerations back into (4.52), the claim is proved. □

This lemma implies that the electromagnetic energy of the solution of the discrete problem (4.50) is conserved in the homogeneous case, i.e. for $f=0$ we have

$$\partial_t\,\|u_h\|_V^2=0.$$

We have seen that the same property holds in the continuous case too.

In the convergence proof of the method, we also need the following integration by parts formula for the discrete operator A_h^{cf}. It allows to move all derivatives and tangential jumps to the test functions.

Lemma 4.39. *For $u=(\mathbf{H},\mathbf{E})\in V_h+\big(D(A_M)\cap H^1(\mathcal{T}_h)^6\big)$ and $w_h=(\phi_h,\psi_h)\in V_h$ the following equation holds:*

$$\begin{aligned}
(A_h^{\mathrm{cf}}u,w_h)_V&=\sum_K\Big((\mathbf{E},\nabla\times\phi_h)_{0,K}-(\mathbf{H},\nabla\times\psi_h)_{0,K}\Big)\\
&\quad+\sum_{F\in\mathcal{F}_h^i}\Big((\{\{\mathbf{E}\}\}_F,\mathbf{n}_F\times[\![\phi_h]\!])_{0,F}-(\{\{\mathbf{H}\}\}_F,\mathbf{n}_F\times[\![\psi_h]\!])_{0,F}\Big)\\
&\quad+\sum_{F\in\mathcal{F}_h^b}(\mathbf{H},\mathbf{n}_F\times\psi_h)_{0,F}.
\end{aligned}\tag{4.53}$$

Proof. We start from (4.48) and integrate by parts in the first sum

$$\begin{aligned}(A_h^{\mathrm{cf}}u, w_h) = &\sum_K \Big((\mathbf{E}, \nabla\times\phi_h)_{0,K} - (\mathbf{H}, \nabla\times\psi_h)_{0,K}\Big)\\ &+ \sum_{F\in\mathcal{F}_h^i}\Big((\mathbf{n}_F\times\mathbf{E}_K, \phi_K)_{0,F} + (\mathbf{n}_{K_F}\times\mathbf{E}_{K_F}, \phi_{K_F})_{0,F}\\ &\qquad - (\mathbf{n}_F\times\mathbf{H}_K, \psi_K)_{0,F} - (\mathbf{n}_{K_F}\times\mathbf{H}_{K_F}, \psi_{K_F})_{0,F}\Big)\\ &+ \sum_{F\in\mathcal{F}_h^b}\Big((\mathbf{n}_F\times\mathbf{E}, \phi_h)_{0,F} - (\mathbf{n}_F\times\mathbf{H}, \psi_h)_{0,F}\Big)\\ &+ \sum_{F\in\mathcal{F}_h^i}\Big((\mathbf{n}_F\times[\![\mathbf{E}]\!]_F, \{\{\phi_h\}\}_F)_{0,F} - \sum_{F\in\mathcal{F}_h^i}(\mathbf{n}_F\times[\![\mathbf{H}]\!]_F, \{\{\psi_h\}\}_F)_{0,F}\Big)\\ &+ \sum_{F\in\mathcal{F}_h^b}(-\mathbf{n}_F\times\mathbf{E}), \phi_h)_{0,F}.\end{aligned}$$

The claim now follows by a straightforward computation. □

Remark. It is easy to see that if we use the weak form in construction of the method, the operator A_h^{cf} will be defined by (4.53) for $u = u_h \in V_h$.

Convergence

Applying the L^2-projection π_h to the continuous problem (3.20) and using the consistency property from Lemma 4.37, shows that the exact solution satisfies

$$\partial_t\pi_h u + A_h^{\mathrm{cf}}u = f_h, \quad \pi_h u(0) = \pi_h u_0. \tag{4.54}$$

We define errors

$$e(t) = u_h(t) - u(t) = e_h(t) - e_\pi(t),$$

where

$$e_h(t) := u_h(t) - \pi_h u(t), \qquad e_\pi(t) := u(t) - \pi_h u(t).$$

Subtracting (4.54) from (4.50) yields the equation for the error

$$\partial_t e_h + A_h^{\mathrm{cf}}e_h = A_h^{\mathrm{cf}}e_\pi, \quad \pi_h u(0) = \pi_h u_0. \tag{4.56}$$

The bounds for the projection error are known from the previous section, see Lemmas 4.33 and 4.34. We summarize these results in the form we need.

Lemma 4.40. *Let $u \in H^{k'+1}(K)^6$ for some $k' \le k$. The following error bounds hold:*

$$\|e_\pi\|_{0,K} \le Ch^{k'+1}\,|u|_{H^{k'+1}(K)^6} \tag{4.57}$$

and

$$\|e_{\pi,K}\|_{0,F} \le Ch^{k'+1/2}\,|u|_{H^{k'+1}(K)^6}\,, \tag{4.58}$$

where C is independent of both h and K.

The following lemma will be of a great importance.

Lemma 4.41. *Let $u \in H^{k'+1}(\mathcal{T}_h)^6$ for some $k' \leq k$. Then for all $w_h = \begin{pmatrix} \phi_h \\ \psi_h \end{pmatrix} \in V_h$ and for all $\gamma > 0$ holds*

$$(A_h^{\mathrm{cf}} e_\pi, w_h)_V \leq C\Big(\gamma\, h^{2k'}\, |u|^2_{H^{k'+1}(\mathcal{T}_h)^6} + \frac{1}{\gamma}\, \|w_h\|_V^2\Big), \tag{4.59}$$

where C is independent of h.

Proof. We use the representation of A_h^{cf} from Lemma 4.39. Since $e_\pi = \begin{pmatrix} e_\pi^{\mathbf{H}} \\ e_\pi^{\mathbf{E}} \end{pmatrix}$ is the projection error and $\nabla_h \times \phi_h, \nabla_h \times \psi_h \in V_h$ we have

$$(e_\pi^{\mathbf{E}}, \nabla \times \phi_h)_{0,K} = 0, \quad (e_\pi^{\mathbf{H}}, \nabla \times \psi_h)_{0,K} = 0,$$

so the first sum in (4.53) vanishes. We have

$$\begin{aligned}(A_h^{\mathrm{cf}} e_\pi, w_h)_V = &\sum_{F \in \mathcal{F}_h^i} \Big((\{\{e_\pi^{\mathbf{E}}\}\}_F, \mathbf{n}_F \times [\![\phi_h]\!]_F)_{0,F} + (\{\{e_\pi^{\mathbf{H}}\}\}_F, \mathbf{n}_F \times [\![\psi_h]\!]_F)_{0,F} \Big) \\ &+ \sum_{F \in \mathcal{F}_h^b} (e_\pi^{\mathbf{H}}, \mathbf{n}_F \times \psi_h)_{0,F}.\end{aligned}$$

Each term can now be bounded by using the Cauchy-Schwarz inequality (2.3), the triangle inequality, Lemma 4.40 and the discrete trace inequality (4.37) as follows

$$\begin{aligned}(\{\{e_\pi^{\mathbf{E}}\}\}, \mathbf{n}_F \times [\![\phi_h]\!])_{0,F} &= \frac{1}{2} \left\| e_{\pi,K}^{\mathbf{E}} + e_{\pi,K_F}^{\mathbf{E}} \right\|_{0,F} \|\mathbf{n}_F \times (\phi_{K_F} - \phi_K)\|_{0,F} \\ &\leq \frac{1}{2} \Big(\left\| e_{\pi,K}^{\mathbf{E}} \right\|_{0,F} + \left\| e_{\pi,K_F}^{\mathbf{E}} \right\|_{0,F} \Big) \Big(\|\phi_{K_F}\|_{0,F} + \|\phi_K\|_{0,F} \Big) \\ &\leq C h_K^{k'+1/2} \Big(|\mathbf{E}|_{H^{k'+1}(K)^3} + |\mathbf{E}|_{H^{k'+1}(K_F)^3} \Big) \\ &\qquad \cdot \Big(h_K^{-1/2} \|\phi_K\|_{0,K} + h_{K_F}^{-1/2} \|\phi_{K_F}\|_{0,K_F} \Big)\end{aligned}$$

By using (4.34), $h = \max_K h_K$, and applying Young's inequality to each of the products we get

$$(\{\{e_\pi^{\mathbf{E}}\}\}, \mathbf{n}_F \times [\![\phi_h]\!])_{0,F} \quad \leq C\gamma h^{2k'} (|\mathbf{E}|^2_{H^{k'+1}(K)^3} + |\mathbf{E}|^2_{H^{k'+1}(K_F)^3}) + C\frac{1}{\gamma} \|\phi_K\|^2_{0,K \cup K_F}.$$

After bounding the other terms analogously and summing over all faces, the lemma is proved. □

The error bound is given in the following theorem.

Theorem 4.42. *Let u be the solution of the continuous problem (3.20) and u_h be the solution of the semidiscrete problem (4.50). Assume that $u \in L^2((0,T), H^{k'+1}(\mathcal{T}_h)^6)$ for some $k' \leq k$. Then, the error $e_h = u_h - u$ satisfies*

$$\|e_h(T)\|_V^2 \leq Ch^{2k'} T \int_0^T |u(t)|^2_{H^{k'+1}(\mathcal{T}_h)^6} \, dt,$$

where C is independent of u, h and T.

Proof. We start from the error equation (4.56). Taking the V-inner product with e_h and integrating from 0 to T we obtain

$$\frac{1}{2} \int_0^T \frac{d}{dt} \|e_h(t)\|_V^2 \, dt = \int_0^T (A_h^{\mathrm{cf}} e_\pi(t), e_h(t))_V dt,$$

since the second term equals zero due to Lemma 4.38. For the right-hand side we use Lemma 4.41. Since $e_h(0) = 0$ this yields

$$\|e_h(T)\|_V^2 \leq C \int_0^T \frac{1}{\gamma} \|e_h(t)\|_V^2 \, dt + C\gamma h^{2k'} \int_0^T |u(t)|^2_{H^{k'+1}(\mathcal{T}_h)^6} \, dt.$$

Applying the continuous Gronwall Lemma 2.6 with $\gamma = T$, proves the claim. □

Corollary 4.43. *If the assumptions of Theorem (4.42) are satisfied, then the semidiscrete error $e = e_h - e_\pi$ is bounded by*

$$\|e(T)\|_V^2 \leq Ch^{2k'} T \int_0^T |u(t)|^2_{H^{k'+1}(\mathcal{T}_h)^6} \, dt + Ch^{2k'+2} \, |u(T)|^2_{H^{k'+1}(\mathcal{T}_h)^6}, \tag{4.60}$$

where C is independent of u, h and T.

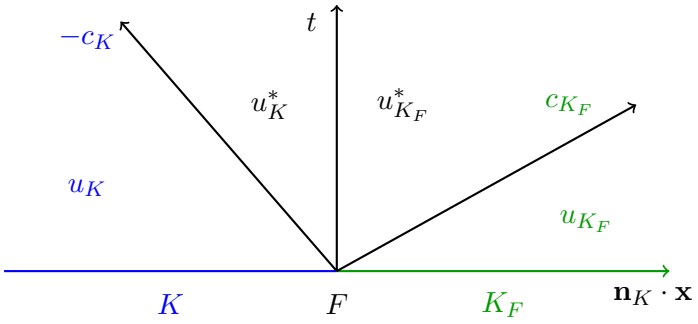
4.3.2 Upwind flux

Construction of upwind fluxes using theory of Riemann solvers

As in the one dimensional example from Section 4.1, upwind fluxes are constructed by solving the Riemann problem. Nevertheless, the situation is more complicated here. For an arbitrary face $F = \partial K \cap \partial K_F$, we consider Maxwell's equations in the conservative form (4.3) with the initial data

$$u_0(x) = \begin{cases} u_K, & x \in K, \\ u_{K_F}, & x \in K_F. \end{cases}$$

Figure 4.4: Construction of upwind fluxes



The solution is known to be piecewise constant [45]. Let us define matrix $A_{\mathbf{n}} := n_x A_x + n_y A_y + n_z A_z$, where n_i, $i = x, y, z$, are the components of the normal vector $\mathbf{n}_K$. The jump occurs for $\lambda t - \mathbf{n}_K \cdot \mathbf{x} = 0$, where λ is an eigenvalue of a generalized eigenvalue problem $A_{\mathbf{n}} w = \lambda M w$, and it has directions of a corresponding eigenvector w, see Figure 4.4.

In our case three double eigenvalues are $-c$, 0 and c, where $c = (\epsilon\mu)^{-1/2}$ denotes the speed of light. We use the values of coefficients from element K and K_F for positive and negative eigenvalues, respectively. Finally, let u_K^* and $u_{K_F}^*$ be the values of the solution as shown in Figure 4.4. Since $A_{\mathbf{n}} u_K^* = A_{\mathbf{n}} u_{K_F}^*$ by construction and since for $u = (\mathbf{H}, \mathbf{E})$ we have $A_{\mathbf{n}} u = (\mathbf{n} \times \mathbf{E}, -\mathbf{n} \times \mathbf{H})$, numerical fluxes are chosen as

$$\begin{pmatrix} (\mathbf{n}_K \times \mathbf{E})^* \\ -(\mathbf{n}_K \times \mathbf{H})^* \end{pmatrix} = A_{\mathbf{n}} u_K^*.$$

The detailed construction of upwind fluxes and the discrete operator can be found in [39]. For $u_h = (\mathbf{H}_h, \mathbf{E}_h)$, $w_h = (\phi_h, \psi_h) \in V_h$ the discontinuous Galerkin operator $A_h^{\text{upw}} : V_h \to V_h$ is given as

$$\begin{aligned}
(A_h^{\text{upw}} u_h, w_h)_V := & \sum_K \Big((\nabla \times \mathbf{E}_h, \phi_h)_{0,K} - (\nabla \times \mathbf{H}_h, \psi_h)_{0,K} \Big) \\
& + \sum_{F \in \mathcal{F}_h^i} \Big((\mathbf{n}_F \times [\![\mathbf{E}_h]\!]_F, \alpha_K \phi_K + \alpha_{K_F} \phi_{K_F})_{0,F} \\
& \qquad - (\mathbf{n}_F \times [\![\mathbf{H}_h]\!]_F, \beta_K \psi_K + \beta_{K_F} \psi_{K_F})_{0,F} \\
& \qquad + \gamma_F (\mathbf{n}_F \times [\![\mathbf{E}_h]\!]_F, \mathbf{n}_F \times [\![\psi_h]\!]_F)_{0,F} \\
& \qquad + \delta_F (\mathbf{n}_F \times [\![\mathbf{H}_h]\!]_F, \mathbf{n}_F \times [\![\phi_h]\!]_F)_{0,F} \Big) \\
& + \sum_{F \in \mathcal{F}_h^b} \Big(- (\mathbf{n}_F \times \mathbf{E}_h, \phi_h)_{0,F} + 2\gamma_F (\mathbf{n}_F \times \mathbf{E}_h, \mathbf{n}_F \times \psi_h)_{0,F} \Big),
\end{aligned} \tag{4.61}$$

with the coefficients

$$\begin{aligned}\alpha_{K,F} &= \frac{c_{K_F}\epsilon_{K_F}}{c_{K_F}\epsilon_{K_F}+c_K\epsilon_K} = \frac{1}{1+\left(\frac{\epsilon_K\mu_{K_F}}{\mu_K\epsilon_{K_F}}\right)^{1/2}},\\ \beta_{K,F} &= \frac{c_{K_F}\mu_{K_F}}{c_{K_F}\mu_{K_F}+c_K\mu_K} = \frac{1}{1+\left(\frac{\mu_K\epsilon_{K_F}}{\epsilon_K\mu_{K_F}}\right)^{1/2}},\\ \gamma_F &= \frac{1}{c_{K_F}\mu_{K_F}+c_K\mu_K}, \qquad \delta_F = \frac{1}{c_{K_F}\epsilon_{K_F}+c_K\epsilon_K}.\end{aligned} \tag{4.62}$$

Note that

$$\alpha_{K,F}+\alpha_{K_F,F}=1, \qquad \beta_{K,F}+\beta_{K_F,F}=1, \qquad \alpha_{K,F}=\beta_{K_F,F}. \tag{4.63}$$

The obtained space semidiscrete problem is

$$\partial_t u_h + A_h^{\mathrm{upw}} u_h = f_h, \quad u_h(0)=\pi_h u_0, \tag{4.64}$$

where $u_h \in C^1(0,T;V_h)$ is the semidiscrete solution and $f_h=\pi_h f$.

Properties of the discrete operator A_h^{upw}

Note that by (4.61), A_h^{upw} is also well defined as an operator from $V_h + D(A_M)$ to V_h. For $u \in D(A_M)$, the following consistency property is satisfied.

Lemma 4.44. *For $u \in D(A_M)$ we have*

$$A_h^{\mathrm{upw}} u = \pi_h A_M u. \tag{4.65}$$

Proof. Same as for Lemma 4.37. □

Lemma 4.45. *For all $u_h=(\mathbf{H}_h,\mathbf{E}_h)\in V_h$ we have*

$$\begin{aligned}(A_h^{\mathrm{upw}}u_h,u_h)_V = &\sum_{F\in\mathcal{F}_h^i}\left(\gamma_F\left\|\mathbf{n}_F\times[\![\mathbf{E}_h]\!]_F\right\|_{0,F}^2+\delta_F\left\|\mathbf{n}_F\times[\![\mathbf{H}_h]\!]_F\right\|_{0,F}^2\right)\\ &+\sum_{F\in\mathcal{F}_h^b}2\gamma_F\left\|\mathbf{n}_F\times\mathbf{E}_h\right\|_{0,F}^2\geq 0.\end{aligned} \tag{4.66}$$

In particular, $-A_h^{\mathrm{upw}}$ is dissipative on V_h.

Proof. Analogously to the proof of Lemma 4.38, we integrate by parts in the first term of $(A_h^{\mathrm{upw}}u_h,u_h)_V$. By using $\alpha_K+\beta_K=1$, we obtain

$$\begin{aligned}(A_h^{\mathrm{upw}}u_h,u_h)_V = \sum_{F\in\mathcal{F}_h^i}\Big(&\alpha_K(\mathbf{n}_F\times\mathbf{E}_{K_F},\mathbf{H}_K)_{0,F}-\alpha_{K_F}(\mathbf{n}_F\times\mathbf{E}_K,\mathbf{H}_{K_F})_{0,F}\\ &-\beta_K(\mathbf{n}_F\times\mathbf{H}_{K_F},\mathbf{E}_K)_{0,F}+\beta_{K_F}(\mathbf{n}_F\times\mathbf{H}_K,\mathbf{E}_{K_F})_{0,F}\\ &+\gamma_F\left\|\mathbf{n}_F\times[\![\mathbf{E}_h]\!]_F\right\|_{0,F}^2+\delta_F\left\|\mathbf{n}_F\times[\![\mathbf{H}_h]\!]_F\right\|_{0,F}^2\Big)\\ &+\sum_{F\in\mathcal{F}_h^b}2\gamma_F\left\|\mathbf{n}_F\times\mathbf{E}_h\right\|_{0,F}^2.\end{aligned}$$

By (4.63), the first and the fourth term sum to zero, and so do the second and the third term. This shows (4.66). □

The previous lemma implies that the electromagnetic energy is non-increasing if $f = 0$:

$$\partial_t \|u_h\|_V^2 \leq 0. \tag{4.67}$$

The following integration by parts formula will be used frequently later. It allows to move all derivatives and tangential jumps to the test functions.

Lemma 4.46. *For $u = (\mathbf{H}, \mathbf{E}) \in V_h + \big(D(A_M) \cap H^1(\mathcal{T}_h)\big)$ and $w_h = (\phi_h, \psi_h) \in V_h$ the following relation holds:*

$$\begin{aligned}(A_h^{\mathrm{upw}} u, w_h)_V = &\sum_K \Big((\mathbf{E}, \nabla \times \phi_h)_{0,K} - (\mathbf{H}, \nabla \times \psi_h)_{0,K} \Big) \\ &+ \sum_{F \in \mathcal{F}_h^i} \Big((\alpha_K \mathbf{E}_{K_F} + \alpha_{K_F} \mathbf{E}_K + \delta_F \mathbf{n}_F \times [\![\mathbf{H}]\!]_F, \mathbf{n}_F \times [\![\phi_h]\!]_F)_{0,F} \\ &+ (-\beta_K \mathbf{H}_{K_F} - \beta_{K_F} \mathbf{H}_K + \gamma_F \mathbf{n}_F \times [\![\mathbf{E}]\!]_F, \mathbf{n}_F \times [\![\psi_h]\!]_F)_{0,F} \Big) \\ &+ \sum_{F \in \mathcal{F}_h^b} (\mathbf{H}, \mathbf{n}_F \times \psi_h)_{0,F} + 2\gamma_F (\mathbf{n}_F \times \mathbf{E}, \mathbf{n}_F \times \psi_h)_{0,F}.\end{aligned} \tag{4.68}$$

Proof. We start from (4.61) and integrate by parts in the first sum

$$\begin{aligned}(A_h^{\mathrm{upw}} u, w_h)_V := &\sum_K \Big((\mathbf{E}, \nabla \times \phi_h)_{0,K} - (\mathbf{H}, \nabla \times \psi_h)_{0,K} \Big) \\ &+ \sum_{F \in \mathcal{F}_h^i} \Big((\mathbf{n}_F \times \mathbf{E}_K, \phi_K)_{0,F} + (\mathbf{n}_{K_F} \times \mathbf{E}_{K_F}, \phi_{K_F})_{0,F} - (\mathbf{n}_F \times \mathbf{H}_K, \psi_K)_{0,F} \\ &\quad - (\mathbf{n}_{K_F} \times \mathbf{H}_{K_F}, \psi_{K_F})_{0,F} \Big) + \sum_{F \in \mathcal{F}_h^b} \Big((\mathbf{n}_F \times \mathbf{E}, \phi_h)_{0,F} - (\mathbf{n}_F \times \mathbf{H}, \psi_h)_{0,F} \Big) \\ &+ \sum_{F \in \mathcal{F}_h^i} \Big((\mathbf{n}_F \times [\![\mathbf{E}]\!]_F, \alpha_K \phi_K + \alpha_{K_F} \phi_{K_F})_{0,F} - (\mathbf{n}_F \times [\![\mathbf{H}]\!]_F, \beta_K \psi_K + \beta_{K_F} \psi_{K_F})_{0,F} \\ &\quad + \gamma_F (\mathbf{n}_F \times [\![\mathbf{E}]\!]_F, \mathbf{n}_F \times [\![\psi_h]\!]_F)_{0,F} + \delta_F (\mathbf{n}_F \times [\![\mathbf{H}]\!]_F, \mathbf{n}_F \times [\![\phi_h]\!]_F)_{0,F} \Big) \\ &+ \sum_{F \in \mathcal{F}_h^b} \Big(- (\mathbf{n}_F \times \mathbf{E}, \phi_h)_{0,F} + 2\gamma_F (\mathbf{n}_F \times \mathbf{E}, \mathbf{n}_F \times \psi_h)_{0,F} \Big).\end{aligned}$$

The sum over all elements and the sum over all exterior faces are ok. It remains to check what happens on interior faces. Using (4.63) gives

$$\begin{aligned}&(\mathbf{n}_F \times \mathbf{E}_K, \phi_K)_{0,F} + (\mathbf{n}_{K_F} \times \mathbf{E}_{K_F}, \phi_{K_F})_{0,F} + (\mathbf{n}_F \times [\![\mathbf{E}]\!]_F, \alpha_K \phi_K + \alpha_{K_F} \phi_{K_F})_{0,F} \\ &= \alpha_K (\mathbf{n}_F \times \mathbf{E}_K, \phi_K)_{0,F} + \alpha_{K_F} (\mathbf{n}_F \times \mathbf{E}_K, \phi_K)_{0,F} + \alpha_K (\mathbf{n}_{K_F} \times \mathbf{E}_{K_F}, \phi_{K_F})_{0,F} \\ &+ \alpha_{K_F} (\mathbf{n}_{K_F} \times \mathbf{E}_{K_F}, \phi_{K_F})_{0,F} + \alpha_K (\mathbf{n}_F \times \mathbf{E}_{K_F}, \phi_K)_{0,F} + \alpha_{K_F} (\mathbf{n}_F \times \mathbf{E}_{K_F}, \phi_{K_F})_{0,F} \\ &- \alpha_K (\mathbf{n}_F \times \mathbf{E}_K, \phi_K)_{0,F} - \alpha_{K_F} (\mathbf{n}_F \times \mathbf{E}_K, \phi_{K_F})_{0,F} \\ &= \alpha_K (\mathbf{E}_{K_F}, \mathbf{n}_F \times [\![\phi_h]\!]_F)_{0,F} + \alpha_{K_F} (\mathbf{E}_K, \mathbf{n}_F \times [\![\phi_h]\!]_F)_{0,F}.\end{aligned}$$

Analogously,

$$(\mathbf{n}_F \times \mathbf{H}_K, \psi_K)_{0,F} + (\mathbf{n}_{K_F} \times \mathbf{H}_{K_F}, \psi_{K_F})_{0,F} + (\mathbf{n}_F \times [\![\mathbf{H}]\!]_F, \beta_K \psi_K + \beta_{K_F} \psi_{K_F})_{0,F} =$$
$$= \beta_K (\mathbf{H}_{K_F}, \mathbf{n}_F \times [\![\psi_h]\!]_F)_{0,F} + \beta_{K_F} (\mathbf{H}_K, \mathbf{n}_F \times [\psi_h]_F)_{0,F},$$

which proves the claim. □

Convergence

By applying the L^2-projection π_h to the continuous problem (3.20) and using the consistency property from Lemma 4.44, the exact solution satisfies

$$\partial_t \pi_h u + A_h^{\mathrm{upw}} u = f_h, \quad \pi_h u(0) = \pi_h u_0. \tag{4.69}$$

The following lemma is useful for proving convergence.

Lemma 4.47. *Let $u \in H^{k'+1}(\mathcal{T}_h)^6$ for some $k' \leq k$. Then for all $w_h = \begin{pmatrix} \phi_h \\ \psi_h \end{pmatrix} \in V_h$ and for all $\gamma > 0$ the following estimate holds*

$$(A_h^{\mathrm{upw}} e_\pi, w_h)_V \leq \frac{1}{2\gamma} (A_h^{\mathrm{upw}} w_h, w_h)_V + C\gamma h^{2k'+1} \, |u|^2_{H^{k'+1}(\mathcal{T}_h)^6}, \tag{4.70}$$

where C is independent of h.

Proof. We use Lemma 4.46. Since e_π is the projection error we have

$$(e_\pi^{\mathbf{E}}, \nabla \times \phi_h)_{0,K} = 0, \quad (e_\pi^{\mathbf{H}}, \nabla \times \psi_h)_{0,K} = 0,$$

so the first sum in (4.68) equals zero. We have

$$\begin{aligned}(A_h^{\mathrm{upw}} e_\pi, e_h)_V = \sum_{F \in \mathcal{F}_h^i} \Big(& (\alpha_K e^{\mathbf{E}}_{\pi,K_F} + \alpha_{K_F} e^{\mathbf{E}}_{\pi,K} + \delta_F \mathbf{n}_F \times [\![e_\pi^{\mathbf{H}}]\!]_F, \mathbf{n}_F \times [\![\phi_h]\!]_F)_{0,F} \\ & + (-\beta_K e^{\mathbf{H}}_{\pi,K_F} - \beta_{K_F} e^{\mathbf{H}}_{\pi,K} + \gamma_F \mathbf{n}_F \times [\![e_\pi^{\mathbf{E}}]\!]_F, \mathbf{n}_F \times [\![\psi_h]\!]_F)_{0,F} \Big) \\ & + \sum_{F \in \mathcal{F}_h^b} (e_\pi^{\mathbf{H}}, \mathbf{n}_F \times \psi_h)_{0,F} + 2\gamma_F (\mathbf{n}_F \times e_\pi^{\mathbf{E}}, \mathbf{n}_F \times \psi_h)_{0,F}.\end{aligned}$$

We bound each of the terms as in Lemma 4.41. Applying Young's inequality with a carefully chosen parameter gives the result. □

The error e_h satisfies the following error bound:

Theorem 4.48. *Let u be the solution of the continuous problem (3.20) and u_h be the solution of the semidiscrete problem (4.64). Assume that $u \in L^2((0,T), H^{k'+1}(\mathcal{T}_h)^6)$ for some $k' \leq k$. Then, the error $e_h = u_h - u$ satisfies*

$$\|e_h(T)\|_V^2 + \int_0^T (A_h^{\mathrm{upw}} e_h(t), e_h(t))_V dt \leq C h^{2k'+1} \int_0^T |u(t)|^2_{H^{k'+1}(\mathcal{T}_h)^6} \, dt, \tag{4.71}$$

where C is independent of u, h and T.

Proof. By subtracting (4.69) from (4.64), we obtain

$$\partial_t e_h(t) + A_h^{\mathrm{upw}} e_h(t) = A_h^{\mathrm{upw}} e_\pi(t). \tag{4.72}$$

Taking the V-inner product with e_h and integrating from 0 to T we obtain

$$\frac{1}{2}\int_0^T \frac{d}{dt}\,\|e_h(t)\|_V^2\,dt + \int_0^T (A_h^{\mathrm{upw}} e_h(t), e_h(t))_V dt = \int_0^T (A_h^{\mathrm{upw}} e_\pi(t), e_h(t))_V dt.$$

Applying Lemma 4.47 with $\gamma = 1$ on the right-hand side and using $e_h(0) = 0$ gives (4.71). □

Corollary 4.49. *If the assumptions of Theorem 4.48 are satisfied, then the semidiscrete error $e = e_h - e_\pi$ is bounded by*

$$\|e(T)\|_V^2 + \int_0^T (A_h^{\mathrm{upw}} e(t), e(t))_V dt \le C h^{2k'+1} \left(\int_0^T |u(t)|^2_{H^{k'+1}(\mathcal{T}_h)^6}\,dt + h\,|u(T)|^2_{H^{k'+1}(\mathcal{T}_h)^6} \right)$$

where C is independent of u and h.

Proof. We have $(A_h^{\mathrm{upw}} v, e_\pi)_V = 0$ for all $v \in V_h + D(A_M)$ by (4.45). This yields the identity

$$(A_h^{\mathrm{upw}}(e_h - e_\pi), e_h - e_\pi)_V = (A_h^{\mathrm{upw}} e_h, e_h)_V - (A_h^{\mathrm{upw}} e_\pi, e_h)_V.$$

from which the estimate follows directly from Lemma 4.40 and Theorem 4.48. □

CHAPTER 5

Time integration

We have seen in Chapter 3 that Maxwell's equations can be written in the form of an abstract ordinary differential equation

$$\partial_t u(t) + A_M u(t) = f(t), \quad u(0) = u_0, \tag{5.1}$$

where A_M is the continuous operator with domain dense in $L^2(\Omega)^6$. In the last chapter we have discretized the problem in space and obtained the system of ordinary differential equations

$$\partial_t u_h(t) + A_h u_h(t) = \pi_h f(t), \quad u_h(0) = \pi_h u_0, \tag{5.2}$$

where $A_h \in \left\{A_h^{\mathrm{cf}}, A_h^{\mathrm{upw}}\right\}$. What we still have not considered until now is the question of solving (abstract) ordinary differential equations that appear here. This issue will be dealt with in this chapter where we will see how the abstract Cauchy problem (5.1) can be handled numerically and continued in the next chapter where fully discrete methods are presented. We consider only one-step methods.

This chapter consists of four sections. In the first one we give a short course on classical Runge–Kutta methods for solving ordinary differential equations, in which we mostly concentrate on a special class of implicit Runge–Kutta methods, namely a collocation methods. We introduce Gauss and Radau collocation methods and state some of their properties. In the second section, we consider some known results for Gauss and Radau collocation methods that are applicable to the case of Maxwell's equations. In Section 5.3 we provide the error analysis for collocation methods applied to (5.1) by using different approach, namely energy techniques. In the last section we consider exponential Runge–Kutta methods which simplify to the exponential quadrature rule in linear case. The convergence result is applicable to the case of continuous Maxwell's equations.

5.1 Runge–Kutta methods

In this section we give a short introduction on Runge–Kutta methods for solving a system of ordinary differential equation

$$\begin{aligned} u'(t) &= F(t, u(t)), \qquad t \in [0, T], \\ u(0) &= u_0, \end{aligned} \tag{5.3}$$

for $F : U \to \mathbb{R}^n$, where $U \subset \mathbb{R} \times \mathbb{R}^n$ is an open, simply connected set and $(0, u_0) \in U$. In our case the system is linear, i. e. $F(t, u(t)) = Au + f(t)$, where A is an operator in the space-continuous case (5.1) (and then strictly speaking does not fit into the setting from above) or a matrix in the space discrete case (5.2). Runge–Kutta methods are one-step methods, i. e they use the current approximation $u_n \approx u(t_n)$ only to construct a new approximation $u_{n+1} \approx u(t_n+\tau)$. Special attention is given to collocation methods, which are a special class of implicit Runge–Kutta methods. We mainly omit proofs here but refer to lecture notes for the course on innovative integrators [35] and for the numerical analysis course [34], as well as to the following well-known books on time integration methods [28], [29] and [27].

5.1.1 Construction, local error, stability

The construction of Runge–Kutta methods relies on the following representation of the exact solution u of (5.3)

$$u(t_n + \tau) = u(t_n) + \int_0^\tau F\big(t_n + \theta, u(t_n + \theta)\big) d\theta, \qquad t_{n+1} = t_n + \tau, \qquad n = 0, 1, \dots \tag{5.4}$$

The idea is to approximate the integral on the right-hand side by a quadrature formula defined by nodes $0 \le c_1, \dots, c_s \le 1$ and weights $b_1, \dots, b_s$. Assume we are given approximations $u_n \approx u(t_n)$ and $U_{ni} \approx u(t_n + c_i\tau)$. Then we have

$$u(t_{n+1}) \approx u_n + \tau \sum_{i=1}^{s} b_i U'_{ni}, \qquad \text{where} \qquad U'_{ni} = F\big(t_n + c_i\tau, U_{ni}\big), \qquad i = 1, \dots, s.$$

The approximations U_{ni}, $i = 1, \dots, s$, can be obtained by yet other quadrature formulas via

$$u(t_n + c_i\tau) = u(t_n) + \int_{t_n}^{t_n + c_i\tau} u'(\theta) d\theta \approx u_n + \tau \sum_{j=1}^{s} a_{ij} U'_{nj}.$$

A general s-stage Runge–Kutta method is defined as

$$\begin{aligned} u_{n+1} &= u_n + \tau \sum_{i=1}^{s} b_i U'_{ni}, \\ U'_{ni} &= F(t_n + c_i\tau, U_{ni}), \qquad i = 1, \dots, s, \\ U_{ni} &= u_n + \tau \sum_{j=1}^{s} a_{ij} U'_{nj}, \quad i = 1, \dots, s. \end{aligned} \tag{5.5}$$

It is convenient to represent it in a so-called Butcher tableau as

$$\begin{array}{c|c} c_i & a_{ij} \\ \hline & b_j \end{array}$$

By $\mathcal{A} = (a_{ij})_{i,j=1}^{s}$ we denote the Runge–Kutta matrix.

Definition 5.1. *The* **local error** *of a one-step method for solving the initial value problem* (5.3) *is defined as*

$$u_1 - u(t_0 + \tau)$$

where u_1 is the approximation obtained from $u_0 = u(t_0)$ after one step with step size τ.

An important property of methods for solving initial value problems is the order.

Definition 5.2. *A numerical scheme for solving the initial value problem* (5.3) *is of* **order** *p or* **accurate of order** *p if for any $F \in C^{p+1}$ the local error is of size $\mathcal{O}(\tau^{p+1})$. If $p \geq 1$, the method is called* **consistent**.

The stability of numerical schemes is studied via the test equation

$$u' = \lambda u, \qquad u(0) = u_0. \tag{5.6}$$

For a motivation of this test equation, see [34, Section 10.1]. The exact solution, $u(t) = \exp(t\lambda)u_0$, remains bounded for all $t \geq 0$ if $\operatorname{Re}\lambda \leq 0$. This motivates the following definition.

Definition 5.3. *The* **stability region** *of a numerical scheme is defined as*

$$\mathcal{S} = \{z \in \mathbb{C} \mid z = \tau\lambda; \quad \text{the scheme yields bounded solutions } \{u_n\}_{n\geq 0}, \text{ when it is applied to (5.6) with step size } \tau\}.$$

To study stability, we apply a Runge–Kutta method to the test equation (5.6):

$$U_{ni} = u_n + \tau\lambda \sum_{j=1}^{s} a_{ij} U_{nj}, \qquad i = 1, \ldots, s,$$

$$u_{n+1} = u_n + \tau\lambda \sum_{i=1}^{s} b_i U_{ni}.$$

With

$$\mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ \vdots \\ b_s \end{bmatrix}, \qquad U_n = \begin{bmatrix} U_{n1} \\ \vdots \\ U_{ns} \end{bmatrix},$$

we have the compact form

$$(I - z\mathcal{A})U_n = u_n \mathbb{1}.$$

If $(I - z\mathcal{A})$ is nonsingular, then $U_n = (I - z\mathcal{A})^{-1}\mathbb{1}u_n$, so that

$$u_{n+1} = u_n + zb^T U_n = [1 + zb^T(I - z\mathcal{A})^{-1}\mathbb{1}]u_n.$$

Definition 5.4. *$R(z) := [1 + zb^T(I - z\mathcal{A})^{-1}\mathbb{1}]$ is called the* **stability function** *of a Runge–Kutta method. It can be interpreted as the numerical solution after one step for the test equation $u' = \lambda u$ with initial data $u_0 = 1$ and $z = \lambda\tau$.*

From $u_{n+1} = R(z)u_n = \ldots = R(z)^{n+1}u_0$ it is clear that the numerical solution remains bounded if and only if $|R(z)| \leq 1$. Hence, we have

$$\mathcal{S} = \{z \in \mathbb{C} \mid (I - z\mathcal{A}) \text{ invertible and } |R(z)| \leq 1\}.$$

One can prove [29, Propostion 3.2] that

$$R(z) = \frac{\det(I - z\mathcal{A} + z\mathbb{1}b^T)}{\det(I - z\mathcal{A})}.$$

Therefore, the stability function is a rational function with numerator and denominator of degree $\leq s$, i. e.

$$R(z) = \frac{P(z)}{Q(z)}, \qquad \deg P,\ \deg Q \leq s.$$

5.1.2 Explicit Runge–Kutta methods

For explicit Runge–Kutta methods we have $a_{ij} = 0$ for $j \geq i$, i. e. the Runge-Kutta matrix $\mathcal{A}$ is strictly lower triangular. Then, the nonlinear system in (5.5) can be solved directly, i. e. the stages U_{ni}, $i = 1, \ldots, s$, can be computed explicitly just by evaluating the function F at already computed approximations. Their obvious advantage is that they are easy to implement and computationally cheap.

The main disadvantage is the stability issue. Indeed, since $\mathcal{A}$ is strictly lower triangular, we have $\det(I - z\mathcal{A}) = 1$. This implies

$$R(z) = P(z),$$

i. e., the stability function is a polynomial of degree at most s and the stability region is given by

$$\mathcal{S} = \{z \in \mathbb{C} \mid |P(z)| \leq 1\}.$$

Therefore, $\mathcal{S}$ is necessarily bounded since $|P(z)| \to \infty$ for $|z| \to \infty$. If the method is of order p then $\mathrm{e}^z - P(z) = \mathcal{O}(\tau^{p+1})$. In the special case of $s = p$ we have

$$P(z) = 1 + z + \ldots + \frac{z^p}{p!}.$$

5.1.3 Implicit Runge–Kutta methods

We have seen that the exact solution of the test equation (5.6) remains bounded for $\lambda \in \mathbb{C}^-$. We would like to have numerical methods with the same property, i. e., that the numerical solution remains bounded on the entire $\mathbb{C}^-$.

Definition 5.5. *A numerical method is A-stable, if*

$$\mathbb{C}^- := \{z \in \mathbb{C} \mid \operatorname{Re} z \leq 0\} \subseteq \mathcal{S}$$

and a numerical scheme is 0*-stable, if* $0 \in \mathcal{S}$*.*

We have seen in Subsection 5.1.2 that explicit methods can not be A-stable. If the Runge-Kutta matrix $\mathcal{A}$ is not strictly lower triangular, we are speaking of an implicit Runge–Kutta method. In this case $\det(I - z\mathcal{A})$ is a polynomial of degree at least 1 and the stability function $R(z)$ really is a rational function. For a method of order p, we have

$$\mathrm{e}^z - R(z) = Cz^{p+1} + \mathcal{O}(z^{p+1}), \qquad C \neq 0.$$

Therefore, we are interested in rational functions which are good approximations to the exponential function.

Definition 5.6. ***Padé approximations*** *of exponential are rational functions which, for a given degree of numerator and denominator, have the highest order of approximation in* $z = 0$*. By* $R_{kj}(z)$ *we denote the Padé approximation with the numerator degree* k *and the denominator degree* j *and call it the* (k, j)*-Padé approximation.*

It can be proven that $R_{kj}(z)$ is the unique rational approximation to e^z of order $j + k$, with degrees of numerator and denominator equal k and j respectively, see [29, Theorem 3.11]. Moreover, the error satisfies

$$\mathrm{e}^z - R_{kj}(z) = (-1)^j \frac{j!\,k!}{(j+k)!\,(j+k+1)!} z^{j+k+1} + \mathcal{O}(z^{j+k+2}).$$

We say that the rational function $R(z)$ is A-stable if $|R(z)| \leq 1$, i. e., if the underlying numerical method is A-stable.

Theorem 5.7 (cf. Theorem IV.4.12. of [29])**.** *The Padé approximation* R_{kj} *is A-stable if and only if* $k \leq j \leq k + 2$*.*

We can construct implicit, A-stable Runge–Kutta methods of arbitrarily high order by a collocation approach. Again, we start from (5.4) and choose nodes $0 \leq c_1 < \ldots < c_s \leq 1$. Then we approximate the solution u on the interval $[t_n, t_{n+1}]$ by a polynomial p_n satisfying the collocation conditions

$$\begin{aligned} p_n(t_n) &= u_n, \\ p_n'(t_n + c_i\tau) &= F(t_n + c_i\tau, p_n(t_n + c_i\tau)), \qquad i = 1, \ldots, s. \end{aligned} \tag{5.7a}$$

Thus the collocation polynomial satisfies the differential equation (5.3) in the collocation points (but in general not between them). The new approximation is defined as

$$u_{n+1} = p_n(t_n + \tau). \tag{5.7b}$$

One can prove (cf. Theorem II.7.7 of [28]) that a collocation method is equivalent to an implicit Runge–Kutta method with coefficients

$$a_{ij} = \int_0^{c_i} \ell_j(x)dx, \qquad b_j = \int_0^1 \ell_j(x)dx, \qquad i,j = 1,\dots,s,$$

where

$$\ell_j(x) = \frac{\prod_{k\neq j}(x-c_k)}{\prod_{k\neq j}(c_j-c_k)}$$

are the Lagrange interpolation polynomials. There is a connection between the order of a collocation method and the corresponding quadrature formula. Recall that a quadrature formula is of order p if it is exact for all polynomials of degree at most $p-1$, i. e.

$$\int_0^1 g(x)dx = \sum_{j=1}^{s} b_j g(c_j), \quad \text{for all } g \in \mathbb{P}^{p-1}.$$

There is an equivalent characterization which is easy to check: a quadrature formula is of order p if and only if

$$\sum_{i=1}^{s} b_i c_i^{j-1} = \frac{1}{j}, \qquad j = 1,\dots,p. \tag{5.8}$$

Theorem 5.8. *Collocation methods are of the same order p as the quadrature formula $(b_i, c_i)_{i=1}^s$.*

Gauss collocation methods

Gauss collocation methods are based on Gaussian quadrature formulas, i. e. $c_1,\dots,c_s$ are the zeros of the shifted Legendre polynomial of degree s (see Section IV.5 of [29])

$$\frac{d^s}{dx^s}\Big(x^s(x-1)^s\Big).$$

Since Gaussian quadrature formulas are known to be of order $2s$, Theorem 5.8 implies the same order for Gauss collocation methods. For any s-stage Runge–Kutta method, the degree of numerator and denominator of its stability function can not be larger than s. Therefore, the order of $2s$ implies that the stability function of the s-stage Gauss collocation method is the (s,s)-Padé approximation. By Theorem 5.7 Gauss collocation methods are A-stable.

For $s = 1$, the Gauß collocation method is the implicit midpoint rule

$$u_{n+1} = u_n + F\left(t_0 + \frac{\tau}{2}, \frac{u_n + u_{n+1}}{2}\right), \tag{5.9}$$

with the stability function

$$R(z) = \frac{1+\frac{z}{2}}{1-\frac{z}{2}}.$$

For all Gauß collocation methods we have that the stability function satisfies

$$\lim_{z\to\infty} |R(z)| = 1. \tag{5.10}$$

Radau collocation methods

We distinguish two types of methods based on Radau quadrature formulas. For the s-stage Radau IA method quadrature points $c_1, \dots c_s$ are the zeros of

$$\frac{d^{s-1}}{dx^{s-1}}\Big(x^s(x-1)^{s-1}\Big),$$

while for the s-stage Radau IIA method quadrature points are the zeros of

$$\frac{d^{s-1}}{dx^{s-1}}\Big(x^{s-1}(x-1)^{s}\Big).$$

For Radau IA methods, $c_1 = 0$ is a collocation point, while for Radau IIA methods $c_s = 1$ is a collocation point. The following result holds, cf. [29, Theorem IV.5.3].

Theorem 5.9. *The s-stage Radau IA and Radau IIA methods are of order $2s-1$. Their stability function is the $(s-1, s)$-Padé approximations and therefore they are A-stable.*

For $s = 1$, the Radau IA method is the implicit Euler scheme, with stability function

$$R(z) = \frac{1}{1-z},$$

which is $(0, 1)$-Padé approximation of exponential.

Note that because of the choice of $c_s = 1$ for Radau IIA methods we have

$$a_{sj} = b_j, \quad j = 1, \dots, s \quad \text{or} \quad b^T = e_s^T \mathcal{A},$$

which implies $\lim_{z\to\infty} R(z) = 0$

Algebraic stability and coercivity property

In our numerical analysis we consider algebraically stable collocation methods which also satisfy a coercivity condition. Here we briefly recall these two concepts and the corresponding results for Gauß and Radau methods.

Definition 5.10. *A Runge–Kutta method with s distinct nodes $0 \le c_i \le 1$ and weights $\mathcal{A} = (a_{ij})_{i,j=1}^s$ and $b = (b_i)_{i=1}^s$ is called **algebraically stable**, if $b_i \ge 0$ for $i = 1, \dots, s$ and*

$$\mathcal{M} = (m_{ij})_{i,j=1}^s, \qquad \text{with} \qquad m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j \tag{5.11}$$

is positive semidefinite.

It can be shown that algebraic stability implies A-stability, see [29, Section IV.12]. The following Lemma shows that Gauß and Radau quadrature formulas have positive weights b_i, $i = 1, \dots s$ and therefore satisfy the first condition needed for algebraic stability.

Lemma 5.11. *The weights of an s-stage quadrature formula of order $p \ge 2s-1$ are positive: $b_j > 0$ for $j = 1, \dots, s$.*

Using the so-called W-transformation [29, Section IV.5], the second condition can bee proved too. The result is given in [29, Theorem IV.12.9].

Theorem 5.12. *Gauß and Radau IA and Radau IIA collocation methods are algebraically stable.*

The following theorem can be found in [29, Corollary IV.13.10].

Theorem 5.13. *An s-stage algebraically stable collocation method is of order $p \geq 2s-1$.*

For the definition of the coercivity condition we need the existence of the inverse of $\mathcal{A}$.

Theorem 5.14. *For Gauß and Radau collocation methods, the Runge–Kutta matrix $\mathcal{A}$ is invertible.*

Definition 5.15. *We say that Runge–Kutta method satisfies the* ***coercivity condition*** *if there exists a diagonal, positive definite matrix $\mathcal{D}$ and a positive scalar α such that*

$$u^T \mathcal{D} \mathcal{A}^{-1} u \geq \alpha u^T \mathcal{D} u \qquad \text{for all} \qquad u \in \mathbb{R}^s. \tag{5.12}$$

This condition also plays an important role in proving the existence of Runge–Kutta approximations, cf. [29, Section IV.14]. For Gauß collocation methods, (5.12) is satisfied for $\mathcal{D} = \mathcal{B}(\mathcal{C}^{-1} - I)$, where $\mathcal{B} := \operatorname{diag}(b_1, \ldots, b_s)$ and $\mathcal{C} := \operatorname{diag}(c_1, \ldots, c_s)$. For Radau IA collocation methods it is satisfied for $\mathcal{D} = \mathcal{B}(I - \mathcal{C})$. For Radau IIA collocation methods it is satisfied for $\mathcal{D} = \mathcal{B}\mathcal{C}^{-1}$.

Defects

Inserting the exact solution of (5.3) in the numerical scheme (5.5) yields

$$\begin{aligned} u(t_n + c_i\tau) &= u(t_n) + \tau \sum_{j=1}^{s} a_{ij} u'(t_n + c_j\tau) + \Delta^{ni}, \quad i = 1, \ldots, s, \\ u(t_{n+1}) &= u(t_n) + \tau \sum_{i=1}^{s} b_i u'(t_n + c_i\tau) + \delta^{n+1}, \end{aligned} \tag{5.13}$$

where Δ^{ni} and δ^{n+1} are the defects of the inner and outer stages, respectively. The goal of this section is to derive formulas for the defects.

Definition 5.16. *The* ***stage order*** *q is the largest number such that*

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad \text{for } k = 1, \ldots, q, \text{ and all } i.$$

Lemma 5.17. *The stage order of an s-stage collocation method equals s.*

Theorem 5.18. *For the defects of an s-stage collocation method of order p we have*

$$\Delta^{ni} = \tau^s \int_{t_n}^{t_{n+1}} u^{(s+1)}(t)\kappa_i\left(\frac{t-t_n}{\tau}\right) dt \tag{5.14}$$

with

$$\kappa_i(\theta) = \frac{1}{s!}(c_i-\theta)_+^s - \frac{1}{(s-1)!}\sum_{j=1}^{s} a_{ij}(c_j-\theta)_+^{s-1}$$

and

$$\delta^{n+1} = \tau^p \int_{t_n}^{t_{n+1}} u^{(p+1)}(t)\kappa\left(\frac{t-t_n}{\tau}\right) dt \tag{5.15}$$

with

$$\kappa(\theta) = \frac{1}{p!}(1-\theta)_+^p - \frac{1}{(p-1)!}\sum_{i=1}^{s} b_i(c_i-\theta)_+^{p-1}.$$

Remark. Functions κ and κ_i, for $i = 1,\dots,s$ are known as **Peano kernels** corresponding to the quadrature rules and they are uniformly bounded on $[t_n, t_{n+1}]$ with constants depending on Runge–Kutta coefficients only.

Proof. From (5.13) we have

$$\Delta^{ni} = \int_{t_n}^{t_n+c_i\tau} u'(t)dt - \tau\sum_{j=1}^{s} a_{ij}u'(t_n+c_j\tau), \quad i=1,\dots,s,$$

$$\delta^{n+1} = \int_{t_n}^{t_{n+1}} u'(t)dt - \tau\sum_{i=1}^{s} b_i u'(t_n+c_i\tau).$$

We derive the expression for Δ^{ni}, for δ^{n+1} the procedure is analogous. Taylor's theorem gives us

$$u'(t) = \sum_{k=0}^{s-1}\frac{u^{(k+1)}(t_n)}{k!}(t-t_n)^k + \int_{t_n}^{t} u^{(s+1)}(\theta)\frac{(t-\theta)^{s-1}}{(s-1)!}d\theta. \tag{5.16}$$

The stage order s implies that the quadrature formulas (a_{ij}, c_j), for $i = 1,\dots,s$ are of order s and therefore all polynomials of order $\le s-1$ are integrated exactly. The sum in the above's expansion is treated exactly and we have

$$\begin{aligned}
\Delta^{ni} &= \\
&= \int_{t_n}^{t_n+c_i\tau}\left(\int_{t_n}^{t} u^{(s+1)}(\theta)\frac{(t-\theta)^{s-1}}{(s-1)!}d\theta\right)dt - \tau\sum_{j=1}^{s} a_{ij}\int_{t_n}^{t_n+c_j\tau} u^{(s+1)}(\theta)\frac{(t_n+c_j\tau-\theta)^{s-1}}{(s-1)!}d\theta \\
&= \int_{t_n}^{t_n+c_i\tau}\left(\int_{t_n}^{t_{n+1}} u^{(s+1)}(\theta)\frac{(t-\theta)_+^{s-1}}{(s-1)!}d\theta\right)dt - \tau\sum_{j=1}^{s} a_{ij}\int_{t_n}^{t_{n+1}} u^{(s+1)}(\theta)\frac{(t_n+c_j\tau-\theta)_+^{s-1}}{(s-1)!}d\theta.
\end{aligned}$$

By using the Fubini's theorem in first term we get

$$\Delta^{ni} = \int_{t_n}^{t_{n+1}} \frac{u^{(s+1)}(\theta)}{(s-1)!} \left(\int_{t_n}^{t_n+c_i\tau} (t-\theta)_+^{s-1} dt - \tau \sum_{j=1}^{s} a_{ij} (t_n + c_j\tau - \theta)_+^{s-1} \right) d\theta.$$

Further on,

$$\int_{t_n}^{t_n+c_i\tau} (t-\theta)_+^{s-1}\, dt = \frac{\tau^s}{s} \left(c_i - \frac{\theta - t_n}{\tau} \right)_+^s$$

implies

$$\Delta^{ni} = \int_{t_n}^{t_{n+1}} \frac{u^{(s+1)}(\theta)}{(s-1)!} \left(\frac{\tau^s}{s} \left(c_i - \frac{\theta - t_n}{\tau} \right)_+^s - \tau^s \sum_{j=1}^{s} a_{ij} \left(c_j - \frac{\theta - t_n}{\tau} \right)_+^{s-1} \right) d\theta.$$

Writing t instead of θ gives (5.14).

□

Remark. For Gauss methods (5.15) is satisfied for all $p \leq 2s$ and for Radau methods for all $p \leq 2s-1$ providing the solution is regular enough.

5.2 Implicit Runge–Kutta methods: known results for Maxwell's equations

Implicit time integration for linear, abstract initial value problems

$$\partial_t u(t) + Au(t) = f(t), \quad u(0) = u_0 \tag{5.17}$$

where the operator A is a generator of a bounded C_0 semigroup, has been studied in [4] generalizing earlier work in [5, 6] for the homogeneous case $f \equiv 0$. The analysis is based on a Hille-Phillips operational calculus which uses Laplace transformations. In this section we explain some basic ideas of this technique, state the main results proven in this work and see if the application to the Maxwell's case is possible.

Since there is a substantial difference between the homogeneous and the inhomogeneous case, we present them separately. In the homogeneous case, the methods do not suffer from the order reduction phenomena. Provided that the initial data is smooth enough, the classical order is obtained. In the inhomogeneous case, the classical order can be achieved only under some unnatural assumptions on the solution, and the expected convergence order is $s+1$ for s-stage collocation methods, in particular for Gauß and Radau methods.

Throughout this subsection $(X, \|.\|_X)$ denotes a Banach space.

Homogeneous case

Here, we need the following condition on the operator A.

Assumption 5.19. *$-A$ generates a strongly continuous semigroup $S(t) := e^{-tA}$ on X of type (C_A, ω).*

The solution of the homogeneous initial value problem

$$\partial_t u(t) + Au(t) = 0, \quad u(0) = u_0 \tag{5.18}$$

is given by $u(t) = S(t)u_0$. Let r be an A-stable rational function which approximates e^z up to order p. An approximate solution of (5.18) is given by

$$u_n = r(-\tau A)^n u_0, \quad n = 0, 1, \dots \tag{5.19}$$

The following stability result holds, cf. [5, Theorem 1].

Theorem 5.20. *Let r be an A-stable rational approximation of the exponential of order p and let the operator A satisfy Assumption 5.19. Then there exist constants $C_1 \geq 0$ and $\omega_1 \geq 0$ such that*

$$\|r(-\tau A)^N\|_{X \leftarrow X} \leq C_A C_1 N^{1/2} \mathrm{e}^{\omega \omega_1 T},$$

where $T = N\tau$.

The convergence result is given in [5, Theorem 3].

Theorem 5.21. *Let r be an A-stable rational approximation of the exponential of order p and let the operator A satisfy Assumption 5.19. Then there exist constants $C_1 \geq 0$ and $\omega_1 \geq 0$ such that*

$$\left\| r^N(-\tau A)u_0 - S(T)u_0 \right\|_X \leq C_A C_1 T \tau^p e^{\omega \omega_1 T} \left\| A^{p+1} u_0 \right\|_X, \quad u_0 \in D(A^{p+1}) \tag{5.20}$$

where $T = N\tau$.

Remark. Maxwell's equations fit into the setting since the Maxwell's operator A_M defines the unitary C_0 semigroup. Therefore, for Gauss collocation method with s stages applied to linear Maxwell's equations, the order of convergence is $2s$ if the initial data is in $D(A_M^{2s+1})$. For Radau collocation method with s stages applied to linear Maxwell's equations, we get that order of convergence of $2s - 1$ if the initial data is in $D(A_M^{2s})$.

As it has been said, the proofs of these results rely on the Hille-Phillips calculus, which we present in the inhomogeneous case.

Inhomogeneous case

In contrast to the homogeneous case, here we need a stronger assumption on the operator A.

Assumption 5.22. *$-A$ generates a bounded semigroup $S(t) := e^{-tA}$ on X, i. e. $\exists\, C_A > 0$ such that*

$$\|S(t)\|_{X \leftarrow X} \leq C_A \quad \forall t > 0. \tag{5.21}$$

According to Theorem 2.5, the exact solution of (5.17) is given by

$$u(t) = S(t)u_0 + \int_0^t S(t-s)f(s)ds. \tag{5.22}$$

Discretizing the problem (5.17) in time by using implicit Runge–Kutta methods with s stages gives the numerical scheme

$$u_{n+1} = r(-\tau A)u_n + \tau \sum_{j=1}^{s} q_j(-\tau A) f(t_n + c_j\tau) \tag{5.23}$$

where $\{c_j\}$ are distinct quadrature points on $[0,1]$ and r, $q_1, \ldots, q_s$ are rational functions which satisfy the following assumption.

Assumption 5.23. *r, $q_1, \ldots, q_s$ are **bounded** for $\operatorname{Re} z \leq 0$.*

Since the spectrum of the operator $-A$ is contained in the left half-plane, $r(-\tau A)$ and $q_j(-\tau A)$ are well defined.
Remark. r is the stability function of the numerical method. For Runge–Kutta method with coefficients $(\mathcal{A}, b)$ we have

$$\begin{aligned} r(z) &= 1 + zb^T(I - z\mathcal{A})^{-1}\mathbb{1}, \\ q_j(z) &= b^T(I - z\mathcal{A})^{-1}e_j, \end{aligned} \tag{5.24}$$

where e_j is the j-th coordinate vector. Gauss and Radau collocation methods satisfy Assumption 5.23.

By following [4], we now present the analysis for the numerical method (5.23). To simplify the notation, we write

$$S_\tau := r(-\tau A) \qquad \text{and} \qquad (Q_\tau f)(t) := \sum_{j=1}^{s} q_j(-\tau A) f(t + c_j\tau).$$

Assumption 5.24. *S_τ is stable, i.e. $\|S_\tau^n\|_{X \leftarrow X} \leq C$ for all $n \in \mathbb{N}$.*

The local error ρ_n according to Definition 5.1 is the error after one time-step starting from the exact solution. Hence it can be written as

$$\rho_n := u(t_{n+1}) - S_\tau u(t_n) - \tau(Q_\tau f)(t_n). \tag{5.25}$$

For the global error $e_n = u_n - u(t_n)$ we then have

$$e_{n+1} = S_\tau e_n - \rho_n, \quad n = 0, 1, \ldots, \quad \text{with} \quad e_0 = 0. \tag{5.26}$$

Solving this recursion gives $e_N = -\sum_{j=0}^{N-1} S_\tau^{N-1-j} \rho_j$. Taking the norm and using Assumption 5.24 yields

$$\|e_N\|_X \leq C \sum_{j=0}^{N-1} \|\rho_j\|_X . \tag{5.27}$$

We assume that the scheme (5.23) is accurate of order p. This means that if the method is applied to an ordinary differential equation with sufficiently regular f, the local error will satisfy

$$\rho_n = \mathcal{O}(\tau^{p+1}) \quad \text{as } \tau \to 0. \tag{5.28}$$

The case of an ordinary differential equation is in this context equivalent to the case of a **bounded** operator A. Then, according to (5.27), the global error estimate is of order $\mathcal{O}(\tau^p)$. But the case that interests us is, of course, the one of an **unbounded** operator A.

The concept of strictly accurate order is crucial for the analysis presented here.

Definition 5.25. *The scheme is strictly accurate of order $p_0 \leq p$ if the local error vanishes for all f and u_0 such that the solution is polynomial in t of degree at most $p_0 - 1$.*

Remark. The Radau method with s stages is strictly accurate of order s. The Gauss method with s stages is strictly accurate of order $s + 1$. A collocation method with s stages cannot be strictly accurate of order $s + 2$. These results can be found in [4, Section 5].

If u and f are sufficiently smooth we can expand ρ_n in a Taylor series with respect to τ. We get

$$\rho_n = \sum_{l=0}^{p} \frac{\tau^l}{l!} u^{(l)}(t_n) - r(-\tau A) u(t_n) - \tau \sum_{j=1}^{s} q_j(-\tau A) \sum_{l=0}^{p-1} \frac{(c_j \tau)^l}{l!} f^{(l)}(t_n) + R_{n,p},$$

where

$$R_{n,p} = \int_{t_n}^{t_{n+1}} \frac{(t_{n+1} - s)^p}{p!} u^{(p+1)} ds - \tau \sum_{j=1}^{s} q_j(-\tau A) \int_{t_n}^{t_n + c_j \tau} \frac{(t_n + c_j \tau - s)^{p-1}}{(p-1)!} f^{(p)}(s) ds.$$

We have $R_{n,p} = \mathcal{O}(\tau^{p+1})$. From (5.17), for sufficiently smooth u and f holds

$$f^{(l)} = u^{(l+1)} + A u^{(l)}. \tag{5.29}$$

By using this, we get rid of the function f and its derivatives in the Taylor expansion above, and obtain

$$\rho_n = \sum_{l=0}^{p} \frac{\tau^l}{l!} h_l(-\tau A)\, u^{(l)}(t_n) + R_{n,p} \tag{5.30}$$

with

$$\begin{aligned} h_0(z) &= 1 - r(z) + z \sum_{j=1}^{s} q_j(z), \\ h_l(z) &= 1 - l \sum_{j=1}^{s} c_j^{l-1}\, q_j(z) + z \sum_{j=1}^{s} c_j^l\, q_j(z), \quad 1 \le l \le p-1, \\ h_p(z) &= 1 - p \sum_{j=1}^{s} c_j^{p-1}\, q_j(z). \end{aligned}$$

In [4, Lemma 1] the order conditions in terms of h_l, $l = 0, \ldots, p$, are given.

Lemma 5.26. *The scheme (5.23) is accurate of order p if and only if*

$$h_l(z) = \mathcal{O}\left(z^{p+1-l}\right) \quad \textit{for} \quad l = 0, \ldots, p. \tag{5.31}$$

It is strictly accurate of order $p_0 \le p$ if and only if

$$h_l(z) = 0 \quad \textit{for} \quad l = 0, \ldots, p_0 - 1. \tag{5.32}$$

In the case of a **bounded** operator A, the local error ρ_n is then indeed in $\mathcal{O}(\tau^{p+1})$

$$\|\rho_n\|_X \le \sum_{l=p_0+1}^{p} \frac{\tau^{p+1}}{l!} \|A^{p+1-l}\|_{X \leftarrow X} \left\|u^{(l)}(t_n)\right\|_X + \|R_{n,p}\|_X .$$

Now let A be an **unbounded operator** satisfying Assumption 5.22. For this purpose we shall first briefly discuss the representation of functions of $-A$ in terms of the semigroup $S(t)$. Let

$$\tilde{M} := \left\{ g : \mathbb{C} \to \mathbb{C} \mid g(z) = \tilde{\mu}(z) = \int_{\mathbb{R}_+} e^{zt} d\mu(t),\ \mu \text{ bounded measure} \right\}$$

be a set of Laplace transforms of bounded measures on $\mathbb{R}_+$. Then $g(-A)$ may be represented as

$$g(-A) = \int_{\mathbb{R}_+} S(t) d\mu(t).$$

Since μ is uniquely determined by g, we may set $m(g) = \int d\,|\mu|\,(t)$ and obtain that $g(-\tau A)$ is a bounded operator on X:

$$\|g(-\tau A)\|_{X \leftarrow X} \le \int_{\mathbb{R}_+} \|S(\tau t)\|_{X \leftarrow X}\, d\,|\mu|\,(t) \le C_A m(g). \tag{5.33}$$

Any rational function which is bounded for $\operatorname{Re} z \le 0$ belongs to $\tilde{M}$. Also, for $f, g \in \tilde{M}$ it holds $fg \in \tilde{M}$ and $(fg)(-A) = f(-A)g(-A)$. Further, if $f, g \in \tilde{M}$ and $g(z) = f(z)\, z^l$, then $g(-A)v = f(-A)(-A)^l v$ for $v \in D(A^l)$, see [5, Lemma 4].

From the definition of h_l, taking into account Assumption 5.23, one can prove that $h_l \in \tilde{M}$ for $l = 0, \dots, p$. We define

$$\tilde{h}_l(z) := z^{-(p+1-l)} h_l(z) \quad \text{for} \quad l = 0, \dots, p. \tag{5.34}$$

If the scheme is accurate of order p then from Lemma 5.26 it also follows that $\tilde{h}_l \in \tilde{M}$ for $l = 0, \dots, p$. Moreover, for every $j = 0, \dots, p+1-l$, is $z^j h_l(z) \in \tilde{M}$.

The local error formula (5.30) can now be rewritten as

$$\rho_n = \tau^{p+1} \sum_{l=p_0}^{p} \frac{1}{l!}\, \tilde{h}_l(-\tau A)\, (-A)^{p+1-l}\, u^{(l)}(t_n) + R_{n,p}, \tag{5.35}$$

provided that $u^{(l)} \in D(A^{p+1-l})$ for $l = p_0, \dots, p$. The convergence result can be given now, cf. [4, Theorem 1].

Theorem 5.27. *Assume that Assumptions 5.22, 5.23 and 5.24 are fulfilled and that the scheme (5.23) is accurate of order p and strictly accurate of order p_0. Then, if $u^{(l)} \in L^1(0, T; D(A^{p+1-l}))$ for $l = p_0, \dots, p+1$, we have*

$$\|u_N - u(T)\|_X \le C\, \tau^p \sum_{l=p_0}^{p+1} \int_0^T \left\| A^{p+1-l} u^{(l)}(\theta) \right\|_X d\theta. \tag{5.36}$$

Proof. We use the following expression for $\phi \in C^1(t_n, t_{n+1})$

$$\tau \phi(t_n) = \int_{t_n}^{t_{n+1}} \big(\phi(\theta) - (t_{n+1} - s)\phi'(\theta) \big)\, d\theta$$

to obtain for $l = p_0, \dots, p$

$$\begin{aligned} \tau \left\| \tilde{h}_l(-\tau A)(-A)^{p+1-l} u^{(l)}(t_n) \right\|_X \le \int_{t_n}^{t_{n+1}} \Big(& \left\| \tilde{h}_l(-\tau A) A^{p+1-l} u^{(l)}(\theta) \right\|_X \\ & + \left\| -\tau A \tilde{h}_l(-\tau A) A^{p-l} u^{(l+1)}(\theta) \right\|_X \Big) d\theta. \end{aligned}$$

Since $\tilde{h}_l(z)$ and $z\tilde{h}_l(z)$ are in $\tilde{M}$, we have

$$\tau \left\| \tilde{h}_l(-\tau A) A^{p+1-l} u^{(l)}(t_n) \right\|_X \le C \int_{t_n}^{t_{n+1}} \left(\left\| A^{p+1-l} u^{(l)}(\theta) \right\|_X + \left\| A^{p-l} u^{(l+1)}(\theta) \right\|_X \right) d\theta.$$

For the reminder term, after using (5.29) we have

$$\|R_{n,p}\|_X \le C\, \tau^p \int_{t_n}^{t_{n+1}} \left(\left\| u^{(p+1)}(\theta) \right\|_X + \left\| A u^{(p)}(\theta) \right\|_X \right) d\theta.$$

From (5.35) now follows

$$\|\rho_n\|_X \leq C\,\tau^p \sum_{l=p_0}^{p+1} \int_{t_n}^{t_{n+1}} \left\|(-A)^{p+1-l} u^{(l)}(\theta)\right\|_X d\theta. \tag{5.37}$$

Inserting this into (5.27) proves the claim. □

Remark. The error estimate in Theorem 5.27 requieres $u^{(l)} \in D(A^{p+1-l})$ for $l = p_0, \ldots, p+1$. In the context of partial differential equations this requires not just smoothness of the solution, but also that its time derivatives satisfy certain boundary conditions. Therefore, these conditions are undesirable. **Appropriate assumptions are of the form $u^{(q)} \in D(A)$, $q = 0, \ldots p$.**

We know that if the method is accurate of order p, it is also accurate of order q, for any $q \leq p$. Therefore, the following result holds.

Corollary 5.28. *Assume that Assumptions 5.22, 5.23 and 5.24 are fulfilled and that the scheme (5.23) is accurate and strictly accurate of order p_0. For $u^{(p_0)} \in L^1(0,T;D(A))$ and $u^{(p_0+1)} \in L^1(0,T;X)$ then holds*

$$\|u_N - u(T)\|_X \leq C\,\tau^{p_0} \int_0^T \left\|Au^{(p_0)}(\theta)\right\|_X + \left\|u^{(p_0+1)}(\theta)\right\|_X d\theta.$$

This phenomenon according to which the order of the method is not equal to the classical order, is called **order reduction**. It has been studied for the case of implicit Runge–Kutta methods in the context of partial differential equations, for example, in [60], [51] and [52]. For the work on avoiding order reduction, we refer the reader to [1] and [2] for the explicit and implicit Runge–Kutta methods, respectively.

Corollary 5.28 shows the convergence of order $s+1$ for Gauss methods, but convergence of order s only for Radau methods. This can be improved by a small trick presented in [4, Theorem 2]. Here is the result.

Theorem 5.29. *Assume that Assumptions 5.22, 5.23 and 5.24 are fulfilled and that the scheme (5.23) is strictly accurate of order p_0 and accurate of order $p_0 + 1$. Further on we suppose that the following condition holds*

$$\sigma(z) = h_{p_0}(z)/(z(1 - r(z))) \in \tilde{M}. \tag{5.38}$$

Then, under the appropriate regularity assumptions we have

$$\|u_N - u(T)\|_X \leq C\tau^{p_0+1} \left\{ \left\|Au^{(p_0)}(0)\right\|_X + \int_0^T \left(\left\|Au^{(p_0+1)}(\theta)\right\|_X + \left\|u^{(p_0+2)}(\theta)\right\|_X \right) d\theta \right\}.$$

Condition (5.38) is valid for the first and the second subdiagonal Padé approximations, and also for the diagonal approximations r_{11} and r_{22} but not for r_{33}. This means the theorem applies for Radau collocation method, providing convergence of order $s + 1$.

Remark. In the case of order of accuracy $p = p_0 + 2$ it is impossible, in general, to infer an $\mathcal{O}(\tau^p)$ global error estimate without making unnatural assumptions of the form $u^{(l)} \in D(A^{p-l})$.

Application to Maxwell's equations

It remains to check that Assumptions 5.22 and 5.24 are fulfilled in the case of Maxwell's equations and Gauss and Radau collocation methods for $X = V$.

From Theorem 3.3 we know that the Maxwell operator A_M generates a unitary C_0 group, i. e.

$$\|\mathrm{e}^{-tA_M}\|_{V\leftarrow V} = 1, \quad t \in \mathbb{R}.$$

Assumption 5.22 is therefore satisfied with $C_A = 1$. According to the following theorem [30, Theorem 6], Assumption 5.24 is satisfied too.

Theorem 5.30. *If $-A$ is dissipative and r is A-stable, then $\|r(-\tau A)^n\|_{X\leftarrow X} \leq 1$, for all $n = 1, 2, \ldots$.*

Therefore, both Corollary 5.28 and Theorem 5.29 apply to the case of Maxwell's equations and give convergence of order $s+1$ for Gauss and Radau methods, respectively.

5.3 Our results obtained using the energy technique

This section is an extended version of our paper [38, Section 3].

5.3.1 Implicit Euler method for continuous Maxwell's equations

In this section we consider the implicit Euler method for the time integration of the abstract problem (5.1)

$$u^{n+1} = u^n + \tau(-A_M u^{n+1} + f^{n+1}). \tag{5.39}$$

We treat this scheme separately because its analysis simplifies considerably compared to general higher order methods. Moreover, since the stage order the and order of the implicit Euler scheme coincide, it does not fit into the assumptions of Theorem 5.35 below.

Theorem 5.31. *Let $u \in C\big(0, T; D(A_M)\big)$ denote the solution of (5.1) and assume that $u'' \in L^2(0, T, V)$. Then, for τ sufficiently small (depending on T only), the error $e^n = u^n - u(t_n)$ of the implicit Euler method is bounded by*

$$\|e^N\|_V \leq C(T+1)^{1/2}\tau\Big(\int_0^T \|u''(t)\|_V^2\, dt\Big)^{1/2},$$

where the constant $C = C(\Omega, b)$ is independent of u.

Proof. The exact solution satisfies

$$u(t_{n+1}) = u(t_n) + \tau(-A_M u(t_{n+1}) + f^{n+1}) + \delta^{n+1},$$

where δ^{n+1} is given in (5.15) for $p = 1$. Subtracting this from (5.39) yields the error recursion

$$e^{n+1} = e^n - \tau A_M e^{n+1} - \delta^{n+1}. \tag{5.40}$$

Taking the V-inner product with e^{n+1} and using Corollary 3.2 we obtain

$$(e^{n+1} - e^n, e^{n+1})_V = (\delta^{n+1}, e^{n+1})_V. \tag{5.41}$$

We sum from 0 to $N-1$ and use the following representation of the left-hand side

$$\begin{aligned}\sum_{n=0}^{N-1}(e^{n+1} - e^n, e^{n+1})_V &= \frac{1}{2}\left\|e^N\right\|_V^2 + \frac{1}{2}\left\|e^N\right\|_V^2 - (e^{N-1}, e^N)_V + \frac{1}{2}\left\|e^{N-1}\right\|_V^2 \\ &+ \frac{1}{2}\left\|e^{N-1}\right\|_V^2 - (e^{N-2}, e^{N-1})_V + \frac{1}{2}\left\|e^{N-2}\right\|_V^2 + \dots \\ &+ \frac{1}{2}\left\|e^1\right\|_V^2 - (e^0, e^1)_V + \frac{1}{2}\left\|e^0\right\|_V^2 - \frac{1}{2}\left\|e^0\right\|_V^2 \\ &\geq \frac{1}{2}\left\|e^N\right\|_V^2 - \frac{1}{2}\left\|e^0\right\|_V^2.\end{aligned}$$

For the right-hand side of (5.41) we have

$$\sum_{n=0}^{N-1}(\delta^{n+1}, e^{n+1})_V \leq \frac{\tau}{2}\sum_{n=0}^{N-1}\left((T+1)\left\|\frac{\delta^{n+1}}{\tau}\right\|_V^2 + \frac{1}{T+1}\left\|e^{n+1}\right\|_V^2\right).$$

The result follows by a discrete Gronwall inequality from Corollary 2.9. □

5.3.2 Higher order collocation methods for continuous Maxwell's equations

We will present error bounds for algebraically stable Runge–Kutta methods which satisfy the coercivity condition (5.12). We start by discretizing the abstract Cauchy problem (5.1) in time by using implicit s-stage Runge–Kutta methods. This yields approximations $U^{ni} \approx u(t_n + c_i\tau)$ and $u^{n+1} \approx u(t_{n+1})$ defined by

$$\begin{aligned}&\dot{U}^{ni} + A_M U^{ni} = f^{ni}, \qquad f^{ni} = f(t_n + c_i\tau), \qquad i = 1, \dots s, \\ &U^{ni} = u^n + \tau\sum_{j=1}^{s} a_{ij}\dot{U}^{nj}, \quad i = 1, \dots s, \\ &u^{n+1} = u^n + \tau\sum_{i=1}^{s} b_i\dot{U}^{ni}.\end{aligned} \tag{5.42}$$

Remark. For A-stable collocation methods such as Gauß- and Radau methods, a unique solution of the linear system defining the interior Runge-Kutta approximations U^{ni}, $i = 1, \dots, s$ exists because A_M is skew-adjoint. This can be easily seen by using the coercivity condition (5.12).

Defects

We start by inserting the exact solution of (5.1) into the numerical scheme using the notation

$$\tilde{u}^n = u(t_n), \qquad \tilde{U}^{ni} = u(t_n + c_i\tau), \qquad \dot{\tilde{U}}^{ni} = u'(t_n + c_i\tau).$$

This yields

$$\begin{aligned}\dot{\tilde{U}}^{ni} + A_M \tilde{U}^{ni} &= f^{ni},\\ \tilde{U}^{ni} &= \tilde{u}^n + \tau \sum_{j=1}^{s} a_{ij} \dot{\tilde{U}}^{nj} + \Delta^{ni},\\ \tilde{u}^{n+1} &= \tilde{u}^n + \tau \sum_{i=1}^{s} b_i \dot{\tilde{U}}^{ni} + \delta^{n+1}.\end{aligned} \tag{5.43}$$

For Gauss collocation methods with $s \geq 1$, Radau collocation methods with $s \geq 2$ and $u^{(s+1)} \in L^2(0,T;D(A_M))$, $u^{(s+2)} \in L^2(0,T;V)$ the defects are given in the Theorem 5.18 with $p = s+1$.

Lemma 5.32. *We have*

$$\tau \sum_{n=1}^{N} \left(\sum_{i=1}^{s} \left\| \Delta^{ni} \right\|_{D(A_M)}^2 + \left\| \frac{1}{\tau} \delta^{n+1} \right\|_V^2 \right) \leq C\tau^{2(s+1)} B(u,s,T), \tag{5.44}$$

where

$$B(u,s,T) = \int_0^T \left\| u^{(s+1)}(t) \right\|_{D(A_M)}^2 dt + \int_0^T \left\| u^{(s+2)}(t) \right\|_V^2 dt. \tag{5.45}$$

Proof. Using the uniform boundedness of Peano kernels and the Cauchy-Schwarz inequality we have

$$\begin{aligned}\left\| \Delta^{ni} \right\|_{D(A_M)}^2 &\leq C\tau^{2s+1} \int_{t_n}^{t_{n+1}} \left\| u^{s+1}(t) \right\|_{D(A_M)}^2 dt,\\ \left\| \delta^{n+1} \right\|_V^2 &\leq C\tau^{2s+3} \int_{t_n}^{t_{n+1}} \left\| u^{s+2}(t) \right\|_V^2 dt,\end{aligned}$$

and therefore

$$\begin{aligned}\sum_{i=1}^{s} \left\| \Delta^{ni} \right\|_{D(A_M)}^2 &\leq C\tau^{2s+1} \int_{t_n}^{t_{n+1}} \left\| u^{s+1}(t) \right\|_{D(A_M)}^2 dt,\\ \left\| \frac{1}{\tau} \delta^{n+1} \right\|_V^2 &\leq C\tau^{2s+1} \int_{t_n}^{t_{n+1}} \left\| u^{s+2}(t) \right\|_V^2 dt.\end{aligned}$$

Summing over all n and multiplying with τ gives the result. □

Remark. The order of the defect δ^{n+1} is not sharp if the solution is more regular. More precisely, for Gauß collocation methods and $u^{(2s+1)} \in L^2(0,T,V)$ we have $\delta^{n+1} = \mathcal{O}(\tau^{2s+1})$, while for Radau collocation methods and $u^{(2s)} \in L^2(0,T,V)$ we have $\delta^{n+1} = \mathcal{O}(\tau^{2s})$. However, we cannot exploit this in our convergence analysis, since the global order is determined by the stage order, which is s for all collocation methods.

By subtracting (5.43) from (5.42), the time integration errors

$$e^n := u^n - \tilde{u}^n, \quad E^{ni} := U^{ni} - \tilde{U}^{ni}$$

satisfy

$$\dot{E}^{ni} + A_M E^{ni} = 0, \quad i = 1, \dots s, \tag{5.46a}$$

$$E^{ni} = e^n + \tau \sum_{j=1}^{s} a_{ij} \dot{E}^{nj} - \Delta^{ni}, \quad i = 1, \dots s, \tag{5.46b}$$

$$e^{n+1} = e^n + \tau \sum_{i=1}^{s} b_i \dot{E}^{ni} - \delta^{n+1}. \tag{5.46c}$$

Let $\Delta^n = \left(\Delta^{n1} \dots \Delta^{ns}\right)^T$, $E^n = \left(E^{n1} \dots E^{ns}\right)^T$, $\dot{E}^n = \left(\dot{E}^{n1} \dots \dot{E}^{ns}\right)^T$. Then, (5.46) can be written in a more compact form as

$$\begin{aligned} &\dot{E}^n + (I \otimes A_M) E^n = 0, \\ &E^n = \mathbb{1} \otimes e^n + \tau (\mathcal{A} \otimes I) \dot{E}^n - \Delta^n, \\ &e^{n+1} = e^n + \tau (b^T \otimes I) \dot{E}^n - \delta^{n+1}, \end{aligned} \tag{5.47}$$

where $\mathbb{1} = [1 \dots 1]^T$.

Energy techniques

The following analysis uses an energy technique motivated by [46].

Lemma 5.33. *The error $e^n = u^n - u(t_n)$ satisfies*

$$\begin{aligned} \left\|e^{n+1}\right\|_V^2 - \left\|e^n\right\|_V^2 \leq \frac{C}{T+1} \tau \left(\left\|e^n\right\|_V^2 + \sum_{i=1}^{s} \left\|E^{ni}\right\|_V^2 \right) \\ + C(T+1)\tau \left(\sum_{i=1}^{s} \left\|\Delta^{ni}\right\|_{D(A_M)}^2 + \left\| \frac{1}{\tau} \delta^{n+1} \right\|_V^2 \right). \end{aligned} \tag{5.48}$$

Here, the constant C only depends on $\mathcal{A}$, b, and s.

Proof. Taking the inner product of (5.46c) with itself we obtain

$$\left\|e^{n+1}\right\|_V^2 = \left\| e^n + \tau \sum_{i=1}^{s} b_i \dot{E}^{ni} \right\|_V^2 - 2(\delta^{n+1}, e^n + \tau \sum_{i=1}^{s} b_i \dot{E}^{ni})_V + \left\|\delta^{n+1}\right\|_V^2. \tag{5.49}$$

We estimate each of these three terms separately. For the first term we have

$$\left\| e^n + \tau \sum_{i=1}^{s} b_i \dot{E}^{ni} \right\|_V^2 = \left\|e^n\right\|_V^2 + 2\tau \sum_{i=1}^{s} b_i (e^n, \dot{E}^{ni})_V + \tau^2 \sum_{i,j=1}^{s} b_i b_j (\dot{E}^{ni}, \dot{E}^{nj})_V \tag{5.50}$$

By using (5.46b) we write e^n as

$$e^n = E^{ni} - \tau \sum_{j=1}^{s} a_{ij} \dot{E}^{nj} + \Delta^{ni}.$$

Inserting this identity into the second term of (5.50) we obtain

$$\left\| e^n + \tau \sum_{i=1}^{s} b_i \dot{E}^{ni} \right\|_V^2 = \|e^n\|_V^2 + 2\tau \sum_{i=1}^{s} b_i (E^{ni} + \Delta^{ni}, \dot{E}^{ni})_V + \tau^2 \sum_{i,j=1}^{s} (b_i b_j - b_i a_{ij} + b_j a_{ji})(\dot{E}^{ni}, \dot{E}^{nj})_V.$$

Since the method is algebraically stable, the last term is not positive and we end up with

$$\left\| e^n + \tau \sum_{i=1}^{s} b_i \dot{E}^{ni} \right\|_V^2 \leq \|e^n\|_V^2 + 2\tau \sum_{i=1}^{s} b_i (\dot{E}^{ni}, E^{ni} + \Delta^{ni})_V. \tag{5.51}$$

The skew-symmetry (3.22) of the operator A_M implies

$$(\dot{E}^{ni}, E^{ni})_V = -(A_M E^{ni}, E^{ni})_V = 0,$$
$$(\dot{E}^{ni}, \Delta^{ni})_V = -(A_M E^{ni}, \Delta^{ni})_V = (E^{ni}, A_M \Delta^{ni})_V \leq \left\| E^{ni} \right\|_V \left\| A_M \Delta^{ni} \right\|_V .$$

$A_M \Delta^{ni}$ is well defined because of $u^{(s+1)} \in L^2(0, T; D(A_M))$. For arbitrary $\gamma > 0$, Young's inequality gives

$$\left\| e^n + \tau \sum_{i=1}^{s} b_i \dot{E}^{ni} \right\|_V^2 \leq \|e^n\|_V^2 + \tau \sum_{i=1}^{s} b_i \left(\frac{1}{\gamma} \left\| E^{ni} \right\|_V^2 + \gamma \left\| A_M \Delta^{ni} \right\|_V \right). \tag{5.52}$$

To bound the second term in (5.49) first observe that

$$(\delta^{n+1}, e^n)_V \leq \tau \left\| \frac{1}{\tau} \delta^{n+1} \right\|_V \|e^n\|_V \leq \frac{1}{2\gamma} \tau \|e^n\|_V^2 + \frac{\gamma}{2} \tau \left\| \frac{1}{\tau} \delta^{n+1} \right\|_V^2. \tag{5.53}$$

In order to bound $(\delta^{n+1}, \dot{E}^{ni})_V$ we use the compact form (5.47). From the second equation of (5.47) we have

$$\dot{E}^n = \frac{1}{\tau} (\mathcal{A}^{-1} \otimes I)(E^n + \Delta^n - \mathbb{1} \otimes e^n).$$

If we denote the inverse of the Runge–Kutta matrix by

$$\mathcal{A}^{-1} = (\omega_{ij})_{i,j}, \tag{5.54}$$

then

$$\dot{E}^{ni} = \frac{1}{\tau} \sum_{j=1}^{s} \omega_{ij} (E^{nj} + \Delta^{nj} - e^n). \tag{5.55}$$

Hence we have

$$\tau(\delta^{n+1}, \dot{E}^{ni})_V \leq \tau \left\| \frac{1}{\tau} \delta^{n+1} \right\|_V \sum_{j=1}^{s} |\omega_{ij}| \left(\|e^n\|_V + \left\| E^{nj} \right\|_V + \left\| \Delta^{nj} \right\|_V \right)$$
$$\leq \frac{\gamma}{2} \tau \left\| \frac{1}{\tau} \delta^{n+1} \right\|_V^2 + \frac{C}{\gamma} \tau \Big(\|e^n\|_V^2 + \sum_{j=1}^{s} \left\| E^{nj} \right\|_V^2 + \sum_{j=1}^{s} \left\| \Delta^{nj} \right\|_V^2 \Big),$$

where $C = C(\mathcal{A}, b, s)$. Since the right-hand side does not depend on i and $\sum_i b_i = 1$, we conclude from (5.53) that

$$(\delta^{n+1}, e^n + \tau \sum_{i=1}^{s} b_i \dot{E}^{ni})_V \leq \frac{C}{\gamma}\tau\Big(\|e^n\|_V^2 + \sum_{j=1}^{s} \left\|E^{nj}\right\|_V^2 + \sum_{j=1}^{s} \left\|\Delta^{nj}\right\|_V^2\Big) + \gamma\tau \left\|\frac{1}{\tau}\delta^{n+1}\right\|_V^2. \tag{5.56}$$

Inserting (5.52) and (5.56) for $\gamma = 1+T$ into (5.49), and writing $\left\|\delta^{n+1}\right\|_V = \tau \left\|\delta^{n+1}/\tau\right\|_V$ finally proves the Lemma. □

In order to use the Gronwall inequality, we have to bound $\left\|E^{ni}\right\|_V$ in terms of $\|e^n\|_V$ and the defects.

Lemma 5.34. *The error of the inner stages satisfies*

$$\sum_{i=1}^{s} \left\|E^{ni}\right\|_V^2 \leq C \left(\|e^n\|_V^2 + \sum_{i=1}^{s} \left\|\Delta^{ni}\right\|_V^2\right), \tag{5.57}$$

where, $C = C(\mathcal{A}, s, \mathcal{D}, \alpha)$.

Proof. From (5.47) we conclude

$$E^n = \mathbb{1} \otimes e^n - \tau(\mathcal{A} \otimes A_M)E^n - \Delta^n.$$

Multiplying by $\mathcal{D}\mathcal{A}^{-1} \otimes I$, where $\mathcal{D} = \operatorname{diag}(d_1, \ldots, d_s)$ is the diagonal matrix arising in the coercivity property (5.12), and taking the inner product with E^n yields

$$(E^n, (\mathcal{D}\mathcal{A}^{-1} \otimes I)E^n)_{V^s} = -\tau(E^n, (\mathcal{D} \otimes A_M)E^n)_{V^s} + (E^n, (\mathcal{D}\mathcal{A}^{-1} \otimes I)(\mathbb{1} \otimes e^n - \Delta^n))_{V^s}.$$

The coercivity implies the lower bound

$$(E^n, (\mathcal{D}\mathcal{A}^{-1} \otimes I)E^n)_{V^s} \geq \alpha \sum_{i=1}^{s} d_i \left\|E^{ni}\right\|_V^2.$$

Note that

$$(E^n, (\mathcal{D} \otimes A_M)E^n)_{V^s} = \sum_{i=1}^{s} d_i (E^{ni}, A_M E^{ni})_V = 0$$

since A_M is skew-symmetric, cf. (3.22). Using the notation (5.54) for the entries of $\mathcal{A}^{-1}$ we obtain

$$\begin{aligned}(E^n, (\mathcal{D}\mathcal{A}^{-1} \otimes I)(\mathbb{1} \otimes e^n - \Delta^n))_{V^s} &= \sum_{i,j=1}^{s} d_i \omega_{ij} (E^{ni}, e^n - \Delta^{nj})_V \\ &\leq \gamma \sum_{i=1}^{s} d_i \left\|E^{ni}\right\|_V^2 + \frac{C}{\gamma}\left(\|e^n\|_V^2 + \sum_{i=1}^{s} \left\|\Delta^{ni}\right\|_V^2\right).\end{aligned} \tag{5.58}$$

Choosing $\gamma = \frac{\alpha}{2}$ completes the proof. □

Main result (time discretization error)

From the bounds on the defects given in Lemma 5.32 we are now able to state and prove our main result for the error of the time integration scheme.

Theorem 5.35. *Let u be the solution of (5.1). Assume that $u^{(s+1)} \in L^2(0,T;D(A_M))$ and $u^{(s+2)} \in L^2(0,T;V)$. Then for $\tau > 0$ sufficiently small (depending on the coefficients of the Runge–Kutta method and T only), the error of an s-stage algebraically stable and coercive Runge–Kutta method of order at least two satisfies*

$$\left\|e^N\right\|_V \leq C(T+1)^{1/2}\tau^{s+1}B(u,s,T)^{1/2},$$

where B is defined in (5.45). The constant $C = C(\mathcal{Q},b)$ is independent of u.

Proof. Inserting (5.57) into (5.48) we get

$$\left\|e^{n+1}\right\|_V^2 - \left\|e^n\right\|_V^2 \leq \frac{C\tau}{1+T}\left\|e^n\right\|_V^2 + C(T+1)\tau\left(\sum_{i=1}^{s}\left\|\Delta^{ni}\right\|_{D(A_M)}^2 + \left\|\frac{1}{\tau}\delta^{n+1}\right\|_V^2\right).$$

Summing over n and applying the discrete Gronwall lemma from Corollary 2.9, we have for τ sufficiently small

$$\left\|e^N\right\|_V^2 \leq Ce^{N\frac{C}{1+T}\tau}(T+1)\tau\sum_{n=1}^{N}\left(\sum_{i=1}^{s}\left\|\Delta^{ni}\right\|_{D(A_M)}^2 + \left\|\frac{1}{\tau}\delta^{n+1}\right\|_V^2\right)$$

The assumptions on the regularity of u and the bound (5.44) for the defects prove the desired result. □

Remark. The result proven here is equivalent to results proven in Subsection 5.2. Nevertheless, this technique enables us to obtain better results in the analysis of the fully discrete problem.

Lemma 5.36. *For $f \equiv 0$, the Runge–Kutta solution preserves the divergence, i.e,*

$$\nabla\cdot(\epsilon\mathbf{E}^n) = \nabla\cdot(\epsilon\mathbf{E}^0), \qquad \nabla\cdot(\mu\mathbf{H}^n) = \nabla\cdot(\mu\mathbf{H}^0), \qquad n = 1,2,\ldots.$$

Proof. From (5.42) it follows that

$$\begin{pmatrix}\mathbf{H}^{n+1}\\ \mathbf{E}^{n+1}\end{pmatrix} = \begin{pmatrix}\mathbf{H}^{n}\\ \mathbf{E}^{n}\end{pmatrix} + \sum_{i=1}^{s} b_i \begin{pmatrix}\mu^{-1}\nabla\times\mathbf{E}^{ni}\\ -\epsilon^{-1}\nabla\times\mathbf{H}^{ni}\end{pmatrix}$$

or equivalently

$$\mu\mathbf{H}^{n+1} = \mu\mathbf{H}^n + \sum_{i=1}^{s} b_i\nabla\times\mathbf{E}^{ni},$$

$$\epsilon\mathbf{E}^{n+1} = \epsilon\mathbf{E}^n - \sum_{i=1}^{s} b_i\nabla\times\mathbf{H}^{ni}.$$

Taking the divergence of each equation and using (2.10) gives $\nabla\cdot(\epsilon\mathbf{E}^{n+1}) = \nabla\cdot(\epsilon\mathbf{E}^n)$ and $\nabla\cdot(\mu\mathbf{H}^{n+1}) = \nabla\cdot(\mu\mathbf{H}^n)$. □

5.4 Exponential Runge–Kutta methods

In contrast to Runge–Kutta methods, the exponential integrators are introduced directly for an abstract evolution equation. X is again a Banach space with norm $\|\cdot\|_X$. We follow [37, Section 2.2] to derive error bounds for exponential Runge–Kutta discretizations of (5.17) with A satisfying Assumption 5.19. The exact solution at time $t_{n+1} = t_n + \tau$ is given by the variation-of-constants formula

$$u(t_{n+1}) = \mathrm{e}^{-\tau A} u(t_n) + \int_0^\tau \mathrm{e}^{-(\tau-\theta)A} f(t_n + \theta) d\theta. \tag{5.59}$$

The main idea of exponential integrators it to handle the first term, i. e. the stiff part, exactly. For linear problems we can approximate the integral by using an *exponential quadrature rule*

$$u^{n+1} = \mathrm{e}^{-\tau A} u^n + \tau \sum_{i=1}^{s} b_i(-\tau A) f(t_n + c_i \tau), \tag{5.60a}$$

where the weights are chosen such that the quadrature rule is exact for all polynomials of degree $s-1$, i. e.

$$b_i(-\tau A) = \int_0^1 \mathrm{e}^{-\tau(1-\theta)A} \ell_i(\theta) d\theta. \tag{5.60b}$$

Here, ℓ_i are the familiar Lagrange interpolation polynomials, see Section 5.1. Obviously, the weights $b_i(z)$ are linear combinations of the entire functions

$$\varphi_k(z) = \int_0^1 \mathrm{e}^{(1-\theta)z} \frac{\theta^{k-1}}{(k-1)!} d\theta, \quad k \geq 1. \tag{5.61}$$

Assumption 5.19 enables us to define the operators

$$\varphi_k(-\tau A) = \int_0^1 \mathrm{e}^{-\tau(1-\theta)A} \frac{\theta^{k-1}}{(k-1)!} d\theta, \quad k \geq 1.$$

The following lemma turns out to be crucial.

Lemma 5.37. *The operators $b_i(-\tau A)$ are bounded uniformly in $\tau \in [0, \tau_*]$, for every $\tau_* > 0$.*

Proof. Follows directly from [37, Lemma 2.4]. □

Without loss of generality we take $\tau_* = 1$, since we do not use larger time steps. The last Lemma implies that the weights $b_i(-\tau A)$ defined in (5.60b) are uniformly bounded in $\tau \in [0, 1]$.

In order to analyse (5.60), we expand the right-hand side of (5.59) in a Taylor series with remainder in integral form:

$$\begin{aligned} u(t_{n+1}) = \mathrm{e}^{-\tau A} u(t_n) + \tau \sum_{k=1}^{p} \varphi_k(-\tau A) \tau^{k-1} f^{(k-1)}(t_n) \\ + \int_0^\tau \mathrm{e}^{-(\tau-\theta)A} \int_0^\theta \frac{(\theta-\xi)^{p-1}}{(p-1)!} f^{(p)}(t_n + \xi) d\xi d\theta. \end{aligned} \tag{5.62}$$

This has to be compared with the Taylor series of the numerical solution (5.60):

$$\begin{aligned} u^{n+1} = \mathrm{e}^{-\tau A}u^n + \tau\sum_{i=1}^{s} b_i(-\tau A)\sum_{k=0}^{p-1}\frac{\tau^k c_i^k}{k!}f^{(k)}(t_n) \\ + \tau\sum_{i=1}^{s} b_i(-\tau A)\int_0^{c_i\tau}\frac{(c_i\tau-\theta)^{p-1}}{(p-1)!}f^{(p)}(t_n+\theta)d\theta. \end{aligned} \tag{5.63}$$

Obviously the error $e_n = u^n - u(t_n)$ satisfies

$$e_{n+1} = \mathrm{e}^{-\tau A}\, e_n - \delta_{n+1} \tag{5.64}$$

with

$$\delta_{n+1} = \sum_{j=1}^{p}\tau^j\psi_j(-\tau A)f^{(j-1)}(t_n) + \delta_{n+1}^{[p]}, \tag{5.65}$$

where

$$\psi_j(-\tau A) = \varphi_j(-\tau A) - \sum_{i=1}^{s} b_i(-\tau A)\frac{c_i^{j-1}}{(j-1)!} \tag{5.66}$$

and

$$\begin{aligned} \delta_{n+1}^{[p]} = \int_0^{\tau}\mathrm{e}^{-(\tau-\theta)A}\int_0^{\theta}\frac{(\theta-\xi)^{p-1}}{(p-1)!}f^{(p)}(t_n+\xi)d\xi d\theta \\ - \tau\sum_{i=1}^{s} b_i(-\tau A)\int_0^{c_i\tau}\frac{(c_i\tau-\theta)^{p-1}}{(p-1)!}f^{(p)}(t_n+\theta)d\theta. \end{aligned}$$

Conditions

$$\psi_j(-\tau A) = 0, \quad j = 1,\dots,p, \tag{5.67}$$

are called the **order conditions**. We can now state the convergence result, cf. [37, Theorem 2.7].

Theorem 5.38. *Let Assumption 5.19 be fulfilled and let $f^{(s)} \in L^1(0,T;X)$. For the numerical solution of (5.17), consider the exponential quadrature rule (5.60). The error bound*

$$\|u^n - u(t_n)\|_X \le C\tau^s\int_0^T\left\|f^{(s)}(\theta)\right\|_X d\theta$$

then holds, uniformly on $0 \le t_n \le T$, with a constant C that depends on T, but it is independent of the chosen step size sequence.

Proof. The weights b_i of the exponential quadrature rule (5.60) satisfy the order conditions (5.67) for $p = s$ by construction. It holds $\delta_{j+1} = \delta_{j+1}^{[s]}$. Solving the error recursion (5.64) yields the estimate

$$\|e_n\|_X \le \sum_{j=0}^{n-1}\|\mathrm{e}^{-(t_n-t_j)A}\|_{X\leftarrow X}\left\|\delta_{j+1}^{[s]}\right\|_X.$$

By straightforward computations, using (2.4), the bound

$$\left\|\delta_{j+1}^{[s]}\right\|_X \leq \frac{1}{(s-1)!}\left(C_A \mathrm{e}^{\omega\tau} + C_B\right)\tau^s \int_0^\tau \left\|f^{(s)}(t_j+\theta)\right\|_X d\theta,$$

follows, where $C_B := \sup_{\tau\in[0,1]} \sum_{i=1}^s \|b_i(-\tau A)\|_{X\leftarrow X}$. Using (2.4) again, we obtain

$$\|e_n\|_X \leq \frac{1}{(s-1)!}\left(C_A \mathrm{e}^{\omega\tau} + C_B\right)\tau^s \sum_{j=0}^{n-1} \mathrm{e}^{\omega(t_n-t_j)} \int_0^\tau \left\|f^{(s)}(t_j+\theta)\right\|_X d\theta.$$

By bounding $\mathrm{e}^{\omega(t_n-t_j)} \leq \mathrm{e}^{\omega T}$ the claim follows with

$$C = \frac{1}{(s-1)!}\left(C_A \mathrm{e}^{\omega\tau} + C_B\right) e^{\omega T}. \tag{5.68}$$

□

The analysis done here is the analysis in the terms of data (here the right-hand side function f, but also the initial data u_0 can be included). In the case of collocation methods we performed the analysis in the terms of the exact solution u. Here we can also get a result in the terms of the exact solution. Indeed, if we suppose that $u^{(s)} \in L^1(0,T;D(A))$ and $u^{(s+1)} \in L^1(0,T;X)$ then from (5.17) $f^{(s)} = u^{(s+1)} + Au^{(s)}$ follows.

Corollary 5.39. *Under the assumptions of Theorem 5.38, if $u^{(s)} \in L^1(0,T;D(A))$ and $u^{(s+1)} \in L^1(0,T;X)$, the following error bound holds*

$$\|u^n - u(t_n)\|_X \leq C\tau^s \int_0^T \left\|Au^{(s)}(\theta)\right\|_X + \left\|u^{(s+1)}(\theta)\right\|_X d\theta$$

5.4.1 Exponential integrators for Maxwell's equations

Thanks to this abstract setting in which exponential integrators have been introduced, now it is quite simple to apply the exponential quadrature rule to the Maxwell's equations (5.1). In this case we have $(X, \|.\|_X) = (V, \|.\|_V)$.

From Theorem 3.3 we know that the Maxwell operator A_M satisfies Assumption 5.19 with $C_A = 1$ and $\omega = 0$. The following result holds.

Theorem 5.40. *Let u be exact solution of Maxwell's equations (5.1) such that $u^{(s)} \in L^1(0,T;D(A))$ and $u^{(s+1)} \in L^1(0,T;V)$. The numerical solution $(u^n)_{n\geq 0}$ defined with the exponential quadrature rule (5.60) satisfy*

$$\|u^n - u(t_n)\|_V \leq C\tau^s \int_0^T \left\|Au^{(s)}(\theta)\right\|_V + \left\|u^{(s+1)}(\theta)\right\|_V d\theta$$

uniformly on $0 \leq t_n \leq T$, with a constant $C = \frac{1+C_B}{(s-1)!}$.

CHAPTER 6

Fully discrete schemes for Maxwell's equations

In this chapter we present convergence results for fully discrete schemes for Maxwell's equations. The chapter consists of 4 sections. In the first section we study the explicit Euler method, which converges only under a CFL condition. This is a common restriction for all explicit methods. In Section 6.2 we extend the results from Section 5.2, which are valid for Gauss and Radau methods, to space discrete problems and get convergence results for fully discrete schemes. Our main result is contained in Section 6.3. Using an energy technique, we prove convergence of the fully discrete scheme which uses the dG method in space and Gauss or Radau methods in time. The error bounds obtained are better than in previous section. In the last section, we consider the application of exponential quadrature rules to a space discrete problem.

We start with a short recapitulation. In Section 3.2 we have formulated the linear Maxwell's equations as the abstract Cauchy problem

$$\partial_t u + A_M u = f, \qquad u(0) = u_0 \tag{6.1}$$

where A_M is the Maxwell operator defined by

$$A_M \begin{pmatrix} \mathbf{H} \\ \mathbf{E} \end{pmatrix} := \begin{pmatrix} -\mu^{-1}\nabla \times \mathbf{E} \\ \epsilon^{-1}\nabla \times \mathbf{H} \end{pmatrix}, \quad D(A_M) = H(\mathrm{curl}, \Omega) \times H_0(\mathrm{curl}, \Omega).$$

It has been shown that the operator A_M is skew-adjoint in the V-inner product, which implies that it generates a unitary C_0 semigroup, $S_M(t) = \mathrm{e}^{-tA_M}$, i. e.

$$\|S_M(t)\|_{V \leftarrow V} = 1. \tag{6.2}$$

This problem has been discretized in space in Section 4.3 by using the dG method with both central and upwind fluxes. We have obtained the semidiscrete problem

$$\partial_t u_h + A_h u_h = f_h, \qquad u_h(0) = \pi_h u_0, \tag{6.3}$$

on a discrete space $V_h = \mathbb{P}_3^k(\mathcal{T}_h)^6$, where the dG operator A_h was given by

$$\begin{aligned}(A_h \begin{pmatrix}\mathbf{H}_h\\ \mathbf{E}_h\end{pmatrix}, \begin{pmatrix}\phi_h\\ \psi_h\end{pmatrix})_V := &\sum_K \Big((\nabla\times\mathbf{E}_h,\phi_h)_{0,K} - (\nabla\times\mathbf{H}_h,\psi_h)_{0,K}\Big)\\ &+ \sum_{F\in\mathcal{F}_h^i}\Big((\mathbf{n}_F\times[\![\mathbf{E}_h]\!]_F, \alpha_K\phi_K + \alpha_{K_F}\phi_{K_F})_{0,F}\\ &\qquad - (\mathbf{n}_F\times[\![\mathbf{H}_h]\!]_F, \beta_K\psi_K + \beta_{K_F}\psi_{K_F})_{0,F}\\ &\qquad + \gamma_F(\mathbf{n}_F\times[\![\mathbf{E}_h]\!]_F, \mathbf{n}_F\times[\![\psi_h]\!]_F)_{0,F}\\ &\qquad + \delta_F(\mathbf{n}_F\times[\![\mathbf{H}_h]\!]_F, \mathbf{n}_F\times[\![\phi_h]\!]_F)_{0,F}\Big)\\ &+ \sum_{F\in\mathcal{F}_h^b}\Big(-(\mathbf{n}_F\times\mathbf{E},\phi_h)_{0,F} + 2\gamma_F(\mathbf{n}_F\times\mathbf{E},\mathbf{n}_F\times\psi_h)_{0,F}\Big).\end{aligned} \tag{6.4}$$

For $A_h = A_h^{\mathrm{cf}}$ the coefficients are given by $\alpha_K = \beta_K = \frac{1}{2}$ for all K and $\gamma_F = \delta_F = 0$ for all faces F, while for $A_h = A_h^{\mathrm{upw}}$ the coefficients are given in (4.62).

Next, we collect some properties of the space-discrete problem (6.3) and the discrete dG operator $A_h \in \{A_h^{\mathrm{cf}}, A_h^{\mathrm{upw}}\}$, which we need for the analysis of fully discrete methods. With $S_h(t) := \mathrm{e}^{-tA_h}$ we denote the semigroup on a finite dimensional space V_h. More precisely, we use S_h^{cf} and S_h^{upw} for the semigroup generated with $-A_h^{\mathrm{cf}}$ and $-A_h^{\mathrm{upw}}$, respectively.

i) From Lemmas 4.38 and 4.45 we have that $-A_h$ is dissipative:

$$(-A_h u_h, u_h)_V \le 0, \quad \forall u_h \in V_h. \tag{6.5}$$

For $A_h = A_h^{\mathrm{cf}}$ equality holds. It is easy to check that the operators $I + A_h : V_h \to V_h$ are injective, and therefore also surjective for both $A_h = A_h^{\mathrm{cf}}$ and $A_h = A_h^{\mathrm{upw}}$. By the Lumer-Phillips Theorem 2.2 it follows that the semigroups S_h are contractive, i. e.

$$\|S_h(t)\|_{V\leftarrow V} \le 1 \quad \forall t \ge 0. \tag{6.6}$$

Again, for $S_h = S_h^{\mathrm{cf}}$ equality holds, i. e. the semigroup is unitary, as in the continuous case.

ii) Since A_h is the discrete version of a first order differential operator, we expect that its norm scales as h^{-1}. Indeed, we can prove:

Theorem 6.1. *If a mesh sequence $\mathcal{T}_{\mathcal{H}}$ is quasi-uniform, i. e. (4.35) holds, then for $A_h \in \{A_h^{\mathrm{cf}}, A_h^{\mathrm{upw}}\}$ we have*

$$\|A_h u_h\|_V \le C h^{-1} \|u_h\|_V, \quad u_h \in V_h,$$

where C is independent of h.

Proof. We use

$$\|A_h u_h\|_V = \max_{\|w_h\|_V=1} (A_h u_h, w_h)_V.$$

and bound all terms in (6.4) separately. With the help of the Cauchy-Schwarz inequality (2.3) and the inverse inequality (4.36) we have

$$\sum_K (\nabla \times \mathbf{E}_h, \phi_h)_{0,K} \leq \sum_K \|\nabla \times \mathbf{E}_h\|_{0,K} \|\phi_h\|_{0,K} \leq C \sum_K h_K^{-1} \|\mathbf{E}_h\|_{0,K} \|\phi_h\|_{0,K}.$$

By using the quasi-uniformity (4.35) and the Cauchy-Schwarz inequality for vectors we get

$$\sum_K (\nabla \times \mathbf{E}_h, \phi_h)_{0,K} \leq C \|\mathbf{E}_h\|_{0,\Omega} \|\phi_h\|_{0,\Omega}.$$

We can estimate the second term in (6.4) analogously. The first interface term can be bounded as follows:

$$\begin{aligned}
&\sum_{F\in\mathcal{F}_h^i} (\mathbf{n}_F \times [\![\mathbf{E}_h]\!]_F, \alpha_K \phi_K + \alpha_{K_F} \phi_{K_F})_{0,F} \\
&\leq C \sum_{F\in\mathcal{F}_h^i} \left(h_{K_F}^{-1/2} \|E_h\|_{0,K_F} + h_K^{-1/2} \|E_h\|_{0,K} \right) \left(h_{K_F}^{-1/2} \|\phi_h\|_{0,K_F} + h_K^{-1/2} \|\phi_h\|_{0,K} \right) \\
&\leq C h^{-1} \Big(\sum_{F\in\mathcal{F}_h^i} \|E_h\|_{0,K_F}^2 + \|E_h\|_{0,K}^2 \Big)^{1/2} \Big(\sum_{F\in\mathcal{F}_h^i} \|\phi_h\|_{0,K_F}^2 + \|\phi_h\|_{0,K}^2 \Big)^{1/2} \\
&\leq C h^{-1} \|E_h\|_{0,\Omega} \|\phi_h\|_{0,K}.
\end{aligned}$$

Here, the Cauchy-Schwarz inequality, the triangle inequality and the discrete trace inequality (4.37) are used for the first inequality, and (4.35) and the Cauchy-Schwarz inequality for vectors are used for the second inequality. Other interface and boundary terms can be bounded analogously. The claim now follows by recalling the equivalence between the L^2-norm and the V-norm. □

iii) We have seen in Section 4.3 that the dG operators A_h are consistent in the sense that $\pi_h A_M u = A_h u$ for all $u \in D(A_M)$. In what follows we also need consistency in the following sense

$$\lim_{h\to 0} \|A_h \pi_h u - \pi_h A_M u\|_V = 0 \quad \forall u \in D(A_M). \tag{6.7}$$

Theorem 6.2. *Let $A_h \in \{A_h^{\mathrm{cf}}, A_h^{\mathrm{upw}}\}$. For every $k' \leq k$ and for all $u \in D(A_M) \cap H^{k'+1}(\mathcal{T}_h)^6$ it holds*

$$\|A_h \pi_h u - \pi_h A_M u\|_V \leq C h^{k'} |u|_{H^{k'+1}(\mathcal{T}_h)^6}. \tag{6.8}$$

Proof. Since $\mathbf{n}_F \times [\![u]\!]_F = 0$ for $u \in D(A_M)$ we have

$$\begin{aligned}
&\|A_h\pi_h u - \pi_h A_M u\|_V = \max_{\|w_h\|_V=1} (A_h\pi_h u - \pi_h A_M u, w_h)_V = \\
&\max_{\|w_h\|_V=1} \Big\{ \sum_K \Big((\nabla \times (\pi_h \mathbf{E} - \mathbf{E}), \phi_h)_{0,K} - (\nabla \times (\pi_h \mathbf{H} - \mathbf{H}), \psi_h)_{0,K} \Big) \\
&+ \sum_{F \in \mathcal{F}_h^i} \Big((\mathbf{n}_F \times [\![\pi_h \mathbf{E} - \mathbf{E}]\!]_F, \alpha_K \phi_K + \alpha_{K_F} \phi_{K_F})_{0,F} \\
&\qquad - (\mathbf{n}_F \times [\![\pi_h \mathbf{H} - \mathbf{H}]\!]_F, \beta_K \psi_K + \beta_{K_F} \psi_{K_F})_{0,F} \\
&\qquad + \gamma_F (\mathbf{n}_F \times [\![\pi_h \mathbf{E} - \mathbf{E}]\!]_F, \mathbf{n}_F \times [\![\psi_h]\!]_F)_{0,F} \\
&\qquad + \delta_F (\mathbf{n}_F \times [\![\pi_h \mathbf{H} - \mathbf{H}]\!]_F, \mathbf{n}_F \times [\![\phi_h]\!]_F)_{0,F} \Big) \\
&+ \sum_{F \in \mathcal{F}_h^b} \Big(- (\mathbf{n}_F \times (\pi_h \mathbf{E} - \mathbf{E}), \phi_h)_{0,F} + 2\gamma_F (\mathbf{n}_F \times (\pi_h \mathbf{E} - \mathbf{E}), n \times \psi_h)_{0,F} \Big) \Big\}.
\end{aligned}$$

By applying the polynomial approximation results from Lemmas 4.33 and 4.34 and the discrete trace inequality (4.37), one proves the desired result.

□

The last statement can be generalized to powers of A_h.

Theorem 6.3. *Let $\mathcal{T}_\mathcal{H}$ be quasi-uniform and let $A_h \in \{A_h^{\mathrm{cf}}, A_h^{\mathrm{upw}}\}$. For $p = 0, \dots k$ and $v \in D(A_M^{p+1}) \cap H^{k+1}(\mathcal{T}_h)^6$ it holds*

$$\left\| A_h^{p+1} \pi_h v - \pi_h A_M^{p+1} v \right\|_V \leq C h^{k-p} \left| v \right|_{H^{k+1}(\mathcal{T}_h)^6}. \tag{6.9}$$

In particular, we have

$$\begin{aligned}
\lim_{h \to 0} \left\| A_h^{p+1} \pi_h v \right\|_V &\leq \left\| A_M^{p+1} v \right\|_V, \quad \text{for } p < k, \text{ and} \\
\lim_{h \to 0} \left\| A_h^{k+1} \pi_h v \right\|_V &\leq \left\| A_M^{k+1} v \right\|_V + C \left| v \right|_{H^{k+1}(\mathcal{T}_h)^6}.
\end{aligned}$$

Proof. We use an induction argument. Theorem 6.2 gives the basis ($p = 0$). Let us suppose that (6.9) holds for some $0 \leq p < k$. Then we have

$$\begin{aligned}
&\left\| A_h^{p+1} \pi_h v - \pi_h A_M^{p+1} v \right\|_V = \left\| A_h \pi_h (A_h^p \pi_h v) - \pi_h A_M (A_M^p v) \right\|_V \\
&\leq \left\| A_h \pi_h (A_h^p \pi_h v) - A_h \pi_h (A_M^p v) \right\|_V + \left\| A_h \pi_h (A_M^p v) - \pi_h A_M (A_M^p v) \right\|_V \\
&= \left\| A_h \left(A_h^p \pi_h v - \pi_h A_M^p v \right) \right\|_V + \left\| A_h \pi_h (A_M^p v) - \pi_h A_M (A_M^p v) \right\|_V \\
&\leq C h^{-1} \left\| A_h^p \pi_h v - \pi_h A_M^p v \right\|_V + C h^{k'} \left| A_M^p v \right|_{H^{k+1}(\mathcal{T}_h)^6}.
\end{aligned}$$

for every $k' \leq k$. Here we have used Theorem 6.1 for the last inequality. Now by using the assumption of induction we get

$$\left\| A_h^{p+1} \pi_h v - \pi_h A_M^{p+1} v \right\|_V \leq C h^{-1} C h^{k-p+1} \left| v \right|_{H^{k+1}(\mathcal{T}_h)^6} + C h^{k'} \left| A_M^p v \right|_{H^{k'+1}(\mathcal{T}_h)^6}$$

Since $H_0^1(\Omega)^6 \subset D(A_M)$ we have that $\left|A_M^p v\right|_{H^{k'+1}(\mathcal{T}_h)^6} \leq |v|_{H^{k'+p+1}(\mathcal{T}_h)^6}$. Setting $k' = k - p$ yields the claim. □

iv) From Theorems 4.42 and 4.48 we have the following space discretization errors

$$\left\| S_h^{\mathrm{cf}}(T)\pi_h u_0 - \pi_h S(T)u_0 \right\|_V \leq Ch^k T^{1/2} \left\| S(.)u_0 \right\|_{L^2(0,T;H^{k+1}(\mathcal{T}_h)^6)}, \tag{6.10}$$

$$\left\| S_h^{\mathrm{upw}}(T)\pi_h u_0 - \pi_h S(T)u_0 \right\|_V \leq Ch^{k+1/2} \left\| S(.)u_0 \right\|_{L^2(0,T;H^{k+1}(\mathcal{T}_h)^6)}. \tag{6.11}$$

Remark. For the sake of readability, all results in this chapter are stated with the assumption that the exact solution is spatially smooth enough to give the full convergence order in space, i. e. that $u(t) \in H^{k+1}(\mathcal{T}_h)^6$ for all $t \in [0,T]$. All results also hold for $u(t) \in H^{k'+1}(\mathcal{T}_h)^6$ for some $k' \leq k$ with the corresponding convergence order, cf. Section 4.3.

Remark. Theorems 6.1 and 6.3 require the assumption of quasi-uniformity of a mesh sequence. Nevertheless, since we wont use these theorems for our main results on implicit collocation methods for inhomogeneous problem in Section 6.3, this assumption is not required in that case. The assumption is required for the results on explicit Runge–Kuta methods in Section 6.1 and for the results on homogeneous problem in Section 6.2.1.

6.1 Explicit RK methods

As we have mentioned in the introduction, there has been some work on the analysis of full discretizations of Maxwell's equation which use explicit methods in time. In 2010, Burman, Ern and Fernandez [9] have proved an optimal convergence rate of $\mathcal{O}(h^{k+1/2}) + \mathcal{O}(\tau^s)$ for the upwind flux dG method combined with explicit Runge–Kutta methods with s stages for $s = 2, 3$. For the central flux dG method and the leap-frog time integration method, an error bound of order $\mathcal{O}(h^{k+1/2}) + \mathcal{O}(\tau^s)$ has been shown in [23].

In order to see the main difference between the analysis of explicit and implicit schemes, we present a convergence result for the explicit Euler method here. We will also see how the CFL condition plays a role in this case.

6.1.1 Explicit Euler method

Applying the explicit Euler method to (6.3) leads to the numerical scheme

$$u_h^n = u_h^n + \tau(-A_h u_h^n + f_h^n), \quad u_h^0 = \pi_h u_0. \tag{6.12}$$

Theorem 6.4. *Let $\mathcal{T}_\mathcal{H}$ be quasi-uniform, let $A_h = A_h^{\mathrm{upw}}$, and let u be the solution of (6.1) and u_h^n be the discrete solution defined with (6.12). Assume that $u \in C(0,T;H^{k+1}(\mathcal{T}_h)^6)$ and $u'' \in L^2(0,T,V)$. Then there is a constant $C_1 > 0$ such that for $\tau \leq C_1 h^2$, the error $e_h^n := u_h^n - \pi_h u(t_n)$ is bounded by*

$$\left\| e_h^N \right\|_V^2 + \tau \sum_{n=0}^{N} (A_h e_h^n, e_h^n)_V \leq C\Big(\tau^2 \int_0^T \left\| u''(t) \right\|_V^2 dt + h^{2k+1} \max_{t\in[0,T]} |u(t)|^2_{H^{k+1}(\mathcal{T}_h)^6}\Big),$$

where the constant C is independent of h and u.

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

$$u(t_{n+1}) = u(t_n) + \tau \partial_t u(t_n) + \int_{t_n}^{t_{n+1}} (t_{n+1} - t)\partial_t^2 u(t) dt.$$

For $\eta^{n+1} := \int_{t_n}^{t_{n+1}} (t_{n+1} - t)\partial_t^2 u(t) dt$ it holds

$$\left\|\eta^{n+1}\right\|_V \leq \tau \int_{t_n}^{t_{n+1}} \left\|u''(t)\right\|_V dt = \mathcal{O}(\tau^2).$$

Writing $\partial_t u = -Au + f$ and taking the L^2-projection of the whole equation yields

$$\pi_h u(t_{n+1}) = \pi_h u(t_n) + \tau(-A_h u(t_n) + f_h^n) + \pi_h \eta^{n+1}. \tag{6.13}$$

Subtracting (6.13) from (6.12) gives the error recursion

$$e_h^{n+1} = e_h^n - \tau A_h e_h^n + \tau A_h e_\pi^n - \pi_h \eta^{n+1}.s \tag{6.14}$$

After taking the V-inner product with e_h^n and using

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

we end up with

$$\left\|e_h^{n+1}\right\|_V^2 - \|e_h^n\|_V^2 + 2\tau(A_h e_h^n, e_h^n)_V = \left\|e_h^{n+1} - e_h^n\right\|_V^2 + 2\tau(A_h e_\pi^n, e_h^n)_V - 2(\pi_h \eta^{n+1}, e_h^n)_V.$$

For the second term on the right-hand side we use Lemma 4.47 with $\gamma = 1$ to get

$$\begin{aligned} &\left\|e_h^{n+1}\right\|_V^2 - \|e_h^n\|_V^2 + \tau(A_h e_h^n, e_h^n)_V \\ &\qquad \leq \left\|e_h^{n+1} - e_h^n\right\|_V^2 + C\tau h^{2k+1} \left|u(t_n)\right|_{H^{k+1}(\mathcal{T}_h)^6}^2 - 2(\pi_h \eta^{n+1}, e_h^n)_V. \end{aligned} \tag{6.15}$$

For the last term we use Young's inequality and the stability ofthe L^2-projection

$$2(\pi_h \eta^{n+1}, e_h^n)_V \leq \tau \left(\left\| \frac{\eta^{n+1}}{\tau} \right\|_V^2 + \|e_h^n\|_V^2 \right).$$

It remains to bound the first term on the right-hand side of (6.15). From (6.14) we conclude

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

By using Theorem 6.1, we can bound the first term as

$$\|\tau A_h e_h^n\|_V^2 \leq C\tau^2 h^{-2} \left\|e_h^n\right\|_V^2.$$

Since $A_h e_\pi = A_h(u - \pi_h u) = \pi_h A u - A_h \pi_h$, Theorem 6.2 implies

$$\|\tau A_h e_\pi^n\|_V^2 \leq C\tau^2 h^{2k} \, |u(t_n)|^2_{H^{k+1}(\mathcal{T}_h)^6} \, .$$

Gathering all inequalities we have

$$\left\|e_h^{n+1}\right\|_V^2 - \|e_h^n\|_V^2 + \tau(A_h e_h^n, e_h^n)_V \leq$$
$$C\tau(1+\tau h^{-2}) \, \|e_h^n\|_V^2 + C\tau(h^{2k+1} + \tau h^{2k}) \, |u(t_n)|^2_{H^{k+1}(\mathcal{T}_h)^6} + C\tau \left\| \frac{\eta^{n+1}}{\tau} \right\|_V^2 .$$

After taking the CFL condition into account, summing over n and applying the discrete Gronwall lemma 2.7, the theorem is proved. □

For $A_h = A_h^{\mathrm{cf}}$ the proof differs only at one point. We use Lemma 4.41 instead of Lemma 4.47 to get

$$\left\|e_h^{n+1}\right\|_V^2 - \|e_h^n\|_V^2 \leq \left\|e_h^{n+1} - e_h^n\right\|_V^2$$
$$+C\tau h^{2k} \, |u(t_n)|^2_{H^{k+1}(\mathcal{T}_h)^6} + C\tau \, \|e_h^n\|_V^2 - 2(\pi_h \eta^{n+1}, e_h^n)_V$$

Theorem 6.5. *Let $A_h = A_h^{\mathrm{cf}}$ and let the other assumptions of Theorem 6.4 be satisfied. Then, the following bound holds*

$$\left\|e_h^N\right\|_V^2 \leq C\Big(\tau^2 \int_0^T \left\|u''(t)\right\|_V^2 dt + h^{2k} \max_{t\in[0,T]} |u(t)|^2_{H^{k+1}(\mathcal{T}_h)^6}\Big),$$

where the constant C is independent of h and u.

6.2 Implicit Runge–Kutta methods: known results and application to Maxwell's equations

In Subsection 5.2 we have presented already known convergence results for implicit Runge–Kutta methods applied to the Cauchy problem

$$\partial_t u(t) + Au(t) = f(t), \quad u(0) = u_0 \tag{6.16}$$

in a Banach space $(X, \|.\|_X)$. In this section we apply these results to the space discrete problem

$$\partial_t u_h + A_h u_h = \pi_h f, \quad u_h(0) = \pi_h u_0, \tag{6.17}$$

in order to obtain the error of a fully discrete scheme. We show that the results we get are applicable to the case of Maxwell's equations.

Let $X_h \subset X$ be a finite dimensional space with norm $\|.\|_h = \|.\|_X$ and $\pi_h : X \to X_h$ be the orthogonal projection onto X_h. We suppose that there exists a dense subspace $Y \subset X$ with norm $\|.\|_Y$ such that

$$\|\pi_h v - v\|_X \leq \delta_h \, \|v\|_Y \, , \quad \forall v \in Y, \tag{6.18}$$

where $\delta_h \to 0$ for $h \to 0$.

6.2.1 The homogeneous case

The solution of the homogeneous problem

$$\partial_t u_h + A_h u_h = 0, \quad u_h(0) = \pi_h u_0 \tag{6.19}$$

is given by $u_h(t) = S_h(t)\pi_h u_0$, where $S_h(t) = e^{-tA_h}$ is a strongly continuous semigroup on X_h generated by A_h.

Assumption 6.6. *$-A$ generates a strongly continuous semigroup of type (C_A, ω) and, in addition, discrete operators $-A_h$ generate semigroups which satisfy*

$$\|S_h(t)\|_{X \leftarrow X} \leq C_A e^{\omega t}, \quad \forall\, t > 0 \text{ and } \forall\, h.$$

Let r be an A-stable rational function which approximates e^z up to order p. The fully discrete solution is given by

$$u_{h,n} = r(-\tau A_h)^n \pi_h u_0, \quad n = 0, 1, \ldots \tag{6.20}$$

With the help of the triangle inequality the full discretization error can be estimated by

$$\begin{aligned}\left\|r^N(-\tau A_h)\pi_h u_0 - \pi_h S(T)u_0\right\|_X \leq{}& \left\|r^N(-\tau A_h)\pi_h u_0 - S_h(T)\pi_h u_0\right\|_X \\ &+ \left\|S_h(T)\pi_h u_0 - \pi_h S(T)u_0\right\|_X,\end{aligned}$$

where $T = N\tau$. For the first term we use Theorem 5.21:

$$\left\|r^N(-\tau A_h)\pi_h u_0 - S_h(T)\pi_h u_0\right\|_X \leq C_A C_1 T \tau^p e^{\omega\kappa T} \left\|A_h^{p+1}\pi_h u_0\right\|_X. \tag{6.21}$$

Here it remains to check that $\left\|A_h^{p+1}\pi_h u_0\right\|_X$ is bounded for $h \to 0$. The second term is the space semidiscretization error.

Application to Maxwell's equations

In order to bound $\left\|A_h^{p+1}\pi_h u_0\right\|_X$ we use Theorem 6.3 (the quasi-uniformity assumption is required). For the space semidiscretization error we use (6.10) and (6.11) for the dG scheme with central and upwind flux, respectively. We have

$$\begin{aligned}\left\|r^n(-\tau A_h^{\mathrm{cf}})\pi_h u_0 - \pi_h S(T)u_0\right\|_V \leq{}& CT\tau^p \left\|A_M^{p+1}u_0\right\|_V \\ &+ CT^{1/2}h^k \left\|S(.)u_0\right\|_{L^2(0,T;H^{k+1}(\mathcal{T}_h)^6)}, \\ \left\|r^n(-\tau A_h^{\mathrm{upw}})\pi_h u_0 - \pi_h S(T)u_0\right\|_V \leq{}& CT\tau^p \left\|A_M^{p+1}u_0\right\|_V \\ &+ Ch^{k+1/2} \left\|S(.)u_0\right\|_{L^2(0,T;H^{k+1}(\mathcal{T}_h)^6)}\end{aligned}$$

for $p \leq k$ (actually for $p = k$ we have $\left\|A_M^{k+1}v\right\|_V + C\left|v\right|_{H^{k+1}(\mathcal{T}_h)^6}$ instead of $\left\|A_M^{p+1}u_0\right\|_V$, see Theorem 6.3).

6.2.2 The inhomogeneous case

Assumption 6.7. *$-A$ generates a bounded semigroup, i. e. it satisfies Assumption 5.22, and discrete operators $-A_h$ generate semigroups which are bounded uniformly in h*

$$\|S_h(t)\|_X \le C_A, \quad \forall\, t > 0 \text{ and } \forall\, h.$$

Applying the time discretization scheme (5.23) to this problem we get the **fully discrete scheme**

$$\begin{aligned} u_{h,n+1} &= S_{\tau h} u_{h,n} + \tau (Q_{\tau h} \pi_h f)(t_n) \quad \text{for} \quad n = 0, 1, \dots, \\ u_{h,0} &= \pi_h u_0, \end{aligned} \tag{6.22}$$

where

$$S_{\tau h} v_h = r(-\tau A_h) v_h, \quad (Q_{\tau h} f_h)(t) = \sum_{j=1}^{s} q_j(-\tau A_h) f_h(t_n + \tau c_j).$$

We again assume that the rational functions r, $q_1, \dots, q_s$ are bounded for $\operatorname{Re} z \le 0$, see Assumption 5.23.

Assumption 6.8. *$S_{\tau h}$ is uniformly stable, i. e. $\|S_{\tau h}^n\|_{X \leftarrow X} \le C$ for all $n \in \mathbb{N}$ and all h.*

To investigate the error of the fully discrete scheme (6.22) we use Corollary 5.28 for Gauss methods and Theorem 5.29 for Radau methods. In what follows we consider Gauss methods only, since the results for Radau methods follow completely analogously. From Corollary 5.28 we have the following bound for the s-stage Gauss method

$$\|u_N - u(T)\|_X \le C\, \tau^{s+1} \int_0^T \left\| A u^{(s+1)}(\theta) \right\|_X + \left\| u^{(s+2)}(\theta) \right\|_X d\theta. \tag{6.23}$$

To estimate the full discretization error, one can try to proceed as in the homogeneous case. The triangle inequality gives

$$\|u_{h,N} - \pi_h u(T)\|_X \le \|u_{h,n} - u_h(T)\|_X + \|u_h(T) - \pi_h u(T)\|_X .$$

The second term is the space semidiscretization error, hence we have to bound the first term. By applying Corollary 5.28 to the semidiscrete problem (6.17), we get

$$\|u_{h,N} - u_h(T)\|_X \le C\, \tau^{s+1} \int_0^T \left\| A_h u_h^{(s+1)}(\theta) \right\|_X + \left\| u_h^{(s+2)}(\theta) \right\|_X d\theta.$$

It is not clear if we can bound it by using $\left\| A u^{(s+1)}(\theta) \right\|_X$ and $\left\| u^{(s+2)}(\theta) \right\|_X$ only. Therefore, we need to use a different approach. In what follows we present two different ways to estimate the full discretization error.

Approach 1: L^2-projection

We suppose that the discrete operator $A_h : X_h \to X_h$ approximates A in the sense

$$\|A_h \pi_h v - \pi_h A v\|_X = \varepsilon_h \|v\|_Y , \quad \forall v \in D(A) \cap Y, \tag{6.24}$$

where $\varepsilon_h \to 0$ for $h \to 0$. Then the following result holds, cf. [4, Theorem 4].

Theorem 6.9. *Assume that Assumptions 6.7 and 6.8 are fulfilled that the time integration scheme (6.22) is a s-stage Gauss collocation method. Under the appropriate regularity assumptions it holds*

$$\|u_{h,N} - \pi_h u(T)\|_X \le C\varepsilon_h T \sup_{\theta \le T} \|u(\theta)\|_Y + C\tau^{s+1} \left\{ \int_0^T \left(\left\| A u^{(s+1)}(\theta) \right\|_X + \left\| u^{(s+2)}(\theta) \right\|_X \right) d\theta \right\}.$$

Proof. The solution of the continuous problem satisfies

$$\pi_h u'(t) + A_h \pi_h u(t) = \tilde{f}_h(t) := \pi_h f(t) + (\pi_h A u(t) + A_h \pi_h u(t)). \tag{6.25}$$

From (6.24) follows

$$\left\| \tilde{f}_h(t) - \pi_h f(t) \right\|_X \le \varepsilon_h \|u(t)\|_Y .$$

By applying the time integration scheme (6.22) to (6.25) we get

$$\begin{aligned} \tilde{u}_{h,n+1} &= S_{\tau h} \tilde{u}_{h,n} + \tau (Q_{\tau h} \tilde{f}_h)(t_n), \\ \tilde{u}_{h,0} &= \pi_h u_0. \end{aligned} \tag{6.26}$$

Corollary 5.28 yields the error bound:

$$\|\tilde{u}_{h,N} - \pi_h u(T)\|_X \le C\tau^{s+1} \int_0^T \left(\left\| A_h \pi_h u^{(s+1)}(\theta) \right\|_X + \left\| \pi_h u^{(s+2)}(\theta) \right\|_X \right) d\theta.$$

Since $\varepsilon_h \to 0$, there exists a constant C_1 such that for $h \le 1$, $\varepsilon_h \|v\|_Y \le C_1 \|Av\|_X$ holds. Therefore, we have

$$\|A_h \pi_h u\|_X \le C \|Au\|_X .$$

By using $\|\pi_h u\|_X \le C \|u\|_X$, it follows that

$$\|\tilde{u}_{h,N} - \pi_h u(T)\|_X \le C\tau^{s+1} \int_0^T \left(\left\| A u^{(s+1)}(\theta) \right\|_X + \left\| u^{(s+2)}(\theta) \right\|_X \right) d\theta. \tag{6.27}$$

On the other hand, by comparing (6.22) and (6.26) and using that $S_{\tau h}$ is uniformly stable, we find that

$$\begin{aligned} \|\tilde{u}_{h,N} - u_{h,N}\|_X &\le C \left\{ \tau \sum_{l=0}^{N-1} \sum_{j=1}^{s} \left\| q_j(-\tau A_h)(\tilde{f}_h(t_l + \tau c_j) - \pi_h f(t_l + \tau c_j)) \right\|_X \right\} \\ &\le C\varepsilon_h T \sup_{\theta \le T} \|u(\theta)\|_Y . \end{aligned}$$

By using the triangle inequality

$$\|u_{h,N} - \pi_h u(T)\|_X \leq \|u_{h,N} - \tilde{u}_{h,N}\|_X + \|\tilde{u}_{h,N} - \pi_h u(T)\|_X ,$$

the theorem is proved. □

Approach 2: Elliptic projection

The elliptic projection $Q_h : Y \to X_h$ is defined by

$$Q_h v := (I + A_h)^{-1} \pi_h (I + A) v \tag{6.28}$$

and it exists for $v \in D(A)$ since $-A_h$ generates a bounded semigroup. We assume the approximation property

$$\|Q_h v - \pi_h v\|_X \leq \varepsilon_h \|v\|_Y , \quad \forall v \in D(A) \cap Y, \tag{6.29}$$

where $\varepsilon_h \to 0$ as $h \to 0$. Then the following result holds for Gauss methods, cf. [4, Theorem 4] for Radau methods.

Theorem 6.10. *Assume that Assumptions 6.7 and 6.8 are fulfilled and that the time integration scheme (6.22) is a s-stage Gauss collocation method. Under appropriate regularity assumptions it holds*

$$\|u_{h,N} - \pi_h u(T)\|_X \leq C\varepsilon_h \left\{ (1+T) \sup_{\theta \leq T} \|u(\theta)\|_Y + T \sup_{\theta \leq T} \left\|u'(\theta)\right\|_Y \right\} + C\tau^{s+1} \left\{ \int_0^T \left(\left\|(I+A)u^{(s+1)}(\theta)\right\|_X + \left\|(I+A)u^{(s+2)}(\theta)\right\|_X \right) d\theta \right\}.$$

Proof. From the definition of Q_h we get that the solution of the continuous problem satisfies

$$Q_h u'(t) + A_h Q_h u(t) = \tilde{f}_h(t) := \pi_h f(t) + (Q_h - \pi_h)(u'(t) - u(t)), \tag{6.30}$$

where

$$\left\|\tilde{f}_h(t) - \pi_h f(t)\right\|_X \leq \varepsilon_h \left\|u'(t) - u(t)\right\|_Y .$$

By applying the time integration scheme (6.22) to (6.30) we get

$$\begin{aligned} \tilde{u}_{h,n+1} &= S_{\tau h} \tilde{u}_{h,n} + \tau (Q_{\tau h} \tilde{f}_h)(t_n), \\ \tilde{u}_{h,0} &= Q_h u_0. \end{aligned} \tag{6.31}$$

Corollary 5.28 yields the error bound:

$$\|\tilde{u}_{h,N} - Q_h u(T)\|_X \leq C\tau^{s+1} \int_0^T (\left\|A_h Q_h u^{(s+1)}(\theta)\right\|_X + \left\|Q_h u^{(s+2)}(\theta)\right\|_X) d\theta.$$

Form (2.5) we have that $\|(I+A_h)^{-1}\|_{X\leftarrow X} \leq C_A$, which implies

$$\|A_h Q_h u\|_X = \left\|A_h(I+A_h)^{-1}\pi_h(I+A)u\right\|_X \leq C\,\|(I+A)u\|_X\,,$$
$$\|Q_h u\|_X \leq C\,\|(I+A)u\|_X\,.$$

It follows that

$$\|\tilde{u}_{h,N} - Q_h u(T)\|_X \leq C\tau^{s+1}\int_0^T (\left\|(I+A)u^{(s+1)}(\theta)\right\|_X + \left\|(I+A)u^{(s+2)}(\theta)\right\|_X)d\theta.$$

On the other hand, by comparing (6.22) and (6.31) and using that $S_{\tau h}$ is uniformly stable, we find that

$$\begin{aligned}
&\|\tilde{u}_{h,N} - u_{h,N}\|_X \\
&\leq C\left\{\|\pi_h u_0 - Q_h u_0\|_X + \tau\sum_{l=0}^{N-1}\sum_{j=1}^{s}\left\|q_j(-\tau A_h)(\tilde{f}_h(t_l+\tau c_j) - \pi_h f(t_l+\tau c_j))\right\|_X\right\} \\
&\leq C\varepsilon_h\left\{(1+T)\sup_{\theta\leq T}\|u(\theta)\|_Y + T\sup_{\theta\leq T}\left\|u'(\theta)\right\|_Y\right\}.
\end{aligned}$$

The triangle inequality gives

$$\|u_{h,N} - \pi_h u(T)\|_X \leq \|u_{h,N} - \tilde{u}_{h,N}\|_X + \|\tilde{u}_{h,N} - Q_h u(T)\|_X + \|Q_h u(T) - \pi_h u(T)\|_X\,.$$

Inserting the bound (6.29) for the last term proves the claim. □

Application to Maxwell's equations

In the case of Maxwell's equations, Assumption 6.7 is fulfilled for both $A_h = A_h^{\mathrm{cf}}$ and $A_h = A_h^{\mathrm{upw}}$ with $C_A = 1$, cf. (6.5). According to Theorem 5.30, Assumption 6.8 is satisfied too.

Approach 1: L^2-projection

Due to Theorem 6.2 we have that (6.24) is satisfied with $Y = H^{k+1}(\mathcal{T}_h)^6$ and $\varepsilon_h = \mathcal{O}(h^k)$. Now Theorem 6.9 implies the following error estimate for the fully discrete scheme (6.22) with a s-stage Gauss collocation method

$$\begin{aligned}
\|u_{h,N} - \pi_h u(T)\|_V \leq\; & C\,h^k\,T\sup_{\theta\leq T}|u(s)|_{H^{k+1}(\mathcal{T}_h)^6} + \\
& C\tau^{s+1}\left\{\int_0^T\left(\left\|A_M u^{(s+1)}(\theta)\right\|_V + \left\|u^{(s+2)}(\theta)\right\|_V\right)d\theta\right\}.
\end{aligned} \tag{6.32}$$

The result holds for dG schemes with both central and upwind fluxes since Theorem 6.2 gives the same estimate for both A_h^{cf} and A_h^{upw}. For the upwind flux this result is only suboptimal.

Approach 2: Elliptic projection

From $(I + A_h)(Q_h v - \pi_h v) = \pi_h A_M v - A_h \pi_h v$ we get

$$Q_h v - \pi_h v = (I + A_h)^{-1}(\pi_h A_M v - A_h \pi_h v)$$

and therefore we have

$$\|Q_h v - \pi_h v\|_X \leq C \|\pi_h A_M v - A_h \pi_h v\|_X .$$

By Theorem 6.2, we have that (6.29) is satisfied, again with $Y = H^{k+1}(\mathcal{T}_h)^6$ and $\varepsilon_h = \mathcal{O}(h^k)$. Now, by using Theorem 6.10 we get the following estimate

$$\|u_{h,N} - \pi_h u(T)\|_V \leq Ch^k \left\{ (1+T) \sup_{\theta \leq T} |u(\theta)|_{H^{k+1}(\mathcal{T}_h)^6} + T \sup_{\theta \leq T} \left|u'(\theta)\right|_{H^{k+1}(\mathcal{T}_h)^6} \right\} + C\tau^{s+1} \left\{ \int_0^T \left(\left\| (I + A_M) u^{(s+1)}(\theta) \right\|_V + \left\| (I + A_M) u^{(s+2)}(\theta) \right\|_V \right) d\theta \right\},$$

which is valid for dG method with both central and upwind flux. A disadvantage, in comparison to the L^2-projection approach, is obvious since more regularity on the solution is required.

For the upwind flux method, we can improve the convergence order in space at the expense of regularity. If we suppose

$$\|S_M(t) v\|_Y \leq C \|v\|_Y \tag{6.33}$$

then from (6.11) we get

$$\|S_h(T)\pi_h v - \pi_h S_M(T) v\|_V \leq CT^{1/2} h^{k+1/2} \|v\|_{H^{k+1}(\mathcal{T}_h)^6} . \tag{6.34}$$

Remark. Since $\|S_M(t)v\|_V = \|v\|_V$, (6.33) is equivalent to e^{-tA_M} commuting with the spatial derivatives which is again equivalent to $-A_M$ commuting with the spatial derivatives.

From the resolvent integral representation (see [3, Proposition 2.26]) we have

$$R(\lambda, A_h)\pi_h v - \pi_h R(\lambda, A_M) v = \int_0^\infty e^{-\lambda\theta} \left(S_h(\theta)\pi_h v - \pi_h S_M(\theta) v \right) d\theta.$$

It follows that

$$\|R(\lambda, A_h)\pi_h v - \pi_h R(\lambda, A_M) v\|_X \leq \frac{C}{\lambda^2} T^{1/2} h^{k+1/2} \|v\|_{H^{k+1}(\mathcal{T}_h)^6} .$$

This implies

$$\begin{aligned} \|Q_h v - \pi_h v\|_X &= \left\| \left((I + A_h)^{-1}\pi_h - \pi_h (I + A_M)^{-1} \right) (I + A_M) v \right\|_X \\ &\leq CT^{1/2} h^{k+1/2} \|(I + A_M) v\|_{H^{k+1}(\mathcal{T}_h)^6} , \end{aligned}$$

i. e. the estimate (6.29) is satisfied with $\varepsilon_h = \mathcal{O}(h^{k+1/2})$ but in the stronger norm $\|.\|_Y$. According to Theorem 6.10 we have

$$\begin{aligned}\|u_{h,N} - \pi_h u(T)\|_V \le CT^{1/2}h^{k+1/2}\Big\{&(1+T)\sup_{\theta\le T}\|(I+A_M)u(\theta)\|_{H^{k+1}(\mathcal{T}_h)^6}\\ &+T\sup_{\theta\le T}\big\|(I+A_M)u'(\theta)\big\|_{H^{k+1}(\mathcal{T}_h)^6}\Big\}\\ +C\tau^{s+1}\Big\{\int_0^T&\Big(\Big\|(I+A_M)u^{(s+1)}(\theta)\Big\|_V+\Big\|(I+A_M)u^{(s+2)}(\theta)\Big\|_V\Big)\,d\theta\Big\}.\end{aligned}$$

In both approaches, we get worse estimates than the ones we were able to prove by using an energy technique. This main result will be presented next.

6.3 Implicit Runge–Kutta methods: our result

This section is an extended version of what can be found in [38, Section 5]. The main difference is that we consider both central and upwind flux dG schemes here, while in [38] only the upwind flux is analyzed.

6.3.1 Error of full discretization for the implicit Euler method

Again we consider the implicit Euler method for the time integration of (6.3) separately:

$$u_h^{n+1} = u_h^n + \tau(-A_h u_h^{n+1} + f_h^{n+1}), \quad u_h^0 = \pi_h u_0. \tag{6.35}$$

Theorem 6.11. *Let $A_h = A_h^{\mathrm{upw}}$ and let u be a solution of (6.1) and u_h^n the discrete solution defined with (6.35). Assume that $u \in C\big(0,T;H^{k+1}(\mathcal{T}_h)^6\big)$ and $u'' \in L^2(0,T,V)$. Then, for τ sufficiently small (depending on T only), the error of the implicit Euler method is bounded by*

$$\begin{aligned}&\|e_h^N\|_V^2 + \tau\sum_{n=0}^{N}(A_h e_h^{n+1}, e_h^{n+1})_V\\ &\qquad\le C(T+1)\Big(\tau^2\int_0^T\|u''(t)\|_V^2\,dt + h^{2k+1}\max_{t\in[0,T]}|u(t)|^2_{H^{k+1}(\mathcal{T}_h)^6}\Big),\end{aligned}$$

where the constant C is independent of h and u.

Proof. The exact solution satisfies

$$\tilde{u}^{n+1} = \tilde{u}^n + \tau(-A\tilde{u}^{n+1} + f^{n+1}) + \delta^{n+1},$$

where δ^{n+1} is given in (5.15) for $p=1$. Projecting onto L^2 and subtracting from (6.35) yields the error recursion

$$e_h^{n+1} = e_h^n - \tau A_h(e_h^{n+1} - e_\pi^{n+1}) - \pi_h\delta^{n+1}. \tag{6.36}$$

After taking the V-inner product with e_h^{n+1} and using Lemma 4.47 with $\gamma = 1$, we obtain

$$\begin{aligned}(e_h^{n+1} - e_h^n, e_h^{n+1})_V + \frac{1}{2}\tau(A_h e_h^{n+1}, e_h^{n+1})_V \\ \leq C\tau h^{2k+1} \left|u^{n+1}\right|^2_{H^{k+1}(\mathcal{T}_h)^6} - (\pi_h \delta^{n+1}, e_h^{n+1})_V.\end{aligned} \tag{6.37}$$

We sum from 0 to $N-1$ and use the following representation of the left-hand side

$$\begin{aligned}\sum_{n=0}^{N-1}(e_h^{n+1} - e_h^n, e_h^{n+1})_L &= \frac{1}{2}\left\|e_h^N\right\|_V^2 + \frac{1}{2}\left\|e_h^N\right\|_V^2 - (e_h^{N-1}, e_h^N)_L + \frac{1}{2}\left\|e_h^{N-1}\right\|_V^2 \\ &+ \frac{1}{2}\left\|e_h^{N-1}\right\|_V^2 - (e_h^{N-2}, e_h^{N-1})_L + \frac{1}{2}\left\|e_h^{N-2}\right\|_V^2 + \dots \\ &+ \frac{1}{2}\left\|e_h^1\right\|_V^2 - (e_h^0, e_h^1)_L + \frac{1}{2}\left\|e_h^0\right\|_V^2 + \frac{1}{2}\left\|e_h^0\right\|_V^2 \\ &\geq \frac{1}{2}\left\|e_h^N\right\|_V^2 + \frac{1}{2}\left\|e_h^0\right\|_V^2.\end{aligned}$$

For the right-hand side of (6.37) we have

$$\sum_{n=0}^{N-1}(\pi_h \delta^{n+1}, e_h^{n+1})_V \leq \frac{\tau}{2}\sum_{n=0}^{N-1}\left((T+1)\left\|\frac{\delta^{n+1}}{\tau}\right\|_V^2 + \frac{1}{T+1}\left\|e_h^{n+1}\right\|_V^2\right).$$

The result follows by a discrete Gronwall inequality from Corollary 2.9. □

For $A_h = A_h^{\mathrm{cf}}$ we use Lemma 4.41 with $\gamma = T+1$ instead of Lemma 4.47 and get

$$(e_h^{n+1} - e_h^n, e_h^{n+1})_V \leq (T+1)\tau h^{2k}\left|u^{n+1}\right|^2_{H^{k+1}(\mathcal{T}_h)^6} + \frac{C}{1+T}\tau\left\|e_h^{n+1}\right\|_V^2 - (\pi_h\delta^{n+1}, e_h^{n+1})_V$$

instead of (6.37). The rest of the proof stays the same.

Theorem 6.12. *Let $A_h = A_h^{\mathrm{cf}}$ and let the other assumptions of Theorem 6.11 be satisfied. Then it holds*

$$\left\|e_h^N\right\|_V^2 \leq C(T+1)\Big(\tau^2\int_0^T \left\|u''(t)\right\|_V^2 dt + Th^{2k}\max_{t\in[0,T]}|u(t)|^2_{H^{k+1}(\mathcal{T}_h)^6}\Big),$$

where the constant C is independent of h and u.

6.3.2 Error of full discretization for higher order Runge–Kutta methods

An implicit s-stage Runge–Kutta method applied to (6.3) yields the approximations

$$\begin{aligned}\dot{U}_h^{ni} + A_h U_h^{ni} &= f_h^{ni}, \\ U_h^{ni} &= u_h^n + \tau\sum_{j=1}^s a_{ij}\dot{U}_h^{nj}, \\ u_h^{n+1} &= u_h^n + \tau\sum_{i=1}^s b_i \dot{U}_h^{ni}.\end{aligned} \tag{6.38}$$

For A-stable collocation methods such as Gauß- and Radau methods, a unique solution of the linear system defining the interior Runge-Kutta approximations U_h^{ni}, $i = 1, \dots, s$ exists because A_h is dissipative.

Defects

Inserting the exact solution $\tilde{u}^n = u(t_n)$, $\tilde{U}^{ni} = u(t_n + c_i\tau)$, and $\dot{\tilde{U}}^{ni} = u'(t_n + c_i\tau)$ into the numerical scheme yields

$$\begin{aligned} \pi_h \dot{\tilde{U}}^{ni} + A_h \tilde{U}^{ni} &= f_h^{ni}, \\ \tilde{U}^{ni} &= \tilde{u}^n + \tau \sum_{j=1}^{s} a_{ij} \dot{\tilde{U}}^{nj} + \Delta^{ni}, \\ \tilde{u}^{n+1} &= \tilde{u}^n + \tau \sum_{i=1}^{s} b_i \dot{\tilde{U}}^{ni} + \delta^{n+1}, \end{aligned} \tag{6.39}$$

where the defects are given in Theorem 5.18. We define errors as

$$\begin{aligned} e_h^n &:= u_h^n - \pi_h \tilde{u}^n, & e_\pi^n &:= \tilde{u}^n - \pi_h \tilde{u}^n, \\ E_h^{ni} &:= U_h^{ni} - \pi_h \tilde{U}^{ni}, & E_\pi^{ni} &:= \tilde{U}^{ni} - \pi_h \tilde{U}^{ni}, \\ \dot{E}_h^{ni} &:= \dot{U}_h^{ni} - \pi_h \dot{\tilde{U}}^{ni}. \end{aligned} \tag{6.40}$$

Subtracting (6.39) from (6.38) yields

$$\dot{E}_h^{ni} + A_h E_h^{ni} = A_h E_\pi^{ni}, \tag{6.41a}$$

$$E_h^{ni} = e_h^n + \tau \sum_{j=1}^{s} a_{ij} \dot{E}_h^{nj} - \pi_h \Delta^{ni}, \tag{6.41b}$$

$$e_h^{n+1} = e_h^n + \tau \sum_{i=1}^{s} b_i \dot{E}_h^{ni} - \pi_h \delta^{n+1}. \tag{6.41c}$$

For

$$\Delta^n = \begin{pmatrix} \Delta^{n1} \\ \vdots \\ \Delta^{ns} \end{pmatrix}, \quad E_h^n = \begin{pmatrix} E_h^{n1} \\ \vdots \\ E_h^{ns} \end{pmatrix}, \quad E_\pi^n = \begin{pmatrix} E_\pi^{n1} \\ \vdots \\ E_\pi^{ns}, \end{pmatrix}, \quad \dot{E}_h^n = \begin{pmatrix} \dot{E}_h^{n1} \\ \vdots \\ \dot{E}_h^{ns} \end{pmatrix},$$

we can write (6.41) in compact form as

$$\dot{E}_h^n + (I \otimes A_h) E_h^n = (I \otimes A_h) E_\pi^n \tag{6.42a}$$

$$E_h^n = \mathbb{1} \otimes e_h^n + \tau(\mathcal{A} \otimes I) \dot{E}_h^n - \pi_h \Delta^n \tag{6.42b}$$

$$e_h^{n+1} = e_h^n + \tau(b^T \otimes I) \dot{E}_h^n - \pi_h \delta^{n+1}. \tag{6.42c}$$

Energy technique

To analyse the error we use the same technique as in the continuous case. Taking the inner product of e_h^{n+1} with itself using (6.41c) yields

$$\left\|e_h^{n+1}\right\|_V^2 = \left\|e_h^n + \tau \sum_{i=1}^s b_i \dot{E}_h^{ni}\right\|_V^2 - 2(\pi_h \delta^{n+1}, e_h^n + \tau \sum_{i=1}^s b_i \dot{E}_h^{ni})_V + \left\|\pi_h \delta^{n+1}\right\|_V^2. \quad (6.43)$$

Theorem 6.13. *Let $A_h = A_h^{\mathrm{upw}}$. The errors (6.40) of the Runge–Kutta method (6.38) applied to (6.3) satisfy*

$$\begin{aligned}
&\left\|e_h^{n+1}\right\|_V^2 - \|e_h^n\|_V^2 + \tau \sum_{i=1}^s b_i (A_h E_h^{ni}, E_h^{ni})_V \\
&\qquad \leq \frac{C}{T+1}\tau \left(\|e_h^n\|_V^2 + \sum_{i=1}^s \left\|E_h^{ni}\right\|_V^2 \right) + C\tau h^{2k+1} \sum_{i=1}^s b_i \left|\tilde{U}^{ni}\right|_{H^{k+1}(\mathcal{T}_h)^6}^2 \\
&\qquad + C(T+1)\tau \left(\sum_{i=1}^s \left(\left\|\pi_h \Delta^{ni}\right\|_V^2 + \left|\Delta^{ni}\right|_{H^1(\mathcal{T}_h)^6}^2 \right) + \left\| \frac{1}{\tau} \pi_h \delta^{n+1} \right\|_V^2 \right).
\end{aligned}$$

Here the constant $C = C(\mathcal{Q}, b, k, \rho)$ is independent of h and u.

Proof. We estimate each of the three terms in (6.43) separately. The second and the third term can be handled completely analogously to the continuous case. For the second term, (5.56) now reads

$$\begin{aligned}
(\pi_h \delta^{n+1}, e_h^n + \tau \sum_{i=1}^s b_i \dot{E}_h^{ni})_V \leq &\frac{C}{\gamma}\tau \Big(\|e_h^n\|_V^2 + \sum_{j=1}^s \left\|E_h^{nj}\right\|_V^2 + \sum_{j=1}^s \left\|\pi_h \Delta^{nj}\right\|_V^2 \Big) \\
&+ \gamma\tau \left\| \frac{1}{\tau} \pi_h \delta^{n+1} \right\|_V^2.
\end{aligned} \quad (6.44)$$

For the first term of (6.43) we have to work harder. The reason is the additional term in (6.41a) and the fact that A_h is not skew-adjoint. However, the derivation of the estimate (5.51) remains true, so that we can start from

$$\left\| e_h^n + \tau \sum_{i=1}^s b_i \dot{E}_h^{ni} \right\|_V^2 \leq \|e_h^n\|_V^2 + 2\tau \sum_{i=1}^s b_i (\dot{E}_h^{ni}, E_h^{ni} + \pi_h \Delta^{ni})_V.$$

Eliminating $\dot{E}_h^{ni}$ by (6.41a) yields

$$\begin{aligned}
(\dot{E}_h^{ni}, E_h^{ni} + \pi_h \Delta^{ni})_V = &(A_h E_\pi^{ni}, E_h^{ni})_V - (A_h E_h^{ni}, E_h^{ni})_V \\
&+ (A_h E_\pi^{ni}, \pi_h \Delta^{ni})_V - (A_h E_h^{ni}, \pi_h \Delta^{ni})_V.
\end{aligned}$$

For the first two terms we have by Lemma 4.47 with $\gamma = 1$

$$\begin{aligned}(A_h E_\pi^{ni}, E_h^{ni})_V - (A_h E_h^{ni}, E_h^{ni})_V \\ \leq -\frac{1}{2}(A_h E_h^{ni}, E_h^{ni})_V + Ch^{2k+1} \left|\tilde{U}^{ni}\right|^2_{H^{k+1}(\mathcal{T}_h)^6}.\end{aligned} \tag{6.45}$$

The bounds for the last two terms are more involved and their proof is postponed to Lemma 6.24 below. This lemma yields for $\gamma = T + 1$

$$(A_h E_h^{ni}, \pi_h \Delta^{ni})_V \leq \frac{C}{T+1} \left\|E_h^{ni}\right\|_V^2 + C(T+1) \left|\Delta^{ni}\right|^2_{H^1(\mathcal{T}_h)^6},$$

and for $\gamma = 1$ using (4.57)

$$(A_h E_\pi^{ni}, \pi_h \Delta^{ni})_V \leq Ch^{2k+2} \left|\tilde{U}^{ni}\right|^2_{H^{k+1}(\mathcal{T}_h)^6} + C \left|\Delta^{ni}\right|^2_{H^1(\mathcal{T}_h)^6}.$$

This finally gives

$$\begin{aligned}\left\| e_h^n + \tau \sum_{i=1}^s b_i \dot{E}_h^{ni} \right\|_V^2 \leq \|e_h^n\|_V^2 + 2\tau \sum_{i=1}^s b_i \Big(-\frac{1}{2}(A_h E_h^{ni}, E_h^{ni})_V \\ + Ch^{2k+1} \left|\tilde{U}^{ni}\right|^2_{H^{k+1}(\mathcal{T}_h)^6} + \frac{C}{T+1} \left\|E_h^{ni}\right\|_V^2 + C(T+1) \left|\Delta^{ni}\right|^2_{H^1(\mathcal{T}_h)^6} \Big),\end{aligned}$$

which shows the desired bound. □

For $A_h = A_h^{\mathrm{cf}}$ the proof differs only at one point. We use Lemma 4.41 with $\gamma = T + 1$ instead of Lemma 4.47 to get

$$(A_h E_\pi^{ni}, E_h^{ni})_V \leq C \frac{1}{T+1} \left\|E_h^{ni}\right\|_V^2 + C(T+1)h^{2k} \left|\tilde{U}^{ni}\right|^2_{H^{k+1}(\mathcal{T}_h)^6}$$

instead of (6.45).

Theorem 6.14. *Let $A_h = A_h^{\mathrm{cf}}$. The errors (6.40) of the Runge–Kutta method (6.38) applied to (6.3) satisfy*

$$\begin{aligned}\left\|e_h^{n+1}\right\|_V^2 - \|e_h^n\|_V^2 \\ \leq \frac{C}{T+1}\tau \left(\|e_h^n\|_V^2 + \sum_{i=1}^s \left\|E_h^{ni}\right\|_V^2 \right) + C\tau h^{2k}(T+1) \sum_{i=1}^s b_i \left|\tilde{U}^{ni}\right|^2_{H^{k+1}(\mathcal{T}_h)^6} \\ + C(T+1)\tau \left(\sum_{i=1}^s \left(\left\|\pi_h \Delta^{ni}\right\|_V^2 + \left|\Delta^{ni}\right|^2_{H^1(\mathcal{T}_h)^6} \right) + \left\| \frac{1}{\tau} \pi_h \delta^{n+1} \right\|_V^2 \right).\end{aligned}$$

Here the constant $C = C(\Omega, b, k, \rho)$ is independent of h and u.

Bound on the inner stages

As in the continuous case, we need to bound the error of the inner stages E_h^{ni} in order to apply a Gronwall lemma.

Lemma 6.15. *Let $A_h = A_h^{\mathrm{upw}}$. The error of the inner stages satisfies*

$$\sum_{i=1}^{s} \left(\left\| E_h^{ni} \right\|_V^2 + \tau (A_h E_h^{ni}, E_h^{ni})_V \right) \\ \leq C \left(\left\| e_h^n \right\|_V^2 + \sum_i \left\| \pi_h \Delta^{ni} \right\|_V^2 + \tau h^{2k+1} \sum_i \left| \tilde{U}^{ni} \right|_{H^{k+1}(\mathcal{T}_h)^6}^2 \right),$$

where the constant $C = C(\mathcal{A}, b, k, \rho)$ is independent of h and u.

Proof. We start from (6.42) and write

$$E_h^n = \mathbb{1} \otimes e_h^n + \tau (\mathcal{A} \otimes A_h) \left(E_\pi^n - E_h^n \right) - \pi_h \Delta^n.$$

Multiplying by $\mathcal{D}\mathcal{A}^{-1} \otimes I$ and taking the inner product with E_h^n gives

$$\begin{aligned} (E_h^n, (\mathcal{D}\mathcal{A}^{-1} \otimes I) E_h^n)_{V^s} &= \tau \left(E_h^n, (\mathcal{D} \otimes A_h)(E_\pi^n - E_h^n) \right)_{V^s} \\ &\quad + (E_h^n, (\mathcal{D}\mathcal{A}^{-1} \otimes I)(\mathbb{1} \otimes e_h^n - \pi_h \Delta^n))_{V^s}. \end{aligned} \tag{6.46}$$

From the coercivity condition (5.12) we conclude

$$(E_h^n, (\mathcal{D}\mathcal{A}^{-1} \otimes I) E_h^n)_{V^s} \geq \alpha \sum_{i=1}^{s} d_i \left\| E_h^{ni} \right\|_V^2.$$

Since $\mathcal{D}$ is diagonal we have by (6.45)

$$(E_h^n, (\mathcal{D} \otimes A_h)(E_\pi^n - E_h^n))_{V^s} \leq \sum_i \left(-\frac{d_i}{2} (A_h E_h^{ni}, E_h^{ni})_V + C h^{2k+1} \left| \tilde{U}^{ni} \right|_{H^{k+1}(\mathcal{T}_h)^6}^2 \right).$$

Treating the last term as in (5.58) for the continuous case and choosing $\gamma = \frac{\alpha}{2}$ shows the result. □

Lemma 6.16. *Let $A_h = A_h^{\mathrm{cf}}$. The error of the inner stages satisfies*

$$\sum_{i=1}^{s} \left\| E_h^{ni} \right\|_V^2 \leq C \left(\left\| e_h^n \right\|_V^2 + \sum_i \left\| \pi_h \Delta^{ni} \right\|_V^2 + \tau h^{2k} \sum_i \left| \tilde{U}^{ni} \right|_{H^{k+1}(\mathcal{T}_h)^6}^2 \right),$$

where the constant $C = C(\mathcal{A}, b, k, \rho)$ is independent of h and u.

Proof. The proof differs from the previous one in a way that last equation changes. By using Lemma 4.41 we have

$$(E_h^n, (\mathcal{D} \otimes A_h)(E_\pi^n - E_h^n))_{V^s} \leq \sum_i \left(C \frac{1}{\gamma} \left\| E_h^{ni} \right\|_V^2 + C h^{2k} \left| \tilde{U}^{ni} \right|_{H^{k+1}(\mathcal{T}_h)^6}^2 \right).$$

Treating the last term as in (5.58) for the continuous case and choosing γ carefully shows the result. □

Main result (full discretization error)

Our main result is contained in the following theorem.

Theorem 6.17. *Let $A_h = A_h^{\mathrm{upw}}$ and let u be the solution of (6.1). Assume that $u \in C\big(0,T;H^{k+1}(\mathcal{T}_h)^6\big)$ and $u^{(s+1)} \in L^2(0,T;D(A_M)\cap H^1(\mathcal{T}_h)^6)$ and $u^{(s+2)} \in L^2(0,T;V)$. Then for τ sufficiently small (depending on the coefficients of the Runge–Kutta method and T only), we have*

$$\begin{aligned}\big\|e_h^N\big\|_V + \left(\tau\sum_{n=1}^{N}\sum_{i=1}^{s} b_i(A_hE_h^{ni},E_h^{ni})_V\right)^{1/2}\\ \le\ C(T+1)^{1/2}\Big(\tau^{s+1}B_h(u,s,T)^{1/2} + h^{k+1/2}\max_{t\in[0,T]}|u(t)|_{H^{k+1}(\mathcal{T}_h)^6}\Big),\end{aligned}$$

where

$$B_h(u,s,T) = \int_0^T \Big\|u^{(s+1)}(t)\Big\|^2_{H^1(\mathcal{T}_h)^6}\,dt + \int_0^T \Big\|u^{(s+2)}(t)\Big\|_V^2\,dt.$$

The constant $C = C(\mathcal{A},b,k,\rho)$ is independent of h and u.

Proof. The orthogonal projection is stable, i.e.,

$$\big\|\pi_h\Delta^{ni}\big\|_V \le \big\|\Delta^{ni}\big\|_V\,,\quad \big\|\pi_h\delta^{n+1}\big\|_V \le \big\|\delta^{n+1}\big\|_V\,.$$

By Theorem 6.13 and Lemma 6.15 we obtain

$$\begin{aligned}\big\|e_h^{n+1}\big\|_V^2 - \|e_h^n\|_V^2 + \tau\sum_{i=1}^{s} b_i(A_hE_h^{ni},E_h^{ni})_V\\ \le \frac{C}{T+1}\tau\,\|e_h^n\|_V^2 + C\tau h^{2k+1}\sum_{i=1}^{s}\Big|\tilde U^{ni}\Big|^2_{H^{k+1}(\mathcal{T}_h)^6}\\ + C(T+1)\tau\left(\sum_{i=1}^{s}\big\|\Delta^{ni}\big\|^2_{H^1(\mathcal{T}_h)^6} + \left\|\frac{1}{\tau}\delta^{n+1}\right\|_V^2\right).\end{aligned}$$

The regularity assumptions on u imply

$$\tau\sum_{n=1}^{N}\left(\sum_{i=1}^{s}\big\|\Delta^{ni}\big\|^2_{H^1(\mathcal{T}_h)^6} + \left\|\frac{1}{\tau}\delta^{n+1}\right\|_V^2\right) \le C\tau^{2(s+1)}B_h(u,s,T).$$

Summing over n and applying a discrete Gronwall inequality from Corollary 2.9 gives

$$\begin{aligned}\big\|e_h^N\big\|_V^2 + \tau\sum_{n=1}^{N}\sum_{i=1}^{s} b_i(A_hE_h^{ni},E_h^{ni})_V \le C(1+T)\tau^{2s+1}B_h(u,s,T)\\ + CTh^{2k+1}\max_{t\in[0,T]}|u(t)|^2_{H^{k+1}(\mathcal{T}_h)^6}\,,\end{aligned}$$

from which the result is easily obtained. □

Analogously, by using Theorem 6.14 and Lemma 6.16, the convergence result for the central flux dG method can be proven.

Theorem 6.18. *Let $A_h = A_h^{\mathrm{cf}}$ and let u be the solution of (6.1). Assume that $u \in C\big(0,T;H^{k+1}(\mathcal{T}_h)^6\big)$ and $u^{(s+1)} \in L^2(0,T;D(A_M)\cap H^1(\mathcal{T}_h)^6)$ and $u^{(s+2)} \in L^2(0,T;V)$. Then for τ sufficiently small (depending on the coefficients of the Runge–Kutta method and T only), we have*

$$\big\|e_h^N\big\|_V \le\ C(T+1)^{1/2}\Big(\tau^{s+1}B_h(u,s,T)^{1/2} + T^{1/2}h^k \max_{t\in[0,T]}|u(t)|_{H^{k+1}(\mathcal{T}_h)^6}\Big).$$

The constant $C = C(\mathcal{Q},b,k,\rho)$ is independent of h and u.

Corollary 6.19. *Under the assumptions of Theorem 6.17, the error is also bounded by*

$$\sum_{n=1}^{N}\tau\,\|e_h^n\|_V^2 \le CT(T+1)\big(\tau^{2s+2}B_h(u,s,T) + h^{2k+1}\max_{t\in[0,T]}|u(t)|^2_{H^{k+1}(\mathcal{T}_h)^6}\big).$$

Divergence error

Next we study the divergence error of the numerical approximation. From the inverse inequality (4.36) and Theorem 6.17 we immediately obtain

$$\big\|\nabla\cdot e_h^N\big\|_V \le\ C(T+1)^{1/2}\Big(h^{-1}\tau^{s+1}B_h(u,s,T)^{1/2} + h^{k-1/2}\max_{t\in[0,T]}|u(t)|_{H^{k+1}(\mathcal{T}_h)^6}\Big).$$

In a weak sense, we can even prove that the discrete divergence is preserved exactly if $f = 0$. As in [23] we define a test space $X_h \subset H_0^1(\Omega)$ as the space of continuous, elementwise polynomial functions:

$$X_h = \big\{v\in C^0(\bar\Omega)\mid\ v|_K\in\mathbb{P}_{k+1}(K)^6,\ K\in\mathcal{T}_h\big\}\cap H_0^1(\Omega).$$

By $\langle\cdot,\cdot\rangle_{-1}$ we denote the duality product between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$, in which

$$\langle\nabla\cdot u,\psi\rangle_{-1} = -(u,\nabla\psi)_{0,\Omega}\qquad\text{for all}\qquad u\in L^2(\Omega)^3,\ \psi\in H_0^1(\Omega).$$

Theorem 6.20. *Let $f\equiv 0$ and $A_h\in\big\{A_h^{\mathrm{upw}},A_h^{\mathrm{cf}}\big\}$. Then the Runge–Kutta solution (6.38) satisfies*

$$\begin{cases}\langle\nabla\cdot\epsilon\mathbf{E}_h^{n+1},\psi\rangle_{-1} = \langle\nabla\cdot\epsilon\mathbf{E}_h^{n},\psi\rangle_{-1},\\ \langle\nabla\cdot\mu\mathbf{H}_h^{n+1},\psi\rangle_{-1} = \langle\nabla\cdot\mu\mathbf{H}_h^{n},\psi\rangle_{-1},\end{cases}\qquad \textit{for all}\qquad \psi\in X_h.$$

Moreover, if the initial data is divergence free, then

$$\langle\nabla\cdot\epsilon\mathbf{E}_h^n,\psi\rangle_{-1} = \langle\nabla\cdot\mu\mathbf{H}_h^n,\psi\rangle_{-1} = 0,\qquad n=0,1,2,\dots.$$

Proof. For $\psi \in X_h \subset H_0^1(\Omega)$, integration by parts shows

$$\langle \nabla \cdot \big(\epsilon(\mathbf{E}_h^{n+1} - \mathbf{E}_h^n)\big), \psi \rangle_{-1} = -(\epsilon(\mathbf{E}_h^{n+1} - \mathbf{E}_h^n), \nabla\psi)_{0,\Omega} = -(u_h^{n+1} - u_h^n, \begin{pmatrix} 0 \\ \nabla\psi \end{pmatrix})_V.$$

By using (6.38) and Lemma 4.46 for A_h^{upw} or Lemma 4.39 for A_h^{cf} we obtain

$$\langle \nabla \cdot \big(\epsilon(\mathbf{E}_h^{n+1} - \mathbf{E}_h^n)\big), \psi \rangle_{-1} = \tau \sum_{i=1}^{s} b_i (A_h U_h^{ni}, \begin{pmatrix} 0 \\ \nabla\psi \end{pmatrix})_V = 0,$$

since for functions in X_h we have $\nabla \times \nabla\psi = 0$, $\mathbf{n}_F \times [\![\nabla\psi_h]\!]_F = 0$ for $F \in \mathcal{F}_h^i$ and $\mathbf{n} \times \nabla\psi_h = 0$ on $\partial\Omega$. The result for $\mathbf{H}_h^n$ is proved analogously.

The second part follows from

$$\langle \nabla \cdot \epsilon(\mathbf{E}_h^0), \psi \rangle_{-1} = -\int_\Omega \pi_h \mathbf{E}^0 \cdot \epsilon \nabla\psi = -\int_\Omega \epsilon \mathbf{E}^0 \cdot \nabla\psi = \int_\Omega \nabla \cdot (\epsilon \mathbf{E}^0)\psi = 0$$

and similar for $\mathbf{H}_h^n$. □

6.3.3 Auxiliary results

For the proof of Lemma 6.24 below we need some auxiliary results:

Lemma 6.21. *If Assumption 4.32 is satisfied, then for $v \in H^1(\mathcal{T}_h)^3$, $w \in L^2(\Omega)^3$ and arbitrary $\gamma > 0$ we have*

$$\sum_K |(w, \nabla \times \pi_h v)_{0,K}| \le \frac{1}{2\gamma} \|w\|_{0,\Omega}^2 + C\gamma \sum_K |v|_{1,K}^2 ,$$

where $C = C(\rho, k)$ is independent of K.

Proof. The Cauchy-Schwarz inequality and Young's inequality yield

$$|(w, \nabla \times \pi_h v)_{0,K}| \le \frac{1}{2\gamma} \|w\|_{0,K}^2 + \frac{\gamma}{2} \|\nabla \times \pi_h v\|_{0,K}^2 . \tag{6.47}$$

From the inverse inequality (4.36) follows

$$\begin{aligned} \|\nabla \times \pi_h v\|_{0,K} &= \|\nabla \times (v + \pi_h v - v)\|_{0,K} \\ &\le \|\nabla \times v\|_{0,K} + \|\nabla \times (\pi_h v - v)\|_{0,K} \\ &\le C\, |v|_{1,K} \end{aligned}$$

for $v \in H^1(K)^3$. Inserting this bound into (6.47) and summing over all elements K shows the desired bound. □

Lemma 6.22. *Let Assumption 4.32 be satisfied and let $\gamma > 0$ be arbitrarily chosen. If F is an interior face connecting the elements K and K_F and $\mathbf{n}_F$ is the unit normal vector pointing from K to K_F, then for $v \in H(\mathrm{curl}, K \cup K_F) \cap H^1(K)^3 \cap H^1(K_F)^3$, $w \in H^1(K)^3$, we have*

$$|(w, \mathbf{n}_F \times [\![\pi_h v]\!]_F)_{0,F}| \leq \frac{1}{\gamma}\Big(h_K^2 \, |w|_{1,K}^2 + 3 \, \|w\|_{0,K}^2\Big) + C\gamma\Big(|v|_{1,K}^2 + |v|_{1,K_F}^2\Big). \tag{6.48}$$

If $F \subset \partial K$ is an exterior face with outward normal vector $\mathbf{n}_F$, $w \in H^1(K)^3$, and $v \in H_0(\mathrm{curl}, \Omega) \cap H^1(K)^3$ then

$$|(w, \mathbf{n}_F \times \pi_h v)_{0,F}| \leq \frac{1}{2\gamma}\Big(h_K^2 \, |w|_{1,K}^2 + 3 \, \|w\|_{0,K}^2\Big) + C\gamma \, |v|_{1,K}^2 \, . \tag{6.49}$$

In both estimates $C = C(\rho, k)$ is independent of K and K_F.

Proof. By assumption, we have $\mathbf{n}_F \times [\![v]\!]_F = 0$. Thus we can write

$$(w, \mathbf{n}_F \times [\![\pi_h v]\!]_F)_{0,F} = (w, \mathbf{n}_F \times (\pi_h v_{K_F} - v_{K_F}))_{0,F} - (w, \mathbf{n}_F \times (\pi_h v_K - v_K))_{0,F}.$$

By using the continuous trace inequality (4.38) and the polynomial approximation properties on faces (4.41) we have

$$\begin{aligned} |(w, \mathbf{n}_F \times (\pi_h v_K - v_K))_{0,F}| &\leq \|w\|_{0,F} \, \|\mathbf{n}_F \times (\pi_h v_K - v_K)\|_{0,F} \\ &\leq C\Big(|w|_{1,K} \, \|w\|_{0,K} + h_K^{-1} \, \|w\|_{0,K}^2\Big)^{1/2} h_K^{1/2} \, |v|_{1,K} \\ &\leq \frac{1}{2\gamma}\Big(h_K^2 \, |w|_{1,K}^2 + 3 \, \|w\|_{0,K}^2\Big) + C\gamma \, |v|_{1,K}^2 \, . \end{aligned}$$

Analogously, by using (4.34), we obtain

$$|(w_K, \mathbf{n}_F \times (\pi_h v_{K_F} - v_{K_F}))_{0,F}| \leq \frac{1}{2\gamma}\Big(h_K^2 \, |w|_{1,K}^2 + 3 \, \|w\|_{0,K}^2\Big) + C\gamma \, |v|_{1,K_F}^2 \, .$$

This proves (6.48). To prove (6.49), note that $v \in H_0(\mathrm{curl}, \Omega)$ implies $\mathbf{n}_F \times v = 0$ on the exterior face F, so that $\mathbf{n}_F \times \pi_h v = \mathbf{n}_F \times (\pi_h v - v)$. □

Lemma 6.23. *Let Assumption 4.32 be satisfied and let $w \in \mathbb{P}_k(K)^3$ and $\gamma > 0$ be arbitrarily chosen. If F is an interior face connecting the elements K and K_F and n_F is the unit normal vector pointing from K to K_F, then for $v \in H(\mathrm{curl}, K \cup K_F) \cap H^1(K)^3 \cap H^1(K_F)^3$, we have*

$$|(w, \mathbf{n}_F \times [\![\pi_h v]\!]_F)_{0,F}| \leq \frac{1}{\gamma} \, \|w\|_{0,K}^2 + C\gamma\Big(|v|_{1,K}^2 + |v|_{1,K_F}^2\Big). \tag{6.50}$$

If F is an exterior face of K with outward normal vector $\mathbf{n}_F$ and $v \in H_0(\mathrm{curl}, \Omega) \cap H^1(K)$, then

$$|(w, \mathbf{n}_F \times \pi_h v)_{0,F}| \leq \frac{1}{\gamma} \, \|w\|_{0,K}^2 + C\gamma \, |v|_{1,K}^2 \, . \tag{6.51}$$

In both estimates $C = C(\rho, k)$ is independent of K and K_F.

Proof. Analogously to the previous lemma using the discrete trace inequality. □

Lemma 6.24. *Suppose that $\mathcal{T}_h$ satisfies Assumption 4.32. Let $v \in \mathcal{D}(A) \cap H^1(\mathcal{T}_h)^6$ and $\gamma > 0$ be arbitrarily chosen. Then, for $u \in H^1(\mathcal{T}_h)^6$ and $A_h \in \{A_h^{\mathrm{upw}}, A_h^{\mathrm{cf}}\}$, we have*

$$|(A_h u, \pi_h v)_V| \leq \frac{C}{\gamma}\Big(\|u\|_V^2 + \sum_K h_K^2 \, |u|_{1,K}^2\Big) + C\gamma \sum_K |v|_{1,K}^2 \,. \tag{6.52}$$

For $u_h \in V_h$, we have

$$|(A_h u_h, \pi_h v)_V| \leq \frac{C}{\gamma} \, \|u_h\|_V^2 + C\gamma \sum_K |v|_{1,K}^2 \,. \tag{6.53}$$

The constants $C = C(\rho, k)$ are independent h and K.

Proof. We use Lemma 4.47 and Lemma 4.41 for $A_h = A_h^{\mathrm{upw}}$ and $A_h = A_h^{\mathrm{cf}}$, respectively, and bound the terms separately. The bound on the sum over the elements follows from Lemma 6.21 and the bounds on the sums over the interior and exterior faces follow from Lemmas 6.22 and 6.23, respectively. □

6.4 Exponential Runge–Kutta methods

Again, we consider the continuous problem (6.16) on a Banach space $(X, \|\cdot\|)$ and its discretization (6.17), as in Section 6.2. We suppose that Assumption 6.6 is satisfied. Applying the exponential quadrature rule (5.60) to the semidiscrete problem (6.17) gives the fully discrete scheme

$$u_h^{n+1} = \mathrm{e}^{-\tau A_h} u_h^n + \tau \sum_{i=1}^{s} b_i(-\tau A_h)(\pi_h f)(t_n + c_i \tau), \tag{6.54a}$$

with the weights

$$b_i(-\tau A_h) = \int_0^1 \mathrm{e}^{-\tau(1-\theta)A_h} \ell_i(\theta) d\theta. \tag{6.54b}$$

Analogously as in Lemma 5.37 the following result can be proved.

Lemma 6.25. *Under the Assumption 6.6, the weights $b_i(-\tau A_h)$ are bounded uniformly in $h \in \mathbb{R}_+$ and $\tau \in [0, 1]$.*

Then it is not hard to see that the following discrete analog of Theorem 5.38 holds.

Theorem 6.26. *Let Assumption 6.6 be fulfilled and let $f^{(s)} \in L^1(0, T; X)$. If u_h is the solution of (6.17), and (u_h^n) the numerical solution obtained by the exponential quadrature rule (6.54), then*

$$\|u_h^n - u_h(t_n)\| \leq C\tau^s \int_0^T \Big\| f^{(s)}(\theta) \Big\| \, d\theta$$

holds uniformly on $0 \leq t_n \leq T$, with a constant C given by (5.68).

Here we have used that

$$\left\|(\pi_h f)^{(s)}\right\| = \left\|\pi_h f^{(s)}\right\| = \left\|f^{(s)}\right\|.$$

Now we can bound the full discretization error using the triangle inequality

$$\|u_h^n - \pi_h u(t_n)\| = \|u_h^n - u_h(t_n)\| + \|u_h(t_n) - \pi_h u(t_n)\|, \tag{6.55}$$

where the second term is the semidiscretization error and the first term is bounded by Theorem 6.26.

Remark. In the case of collocation methods, we were not able to bound the full discretization error by using this kind of triangle inequality (see Section 6.2.2) and we had to use alternatives such as L^2 or elliptic projections. The reason for that is that the analysis there was done in terms of the exact solution, while here we provided the analysis in terms of the data. The analysis in terms of the data is also possible for collocation methods, see [4, Theorem 3]. Then the inequality (6.55) can be used to get a full discretization error estimate, but this will not improve the bounds obtained in Section 6.2.2, see [4, Theorem 5].

6.4.1 Application to Maxwell's equation

Assumption 6.6 holds for Maxwell's equation with $C_A = 1$ and $\omega = 0$, i.e

$$\|S_h(t)\|_{V \leftarrow V} \leq 1 \quad \forall t \geq 0,\ h > 0,$$

and therefore Theorem 6.26 applies for the first term in (6.55). For the second term we use (6.10) and (6.11) for central and upwind flux scheme, respectively.

Theorem 6.27. *Let u be the exact solution of Maxwell's equations (5.1) such that $u^{(s)} \in L^1(0,T;D(A_M))$ and $u^{(s+1)} \in L^1(0,T;X)$ and let $(u_h^n)_{n\geq 0}$ be a numerical solution defined with the exponential quadrature rule (5.60). For $A_h = A_h^{\mathrm{cf}}$ we have*

$$\|u_h^n - \pi_h u(t_n)\|_V \leq C\tau^s \int_0^T \left\|A_M u^{(s)}\right\|_V + \left\|u^{(s+1)}\right\|_V + Ch^k T^{1/2} \int_0^T \left(|u|^2_{H^{k+1}(\mathcal{T}_h)^6)}\right)^{1/2}$$

uniformly on $0 \leq t_n \leq T$. For $A_h = A_h^{\mathrm{upw}}$ we have

$$\|u_h^n - \pi_h u(t_n)\|_V \leq C\tau^s \int_0^T \left\|A_M u^{(s)}\right\|_V + \left\|u^{(s+1)}\right\|_V + Ch^{k+1/2} \int_0^T \left(|u|^2_{H^{k+1}(\mathcal{T}_h)^6)}\right)^{1/2}$$

uniformly on $0 \leq t_n \leq T$. Constants here are independent of τ, h, T and u. □

CHAPTER 7

Implementation and numerical experiments

In this last chapter we discuss some implementation issues and provide numerical experiments which should enable a better understanding of our theoretical results. In the first section the dG method is considered and some results from Chapter 4 are confirmed numerically. Section 7.2 is devoted to Gauss collocation methods for homogeneous problems, while in Section 7.3 inhomogeneous problems are considered. We end the thesis with a summary and an outlook.

7.1 The dG method

To solve the space semidiscrete problem given in (6.3) on a computer, we have to choose a basis of the dG space V_h. Let $\{\phi_1, ..., \phi_N\}$ be a basis for V_h. By taking the V-inner product of (6.3) with all basis functions, we end up with a system of ordinary differential equations

$$\mathbf{M}\mathbf{u}'(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{f}(t), \quad \mathbf{u}(0) = \mathbf{u}_0 \tag{7.1}$$

with large sparse matrices $\mathbf{M}, \mathbf{A} \in \mathbb{R}^{N\times N}$ defined by

$$\mathbf{M} = \Big((\vec{\phi}_j, \vec{\phi}_k)_V\Big)_{j,k}, \qquad \mathbf{A} = \Big((\vec{\phi}_j, A_h\vec{\phi}_k)_V\Big)_{j,k}.$$

$\mathbf{M}$ is called the **mass matrix** and it is block diagonal and symmetric, and $\mathbf{A}$ is the **stiffness matrix**. Further on, $\mathbf{u}(t) \in \mathbb{R}^N$ is the coefficient vector of the solution at time t with respect to the finite element basis, the right-hand side vector is defined by $\mathbf{f} = \Big((f, \phi_j)_V\Big)_j$ and the initial data by $\mathbf{u}_0 = \Big((u_0, \phi_j)_V\Big)_j$.

The natural choice for the inner-product is the $\mathbf{M}$-inner product defined as

$$(\mathbf{u}, \mathbf{v})_{\mathbf{M}} := (\mathbf{M}\mathbf{u}, \mathbf{v}).$$

The reason for this is that for a discontinuous finite element function $u_h(t) \in V_h$,

$$u_h(t) = \sum_{j=1}^{N} u_j(t)\phi_j,$$

the vector $\mathbf{u}(t)$ is given by $\mathbf{u}(t) = \big(u_1(t), \dots, u_N(t)\big)^T$ and therefore satisfies

$$(u_h(t), v_h(t))_V = \sum_{j,k} u_j(t)v_k(t)(\phi_j, \phi_k)_V = \mathbf{v}(t)^T \mathbf{M}\mathbf{u}(t) = (\mathbf{u}(t), \mathbf{v}(t))_{\mathbf{M}}.$$

In particular, we have

$$\|u_h(t)\|_V = \|\mathbf{u}(t)\|_{\mathbf{M}},$$

which means that the $\mathbf{M}$-norm of the vector $\mathbf{u}(t)$ is equal to the V-norm of the corresponding function u_h in the dG space, which is exactly the norm of interest.

Our numerical experiments are done in two spatial dimensions in Matlab. As a basis for our implementation we have used the Matlab code of Hesthaven and Warburton provided in [33, Chapter 6]. In particular, we have constructed the matrices $\mathbf{M}$ and $\mathbf{A}$ defined above with this code. Basis functions that are used are Lagrange polynomials.

Example

We consider the TM polarization of Maxwell's equations (3.9) on $\Omega = [-1, 1]^2$ in a homogeneous medium with $\epsilon = \mu = 1$:

$$\partial_t \begin{pmatrix} H_x \\ H_y \\ E_z \end{pmatrix} + A_{TM} \begin{pmatrix} H_x \\ H_y \\ E_z \end{pmatrix} = f, \qquad \text{in } \Omega = [-1, 1]^2, \tag{7.2}$$

where

$$\begin{aligned} A_{TM} &= \begin{pmatrix} 0 & 0 & \partial_y \\ 0 & 0 & -\partial_x \\ \partial_y & -\partial_x & 0 \end{pmatrix}, \\ D(A) &= \{(H_x, H_y, E_z)^T \mid H_x, H_y \in H^1(\Omega), E_z \in H_0^1(\Omega)\}. \end{aligned} \tag{7.3}$$

We have discretized the problem in space and computed the eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$. It is important to note that this is the matrix which corresponds to the discrete operator A_h. For $k = 2$ and $h = 0.25$, which gives 2628 degrees of freedom, the eigenvalues are plotted in Figures 7.1 and 7.2, for the upwind and the central flux, respectively. We see that for the upwind flux method, the eigenvalues are in the left half-plane, with a bigger concentration closer to imaginary axis. For the central flux, the computed eigenvalues are close to the imaginary axis. Small positive real part resulting from inaccuracies of the eigensolver can though produce numerical instabilities.

If we choose the source term f as (cf. example from [17])

$$\begin{aligned} f_x &= f_y = 0, \\ f_z(x, y, t) &= e^t \left((1 - x^2)(1 - y^2) + 2(1 - x^2) + 2(1 - y^2)\right), \end{aligned}$$

Figure 7.1: eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$, upwind flux

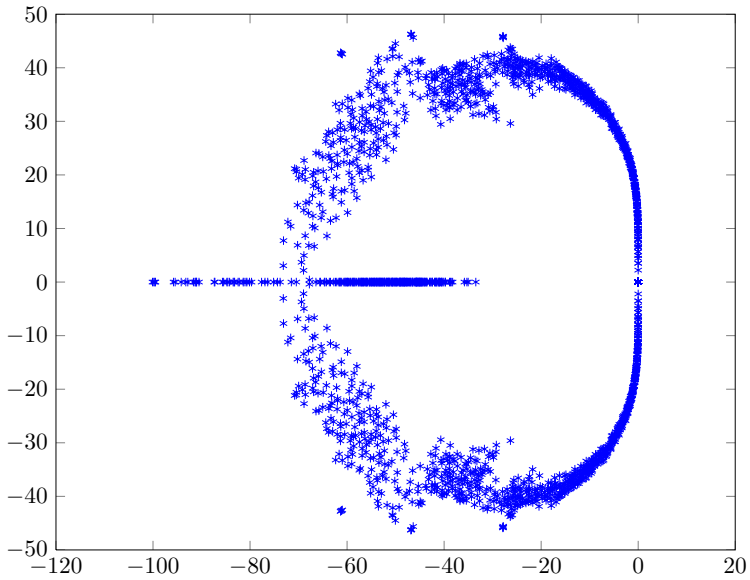
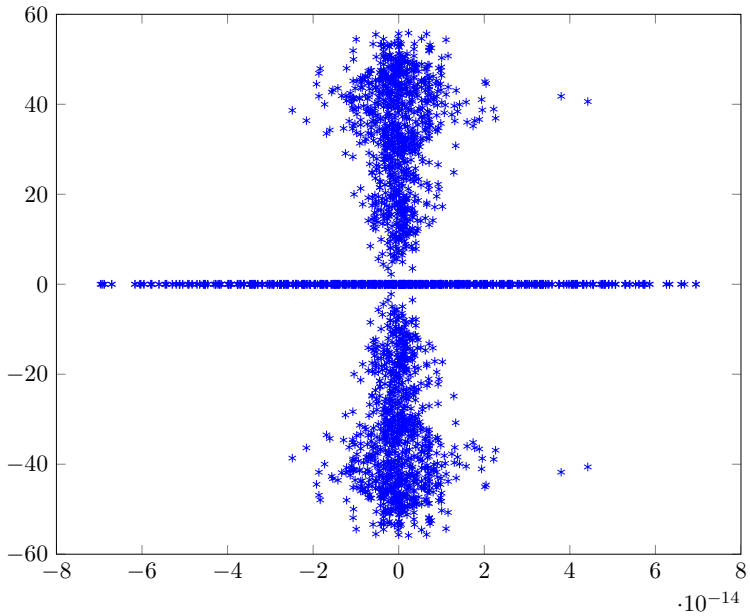


Figure 7.2: eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$, central flux



then the exact solution is given by

$$\begin{aligned} H_x(x,y,t) =& 2e^t y(1-x^2), \\ H_y(x,y,t) =& -2e^t x(1-y^2), \\ E_z(x,y,t) =& e^t(1-x^2)(1-y^2). \end{aligned}$$

Now we investigate the convergence of dG methods. The full discretization errors are plotted against h in Figures 7.3 and 7.4 for upwind and central fluxes, respectively. Time integration is done by a Gauss collocation method with $s = 3$ and $\tau = 0.01$ so that the time integration error is negligible. We observe order h^{k+1} for the upwind flux method and h^k for central flux method. From now on, we work with the upwind flux only.

Figure 7.3: order of convergence, upwind flux

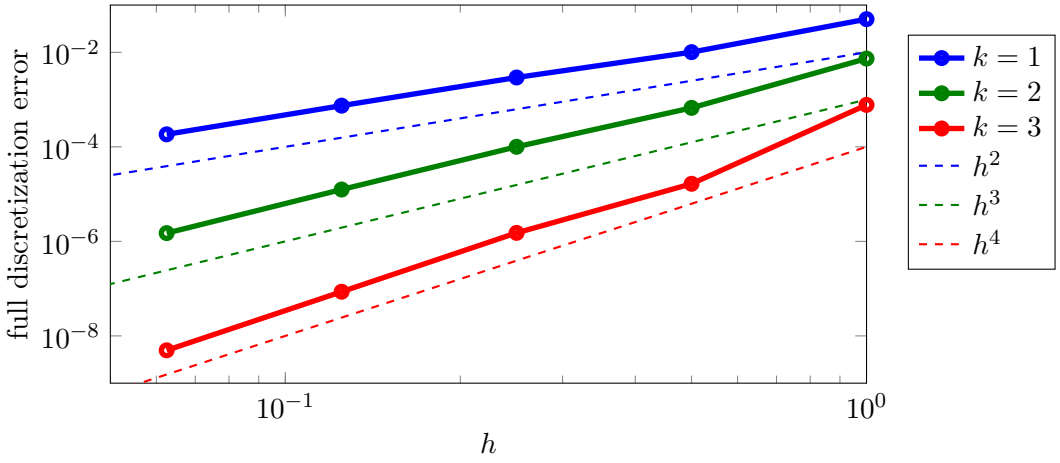
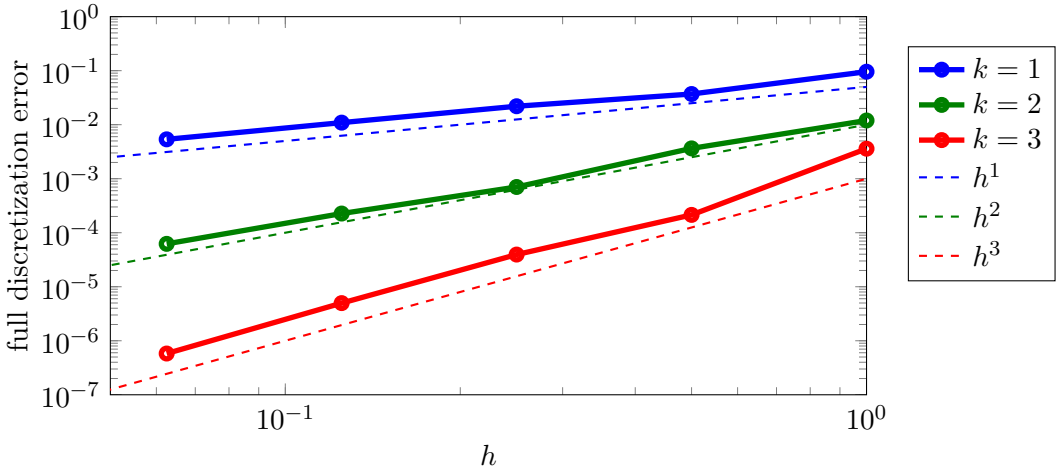


Figure 7.4: order of convergence, central flux



7.2 The homogenous case

In the homogeneous case the exact solution of (7.1) is given by

$$\mathbf{u}(t) = \exp(-t\mathbf{M}^{-1}\mathbf{A})\mathbf{u}_0, \quad t \geq 0. \tag{7.4}$$

We have implemented Gauss collocation methods with $s = 1$ and $s = 3$ stages.

Implicit midpoint rule

As we have seen in Chapter 5 the stability function of the implicit midpoint rule is given by

$$R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}} = 1 + \frac{z}{1 - \frac{z}{2}}.$$

The discrete solution can therefore be computed via

$$\begin{aligned}
\mathbf{u}_{n+1} &= R(-\tau \mathbf{M}^{-1}\mathbf{A})\mathbf{u}_n \\
&= \left(I - (I + \frac{\tau \mathbf{M}^{-1}\mathbf{A}}{2})^{-1} \tau \mathbf{M}^{-1}\mathbf{A} \right) \mathbf{u}_n \\
&= \left(I - \tau (\mathbf{M} + \frac{\tau}{2}\mathbf{A})^{-1}\mathbf{A} \right) \mathbf{u}_n \\
&= \mathbf{u}_n - \tau (\mathbf{M} + \frac{\tau}{2}\mathbf{A})^{-1}\mathbf{A}\mathbf{u}_n.
\end{aligned}$$

Gauss with $s = 3$

The stability function is the $(3, 3)$-Padé approximation

$$R(z) = \frac{1 + \frac{1}{2}z + \frac{1}{10}z^2 + \frac{1}{120}z^3}{1 - \frac{1}{2}z + \frac{1}{10}z^2 - \frac{1}{120}z^3} = \frac{P(z)}{Q(z)}.$$

We can rewrite it as

$$R(z) = 1 + \frac{z + \frac{1}{60}z^3}{Q(z)},$$

and factorize $Q(z)$

$$Q(z) = -\frac{1}{120}(z - z_1)(z - z_2)(z - \bar{z}_2) = (1 - \frac{z}{z_1})(1 - \frac{z}{z_2})(1 - \frac{z}{\bar{z}_2})$$

with

$$\begin{aligned}
z_1 &= 4 - 2\sqrt[3]{\frac{2}{1 + \sqrt{5}}} + 2^{2/3}\sqrt[3]{1 + \sqrt{5}} \approx 4.6444, \\
z_2 &= 4 + (1 - i\sqrt{3})\sqrt[3]{\frac{2}{1 + \sqrt{5}}} - (1 + i\sqrt{3})\sqrt[3]{\frac{1}{2}(1 + \sqrt{5})} \approx 3.6778 - 3.5088i.
\end{aligned}$$

We also have

$$z + \frac{1}{60}z^3 = z(1 - \frac{z}{y_1})(1 - \frac{z}{\bar{y}_1})$$

with $y_1 = i\sqrt{60}$. Therefore, we can write the stability function as

$$R(z) = 1 + (1 - \frac{z}{z_1})^{-1}(1 - \frac{z}{y_1})(1 - \frac{z}{z_2})^{-1}(1 - \frac{z}{\bar{y}_1})(1 - \frac{z}{\bar{z}_2})^{-1}z. \tag{7.5}$$

Our discrete approximation is defined as $\mathbf{u}_{n+1} = R(-\tau\mathbf{M}^{-1}\mathbf{A})\mathbf{u}_n$. Substituting z by $-\tau\mathbf{M}^{-1}\mathbf{A}$ in (7.5) we get

$$\mathbf{u}_{n+1} = \mathbf{u}_n - \tau(\mathbf{M} + \frac{\tau}{z_1}\mathbf{A})^{-1}(\mathbf{M} + \frac{\tau}{y_1}\mathbf{A})(\mathbf{M} + \frac{\tau}{z_2}\mathbf{A})^{-1}(\mathbf{M} + \frac{\tau}{\bar{y}_1}\mathbf{A})(\mathbf{M} + \frac{\tau}{\bar{z}_2}\mathbf{A})^{-1}\mathbf{A}\mathbf{u}_n.$$

In each time step we have to compute three matrix-vector multiplications and solve three linear systems. For solving the linear systems we compute the LU decomposition of matrices at the beginning and then solve triangular systems in each time step. For larger examples we use iterative solvers (for example gmres [59]).

7.2.1 Example I: TM polarization of ME on square in 2d

We consider again the TM polarization of Maxwell's equations on $[-1,1]^2$ defined in (7.2) and (7.3). The final time of the simulation is $T = 1$.

We construct the initial data with different smoothness, i. e. we search for functions u_0 which satisfy $u_0 \in D(A_{TM}^m)$ and $u_0 \notin D(A_{TM}^{m+1})$ for some $m \in \mathbb{N}$. We remind that for an operator A, $D(A^m)$ is defined recursively as

$$D(A^m) = \left\{ v \in D(A^{m-1}) \mid A^{m-1}v \in D(A) \right\}, \quad \text{for } m = 2, 3, \ldots$$

Initial data in $D(A_{TM}^2)$

We set $H_x^0 = H_y^0 = 0$ and

$$E_z^0(x,y) = (x-1)^2(x+1)^2(y-1)^2(y+1)^2.$$

A simple computation shows that $u_0 = (H_x^0, H_y^0, E_z^0) \in D(A_{TM}^2)$ but $u_0 \notin D(A_{TM}^3)$.

According to Theorem 6.3 we expect that $A_h^2\pi_h u_0$ stays bounded for $h \to 0$ if a polynomial degree $k \geq 1$ is used. Figures 7.5 and 7.6 show that norms of $A_h^1\pi_h u$ and $A_h^2\pi_h u$ remain finite for $k = 1$ and $k = 3$, respectively. We also see that for $k = 3$, the norm of $A_h^3\pi_h u$ grows only slowly.

Figure 7.5: Initial value in $D(A_{TM}^2)$, $k = 1$

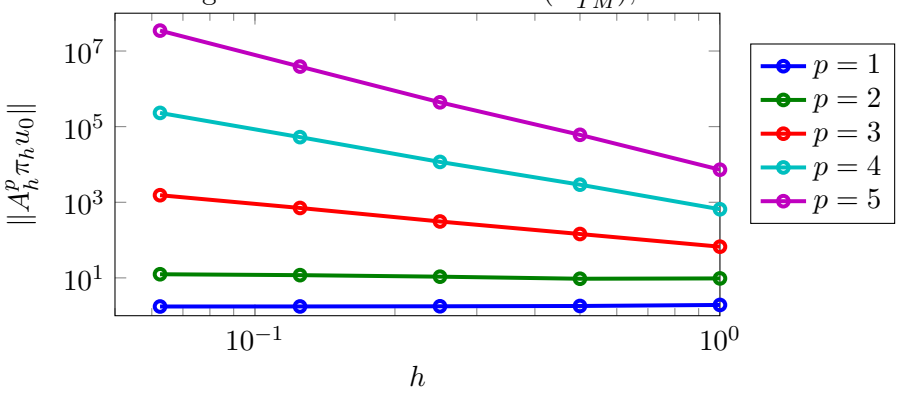
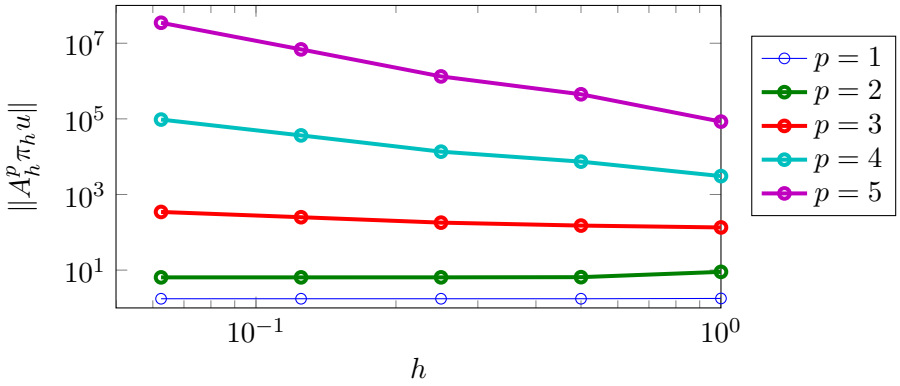


Figure 7.6: Initial value in $D(A_{TM}^2)$, $k = 3$



From Section 6.2.1 we know that the full discretization error behaves like

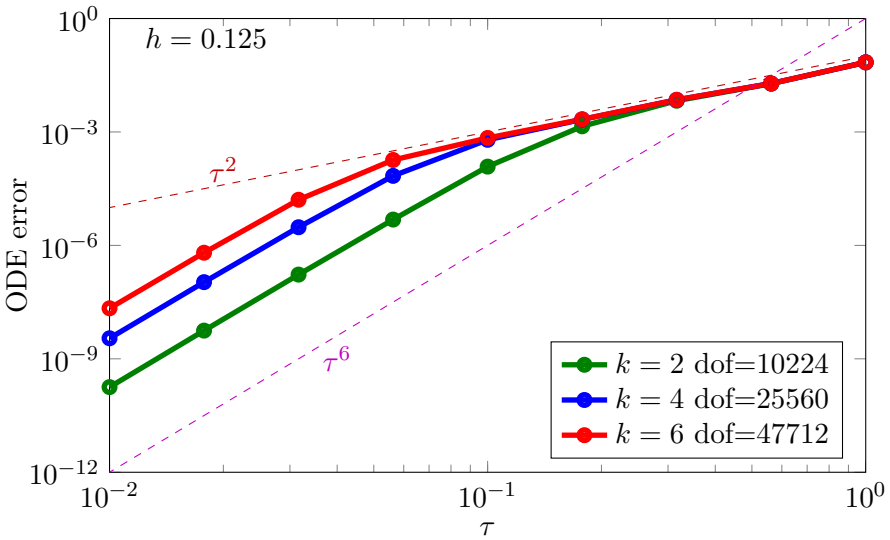
$$C T \tau^{m-1} \left\| A_{TM}^m u_0 \right\| + C h^{k+1/2} \left\| S(.) u_0 \right\|_{L^2(0,T;H^{k+1}(\mathcal{T}_h))^6}$$

for $u_0 \in D(A_{TM}^m)$ and $k \geq m-1$. The first term here is the time integration error which actually comes from

$$C T \tau^{m-1} \left\| A_h^m \pi_h u_0 \right\| .$$

We investigate the time integration error of the Gauss collocation method with $s = 3$. In Figure 7.7 the ODE error is plotted and we can see that the method has full order 6 if we are in the non-stiff region ($\tau \leq Ch$, for some $C > 0$). In the stiff region, we observe order 2, which is actually more than expected (we expect order 1 since u_0 is in $D(A_{TM}^2)$ only). The explanation for this may lay in the fact that the norm of $A_h^3 \pi_h v$ grows also very slowly and therefore is not large enough (for $h = 0.125$) to show order reduction.

Figure 7.7: Convergence of Gauß RK method, $s = 3$, $u_0 \in D(A_{TM}^2)$



Initial data in $D(A_{TM}^4)$

We set $H_x^0 = H_y^0 = 0$ and

$$E_z^0(x,y) = (x-1)^4(x+1)^4(y-1)^4(y+1)^4.$$

A simple computation shows that $u_0 = (H_x^0, H_y^0, E_z^0) \in D(A_{TM}^4)$ but $u_0 \notin D(A_{TM}^5)$. Figure 7.8 shows E_z^0. Again, from Theorem 6.3 we expect that the norm of $A_h^4 \pi_h u_0$ is bounded for $k \geq 3$, which can be seen in Figure 7.11. In Figure 7.10 we see that this is not the case for $k = 1$. For the Gauss collocation method with $s = 3$, we expect convergence of order 3 in the stiff region. We observe order 3.5, see Figure 7.9.

Figure 7.8: Initial data $E_z^0(x,y) = (x-1)^4(x+1)^4(y-1)^4(y+1)^4$

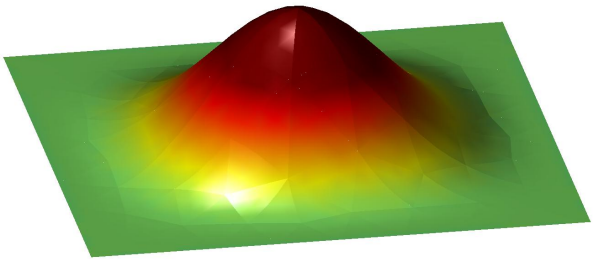


Figure 7.9: Convergence of Gauß RK method, $s = 3$, $u_0 \in D(A_{TM}^4)$

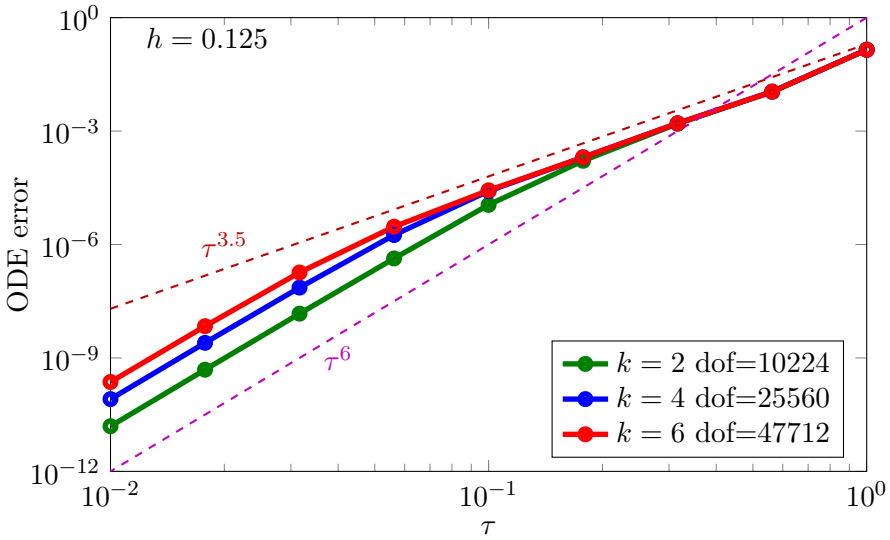


Figure 7.10: Initial value in $D(A_{TM}^4)$, $k = 1$

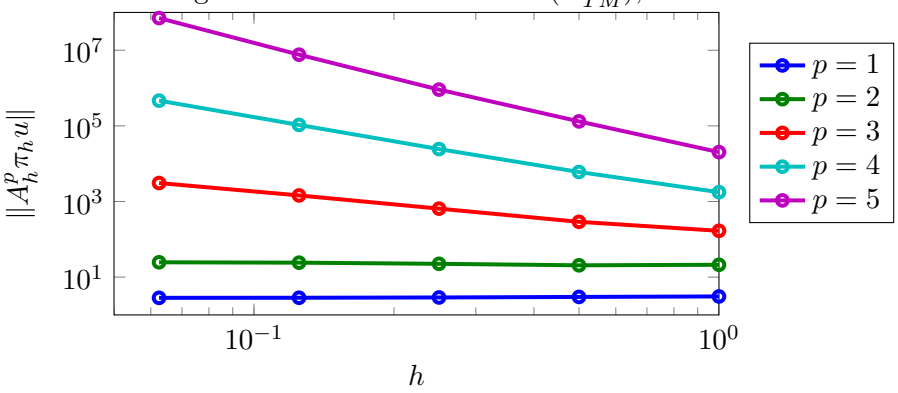
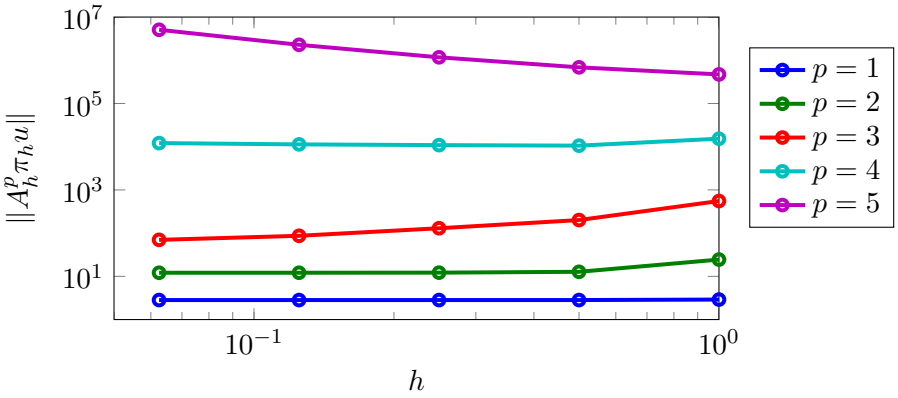


Figure 7.11: Initial value in $D(A_{TM}^4)$, $k = 3$



Initial data in $D(A_{TM}^\infty)$

We set $H_x^0 = H_y^0 = 0$ and

$$E_z^0(x, y) = \sin(\pi x) \sin(\pi y).$$

By noting that

$$u_0 \in D(A_{TM}), \quad A_{TM} u_0 \in D(A_{TM})$$

and

$$A_{TM}^2 u_0 = -\pi^2 u_0$$

we conclude that $u_0 \in D(A_{TM}^\infty)$. From (6.9) we expect that for every p, the norm of $A_h^p \pi_h u$ is bounded if $k \geq p - 1$. We can see in Figure 7.12 that this is the case. We observe the convergence of order 6, i. e. the full order of convergence, see Figure 7.13.

Figure 7.12: Initial value in $D(A_{TM}^{\infty})$, $k = 4$

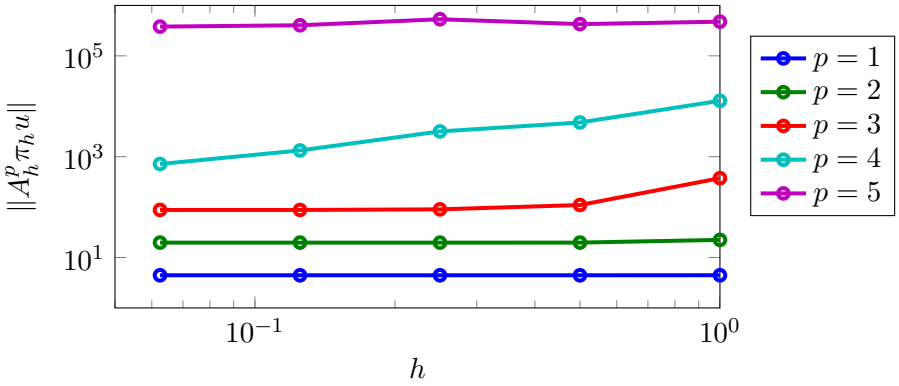


Figure 7.13: Convergence of Gauß RK method, $s = 3$, $u_0 \in D(A_{TM}^{\infty})$

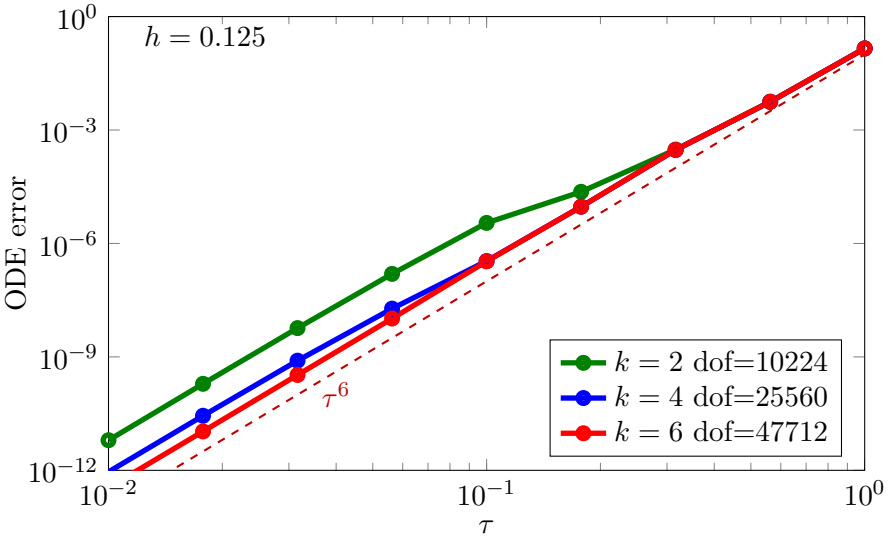
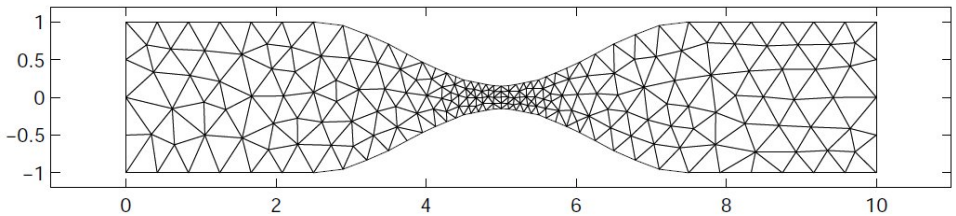


Figure 7.14: Domain and mesh for Example II



7.2.2 Example II: TE polarization of ME on deformed domain in 2d

Here we consider a TE polarization of Maxwell's equations (3.10) on an irregular domain in 2d [39]. The domain and mesh are shown in Figure 7.14. The initial data is equal to 0 for electric field components E_x and E_y. The initial data H_z^0 is plotted in Figure 7.16. The Gauss method with $s = 3$ was applied after discretizing in space by using the dG method with upwind flux. An order reduction occurs as we can see in Figure 7.15. The level here indicates the mesh refinement.

Numerical solution was computed with $\tau = 0.1$ for the dG method with $k = 2$ and mesh refinement level 2, which gives 26784 degrees of freedom. Snapshots of the solution at certain time steps are shown in Figures 7.17 – 7.20.

Figure 7.15: Convergence of Gauß RK method, $s = 3$, $T = 2$, $k = 2$

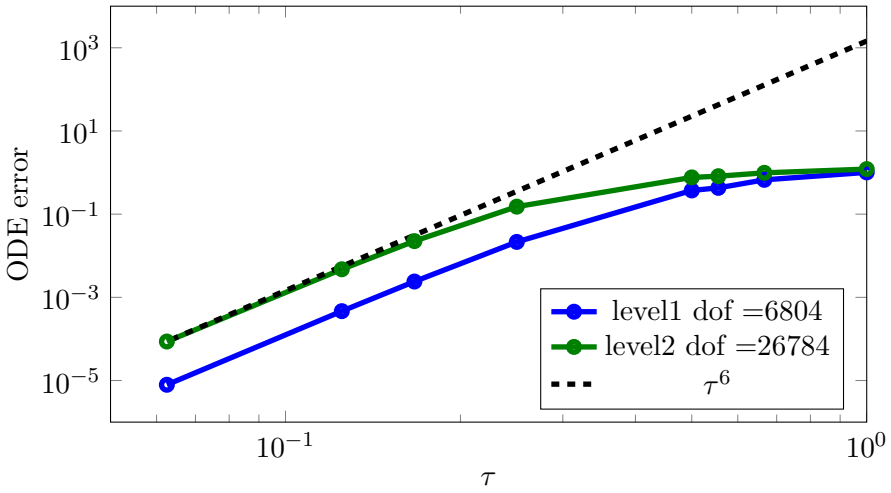


Figure 7.16: Initial data H_z^0

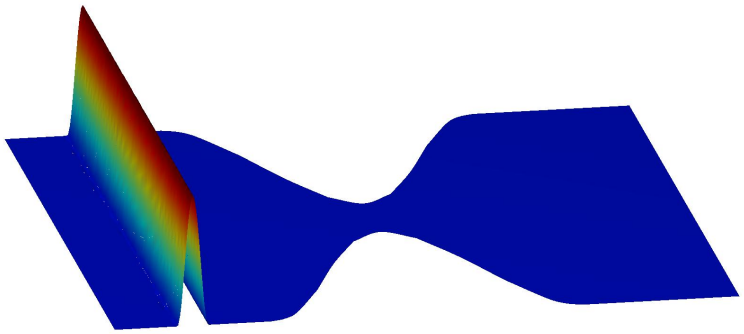


Figure 7.17: H^z component of the solution computed by the 3-stage Gauss method with $\tau = 0.1$ at time $t = 2$

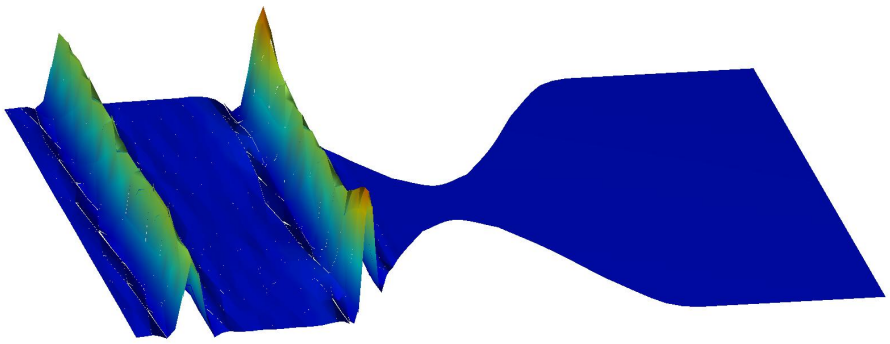


Figure 7.18: H^z component of the solution computed by the 3-stage Gauss method with $\tau = 0.1$ at time $t = 4$

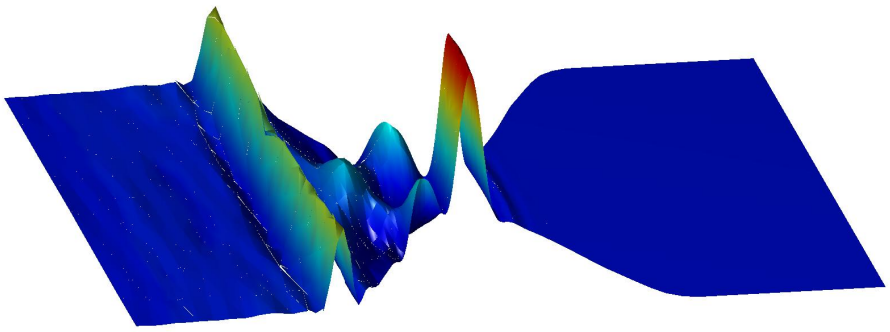


Figure 7.19: H^z component of the solution computed by the 3-stage Gauss method with $\tau = 0.1$ at time $t = 6$

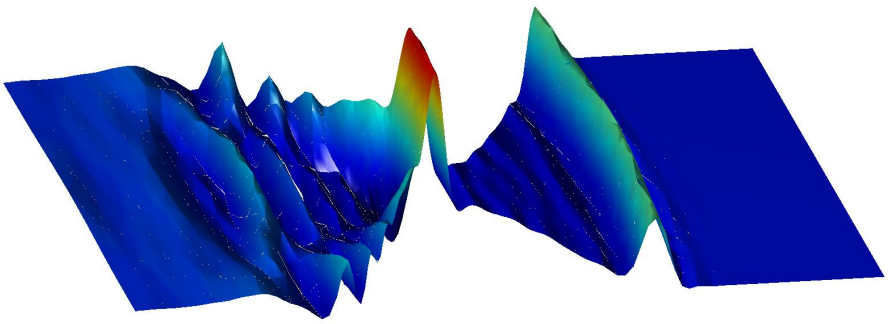
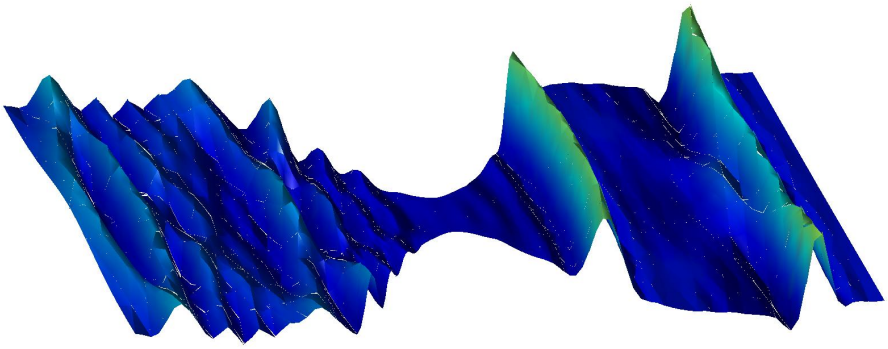


Figure 7.20: H^z component of the solution computed by the 3-stage Gauss method with $\tau = 0.1$ at time $t = 8$



7.3 The inhomogeneous case

The implementation of the 3-stage Gauss method that we presented in the homogeneous case does not work for the inhomogeneous case. The Butcher tableau reads

$$\begin{aligned}\mathcal{A} &= \begin{pmatrix} \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\ \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\ \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \end{pmatrix}, \\ b &= \begin{pmatrix} \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{pmatrix}^T, \\ c &= \begin{pmatrix} \frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{1}{2} & \frac{1}{2}+\frac{\sqrt{15}}{10} \end{pmatrix}^T.\end{aligned}$$

We give the main idea of the implementation for a linear system of ordinary differential equations of the form

$$\mathbf{u}'(t) + \mathbf{L}u(t) = \mathbf{f}(t), \quad \mathbf{u}(0) = \mathbf{u}_0,$$

where $\mathbf{u}, \mathbf{f} : \mathbb{R}_{\geq 0} \to \mathbb{R}^N$, $A \in \mathbb{R}^{N\times N}$. A Runge–Kutta method applied to this system can be written in compact form (similar as in Chapter 5.1) as

$$\begin{aligned}(I + \tau\mathcal{A}\otimes\mathbf{L})\mathbf{U}_n &= \mathbb{1}\otimes\mathbf{u}_n + \tau(\mathcal{A}\otimes I_N)\mathbf{F}(t), \\ \mathbf{u}_{n+1} &= \mathbf{u}_n + \tau(b^T\otimes\mathbf{L})\mathbf{U}_n + \tau(b^T\otimes I_N)\mathbf{F}_n,\end{aligned}$$

where $\mathbf{F}_n$ is defined by

$$\mathbf{F}_n := (\mathbf{f}(t_n + c_1\tau), \mathbf{f}(t_n + c_2\tau), \mathbf{f}(t_n + c_3\tau))^T.$$

One can show that this is equivalent to

$$\begin{aligned}\left(\frac{1}{\tau}\mathbf{\Lambda}\otimes I_N - I_s\otimes\mathbf{L}\right)\mathbf{W}_n &= (T^{-1}\otimes\mathbf{L})\big(\mathbb{1}\otimes\mathbf{u}_n + \tau(\mathcal{A}\otimes I_N)\mathbf{F}_n\big), \\ \mathbf{u}_{n+1} &= \mathbf{u}_n + (b^T\mathbf{T}\mathbf{\Lambda}\otimes I_N)\mathbf{W}_n + \tau(b^T\otimes I_N)\mathbf{F}_n,\end{aligned} \tag{7.6}$$

where

$$\mathbf{T}^{-1}\mathbf{L}^{-1}\mathbf{T} = \mathbf{\Lambda} := \begin{pmatrix} \gamma & & \\ & \alpha+i\beta & \\ & & \alpha-i\beta \end{pmatrix}.$$

To calculate the numerical solution at the next time step, we have to solve 3 linear systems of dimension $N\times N$. This implementation obviously applies to the homogeneous case too.

In the remark after Theorem 5.27 we stated that appropriate assumptions on the solution u of the problem

$$u'(t) + Au(t) = f(t), \quad u(0) = u_0,$$

where A is the generator of bounded C_0-semigroup are of the form $u^{(q)} \in D(A)$, for some $q \geq 0$, and that it is not plausible to request $u^{(q)} \in D(A^p)$ for $p \geq 2$ and $q \geq 0$.

We now give two numerical examples to justify this statement. In both examples we consider the TM polarization of Maxwell's equations on $[-1,1]^2$ defined in (7.2) and (7.3). The material is supposed to be homogeneous and the final time of the simulation is $T = 1$.

7.3.1 Example I

The function $u = (H_x, H_y, E_z)$ defined by

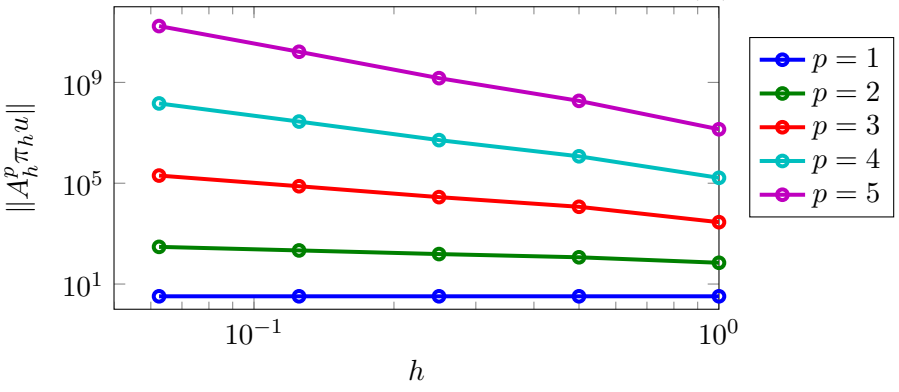
$$\begin{aligned} H_x(t,x,y) &:= (y^2-1)\cos(\omega t), \\ H_y(t,x,y) &:= (x^2-1)\cos(\omega t), \\ E_z(t,x,y) &:= \sin(\pi x)\sin(\pi y)\sin(\omega t) \end{aligned}$$

is a solution of (7.2) with

$$\begin{aligned} f_x &= (-\omega(y^2-1) + \pi\sin(\pi x)\cos(\pi y))\sin(\omega t), \\ f_y &= (-\omega(x^2-1) - \pi\cos(\pi x)\sin(\pi y))\sin(\omega t), \\ f_z &= (\sin(\pi x)\sin(\pi y) + 2y - 2x)\cos(\omega t). \end{aligned}$$

Also, it is easy to see that $u(t) \in \mathcal{D}(A_{TM})$ for all times t. Moreover, $u^{(q)}(t) \in D(A_{TM})$ for all times t, since u can be written in a form with separated spatial and time variables. This is often the case in partial differential equations and therefore assumptions of the form $u^{(q)} \in D(A)$, for some $q \geq 0$, are appropriate.

Figure 7.21: Example I, $k = 6$, solution is in $D(A)$ only



On the other hand, $A_{TM}u(t) \notin D(A_{TM})$, i. e. $u(t) \notin D(A_{TM}^2)$. In general, if $u'(t) \in D(A)$ holds, requesting $u(t) \in D(A^2)$ implies by using

$$Au(t) = f(t) - u'(t),$$

that $f(t) \in D(A)$, which must not be the case. Analogously, for $u^{(q)} \in D(A^2)$ we need $f^{(q)} \in D(A^2)$. For $u^{(q)} \in D(A^p)$, $p \geq 2$ we need even more spatial regularity of f. In physically interesting problems f is most often just in $L^2(\Omega)$ and therefore it is not realistic to assume $u^{(q)} \in D(A^p)$ for $p \geq 2$ and $q \geq 0$.

In Figure 7.21 we can see that norm of $A_h \pi_h u$, which is a discrete version of $A_{TM} u$, stays bounded when $h \to 0$. On the other hand $A_h^2 \pi_h u$ and higher powers of A_h grow as $h \to 0$. We investigate the convergence of the implicit midpoint rule. Figure 7.22 shows that it converges with order 2 which is expected. The convergence is independent of the meshsize h.

Figure 7.22: Convergence of implicit MP rule, $k = 3$

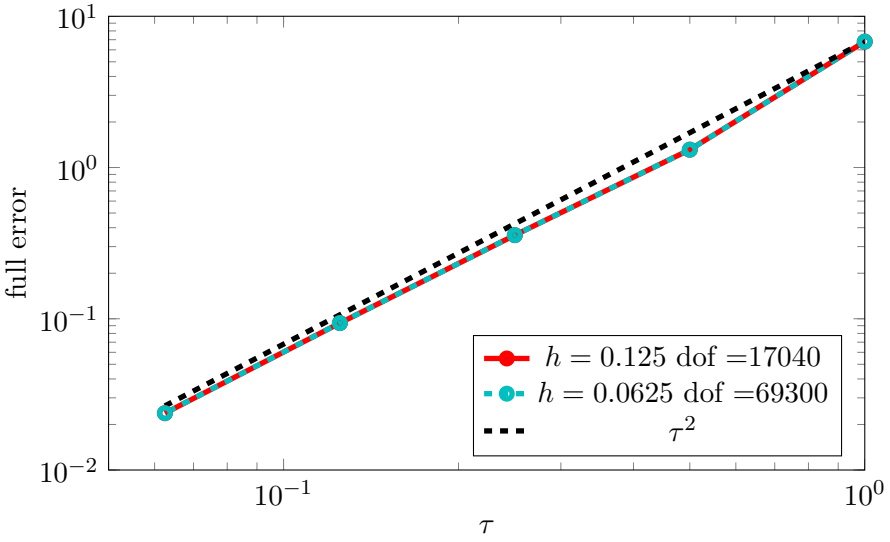


Figure 7.23: Convergence of Gauß RK method, $s = 3$, $h = 0.125$

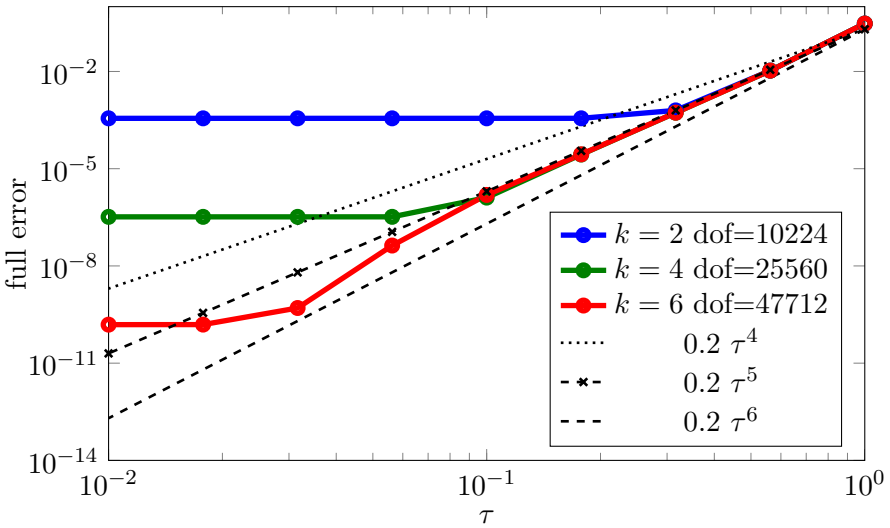
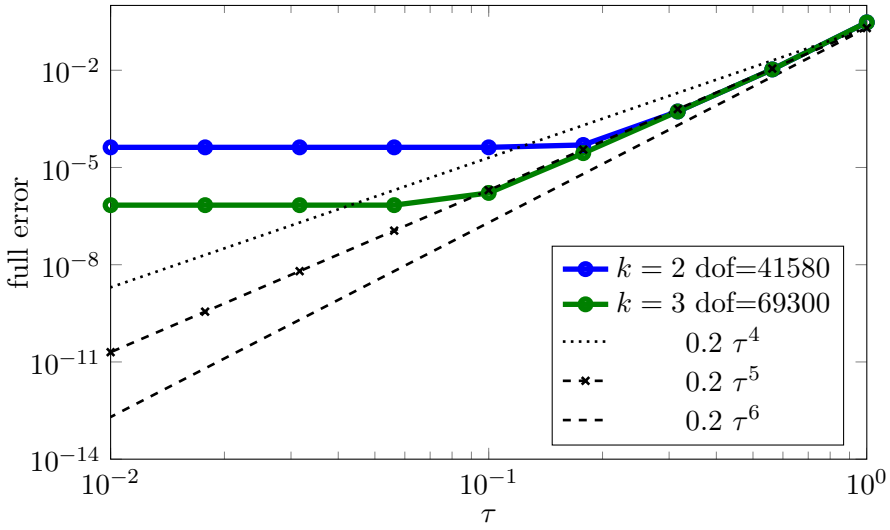


Figure 7.24: Convergence of Gauß RK method, $s = 3$, $h = 0.0625$

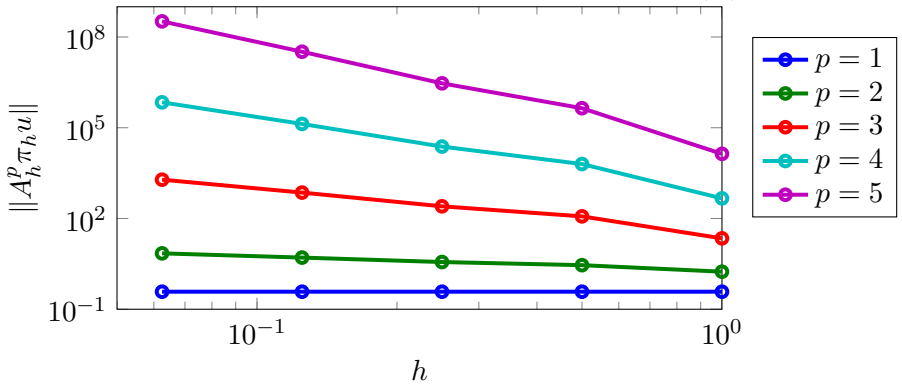


We also investigate the time integration error of 3-stage Gauss collocation methods. As Figures 7.23 and 7.24 show, they converge with order 5, although order 4 was expected. This can be explained by the fact that $A_h^2 \pi_h u$ grows very slowly, and for $h = 0.0625$ it is not large enough to cause additional order reduction.

7.3.2 Example II

We consider the example from Section 7.1. Again the solution $u = (H_x, H_y, E_z)$ satisfies $u^{(q)} \in D(A_{TM})$ for all $q \geq 0$. and $u \notin D(A_{TM}^2)$. Numerical confirmation is given in Figure 7.25.

Figure 7.25: Example II, $k = 6$, solution is in $D(A)$ only



The convergence of the implicit midpoint rule and the 3-stage Gauss method is investigated. The results are shown in Figures 7.26 and 7.27, respectively. They are essentially

the same as in the previous example and can be explained in the same way.

Figure 7.26: Convergence of implicit MP rule, $k = 3$

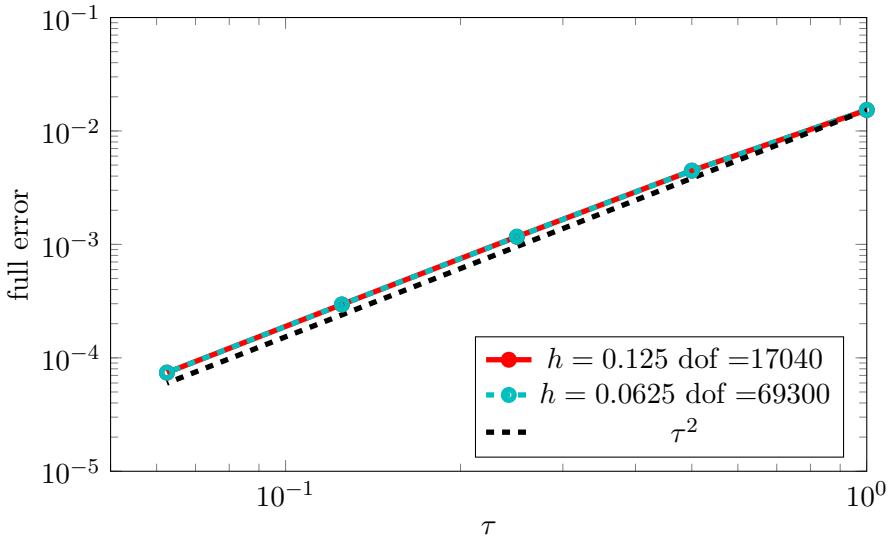
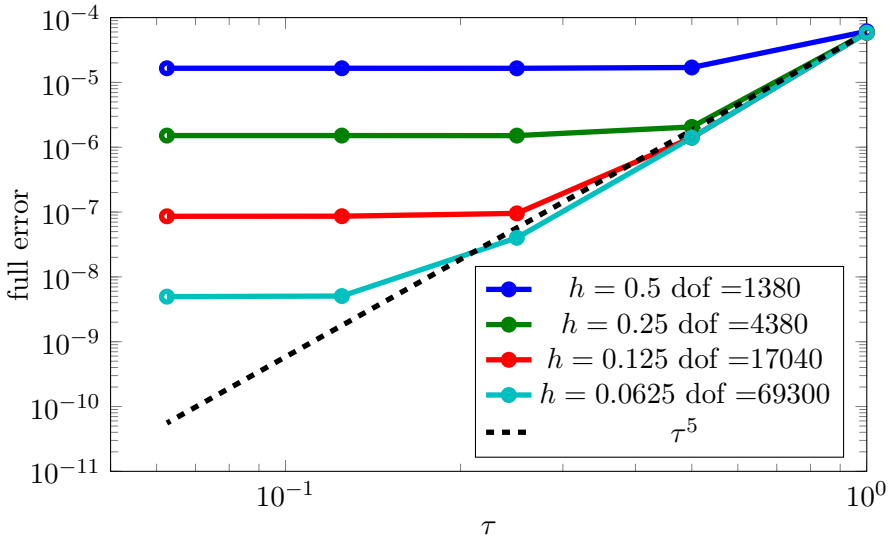


Figure 7.27: Convergence of Gauß RK method, $s = 3$, $k = 3$



7.4 Summary and outlook

In this thesis we provided an error analysis of the following numerical methods for solving Maxwell's equations:

- Gauss and Radau collocation methods for the time discretization
- dG method with both central and upwind flux for the spatial discretization
- full discretization methods with dG combining space and time integration

Our results improve results from earlier work by Brenner, Crouzeix, Thomée for bounded C_0 semigroups [4]. Techniques that we have used are related to work by Lubich, Ostermann for parabolic problems [47].

Our future work is to investigate related nonlinear problems and to generalize our error analysis. Some first steps in the direction of quasilinear Maxwell's equations have already been done. More numerical experiments, especially comparison of large scale problems using a parallel implementation are planed as well.

Bibliography

[1] I. Alonso-Mallo. Explicit single step methods with optimal order of convergence for partial differential equations. *Appl. Numer. Math.*, 31(2):117–131, 1999.

[2] I. Alonso-Mallo. Rational methods with optimal order of convergence for partial differential equations. *Appl. Numer. Math.*, 35(4):265–292, 2000.

[3] A. Bátkai, B. Farkas, P. Csomós, and A. Ostermann. Operator semigroups for numerical analysis. Technical report, 15th Internet Seminar on Evolution Equations, 2012.

[4] P. Brenner, M. Crouzeix, and V. Thomée. Single step methods for inhomogeneous linear differential equations in Banach space. *RAIRO - Analyse numérique*, 12(1):5–26, 1982.

[5] P. Brenner and V. Thomée. On rational approximations of semigroups. *SIAM J. Numer. Anal.*, 16(4):683–694, 1979.

[6] P. Brenner and V. Thomée. On rational approximations of groups of operators. *SIAM J. Numer. Anal.*, 17(1):119–125, 1980.

[7] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[8] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[9] E. Burman, A. Ern, and M. A. Fernández. Explicit Runge–Kutta schemes and finite elements with symmetric stabilization for first-order linear pde systems. *SIAM J. Numerical Analysis*, 48(6):2019–2042, 2010.

[10] M.-H. Chen, B. Cockburn, and F. Reitich. High-order RKDG methods for computational electromagnetics. *J. Sci. Comput.*, 22/23:205–226, 2005.

[11] B. Cockburn, B. Dong, and J. Guzmán. Optimal convergence of the original DG method for the transport-reaction equation on special meshes. *SIAM J. Numer. Anal.*, 46(3):1250–1265, 2008.

[12] 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. *Math. Comp.*, 54(190):545–581, 1990.

[13] B. Cockburn, S. Y. Lin, and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. III. One-dimensional systems. *J. Comput. Phys.*, 84(1):90–113, 1989.

[14] B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. *Math. Comp.*, 52(186):411–435, 1989.

[15] B. Cockburn and C.-W. Shu. The Runge-Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws. *RAIRO Modél. Math. Anal. Numér.*, 25(3):337–361, 1991.

[16] B. Cockburn and C.-W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws. V. Multidimensional systems. *J. Comput. Phys.*, 141(2):199–224, 1998.

[17] S. Descombes, S. Lanteri, and L. Moya. Locally implicit time integration strategies in a discontinuous Galerkin method for Maxwell's equations. *J. Sci. Comput.*, 56(1):190–218, 2013.

[18] V. Dolean, H. Fahs, L. Fezoui, and S. Lanteri. Locally implicit discontinuous Galerkin method for time domain electromagnetics. *J. Comput. Phys.*, 229(2):512–526, 2010.

[19] W. Dörfler, A. Lechleiter, and M. Plum. *Photonic Crystals: Mathematical Analysis and Numerical Approximation.* Oberwolfach Seminars, 42. Birkhauser Verlag GmbH, 2011.

[20] E. Emmrich. Discrete versions of Gronwall's lemma and their application to the numerical analysis of parabolic problems. *Preprint No. 637, Fachbereich Mathematik, TU Berlin*, 1999.

[21] K.-J. Engel and R. Nagel. *One-parameter semigroups for linear evolution equations*, volume 194 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 2000. With contributions by S. Brendle, M. Campiti, T. Hahn, G. Metafune, G. Nickel, D. Pallara, C. Perazzoli, A. Rhandi, S. Romanelli and R. Schnaubelt.

[22] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[23] L. Fezoui, S. Lanteri, S. Lohrengel, and S. Piperno. Convergence and stability of a discontinuous Galerkin time-domain method for the 3D heterogeneous Maxwell equations on unstructured meshes. *M2AN Math. Model. Numer. Anal.*, 39(6):1149–1176, 2005.

[24] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev*, 43:89–112, 2001.

[25] M. J. Grote, A. Schneebeli, and D. Schötzau. Interior penalty discontinuous Galerkin method for Maxwell's equations: energy norm error estimates. *J. Comput. Appl. Math.*, 204(2):375–386, 2007.

[26] M. J. Grote, A. Schneebeli, and D. Schötzau. Interior penalty discontinuous Galerkin method for Maxwell's equations: optimal L^2-norm error estimates. *IMA J. Numer. Anal.*, 28(3):440–468, 2008.

[27] E. Hairer, C. Lubich, and G. Wanner. Geometric numerical integration illustrated by the Störmer/Verlet method. *Acta Numerica*, 12:399–450, 2003.

[28] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 2nd edition, 1993.

[29] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 2nd edition, 1996.

[30] R. Hersh and T. Kato. High-accuracy stable difference schemes for well-posed initial value problems. *SIAM J. Numer. Anal.*, 16(4):670–682, 1979.

[31] J. S. Hesthaven and T. Warburton. Nodal high-order methods on unstructured grids. I. Time-domain solution of Maxwell's equations. *J. Comput. Phys.*, 181(1):186–221, 2002.

[32] J. S. Hesthaven and T. Warburton. High-order nodal discontinuous Galerkin methods for the Maxwell eigenvalue problem. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 362(1816):493–524, 2004.

[33] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. Algorithms, analysis, and applications.

[34] M. Hochbruck. *Skriptum zur Vorlesung Numerik I-IV*. Sommersemester 2010 – Wintersemester 2011/12. Arbeitsgruppe Numerik, Karlsruher Institut für Technologie, 2012.

[35] M. Hochbruck. *Skriptum zur Vorlesung Innovative Integratoren für Evolutionsgleichungen.* Wintersemester 2012/13. Arbeitsgruppe Numerik, Karlsruher Institut für Technologie, 2013.

[36] M. Hochbruck, T. Jahnke, and R. Schnaubelt. Convergence of an ADI splitting for Maxwell's equations. Technical report, Karlsruhe Institute of Technology, 2013.

[37] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numer.*, 19:209–286, 2010.

[38] M. Hochbruck and T. Pažur. Implicit Runge–Kutta methods and discontinuous Galerkin discretizations for linear maxwell's equations. Technical report, Karlsruhe Institute of Technology, 2013.

[39] M. Hochbruck, T. Pažur, A. Schulz, E. Thawinan, and C. Wieners. Efficient time integration for discontinuous Galerkin approximations of linear wave equations. Technical report, Karlsruhe Institute of Technology, 2013.

[40] D. Hundertmark, M. Meyries, L. Machinek, and R. Schnaubelt. Operator semigroups and dispersive equations, 16th internet seminar on evolution equations. Technical report, Karlsruhe Institute for Technology, 2013.

[41] J. D. Jackson. *Classical Electrodynamics Third Edition.* Wiley, third edition, Aug. 1998.

[42] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46(173):1–26, 1986.

[43] A. Kirsch. Mathematical theory of Maxwell's equations. Lecture notes, 2009.

[44] P. Lasaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical aspects of finite elements in partial differential equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974)*, pages 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.

[45] R. J. LeVeque. *Finite volume methods for hyperbolic problems.* Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002.

[46] C. Lubich and A. Ostermann. Runge-Kutta approximation of quasi-linear parabolic equations. *Math. Comp.*, 64(210):601–627, 1995.

[47] C. Lubich and A. Ostermann. Runge-Kutta approximation of quasi-linear parabolic equations. *Mathematics of computation*, 64(210):601–628, 1995.

[48] P. Monk. *Finite element methods for Maxwell's equations.* Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003.

[49] L. Moya. Temporal convergence of a locally implicit discontinuous Galerkin method for Maxwell's equations. *ESAIM Math. Model. Numer. Anal.*, 46(5):1225–1246, 2012.

[50] J. Niegemann. *Higher-Order Methods for Solving Maxwell's Equations in the Time-Domain*. PhD thesis, Department of Physics, Karlsruhe Institute of Technology, 2009.

[51] A. Ostermann and M. Roche. Runge-Kutta methods for partial differential equations and fractional orders of convergence. *Math. Comp.*, 59(200):403–420, 1992.

[52] A. Ostermann and M. Roche. Rosenbrock methods for partial differential equations and fractional orders of convergence. *SIAM J. Numer. Anal.*, 30(4):1084–1098, 1993.

[53] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Number v. 44 in Applied Mathematical Sciences. Springer, 1992.

[54] T. E. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM J. Numer. Anal.*, 28(1):133–140, 1991.

[55] D. A. D. Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Springer-Verlag Berlin Heidelberg, 2012.

[56] S. Piperno. Symplectic local time-stepping in non-dissipative DGTD methods applied to wave propagation problems. *M2AN Math. Model. Numer. Anal.*, 40(5):815–841 (2007), 2006.

[57] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

[58] G. R. Richter. An optimal-order error estimate for the discontinuous Galerkin method. *Math. Comp.*, 50(181):75–88, 1988.

[59] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.

[60] J. M. Sanz-Serna, J. G. Verwer, and W. H. Hundsdorfer. Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations. *Numer. Math.*, 50(4):405–418, 1987.

[61] C. Schwab. *p- and hp-finite element methods*. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 1998. Theory and applications in solid and fluid mechanics.

[62] A. Taflove and S. Hagness. *Computational Electrodynamics: The Finite-Difference Time-Domain Method*. The Artech House antenna and propagation library. Artech House, Incorporated, 2005.

[63] B. Wang, Z. Xie, and Z. Zhang. Error analysis of a discontinuous Galerkin method for Maxwell equations in dispersive media. *J. Comput. Phys.*, 229(22):8552–8563, 2010.

[64] T. Warburton and J. S. Hesthaven. On the constants in *hp*-finite element trace inverse inequalities. *Comput. Methods Appl. Mech. Engrg.*, 192(25):2765–2773, 2003.

[65] K. S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans. Antennas and Propagation*, pages 302–307, 1966.

[66] F. Zhen, Z. Chen, and J. Zhang. Toward the development of a three-dimensional unconditionally stable finite-difference time-domain method. *IEEE Transactions on Microwave Theory and Techniques*, 48(9):1550–1558, Sep 2000.