The Peaceman–Rachford ADI-dG method for linear wave-type problems

Zur Erlangung des akademischen Grades eines

DOKTOR DER NATURWISSENSCHAFTEN

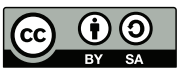
von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
Jonas Köhler

Tag der mündlichen Prüfung: 26. September 2018

Referentin:	Prof. Dr. Marlis Hochbruck
Korreferent:	Prof. Dr. Roland Schnaubelt



Acknowledgment

In German.

An dieser Stelle möchte ich mich bei all jenen bedanken, die dazu beigetragen haben, dass das Projekt Promotion einen erfolgreichen Abschluss finden konnte.

Mein ganz besonderer Dank gilt hierbei meiner Betreuerin Prof. Dr. Marlis Hochbruck. Bereits seit dem Grundstudium profitiere ich nun schon von ihrer Art, Mathematik (und insbesondere Numerik) anschaulich und klar verständlich zu vermitteln, was sicher maßgeblich zu meinem Interesse an numerischer Mathematik beigetragen hat. Auch durch ihre Betreuung während meiner Diplomarbeit wurde dieses Interesse weiter verfestigt, was sicher ein wesentlicher Beitrag zu der Entscheidung war, die Promotion überhaupt in Betracht zu ziehen. Für die hervorragende Betreuung, die vielen ertragreichen Diskussionen und die große Freiheit, in der Art und Weise zu arbeiten, die für mich am besten war, möchte ich mich herzlich bedanken.

Prof. Dr. Roland Schnaubelt möchte ich für das Übernehmen der Korreferenz bedanken. Insbesondere der analytische Teil der Arbeit hätte ohne seine Anmerkungen und Korrekturen sicherlich nicht die Qualität erreicht, die er nun aufweist.

Weiterhin möchte ich meiner Arbeitsgruppe mit all ihren aktuellen und ehemaligen Mitgliedern für die wirklich hervorragende Arbeitsatmosphäre danken. Besonderer Dank geht auch an meine Korrekturleser David, Andreas und Johannes, die viel Zeit investiert haben und mit ihren Anmerkungen und Verbesserungsvorschlägen maßgeblich zur Güte dieser Arbeit beigetragen haben.

Vielen Dank auch an meine Freundin Bettina für die Unterstützung und das Verständnis während der heißen Phase der Promotion. Besonders möchte ich auch Michael und Steffi danken, ohne die die letzten fast $2\,1/2$ Jahre nicht annähernd so gut gewesen wären, wie sie es waren. Auch Nadja, Daniel B. und Sarah, Philip und Tine, Dominique, Daniel S. und allen, die ich jetzt leider vergesse, möchte ich herzlich für ihre Freundschaft danken.

Zuletzt gebührt mein herzlichster Dank natürlich meiner Familie, die mich stets bedingungslos und in jeder denkbaren Form unterstützt hat. Ohne diese Unterstützung wäre mein Studium und die Promotion in der Art und Weise sicher nicht möglich gewesen.

Contents

1 | Introduction

Wave-type problems, such as Maxwell's equations or the acoustic wave equation, play an important role in the description of physical processes. For instance, Maxwell's equations lay the foundation of the field of classical electromagnetism, as they describe the interaction of time-dependent electromagnetic fields with each other and their behavior in different materials. Further examples for such problems are the elastic wave equation or advection-reaction equation.

Because of their widespread applications and importance, such problems have been intensively studied in the past. Despite that, solving them is still a challenging task, especially since analytical solutions can only be found in very few cases. Hence, one usually tackles such problems with the help of numerical simulations. This has led to a wide variety of algorithms, which can be used to approximately solve the wave-type problem under consideration.

In practice, temporal discretization is often achieved via explicit time stepping methods like the leapfrog scheme. Such explicit methods are very popular, see, e.g., [Fahs, 2009, Burman et al., 2010, Diehl et al., 2010] for Maxwell's equations. Their advantage is that they are easy to implement and that one step can be performed at very low cost. However, they might be inefficient if applied to stiff problems, which usually arise from the spatial discretization of partial differential equations. This inefficiency is due to the CFL condition [Courant et al., 1928], which gives a limit on the timestep size under which the scheme is stable. This can cause explicit methods to become impractical as a huge number of timesteps can be necessary to ensure stability, often times many more steps than needed to achieve the desired accuracy.

Implicit schemes like the Crank–Nicolson scheme or implicit Runge-Kutta methods pose alternatives to explicit time integration. Their main advantage is that they can be unconditionally stable, yielding a timestep restriction that is only governed by accuracy, not by stability. However, such schemes usually require the solution of huge linear systems, which can severely limit their efficiency.

Different approaches have been worked out in the last decades to relieve or completely get rid of the CFL condition, while still ending up with an efficient scheme. For instance, the CFL condition can become problematic if the computational domain involves small scale features. In this case, one has to use small elements in the spatial discretization to resolve the geometry, which worsens the CFL condition. However, if only a few elements have to be small to resolve the geometry, local time stepping [Diaz and Grote, 2009, Grote and Mitkova, 2010] or locally implicit [Piperno, 2006, Verwer, 2011, Descombes et al., 2013, Hochbruck and Sturm, 2016, Hochbruck and Sturm, 2018] methods are suitable to relieve the stability issues.

If the computational domain is comprised of rectangular or cuboidal domains and the considered problem admits a certain structure, an **alternating direction implicit (ADI)** scheme can be the method of choice. This is because under these conditions, ADI schemes are unconditionally stable, despite being of roughly the cost of an explicit scheme. The original ADI scheme was introduced by Donald W. Peaceman and Henry H. Rachford, Jr. in [Peaceman and Rachford, 1955]. Their idea was to split the spatial differential operator of a partial differential equation w.r.t. the direction of the occurring partial derivatives, resulting in a dimension splitting. If this splitting is applied to problems of the aforementioned structure, the resulting split operators lead to subproblems that can be tackled more easily by time integration schemes. More precisely, the idea of ADI methods is to approximate these subproblems in an alternating fashion, whereby in each step one subproblem is treated in an implicit and the other in an explicit way.

The original ADI method was proposed for a finite differences discretization of the two-dimensional heat equation. However, the concept is quite general and can be applied to other problems as well, most notably to the full three-dimensional linear Maxwell's equations. This method, known as the finite differences time domain alternating direction implicit (FDTD-ADI), was independently proposed by Takefumi Namiki in [Namiki, 1999] and Fenghua Zheng, Zhizhang Chen and Jiazong Zhang in [Zhen et al., 2000]. However, a rigorous error analysis of this scheme is still an open problem.

The numerical scheme that is applied to the split problem in [Peaceman and Rachford, 1955] to propagate in time is nowadays known as the **Peaceman–Rachford** scheme. However, it is possible to combine the ADI splitting with other time stepping schemes for split problems like the Douglas ADI method [Hundsdorfer and Verwer, 2003, Douglas, 1962, Douglas and Gunn, 1964, Brian, 1961] or variants of the Peaceman–Rachford scheme [Chen et al., 2008, Chen et al., 2010, Gao et al., 2007, Gao et al., 2013, Lee and Fornberg, 2003, Lee and Fornberg, 2004]. In this thesis, we focus on the original Peaceman–Rachford scheme.

Recently, the Peaceman–Rachford ADI scheme applied to Maxwell's equations has been analyzed in the context of abstract Cauchy problems in a series of papers. The first of these papers [Hochbruck et al., 2015a] considered Maxwell's equations in the absence of external currents and damping, while [Eilinghoff and Schnaubelt, 2018, Eilinghoff and Schnaubelt, 2017] included both. In [Eilinghoff et al., 2018], an energy preserving variant of the scheme was analyzed. These results provide a first step towards a rigorous analysis of a fully discrete method achieved by using the Peaceman–Rachford scheme in time.

Independent from these contributions, in [Hansen and Henningsson, 2016], a fully discrete scheme obtained by the Peaceman–Rachford scheme (and the Douglas–Rachford scheme) combined with a generic spatial discretization was analyzed. However, the results therein were derived under assumptions on the spatially discrete operators, which can be tedious to verify in applications if they are fulfilled at all. In contrast, in this thesis, we only pose assumptions on the regularity of the exact solution.

Aims and results

This thesis has three main goals.

1. The construction of a **discontinuous Galerkin (dG)** discretization, which enables an efficient implementation of the Peaceman–Rachford ADI scheme for Maxwell's equations.
2. The derivation of rigorous error bounds for the resulting full discretization.
3. The generalization of the results achieved in the first two goals to a general class of wave-type problems.

The first two goals were the original starting point of the work presented in this thesis. However, it turned out that it is possible to generalize the ADI method to a broader class of problems, which led to the third goal and thereby to the results presented in this thesis.

To achieve the first goal, we have discretized the split operators occurring in the FDTD-ADI method by a dG discretization. Consequently, we have studied the structure of these discrete operators to work out the conditions under which an efficient implementation of the dG-Peaceman–Rachford ADI scheme is possible. We have further identified how such an implementation can be achieved by using two different orderings of the degrees of freedom. In fact, the crucial ingredient for an efficient scheme is to exploit the special structure of Maxwell's equations and the tensorial structure of the computational domain and the chosen grid. This ensures that the flows of the split operators completely decouple as they travel along different directions in the grid.

For the second goal we have exploited that the Peaceman–Rachford scheme can be interpreted as a perturbation of the Crank–Nicolson method. In [Sturm, 2017] techniques to analyze the Crank–Nicolson, leapfrog and a locally implicit scheme applied to dG discretizations of Maxwell's equations were developed. As the leapfrog and the locally implicit schemes are treated as perturbations of the Crank–Nicolson scheme, we have transferred these ideas to the Peaceman–Rachford scheme. However, it turned out that the perturbation caused by the latter can not be treated with the same arguments used in [Sturm, 2017]. We have therefore worked out new techniques that can be used to analyze perturbed Crank–Nicolson schemes including the Peaceman–Rachford method.

The third goal was motivated by the observation that the ADI scheme is efficient for Maxwell's equations because of their special structure. By identifying this structure, we were able to pose precise conditions on more general wave-type problems for which the Peaceman–Rachford scheme can be applied at roughly the cost of an explicit scheme. These problems are such that we can split the corresponding spatial operators into two operators, whose associated flows completely decouple. This enables us to directly transfer the efficient implementation of the dG-Peaceman–Rachford scheme for Maxwell's equations to such problems.

To generalize our error analysis we have developed a Hilbert space framework applicable to a broad class of wave-type problems. This framework is based on the theory of Friedrichs' systems, which is originally due to Kurt Otto Friedrichs [Friedrichs, 1958] and was recently refined by Daniele A. Di Pietro, Alexandre Ern and Jean-Luc Guermond in [Ern and Guermond, 2006a, Ern and Guermond, 2006b, Ern and Guermond, 2008, Di Pietro and Ern, 2012]. However, the results in these publications are mostly given for stationary problems or in a space-time framework, which is why we can not directly apply them to our

setting. We have overcome this problem by using the aforementioned theory to work out the conditions under which the spatial operator is maximal dissipative and therefore generates a contraction semigroup by the Lumer–Phillips theorem. To do so, special treatment has to be given to the material parameters belonging to the temporal derivative, which we incorporate in the inner product of the Hilbert space in which we analyze the abstract problem. Similar ideas for a unified theory for wave-type equations can be found, e.g., in [Benzoni-Gavage and Serre, 2007] and [Burazin and Erceg, 2016]. In fact, the latter is also based on the framework found in the publications of Di Pietro, Ern and Guermond.

Outline

The thesis is organized as follows. In Chapter 2 we specify the class of wave-type problems considered in this thesis. To this end, we study spatial operators called Friedrichs' operators, which govern the temporal evolution of the solution of these problems. After having shown some crucial properties of Friedrichs' operators, we proceed by applying semigroup theory to show the wellposedness of the corresponding wave-type problems. As we are mainly interested in the analysis of the Peaceman–Rachford method, which is a splitting method, we then briefly discuss suitable splittings of a Friedrichs' operator. In Chapter 3 we derive the central flux dG discretization of a general Friedrichs' operator. Then, using these discrete operators, we are able to state the spatially semidiscrete problem and show its wellposedness. We conclude the chapter by showing bounds on the error of the spatial semidiscretization. In Chapter 4 we consider the temporal discretization of abstract wave-type problems. Our analysis of the Peaceman–Rachford method is based upon the fact that it can be considered as a perturbation of the Crank–Nicolson scheme. Hence, as a first step, we study the Crank–Nicolson method, in particular its wellposedness and stability. We then proceed accordingly for the Peaceman–Rachford scheme. Consequently, we perform the temporally semidiscrete analysis of both schemes. Having introduced both spatial as well as temporal semidiscretizations we then use a method of lines approach to obtain fully discrete schemes in Chapter 5. This chapter is structured similarly to Chapter 4, meaning that we start by showing wellposedness and stability of both the dG-Crank–Nicolson as well as the dG-Peaceman–Rachford method. Subsequently, we state the first main result of this thesis, namely we give rigorous bounds on the full error of the dG-Peaceman–Rachford scheme. Chapter 6 is devoted to the second main result, i.e., the construction and implementation of the dG-Peaceman–Rachford method in the context of an ADI splitting. In particular, we identify a class of Friedrichs' operators for which the corresponding wave-type problems can be tackled extremely efficiently. Namely, despite the fact that the dG-Peaceman–Rachford method is an implicit method, we show that we can perform one step of the scheme at roughly the cost of an explicit scheme if applied to the aforementioned problems. We conclude the thesis by showing the results of some numerical experiments in Chapter 7 to back up the theoretical results.

Notation

In this section, we introduce the notation used throughout the thesis. To this end, let $n, k \in \mathbb{N}$ and $K \subset \mathbb{R}^d$ be open.

Miscellaneous

By $\mathbb{R}_+$ we denote the non-negative real numbers, i.e., the interval $[0, \infty)$. We write $\mathbb{N}_0$ for the natural numbers including 0.

Throughout, $d \in \mathbb{N}$ denotes the spatial dimension, and $m \in \mathbb{N}$ is a generic positive integer, usually being the number of components of vector-valued functions.

The support of a function f is denoted by $\operatorname{supp} f$. We denote the Kronecker delta by δ_{ij}, and we denote the indicator function of a set $S \subset \mathbb{R}^n$ as $\mathbb{1}_S$.

Given a countable set S with finitely many elements, we denote the cardinality of this set by $|S|$. Further, we call a set $S \subset \mathbb{R}^n$ tensorial if we have $\overline{S} = \bigtimes_{i=1}^n [a_i^-, a_i^+]$, for $a_i^-, a_i^+ \in \mathbb{R}$, $i = 1, \dots, n$.

Vector algebra

Given a vector $a \in \mathbb{R}^n$, we denote the components of a by $a_1, \dots, a_n$, its Euclidean norm by $\|a\|$ and the transpose of a by a^T. The canonical unit vectors in $\mathbb{R}^n$ are denoted by $e_1, \dots, e_n$. We denote the Euclidian scalar product of two vectors $a, b \in \mathbb{R}^n$ by

$$a \cdot b = \sum_{i=1}^n a_i b_i.$$

For $a, b \in \mathbb{R}^3$, we denote the cross product of a and b by

$$a \times b = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{pmatrix}.$$

Spatial and temporal derivatives

We often consider multivariate functions $u \colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$, where the first variable is the time variable $t \in \mathbb{R}_+$, and the vector-valued variable is the space variable $x \in \mathbb{R}^d$. The partial derivatives of u are denoted by

$$\partial_t u = \frac{\partial}{\partial t} u, \qquad \partial_1 u = \frac{\partial}{\partial x_1} u, \qquad \dots, \qquad \partial_d u = \frac{\partial}{\partial x_d} u.$$

We collect the spatial derivatives in the gradient of u given by

$$\nabla u = \begin{pmatrix} \partial_1 u \\ \vdots \\ \partial_d u \end{pmatrix}.$$

For vector fields $u \colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^m$ the divergence of u is denoted by

$$\nabla \cdot u = \sum_{i=1}^d \partial_i u_i,$$

and for three-dimensional fields $u\colon \mathbb{R}_+ \times \mathbb{R}^3 \to \mathbb{R}^3$ we denote the curl of u by

$$\nabla\times u = \begin{pmatrix} \partial_2 u_3 - \partial_3 u_2 \\ \partial_3 u_1 - \partial_1 u_3 \\ \partial_1 u_2 - \partial_2 u_1 \end{pmatrix}.$$

We frequently interpret $u\colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^m$ as a function $u\colon \mathbb{R}_+ \to X$, where X is a function space containing functions mapping from $\mathbb{R}^d$ to $\mathbb{R}^m$. Hence, in that case we omit the spatial dependence and consequently write $u(t)$ instead of $u(t,x)$. Further, we then denote the temporal derivative of u by

$$\mathrm{d}_t u = \frac{\mathrm{d}}{\mathrm{d}t} u.$$

Hilbert spaces and operators

Let $(X, \left(\,\cdot\,\middle|\,\cdot\,\right)_X)$ and $(Y, \left(\,\cdot\,\middle|\,\cdot\,\right)_Y)$ be Hilbert spaces over $\mathbb{K} = \{\mathbb{R}, \mathbb{C}\}$.

We denote the dual space of a Hilbert space X by X' and the canonical dual pairing between a Hilbert space and its dual space by $\left\langle \,\cdot\, \middle|\, \cdot\, \right\rangle\colon X' \times X \to \mathbb{K}$.

By $\mathcal{B}(X, Y)$ we denote the set of all bounded operators from X to Y.

We usually denote the domain of an operator $\mathcal{A}$ by $D(\mathcal{A})$. The domain of the concatenation of two linear operators $\mathcal{A}$ and $\mathcal{B}$ is then defined by

$$D(\mathcal{A}\mathcal{B}) = \{v \in D(\mathcal{B}) \mid \mathcal{B}v \in D(\mathcal{A})\}$$

and recursively for more factors, e.g., $\mathcal{A}^n$.

Let $D(\mathcal{A}) \subset X$. Given an operator $\mathcal{A}\colon D(\mathcal{A}) \to Y$, we denote the range of $\mathcal{A}$ by $\operatorname{ran}(\mathcal{A})$ and the kernel or null space of $\mathcal{A}$ by $\ker(\mathcal{A})$.

We usually denote the identity operator on a Hilbert space X by $\mathcal{I}$.

Function spaces

For $p \in \mathbb{N} \cup \{\infty\}$ we denote by $L^p(K)$ the standard Lebesgue spaces of real valued L^p-functions on K and by $W^{k,p}(K)$ the L^p-Sobolev spaces of functions in $L^p(K)$ whose weak derivatives up to order k lie in $L^p(K)$.

For vector-valued L^2-functions $u, v \in L^2(K)^m$ we denote the $L^2(K)$-inner product by

$$\left(u \,\middle|\, v\right)_K = \int_K u \cdot v \,\mathrm{d}x,$$

and for $F \subset \partial K$ we write

$$\left(u \,\middle|\, v\right)_F = \int_F u \cdot v \,\mathrm{d}\sigma.$$

We denote the norms induced by these inner products by $\|\cdot\|_K$ and $\|\cdot\|_F$.

Further, we abbreviate $H^k(K) = W^{k,2}(K)$ and denote the corresponding norms and seminorms on $H^k(K)$ by $\|\cdot\|_{k,K}$ and $|\cdot|_{k,K}$, respectively. The vector-valued case is treated analogously to the $L^2(K)$-norm, i.e., by utilizing the Euclidean scalar product.

Lastly, we denote by $C_c^\infty(K)$ the space of infinitely differentiable functions, which have compact support on K.

Matrices and matrix fields

Let $A \in \mathbb{R}^{n\times n}$ be a square matrix. We denote the spectral norm of A by $\|A\|$, the matrix resulting from taking the absolute value of the components of A by $|A|$ and the transpose of A by A^T.

Inequalities between matrices are understood on the associated quadratic forms. More precisely, for $A_1, A_2 \in \mathbb{R}^{n\times n}$, the inequality $A_1 \leq A_2$ means that for all $a \in \mathbb{R}^n$ we have $a^T A_1 a \leq a^T A_2 a$ and accordingly for $<$, $\geq$ and $>$.

We denote the identity matrix by I.

Let $M \in L^\infty(K)^{n\times n}$ be a square matrix-valued field on K. We denote by

$$\|M\|_{\infty,K} = \operatorname*{ess\,sup}_{x\in K} \|M(x)\|$$

the essential supremum of the spectral norm of M.

We say that M is uniformly positive a.e. on K if there exists $\mu > 0$ such that $M \geq \mu I$ a.e. on K.

Further, slightly abusing notation, we identify a matrix-valued field $M \in L^\infty(K)^{n\times n}$ with the associated bounded linear operator

$$(u \mapsto Mu) \in \mathcal{B}(L^2(K)^n,\, L^2(K)^n).$$

Hence, the adjoint of this operator and the transpose of the matrix-valued field M coincide, and we denote it by M^*.

Discretized objects

Discrete objects, i.e., functions in the approximation spaces introduced in Chapter 3 or discrete operators defined on these spaces are denoted by bold letters. In contrast, objects related to infinite-dimensional spaces (like $L^2(K)$ or $H^1(K)$) are denoted by the standard fonts. Combinations of both such objects (like projection errors) are also written in standard fonts.

2 | Linear wave-type equations

In the following, let $\Omega \subset \mathbb{R}^d$ be a bounded, open and connected Lipschitz domain with boundary $\Gamma = \partial\Omega$. Throughout this thesis, we are interested in solving linear wave-type problems of the following form. Seek $u\colon \mathbb{R}_+ \times \Omega \to \mathbb{R}^m$, such that

$$M(x)\partial_t u(t,x) = \widetilde{\mathcal{L}} u(t,x) + g(t,x) \qquad \text{in } \mathbb{R}_+ \times \Omega. \tag{2.1}$$

Here, $M\colon \Omega \to \mathbb{R}^{m\times m}$ is a symmetric material tensor, and $g\colon \mathbb{R}_+ \times \Omega \to \mathbb{R}^m$ is a source term. Moreover, $\widetilde{\mathcal{L}}$ is a first order spatial differential operator of the form

$$\widetilde{\mathcal{L}} u(t,x) = \sum_{i=1}^{d} L_i(x)\partial_i u(t,x) + L_0(x)u(t,x),$$

where $L_0, \dots, L_d\colon \Omega \to \mathbb{R}^{m\times m}$ are matrix-valued coefficients with $L_1, \dots, L_d$ being symmetric. Since problem (2.1) is an initial value problem on a bounded domain, the equation has to be supplied with suitable boundary and initial conditions to be wellposed. We will specify these conditions in the course of this chapter.

Well-known examples for this equation prototype are the wave equation in first order formulation, Maxwell's equations and the advection equation. These problems will also be the three examples accompanying us throughout this thesis to illustrate our results.

The spatial differential operator $\widetilde{\mathcal{L}}$ belongs to a class of operators introduced in [Friedrichs, 1958]. We will review and use results known for such operators, throughout referred to as Friedrichs' operators. These results mostly stem from the series of papers [Ern and Guermond, 2006a, Ern and Guermond, 2006b, Ern and Guermond, 2008] and from Chapter 7 of the book [Di Pietro and Ern, 2012]. However, in there, either the stationary case or a space-time framework, where the temporal derivative is incorporated in the Friedrichs' operator, is considered. Thus, we transfer them to the transient case studied in this thesis. Similar ideas were already pursued in [Burazin and Erceg, 2016].

The overall aim of this chapter is to investigate wellposedness of problems of the form (2.1) supplied with suitable boundary and initial conditions. As such problems can be cast into the form of an abstract Cauchy problem on a Hilbert space, we apply the theory of semigroups, which is concerned with the wellposedness of such problems. Hence, in Section 2.1 we review some results from this theory. Subsequently, we investigate Friedrichs' operators and their properties in Section 2.2. Having introduced the criteria necessary to show wellposedness of the wave-type problem (2.1), we consequently apply them to this class of problems in Section 2.3. We are ultimately interested in the analysis of the Peaceman–Rachford scheme, which is a splitting scheme. Hence, in Section 2.4 we will briefly discuss suitable splittings of the spatial differential operator. Lastly, in Section 2.5, we present the aforementioned examples for wave-type equations.

2.1 Wellposedness of abstract Cauchy problems

Throughout this section, let $\big(X, \big(\cdot \,\big|\, \cdot\big)_X\big)$ be a Hilbert space and $\|\cdot\|_X$ be the norm induced by its inner product. The goal of this section is to review some theory concerned with the wellposedness of an **abstract Cauchy problem** of the form

$$\begin{cases} \mathrm{d}_t u(t) = \mathcal{A}u(t) + f(t), & t \in \mathbb{R}_+, \quad (2.2a) \\ u(0) = u^0, & (2.2b) \end{cases}$$

where $u\colon \mathbb{R}_+ \to X$ is the **solution**, $u^0 \in X$ the **initial value**, $f\colon \mathbb{R}_+ \to X$ is an **inhomogeneity** or **source term**, and $\mathcal{A}\colon X \supset D(\mathcal{A}) \to X$ is a linear operator on the Hilbert space X. The differential equation (2.2a) is called an **abstract evolution equation**.

The material in this section is mostly taken from [Engel and Nagel, 2000], [Jacob and Zwart, 2012, Chapters 5 and 6] and [Pazy, 1983]. Additionally, the lecture notes [Schnaubelt, 2011], [Schnaubelt, 2013] and [Schnaubelt, 2015] were considered, and we closely follow the presentation of this topic in [Sturm, 2017].

Before discussing wellposedness of the abstract Cauchy problem (2.2), we give an introductory example to illustrate some ideas behind the theory we are going to employ.

Example 2.1. In the following, let $u^0 \in \mathbb{C}^m$ and $A \in \mathbb{C}^{m\times m}$. We consider the system of linear homogeneous ordinary differential equations with initial value u^0 given by

$$\begin{cases} \mathrm{d}_t u(t) = Au(t), & t \in \mathbb{R}_+, \\ u(0) = u^0, \end{cases}$$

where A determines the temporal evolution of the solution $u\colon \mathbb{R}_+ \to \mathbb{C}^d$. It is well known that this solution exists, is unique and is given by

$$u(t) = \mathrm{e}^{tA}\, u^0,$$

where $\mathrm{e}^{tA} \in \mathbb{C}^{d\times d}$ is the matrix exponential of tA. Further, we can recover the matrix A from the exponential by

$$\big(\mathrm{d}_t\, \mathrm{e}^{tA}\big)\,|_{t=0} = \big(A\, \mathrm{e}^{tA}\big)\,|_{t=0} = A. \tag{2.3}$$

Let $I \in \mathbb{C}^{m\times m}$ be the identity matrix. Then we further have

$$\lim_{t\to 0+} \mathrm{e}^{tA} = I \qquad \text{and} \qquad \mathrm{e}^{(t+s)A} = \mathrm{e}^{tA}\, \mathrm{e}^{sA}$$

for all $t, s \in \mathbb{R}_+$.

Hence, in the case of homogeneous ordinary differential equations we have at hand an operator—namely the matrix exponential—that returns the solution of a given initial value problem as a function of the initial value u^0.

Further, for $f\colon \mathbb{R}_+ \to \mathbb{C}^m$ smooth enough, the unique solution of the inhomogeneous initial value problem

$$\begin{cases} \mathrm{d}_t u(t) = Au(t) + f(t), & t \in \mathbb{R}_+, \\ u(0) = u^0 \end{cases}$$

is given by the variation-of-constants formula

$$u(t) = \mathrm{e}^{tA}\, u^0 + \int_0^t \mathrm{e}^{(t-s)A}\, f(s)\, \mathrm{d}s. \tag{2.4}$$

We will see that this formula is still valid in the case of an abstract Cauchy problem. ◇

In the next section, the concepts introduced in this example will be generalized to the case of the abstract Cauchy problem (2.2). In particular, the matrix exponential will be generalized by a strongly continuous semigroup and the matrix A by the infinitesimal generator of that semigroup.

2.1.1 Abstract evolution equations and semigroups

We begin this section by introducing the notion of a semigroup on the Hilbert space X.

Definition 2.2. *A one-parameter family of bounded linear operators $(\mathcal{T}(t))_{t\geq 0}$ from X to X is called a* ***semigroup of bounded linear operators on*** *X if*

(i) $\mathcal{T}(0) = \mathcal{I}$.

(ii) $\mathcal{T}(t+s) = \mathcal{T}(t)\mathcal{T}(s)$ *for all* $t, s \geq 0$.

A semigroup is called a ***strongly continuous semigroup*** *or* ***C_0-semigroup*** *if*

$$\lim_{t\to 0+} \|\mathcal{T}(t)x - x\|_X = 0$$

for all $x \in X$.

We will see later that strongly continuous semigroups can be understood as a generalization of the matrix exponential in Example 2.1 in the sense that they yield the solutions of an abstract Cauchy problem for a given initial value. In fact, the matrix exponential fulfills the properties in Definition 2.2 and is consequently a strongly continuous semigroup on the finite dimensional Hilbert space $\mathbb{C}^m$.

The next lemma gives an explicit bound on the operator norm of a strongly continuous semigroup, yielding the stability of the system. Further, it elaborates on the strong continuity.

Lemma 2.3. *Let $(\mathcal{T}(t))_{t\geq 0}$ be a strongly continuous semigroup. Then the following holds.*

(i) *There exist constants $C \geq 1$ and $\omega \geq 0$ such that for all $x \in X$ we have*

$$\|\mathcal{T}(t)x\|_X \leq C\,\mathrm{e}^{\omega t}\,\|x\|_X \qquad \text{for all } t \geq 0. \tag{2.5}$$

(ii) *The mapping $t \mapsto \mathcal{T}(t)$ is strongly continuous on $\mathbb{R}_+$, i.e.,*

$$\lim_{s\to 0} \|\mathcal{T}(t+s)x - \mathcal{T}(t)x\|_X = 0 \qquad \text{for all } t > 0.$$

Next, we define the infinitesimal generator of a strongly continuous semigroup. Note that this definition corresponds to (2.3), and the infinitesimal generator can be seen as a generalization of the matrix A in Example 2.1.

Definition 2.4. *Let $(\mathcal{T}(t))_{t\geq 0}$ be a strongly continuous semigroup and*

$$D(\mathcal{A}) = \big\{x \in X \mid \lim_{t\to 0+} \big(\tfrac{1}{t}(\mathcal{T}(t)x - x)\big) \in X\big\}.$$

The ***infinitesimal generator*** *of $(\mathcal{T}(t))_{t\geq 0}$ is defined as the linear operator $\mathcal{A}\colon D(\mathcal{A}) \to X$ given by*

$$\mathcal{A}x = \lim_{t\to 0+} \frac{\mathcal{T}(t)x - x}{t} \qquad \text{for all } x \in D(\mathcal{A}).$$

The set $D(\mathcal{A})$ is called the ***domain of*** *$\mathcal{A}$.*

The next lemma states some important properties of semigroups and their generators.

Lemma 2.5. *Let $(\mathcal{T}(t))_{t\geq 0}$ be a strongly continuous semigroup with infinitesimal generator $\mathcal{A}$. Then the following holds.*

(i) *For $x \in D(\mathcal{A})$ and $t \geq 0$ we have $\mathcal{T}(t)x \in D(\mathcal{A})$.*

(ii) *For $x \in D(\mathcal{A})$ and $t \geq 0$ we have*

$$\mathrm{d}_t(\mathcal{T}(t)x) = \mathcal{A}\mathcal{T}(t)x = \mathcal{T}(t)\mathcal{A}x. \tag{2.6}$$

(iii) *The domain of $\mathcal{A}$ is dense in X and $\mathcal{A}$ is a closed operator.*

By Definition 2.4 every strongly continuous semigroup has a unique generator. The following corollary of Lemma 2.5 states that the converse also holds true.

Corollary 2.6. *Let $(\mathcal{T}_1(t))_{t\geq 0}$ and $(\mathcal{T}_2(t))_{t\geq 0}$ be strongly continuous semigroups with infinitesimal generators $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. If $\mathcal{A}_1 = \mathcal{A}_2$, then $\mathcal{T}_1(t) = \mathcal{T}_2(t)$ for all $t \in \mathbb{R}_+$.*

Lemma 2.5 further enables us to establish a connection between a strongly continuous semigroup $(\mathcal{T}(t))_{t\geq 0}$ generated by an operator $\mathcal{A}$ and the solution $u\colon \mathbb{R}_+ \to X$ of the corresponding **homogeneous abstract Cauchy problem** given by

$$\begin{cases} \mathrm{d}_t u(t) = \mathcal{A}u(t), & t \in \mathbb{R}_+, \\ u(0) = u^0. \end{cases} \tag{2.7}$$

Namely, the solution of this problem is $u(t) = \mathcal{T}(t)u^0$. This can be seen by replacing x in (2.6) by the initial value u^0.

Thus, the connection between a strongly continuous semigroup and an abstract evolution equation reflects the relation between a matrix exponential and the corresponding system of ordinary differential equations, and from now on we adopt the notation in Example 2.1. Namely, if $\mathcal{A}$ generates a strongly continuous semigroup, we denote this semigroup by $\left(\mathrm{e}^{t\mathcal{A}}\right)_{t\geq 0}$ instead of $(\mathcal{T}(t))_{t\geq 0}$. The next theorem states that (2.7) is wellposed if $\mathcal{A}$ is the generator of a strongly continuous semigroup.

Theorem 2.7. *Let $\mathcal{A}$ be the infinitesimal generator of the strongly continuous semigroup $\left(\mathrm{e}^{t\mathcal{A}}\right)_{t\geq 0}$ and $u^0 \in D(\mathcal{A})$. Then there exists a unique solution $u \in C^1(\mathbb{R}_+; X) \cap C(\mathbb{R}_+; D(\mathcal{A}))$ of the homogeneous abstract Cauchy problem (2.7) given by*

$$u(t) = \mathrm{e}^{t\mathcal{A}}\, u^0.$$

Now, consider the **inhomogeneous abstract Cauchy problem** given by

$$\begin{cases} \mathrm{d}_t u(t) = \mathcal{A}u(t) + f(t), & t \in \mathbb{R}_+, \\ u(0) = u^0. \end{cases} \tag{2.8}$$

Then the variation-of-constants formula (2.4) is still valid as shown in the following well-posedness result.

Theorem 2.8. *Let $\mathcal{A}$ be the infinitesimal generator of the strongly continuous semigroup $\left(\mathrm{e}^{t\mathcal{A}}\right)_{t\geq 0}$ and $u^0 \in D(\mathcal{A})$. Moreover, let either $f \in C^1(\mathbb{R}_+; X)$ or $f \in C(\mathbb{R}_+; D(\mathcal{A}))$. Then there exists a unique solution $u \in C^1(\mathbb{R}_+; X) \cap C(\mathbb{R}_+; D(\mathcal{A}))$ of the inhomogeneous abstract Cauchy problem (2.8) given by*

$$u(t) = \mathrm{e}^{t\mathcal{A}}\, u^0 + \int_0^t \mathrm{e}^{(t-s)\mathcal{A}}\, f(s)\, \mathrm{d}s.$$

To summarize, abstract Cauchy problems of the form (2.7) and (2.8) are wellposed if the corresponding operator $\mathcal{A}$ generates a strongly continuous semigroup, and the initial value u^0 and the inhomogeneity f are smooth enough.

2.1.2 Dissipative operators and the Lumer–Phillips Theorem

We now identify sufficient conditions for an operator $\mathcal{A}$ to generate a strongly continuous semigroup. Before doing so, we introduce the notion of a dissipative operator.

Definition 2.9. *A linear operator $\mathcal{A}\colon D(\mathcal{A}) \to X$ is called **dissipative** if for every $x \in D(\mathcal{A})$ we have*

$$\operatorname{Re}\left(\mathcal{A}x \,\middle|\, x\right)_X \leq 0.$$

We gather two well-known properties of dissipative operators in the next lemma, namely the contractivity of the resolvent $(\mathcal{I} - \lambda\mathcal{A})^{-1}$ and of the transform $(\mathcal{I} + \lambda\mathcal{A})(\mathcal{I} - \lambda\mathcal{A})^{-1}$. The proofs can be found in [Engel and Nagel, 2000, Theorem II.3.14] and [Phillips, 1959, Section 1.1], respectively.

Lemma 2.10. *Let $\mathcal{A}\colon D(\mathcal{A}) \to X$ be dissipative. Then the following holds for all $\lambda > 0$.*

(i) *The resolvent $\mathcal{I} - \lambda\mathcal{A}$ is injective, and for $x \in \operatorname{ran}(\mathcal{I} - \lambda\mathcal{A})$ we have*

$$\|(\mathcal{I} - \lambda\mathcal{A})^{-1}x\|_X \leq \|x\|_X.$$

(ii) *For $x \in \operatorname{ran}(\mathcal{I} - \lambda\mathcal{A})$ we have*

$$\|(\mathcal{I} + \lambda\mathcal{A})(\mathcal{I} - \lambda\mathcal{A})^{-1}x\|_X \leq \|x\|_X.$$

As we will see in the next result, dissipative operators are connected to the following class of strongly continuous semigroups.

Definition 2.11. *A strongly continuous semigroup is called **contractive** or a **contraction semigroup** if (2.5) holds with $C = 1$ and $\omega = 0$.*

Now, we state the famous **Lumer–Phillips Theorem**, which yields the aforementioned sufficient conditions for an operator $\mathcal{A}$ to generate a strongly continuous semigroup and links dissipative operators to contraction semigroups. This result can, e.g., be found in [Jacob and Zwart, 2012, Theorem 6.1.7] or [Engel and Nagel, 2000, Theorem II.3.15 & Corollary II.3.20].

Theorem 2.12. *Let $\mathcal{A}\colon D(\mathcal{A}) \to X$ be a linear operator. Then the following statements are equivalent.*

(i) *$\mathcal{A}$ is dissipative and $\operatorname{ran}(\mathcal{I} - \lambda\mathcal{A}) = X$ for some $\lambda > 0$.*

(ii) *$\mathcal{A}$ is dissipative and $\operatorname{ran}(\mathcal{I} - \lambda\mathcal{A}) = X$ for all $\lambda > 0$.*

(iii) *$\mathcal{A}$ generates a contraction semigroup.*

This motivates the next definition of maximal dissipative operators.

Definition 2.13. *A linear operator $\mathcal{A}\colon D(\mathcal{A}) \to X$ is called **maximal dissipative** if it is dissipative and $\operatorname{ran}(\mathcal{I} - \lambda\mathcal{A}) = X$ for some $\lambda > 0$.*

2.2 Friedrichs' operators

Before applying the semigroup theory to the class of wave-type problems introduced earlier, in this section we introduce the class of Hilbert space operators we call Friedrichs' operators. As stated before, we closely follow ideas from [Ern and Guermond, 2006a] and [Di Pietro and Ern, 2012, Section 7], where a variant of the theory of Friedrichs' systems [Friedrichs, 1958] is introduced. Again, we start with an example to motivate the concepts we are going to employ.

Example 2.14. Let $d = 2$. We consider the two-dimensional **linear homogeneous advection** equation

$$\begin{cases} \partial_t u = \alpha \cdot \nabla u, & \text{in } \mathbb{R}_+ \times \Omega, \\ u(0) = u^0, & \text{in } \Omega \end{cases} \tag{2.9}$$

with **solution** $u\colon \mathbb{R}_+ \times \Omega \to \mathbb{R}$ and $\alpha\colon \Omega \to \mathbb{R}^2$ being the **advection velocity**. This equation can be seen as a prototype for the more general wave-type problem (2.1) as we have

$$\alpha \cdot \nabla u = \alpha_1 \partial_1 u_1 + \alpha_2 \partial_2 u_2.$$

In this example we investigate some properties of this equation, or rather the spatial operator $\alpha \cdot \nabla$. Throughout the example, let $v, w\colon \Omega \to \mathbb{R}$ and α be smooth enough, such that the following expressions make sense.

Then the usual integration by parts formula and subsequently the product rule of differentiation yields

$$\begin{aligned} \big(\alpha \cdot \nabla v \,\big|\, w\big)_\Omega &= -\big(v \,\big|\, \nabla\cdot(\alpha w)\big)_\Omega + \big((\alpha \cdot \mathfrak{n})v \,\big|\, w\big)_\Gamma \\ &= -\big(v \,\big|\, \alpha \cdot \nabla w + (\nabla\cdot\alpha) w\big)_\Omega + \big((\alpha \cdot \mathfrak{n})v \,\big|\, w\big)_\Gamma, \end{aligned} \tag{2.10}$$

where $\mathfrak{n}$ is the outer unit normal vector to Γ. We abbreviate the spatial operator by

$$\widetilde{\mathcal{L}} = \alpha \cdot \nabla, \tag{2.11}$$

define the **formal adjoint** $\widetilde{\mathcal{L}}^\circledast$ of $\widetilde{\mathcal{L}}$ by

$$\widetilde{\mathcal{L}}^\circledast v = -\alpha \cdot \nabla v - (\nabla\cdot\alpha) v \tag{2.12}$$

and the **boundary field corresponding to** $\widetilde{\mathcal{L}}$ by

$$\widetilde{\mathcal{L}}_\partial = \alpha \cdot \mathfrak{n}.$$

Using this, we can write the **integration by parts** formula (2.10) more compactly as

$$\big(\widetilde{\mathcal{L}} v \,\big|\, w\big)_\Omega = \big(v \,\big|\, \widetilde{\mathcal{L}}^{\circledast} w\big)_\Omega + \big(\widetilde{\mathcal{L}}_\partial v \,\big|\, w\big)_\Gamma. \tag{2.13}$$

To obtain a wellposed problem, one has to impose boundary conditions on (2.9). We consider homogeneous **inflow boundary conditions** given by

$$u = 0 \qquad \text{on } \mathbb{R}_+ \times \Gamma^+, \tag{2.14}$$

where the **inflow boundary** Γ^+ is given by $\Gamma^+ = \{x \in \Gamma \mid \widetilde{\mathcal{L}}_\partial(x) > 0\}$. This can be rewritten as an equation on the whole boundary Γ by defining the boundary field

$$\widetilde{\mathcal{L}}_\Gamma = -|\alpha \cdot \mathfrak{n}|. \tag{2.15}$$

With this, (2.14) is equivalent to

$$(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma) u = 0 \qquad \text{on } \mathbb{R}_+ \times \Gamma^+. \tag{2.16}$$

This will be the way we model boundary conditions in the general setting. ◇

2.2.1 Definition of a Friedrichs' operator

Throughout, we consider the (real) Hilbert space $(L^2(\Omega)^m, \big(\,\cdot \,\big|\, \cdot\,\big)_\Omega)$. We start by defining the general version of the operator in (2.11).

Definition 2.15. *Let $F_0, \dots, F_d \colon \Omega \to \mathbb{R}^{m\times m}$ and*

$$H(\mathcal{F}) = \{v \in L^2(\Omega)^m \mid \sum_{i=1}^d F_i \partial_i v \in L^2(\Omega)^m\}. \tag{2.17}$$

We call the operator $\mathcal{F} \colon H(\mathcal{F}) \to L^2(\Omega)^m$ defined by

$$\mathcal{F} v = \sum_{i=1}^d F_i \partial_i v + F_0 v \text{ for all } v \in H(\mathcal{F})$$

*a **Friedrichs' operator with coefficients** $(F_i)_{i=0}^d$ if the following holds.*

(F1) $F_0, \dots, F_d \in L^\infty(\Omega)^{m\times m}$.

(F2) $F_1, \dots, F_d$ *are symmetric a.e. on* Ω.

(F3) $\nabla\cdot \mathcal{F} \coloneqq \sum_{i=1}^d \partial_i F_i \in L^\infty(\Omega)^{m\times m}$.

Throughout the rest of this section, let $\mathcal{F}$ be a Friedrichs' operator with coefficients $(F_i)_{i=0}^d$. We endow the **graph space** $H(\mathcal{F})$ of $\mathcal{F}$ with the **graph norm**

$$\|v\|_{H(\mathcal{F})} = \|v\|_\Omega + \|\mathcal{F} v\|_\Omega.$$

Then, by [Di Pietro and Ern, 2012, Lemma 7.2], the graph space $H(\mathcal{F})$ is a Hilbert space, and by definition we have $\mathcal{F} \in \mathcal{B}(H(\mathcal{F}), L^2(\Omega)^m)$.

Friedrichs' operators are first order differential operators. Hence, if the coefficients are smooth enough, we have the following bound.

Lemma 2.16. *Let $K \subset \Omega$ be open and $v \in H(\mathcal{F})$ with $v|_K \in H^2(K)$. Further, let $F_i \in W^{1,\infty}(K)^{m\times m}$ for all $i = 0, \dots, d$. Then we have*

$$|\mathcal{F}v|_{1,K} \leq C_{1,K,\mathcal{F}} \|v\|_{2,K},$$

where $C_{1,K,\mathcal{F}} = \sqrt{2}\,(d+1)\widetilde{C}_{1,K,\mathcal{F}}$ with $\widetilde{C}_{1,K,\mathcal{F}} = \max\limits_{i=0,\dots,d}\{\max\limits_{j=1,\dots,d} \|\partial_j F_i\|_{\infty,K}, \|F_i\|_{\infty,K}\}$.

Proof. The definition of $\mathcal{F}$ and the $H^1(K)$-seminorm and using the product rule yields

$$\begin{aligned}
|\mathcal{F}v|_{1,K}^2 &= \sum_{j=1}^{d} \|\partial_j(\mathcal{F}v)\|_K^2 \\
&= \sum_{j=1}^{d} \| \sum_{i=1}^{d} \partial_j(F_i\partial_i v) + \partial_j(F_0 v)\|_K^2 \\
&= \sum_{j=1}^{d} \| \sum_{i=1}^{d} \partial_j F_i\, \partial_i v + \sum_{i=1}^{d} F_i\, \partial_j\partial_i v + \partial_j F_0\, v + F_0 \partial_j v\|_K^2 \\
&\leq \sum_{j=1}^{d} \big(\sum_{i=1}^{d} \|\partial_j F_i\, \partial_i v\|_K + \sum_{i=1}^{d} \|F_i\, \partial_j\partial_i v\|_K + \|\partial_j F_0\, v\|_K + \|F_0\partial_j v\|_K \big)^2 \\
&\leq \widetilde{C}_{1,K,\mathcal{F}}^2 \sum_{j=1}^{d} \big(\sum_{i=1}^{d} \|\partial_i v\|_K + \sum_{i=1}^{d} \|\partial_j\partial_i v\|_K + \|v\|_K + \|\partial_j v\|_K \big)^2.
\end{aligned}$$

We now use the equivalence of the 1- and 2-norm on $\mathbb{R}^{2(d+1)}$ to obtain

$$\begin{aligned}
|\mathcal{F}v|_{1,K}^2 &\leq 2(d+1)\,\widetilde{C}_{1,K,\mathcal{F}}^2 \sum_{j=1}^{d} \big(\sum_{i=1}^{d} \|\partial_i v\|_K^2 + \sum_{i=1}^{d} \|\partial_j\partial_i v\|_K^2 + \|v\|_K^2 + \|\partial_j v\|_K^2 \big) \\
&= 2(d+1)\,\widetilde{C}_{1,K,\mathcal{F}}^2 \big((d+1) \sum_{i=1}^{d} \|\partial_i v\|_K^2 + \sum_{j=1}^{d}\sum_{i=1}^{d} \|\partial_j\partial_i v\|_K^2 + d\|v\|_K^2 \big) \\
&\leq 2(d+1)^2\,\widetilde{C}_{1,K,\mathcal{F}}^2 \big(\sum_{i=1}^{d} \|\partial_i v\|_K^2 + \sum_{j=1}^{d}\sum_{i=1}^{d} \|\partial_j\partial_i v\|_K^2 + \|v\|_K^2 \big) \\
&= 2(d+1)^2\,\widetilde{C}_{1,K,\mathcal{F}}^2 \|v\|_{2,K}^2,
\end{aligned}$$

concluding the proof. □

We further show a version of Lemma 2.16 for higher derivatives. For the sake of presentation, we assume the coefficients of $\mathcal{F}$ to be constant. However, the statement is still valid for sufficiently smooth coefficients, which can be verified by using repeatedly the argument in the proof of Lemma 2.16.

Lemma 2.17. *Let $K \subset \Omega$ be open and $q > 0$. Further, let $v \in H(\mathcal{F})$ with $v|_K \in H^{q+1}(K)$ and let the coefficients of $\mathcal{F}$ be constant on K. Then we have*

$$|\mathcal{F}v|_{q,K} \leq \sqrt{d+1}\, C_{\mathcal{F},K} \|v\|_{q+1,K}$$

with $C_{\mathcal{F},K} = \max\limits_{i=0,\dots,d} \|F_i\|_{\infty,K}$.

Proof. By the definition of $\mathcal{F}$ and since the coefficients F_i are constant on K for all $i = 0, \dots, d$, we have

$$|\mathcal{F}v|_{q,K} = |\sum_{i=1}^{d} F_i \partial_i v + F_0 v|_{q,K} \leq C_{\mathcal{F},K} \big(\sum_{i=1}^{d} |\partial_i v|_{q,K} + |v|_{q,K} \big).$$

Using the equivalence of the 1- and 2-norm on $\mathbb{R}^{d+1}$, we thus obtain

$$|\mathcal{F}v|_{q,K} \leq \sqrt{d+1}\, C_{\mathcal{F},K} \big(\sum_{i=1}^{d} |\partial_i v|_{q,K}^2 + |v|_{q,K}^2 \big)^{1/2},$$

which concludes the proof. □

2.2.2 The formal adjoint of a Friedrichs' operator

Next, we define the general version of (2.12), i.e., the formal adjoint of the Friedrichs' operator $\mathcal{F}$. As the adjoint operators of the partial derivatives are given by the integration by parts formula, this can be done explicitly.

Definition 2.18. *We call $\mathcal{F}^{\circledast}\colon H(\mathcal{F}) \to L^2(\Omega)^m$ defined by*

$$\mathcal{F}^{\circledast} v = -\sum_{i=1}^{d} \partial_i (F_i v) + F_0^* v \qquad \text{for all } v \in H(\mathcal{F})$$

*the **formal adjoint of** $\mathcal{F}$.*

By the standard product rule of differentiation we have

$$\mathcal{F}^{\circledast} v = -\sum_{i=1}^{d} F_i \partial_i v + F_0^* v - (\nabla \cdot \mathcal{F}) v \qquad \text{for all } v \in H(\mathcal{F}).$$

Hence, taking Definition 2.15 (F3) into account yields $\mathcal{F}^{\circledast} \in \mathcal{B}(H(\mathcal{F}), L^2(\Omega)^m)$. Further, we have

$$\mathcal{F}v + \mathcal{F}^{\circledast} v = (F_0 + F_0^* - \nabla \cdot \mathcal{F}) v \qquad \text{for all } v \in H(\mathcal{F}). \tag{2.18}$$

2.2.3 Boundary operators

Having defined the formal adjoint of $\mathcal{F}$, we can now define the boundary operator associated with $\mathcal{F}$.

Definition 2.19. *We call $\mathcal{F}_\partial\colon H(\mathcal{F}) \to H(\mathcal{F})'$ defined by*

$$\langle \mathcal{F}_\partial v \,|\, w \rangle = \big(\mathcal{F}v \,\big|\, w \big)_\Omega - \big(v \,\big|\, \mathcal{F}^{\circledast} w \big)_\Omega \qquad \text{for all } v, w \in H(\mathcal{F}) \tag{2.19}$$

*the **boundary operator associated with** $\mathcal{F}$.*

In fact, (2.19) can be seen as a generalization of the integration by parts formula (2.13). Next, we state two properties of this boundary operator, namely self-adjointness and boundedness.

Lemma 2.20. *The boundary operator $\mathcal{F}_\partial$ fulfills the following.*

(i) $\mathcal{F}_\partial \in \mathcal{B}(H(\mathcal{F}), H(\mathcal{F})')$.

(ii) $\langle \mathcal{F}_\partial v \,|\, w \rangle = \langle \mathcal{F}_\partial w \,|\, v \rangle$ *for all* $v, w \in H(\mathcal{F})$.

Proof. (i) This is a straightforward consequence of the definition of $\mathcal{F}_\partial$ and the boundedness of $\mathcal{F}$ and $\mathcal{F}^\circledast$ on the graph space $H(\mathcal{F})$.

(ii) Let $v, w \in H(\mathcal{F})$ and set $B = F_0 + F_0^* - \nabla\cdot\mathcal{F}$. By the definition of $\mathcal{F}_\partial$ and (2.18) we have

$$\begin{aligned}\langle \mathcal{F}_\partial v \,|\, w \rangle - \langle \mathcal{F}_\partial w \,|\, v \rangle &= \big(\mathcal{F}v \,|\, w\big)_\Omega - \big(v \,|\, \mathcal{F}^\circledast w\big)_\Omega - \Big(\big(\mathcal{F}w \,|\, v\big)_\Omega - \big(w \,|\, \mathcal{F}^\circledast v\big)_\Omega\Big) \\ &= \big((\mathcal{F} + \mathcal{F}^\circledast)v \,|\, w\big)_\Omega - \big(v \,|\, (\mathcal{F}^\circledast + \mathcal{F})w\big)_\Omega \\ &= \big(Bv \,|\, w\big)_\Omega - \big(v \,|\, Bw\big)_\Omega \\ &= 0,\end{aligned}$$

where the last equality follows since B is self-adjoint. □

In this chapter, we are ultimately interested in the wellposedness of the wave-type problem (2.1). Since this problem is posed on a bounded domain, we need to pose boundary conditions to obtain uniqueness. We follow the approach in [Ern and Guermond, 2006a], allowing us to use the results obtained therein.

Definition 2.21. *Let $\mathcal{K} \in \mathcal{B}(H(\mathcal{F}), H(\mathcal{F})')$ be such that*

(B1) $\langle \mathcal{K}v \,|\, v \rangle \leq 0$ *for all* $v \in H(\mathcal{F})$.

(B2) $H(\mathcal{F}) = \ker(\mathcal{F}_\partial - \mathcal{K}) + \ker(\mathcal{F}_\partial + \mathcal{K})$.

Then we call $\mathcal{K}$ a ***dissipative boundary condition for*** *$\mathcal{F}$.*

In the following, let $\mathcal{F}_\Gamma$ be a dissipative boundary condition for $\mathcal{F}$. As $\ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma)$ is the kernel of a bounded operator on $H(\mathcal{F})$, it is a Hilbert space if endowed with the graph norm. Further, $\mathcal{F}_\Gamma$ can be seen as the generalization of (2.15). Hence, $\ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma)$ is the (closed) subspace of $H(\mathcal{F})$ incorporating the boundary condition as it implies a general version of (2.16).

For more insight into this approach of modeling boundary conditions, we again refer to [Ern and Guermond, 2006a] and [Di Pietro and Ern, 2012, Chapter 7]. Further details can also be found in [Ern et al., 2007].

2.2.4 Dissipativity and invertibility of a Friedrichs' operator

Having introduced the necessary concepts associated with Friedrichs' operators, we can now directly determine the conditions under which such an operator is maximal dissipative. By the Lumer–Phillips Theorem 2.12 this implies that the operator is the generator of a contractive semigroup, which will be used to show wellposedness of the wave-type equation.

Theorem 2.22. *Let $\mathcal{F}$ be a Friedrichs' operator with coefficients $(F_i)_{i=0}^d$ and dissipative boundary condition $\mathcal{F}_\Gamma$. Further, let*

$$F_0 + F_0^* - \nabla\cdot\mathcal{F} \leq 0 \qquad \text{a.e. on } \Omega. \tag{2.20}$$

Then the following holds.

(i) *The restriction of* $\mathcal{F}$ *to* $\ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma)$ *is dissipative w.r.t.* $\big(\,\cdot\,\big|\,\cdot\,\big)_\Omega$.

(ii) $(\mathcal{I} - \lambda\mathcal{F})\colon \ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma) \to L^2(\Omega)^m$ *is an isomorphism for all* $\lambda > 0$.

Hence, the restriction of $\mathcal{F}$ *to* $\ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma)$ *is maximal dissipative.*

Proof. (i) Let $v \in \ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma)$. By (2.18) and (2.19) we have

$$\begin{aligned} 2\big(\mathcal{F}v \,\big|\, v\big)_\Omega &= \big(\mathcal{F}v \,\big|\, v\big)_\Omega + \big(\mathcal{F}^\circledast v \,\big|\, v\big)_\Omega + \big(\mathcal{F}v \,\big|\, v\big)_\Omega - \big(\mathcal{F}^\circledast v \,\big|\, v\big)_\Omega \\ &= \big((F_0 + F_0^* - \nabla\cdot\mathcal{F})v \,\big|\, v\big)_\Omega + \big\langle \mathcal{F}_\partial v \,\big|\, v\big\rangle \\ &\le \big\langle (\mathcal{F}_\partial - \mathcal{F}_\Gamma)v \,\big|\, v\big\rangle + \big\langle \mathcal{F}_\Gamma v \,\big|\, v\big\rangle \\ &\le 0, \end{aligned}$$

where the first inequality follows because of (2.20) and the second because of $v \in \ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma)$ and Definition 2.21.

(ii) This is a direct consequence of [Ern and Guermond, 2006a, Theorem 2.5]. □

The maximal dissipativity of $\mathcal{F}$ on $\ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma)$ motivates the following definition.

Definition 2.23. *Let* $\mathcal{F}$ *be a Friedrichs' operator with coefficients* $(F_i)_{i=0}^d$ *fulfilling*

$$F_0 + F_0^* - \nabla\cdot\mathcal{F} \le 0 \text{ a.e. on } \Omega. \tag{2.21}$$

Further, let $\mathcal{F}_\Gamma$ *be a dissipative boundary condition for* $\mathcal{F}$ *and*

$$D(\mathcal{F}) = \ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma).$$

Then we call the restriction of $\mathcal{F}$ *to* $D(\mathcal{F})$ *a* ***dissipative Friedrichs' operator with coefficients*** $(F_i)_{i=0}^d$ ***and boundary condition*** $\mathcal{F}_\Gamma$.

Remark 2.24. Using the exact same strategy of proof, we can show that Theorem 2.22 also holds for the restriction of the formal adjoint $\mathcal{F}^\circledast$ to the space $\ker(\mathcal{F}_\partial + \mathcal{F}_\Gamma^*)$, cf., [Ern and Guermond, 2006a]. In fact, this is the adjoint operator of the restriction of $\mathcal{F}$ to $\ker(\mathcal{F}_\partial - \mathcal{F}_\Gamma)$. ◇

2.3 Wellposedness of wave-type equations

We are now able to discuss wellposedness of the wave-type problem stated as follows. Given $u^0\colon \Omega \to \mathbb{R}^m$ and $g\colon \mathbb{R}_+ \times \Omega \to \mathbb{R}^m$, seek $u\colon \mathbb{R}_+ \times \Omega \to \mathbb{R}^m$, such that

$$\begin{cases} M\partial_t u(t,x) = \widetilde{\mathcal{L}}u(t,x) + g(t,x), & t \in \mathbb{R}_+,\ x \in \Omega, & \text{(2.22a)} \\ u(0,x) = u^0(x), & x \in \Omega. & \text{(2.22b)} \end{cases}$$

We begin by stating the assumptions on the operators M and $\widetilde{\mathcal{L}}$ under which we show wellposedness of (2.22). Subsequently, we reformulate (2.22) in the form of the inhomogeneous abstract Cauchy problem (2.8) so that we can make use of the semigroup theory in Section 2.1. Using the results on Friedrichs' operators from Section 2.2 we show that the resulting spatial operator generates a contraction semigroup. Theorem 2.8 then yields wellposedness of the wave-type problem (2.22) in a subspace of the graph space of $\widetilde{\mathcal{L}}$ for suitable initial values and inhomogeneities.

2.3.1 Reformulation as an abstract Cauchy problem

We start by stating the assumptions on M and $\widetilde{\mathcal{L}}$.

Assumption 2.25. *We assume that the following holds.*

(i) $M \in L^\infty(\Omega)^{m\times m}$ *is symmetric and uniformly positive a.e. on* Ω.

(ii) $\widetilde{\mathcal{L}}\colon D(\widetilde{\mathcal{L}}) \to L^2(\Omega)^m$ *is a dissipative Friedrichs' operator with boundary condition* $\widetilde{\mathcal{L}}_\Gamma$.

Here, the first assumption is needed to reformulate the problem into the form of an abstract Cauchy problem. The second assumption will then be used to apply the results from Section 2.2. Recall that by Definition 2.23 we have $D(\widetilde{\mathcal{L}}) = \ker(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma)$.

Let the weighted L^2-inner product $\big(\,\cdot \,\big|\, \cdot\,\big)_M$ be defined by

$$\big(v \,\big|\, w\big)_M = \big(Mv \,\big|\, w\big)_\Omega \qquad \text{for all } v, w \in L^2(\Omega)^m.$$

By Assumption 2.25 (i), M is invertible with $M^{-1} \in L^\infty(\Omega)^{m\times m}$ also being symmetric and uniformly positive a.e. on Ω. Hence, the norm $\|\cdot\|_M$ induced by $\big(\,\cdot \,\big|\, \cdot\,\big)_M$ is equivalent to the standard L^2-norm, namely we have

$$\|M^{-1}\|_{\infty,\Omega}^{-1/2}\, \|v\|_\Omega \le \|v\|_M \le \|M\|_{\infty,\Omega}^{1/2}\, \|v\|_\Omega \tag{2.23}$$

for all $v \in L^2(\Omega)^m$. Therefore, $\big(L^2(\Omega)^m, \big(\,\cdot \,\big|\, \cdot\,\big)_M\big)$ is a Hilbert space. This is the space in which we consider the abstract Cauchy problem.

Let $D(\mathcal{L}) = D(\widetilde{\mathcal{L}})$. To reformulate the wave-type problem (2.22), we define the operator $\mathcal{L}\colon D(\mathcal{L}) \to L^2(\Omega)^m$ such that we have

$$\mathcal{L} = M^{-1}\widetilde{\mathcal{L}}.$$

Multiplying (2.22a) by M^{-1}, we obtain the equivalent problem stated as follows. Seek $u\colon \mathbb{R}_+ \to L^2(\Omega)^m$, such that

$$\begin{cases} \mathrm{d}_t u(t) = \mathcal{L}u(t) + f(t), & t \in \mathbb{R}_+, \qquad (2.24a)\\ u(0) = u^0, \qquad (2.24b) \end{cases}$$

where we abbreviated $f(t) = M^{-1}g(t)$.

2.3.2 The wellposedness result

In this section we show that problem (2.24) is wellposed if supplied with suitable initial conditions and if f is smooth enough. This is due to the fact that $\mathcal{L}$ is maximal dissipative, enabling us to apply the Lumer–Phillips Theorem 2.12. We show this in the next theorem, which is a straightforward consequence of Theorem 2.22, owing to the assumptions on M and $\widetilde{\mathcal{L}}$.

Theorem 2.26. *The following statements hold true.*

(i) *The operator* $\mathcal{L}$ *is dissipative w.r.t.* $\big(\,\cdot \,\big|\, \cdot\,\big)_M$.

(ii) $(\mathcal{I} - \lambda\mathcal{L})\colon D(\mathcal{L}) \to L^2(\Omega)^m$ *is an isomorphism for all* $\lambda > 0$.

Hence, $\mathcal{L}$ *is maximal dissipative.*

Proof. (i) By the definition of $\left(\cdot \,\middle|\, \cdot\right)_M$ and $\mathcal{L}$, we have

$$\left(\mathcal{L}v \,\middle|\, v\right)_M = \left(M\mathcal{L}v \,\middle|\, v\right)_\Omega = \left(\widetilde{\mathcal{L}}v \,\middle|\, v\right)_\Omega \leq 0$$

for all $v \in D(\mathcal{L})$, where the inequality follows by Theorem 2.22 (i).

(ii) Note that by the assumptions on M, the claim is equivalent to $(M - \lambda\widetilde{\mathcal{L}})\colon D(\widetilde{\mathcal{L}}) \to L^2(\Omega)^m$ being an isomorphism for all $\lambda > 0$. This is again a direct consequence of [Ern and Guermond, 2006a, Theorem 2.5], since M is uniformly positive a.e. on Ω. □

By Theorem 2.26 and the Lumer–Phillips Theorem 2.12 it is apparent that $\mathcal{L}$ is the generator of a contraction semigroup. We denote this semigroup by $\left(\mathrm{e}^{t\mathcal{L}}\right)_{t\geq 0}$. Theorem 2.8 now directly yields the following wellposedness result.

Corollary 2.27. *Let $u^0 \in D(\mathcal{L})$ and either $f \in C^1(\mathbb{R}_+; L^2(\Omega)^m)$ or $f \in C(\mathbb{R}_+; D(\mathcal{L}))$. Then there exists a unique solution $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}))$ of (2.24) and hence of (2.22) given by the variation-of-constants formula*

$$u(t) = \mathrm{e}^{t\mathcal{L}}\, u^0 + \int_0^t \mathrm{e}^{(t-s)\mathcal{L}}\, f(s)\,\mathrm{d}s. \tag{2.25}$$

Remark 2.28. We want to point out that $\widetilde{\mathcal{L}}$ fulfilling (2.21) is not necessary to obtain a wellposedness result. If the condition is not fulfilled, the Friedrichs' operator is no longer dissipative but shift-dissipative, leading to new technicalities. However, the results in this thesis can also be generalized to this case. For the sake of presentation, we restrict ourselves to the dissipative case. ◇

2.4 Splitting

The Peaceman–Rachford method we analyze in this thesis is a splitting method. Thus, we assume that there are two dissipative Friedrichs' operators $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ with coefficients $(A_i)_{i=0}^d$ and $(B_i)_{i=0}^d$ and boundary condition $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$, respectively such that we have

$$\widetilde{\mathcal{L}}v = \widetilde{\mathcal{A}}v + \widetilde{\mathcal{B}}v \qquad \text{for all } v \in H(\widetilde{\mathcal{A}}) \cap H(\widetilde{\mathcal{B}}) \subset H(\widetilde{\mathcal{L}}). \tag{2.26}$$

By equating coefficients, this readily implies $L_i = A_i + B_i$ for $i = 0, \ldots, d$. We further assume that the boundary conditions are consistent in the sense that we have

$$\widetilde{\mathcal{L}}_\Gamma v = \widetilde{\mathcal{A}}_\Gamma v + \widetilde{\mathcal{B}}_\Gamma v \qquad \text{for all } v \in H(\widetilde{\mathcal{A}}) \cap H(\widetilde{\mathcal{B}}) \subset H(\widetilde{\mathcal{L}}). \tag{2.27}$$

We define the operators $\mathcal{A}$ and $\mathcal{B}$ analogously to $\mathcal{L}$. Note that the splitting property (2.26) also holds for these operators, i.e., we have

$$\mathcal{L}v = \mathcal{A}v + \mathcal{B}v \qquad \text{for all } v \in D(\mathcal{A}) \cap D(\mathcal{B}) \subset D(\mathcal{L}).$$

Further, using the same arguments as in Section 2.3 we immediately see that these operators are maximal dissipative and thus fulfill the Lumer–Phillips Theorem 2.12.

Theorem 2.29. *The following statements hold.*

(i) *The operators $\mathcal{A}$ and $\mathcal{B}$ are dissipative w.r.t. $\left(\cdot \,\middle|\, \cdot\right)_M$ on their respective domains.*

(ii) *$(\mathcal{I} - \lambda\mathcal{A})\colon D(\mathcal{A}) \to L^2(\Omega)^m$ and $(\mathcal{I} - \lambda\mathcal{B})\colon D(\mathcal{B}) \to L^2(\Omega)^m$ are isomorphisms for all $\lambda > 0$.*

Hence, $\mathcal{A}$ and $\mathcal{B}$ are maximal dissipative.

2.5 Examples

In this section, we have a look at three examples fitting into the framework above. As mentioned before, these examples are the advection equation already encountered in Example 2.14, the wave equation and Maxwell's equations.

In the following we work out under which conditions these equations fit into the framework. To do so, we loosely follow [Di Pietro and Ern, 2012, Section 7.2.5] and [Ern and Guermond, 2006a, Section 3], where additional examples of Friedrichs' operators can be found. Suitable splittings for our examples (or rather special instances thereof) will be discussed in Chapter 6.

Throughout, let $\mathfrak{n}$ be the outward unit normal vector of Ω to Γ.

2.5.1 The advection equation

In Example 2.14 we have already seen that the homogeneous two-dimensional advection equation is a candidate to fit the above framework of Friedrichs' operators. As a matter of fact, this is still the case in higher dimensions and including an inhomogeneity.

Problem formulation

For the d-dimensional **advection equation**, let $m = 1$ and $g \in L^2(\Omega)$. We consider

$$\begin{cases} \partial_t u = \alpha \cdot \nabla u + g & \text{in } \mathbb{R}_+ \times \Omega, \\ u(0) = u^0 & \text{in } \Omega, \end{cases} \tag{2.28}$$

where $\alpha \in L^\infty(\Omega)^d$ is the **advection velocity**. We further assume $\nabla \cdot \alpha \in L^\infty(\Omega)$ and $\nabla \cdot \alpha \geq 0$ a.e. on Ω.

Formulation as a wave-type equation

We see that (2.28) fits the form of the general wave-type problem (2.22) by setting

$$\widetilde{\mathscr{L}} = \alpha \cdot \nabla \qquad \text{and} \qquad M = 1.$$

Further, by the assumptions on the advection velocity, $\widetilde{\mathscr{L}}$ fulfills the conditions of Definition 2.15 and is consequently a Friedrichs' operator with coefficients

$$L_0 = 0, \qquad L_i = \alpha_i, \qquad i = 1, \dots, d$$

and graph space

$$H(\widetilde{\mathscr{L}}) = \{v \in L^2(\Omega) \mid \alpha \cdot \nabla v \in L^2(\Omega)\}.$$

Associated boundary operator

To discuss boundary operators, we assume more regularity on the advection velocity, namely that α is Lipschitz-continuous on Ω. This enables us to define the **inflow boundary** Γ^+ and the **outflow boundary** Γ^- by

$$\Gamma^\pm = \{x \in \Gamma \mid \pm\alpha(x) \cdot \mathfrak{n}(x) > 0\}.$$

We assume that Γ^+ and Γ^- are well-separated, i.e.,

$$\min_{x\in\Gamma^-,\ y\in\Gamma^+} \|x-y\| > 0.$$

In [Di Pietro and Ern, 2012, Section 2.1.3] it is shown that, under these conditions, the boundary operator $\widetilde{\mathcal{L}}_\partial$ associated with $\widetilde{\mathcal{L}}$ can be represented for all $v, w \in H(\widetilde{\mathcal{L}})$ by

$$\big\langle \widetilde{\mathcal{L}}_\partial v \,\big|\, w \big\rangle = \big((\alpha\cdot\mathfrak{n})v \,\big|\, w \big)_\Gamma.$$

Homogeneous inflow boundary conditions

It remains to identify a suitable boundary operator $\widetilde{\mathcal{L}}_\Gamma$. As in Example 2.14 we consider homogeneous **inflow boundary conditions**. Hence, following (2.15), we set

$$\big\langle \widetilde{\mathcal{L}}_\Gamma v \,\big|\, w \big\rangle = -\big(|\alpha\cdot\mathfrak{n}| v \,\big|\, w \big)_\Gamma \qquad \text{for all } v, w \in H(\widetilde{\mathcal{L}}). \tag{2.29}$$

Obviously, $\widetilde{\mathcal{L}}_\Gamma$ fulfills the dissipativity condition Definition 2.21 (B1). We refer to [Di Pietro and Ern, 2012, Section 7.2.5.1] for the confirmation of condition (B2).

Conclusion

In conclusion—as we have assumed $\nabla\cdot\alpha \geq 0$ a.e. on Ω, corresponding to (2.20)—the operator $\widetilde{\mathcal{L}}$ restricted to $\ker(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma)$ is a dissipative Friedrichs' operator. Hence, the advection equation considered in this example fits the above framework.

2.5.2 The acoustic wave equation

The acoustic wave equation considered in this example describes the propagation of acoustic waves in an isotropic medium. It is usually considered in a second order formulation. However, it can be transformed into a first order system, which is the one we consider herein. We refer to [Hochbruck et al., 2015b, Section 2.2] for the derivation of the first order formulation.

Problem formulation

For the d-dimensional **acoustic wave equation**, let $m = d+1$ and $\tilde{g} \in L^2(\Omega)$. We consider the d-dimensional acoustic wave equation in div-grad formulation given by

$$\begin{cases} \rho\partial_t p = \nabla\cdot q + \tilde{g} & \text{in } \mathbb{R}_+\times\Omega, \\ \partial_t q = \nabla p & \text{in } \mathbb{R}_+\times\Omega, \\ p(0) = p^0 \quad q(0) = q^0 & \text{in } \Omega, \end{cases} \tag{2.30}$$

where we seek the **pressure** $p\colon \Omega\to\mathbb{R}$ and the **flux** $q\colon \Omega\to\mathbb{R}^d$. The **density** $\rho\in L^\infty(\Omega)$ is a given scalar field, which we assume to be uniformly positive a.e. on Ω.

Formulation as a wave-type equation

By writing

$$u = \begin{pmatrix} p \\ q \end{pmatrix}, \qquad u^0 = \begin{pmatrix} p^0 \\ q^0 \end{pmatrix}, \qquad g = \begin{pmatrix} \tilde{g} \\ 0 \end{pmatrix}$$

and

$$\widetilde{\mathcal{L}} = \begin{pmatrix} 0 & \nabla\cdot \\ \nabla & 0 \end{pmatrix}, \qquad M = \begin{pmatrix} \rho & 0 \\ 0 & \mathcal{I} \end{pmatrix},$$

equation (2.30) can be written in the form of (2.22), i.e.,

$$\begin{cases} M\partial_t u = \widetilde{\mathcal{L}} u + g & \text{in } \mathbb{R}_+ \times \Omega, \\ u(0) = u^0 & \text{in } \Omega. \end{cases}$$

Further, the div-grad operator $\widetilde{\mathcal{L}}$ is brought into the form of a Friedrichs' operator by defining the coefficients

$$L_0 = 0, \qquad L_i = \begin{pmatrix} 0 & e_i^T \\ e_i & 0 \end{pmatrix}, \qquad i = 1, \dots, d.$$

Since we have $\nabla\cdot \widetilde{\mathcal{L}} = 0 \in L^\infty(\Omega)^{m\times m}$, the conditions of Definition 2.15 are fulfilled, and the operator is indeed a Friedrichs' operator. The graph space of $\widetilde{\mathcal{L}}$ is given by

$$\begin{aligned} H(\widetilde{\mathcal{L}}) &= \{(p,q) \in L^2(\Omega) \times L^2(\Omega)^d \mid \nabla p \in L^2(\Omega)^d,\ \nabla\cdot q \in L^2(\Omega)\} \\ &= H^1(\Omega) \times H(\mathrm{div};\Omega), \end{aligned}$$

where $H(\mathrm{div};\Omega)$ is the graph space of the divergence operator, cf., [Monk, 2003, Section 3.5.2], [Di Pietro and Ern, 2012, Section 1.2.6] or [Dautray and Lions, 1988, Section IX.1.2].

Associated boundary operator

Throughout the rest of this section, let

$$v = \begin{pmatrix} p \\ q \end{pmatrix} \qquad \text{and} \qquad w = \begin{pmatrix} \widetilde{p} \\ \widetilde{q} \end{pmatrix}.$$

Let $H^{1/2}(\Gamma)$ be the vector space spanned by the traces of functions in $H^1(\Omega)$ on Γ. By [Di Pietro and Ern, 2012, Remark 1.26], functions in $H(\mathrm{div};\Omega)$ then have normal traces in $H^{-1/2}(\Gamma)$, the dual space of $H^{1/2}(\Gamma)$. The boundary operator associated with the div-grad operator $\widetilde{\mathcal{L}}$ can thus be represented by

$$\langle \widetilde{\mathcal{L}}_\partial v \,|\, w \rangle = \langle \mathfrak{n}\cdot q \,|\, \widetilde{p} \rangle + \langle \mathfrak{n}\cdot \widetilde{q} \,|\, p \rangle \qquad \text{for all } v, w \in H(\widetilde{\mathcal{L}}).$$

Homogeneous Dirichlet boundary conditions

We consider homogeneous **Dirichlet boundary conditions**, which can be implemented by defining the boundary operator

$$\langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w \rangle = \langle \mathfrak{n}\cdot q \,|\, \widetilde{p} \rangle - \langle \mathfrak{n}\cdot \widetilde{q} \,|\, p \rangle \qquad \text{for all } v, w \in H(\widetilde{\mathcal{L}}). \tag{2.31}$$

It is apparent that this operator fulfills Definition 2.21 (B1), since for all $v, w \in H(\widetilde{\mathcal{L}})$ we have

$$\langle \widetilde{\mathcal{L}}_\Gamma w \,|\, v \rangle = \langle \mathfrak{n}\cdot \widetilde{q} \,|\, p \rangle - \langle \mathfrak{n}\cdot q \,|\, \widetilde{p} \rangle = -\langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w \rangle,$$

i.e., $\widetilde{\mathcal{L}}_\Gamma$ is skew-symmetric. To confirm condition (B2) let $v, w \in H(\widetilde{\mathcal{L}})$. We have

$$\langle (\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma) w \,|\, v \rangle = 2 \langle \mathfrak{n}\cdot \widetilde{q} \,|\, p \rangle$$

and

$$\langle (\widetilde{\mathcal{L}}_\partial + \widetilde{\mathcal{L}}_\Gamma) w \,|\, v \rangle = 2 \langle \mathfrak{n} \cdot q \,|\, \widetilde{p} \rangle.$$

Therefore, we can decompose arbitrary $v \in H(\widetilde{\mathcal{L}})$ into

$$v = \begin{pmatrix} p \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ q \end{pmatrix} \qquad \text{with} \qquad \begin{pmatrix} p \\ 0 \end{pmatrix} \in \ker(\widetilde{\mathcal{L}}_\partial + \widetilde{\mathcal{L}}_\Gamma) \quad \text{and} \quad \begin{pmatrix} 0 \\ q \end{pmatrix} \in \ker(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma),$$

showing Definition 2.21 (B2).

Remark 2.30. Let $v, w \in H(\widetilde{\mathcal{L}})$ be sufficiently smooth. Then the boundary operators can be represented by

$$\langle \widetilde{\mathcal{L}}_\partial v \,|\, w \rangle = (\mathfrak{n} \cdot q \,|\, \widetilde{p})_\Gamma + (p \,|\, \mathfrak{n} \cdot \widetilde{q})_\Gamma$$

and

$$\langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w \rangle = (\mathfrak{n} \cdot q \,|\, \widetilde{p})_\Gamma - (p \,|\, \mathfrak{n} \cdot \widetilde{q})_\Gamma.$$

This amounts to

$$\langle (\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma) v \,|\, w \rangle = 2 (p \,|\, \mathfrak{n} \cdot \widetilde{q})_\Gamma = 0 \qquad \text{for all } w \in C^\infty(\overline{\Omega}) \times C^\infty(\overline{\Omega})^d,$$

since $C^\infty(\overline{\Omega}) \times C^\infty(\overline{\Omega})^d \subset H(\widetilde{\mathcal{L}}) = H^1(\Omega) \times H(\mathrm{div}; \Omega)$, implying

$$p = 0 \qquad \text{a.e. on } \Gamma.$$

Hence, we recover the usual representation for homogeneous Dirichlet boundary conditions.

We want to point out that other boundary conditions like Neumann or Robin conditions can be easily incorporated in this setting. We refer to [Di Pietro and Ern, 2012, Section 7.1.2.2 & 7.1.5.2] for details. ◇

Conclusion

As we have $\nabla \cdot \widetilde{\mathcal{L}} = 0$ and $L_0 = 0$, condition (2.20) is fulfilled. Hence, all in all, the div-grad operator $\widetilde{\mathcal{L}}$ restricted to $\ker(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma)$ is a dissipative Friedrichs' operator by Definition 2.23. As a consequence, the considered wave equation fits the setting from above.

2.5.3 Maxwell's equations

The last example we consider are Maxwell's equations, which are fundamental for describing the propagation of electromagnetic waves.

Problem formulation

Let $d = 3$ and $m = 6$. We consider linear **Maxwell's equations** including external currents and damping given by

$$\begin{cases} \varepsilon \partial_t E = \nabla \times H - \sigma E - J & \text{in } \mathbb{R}_+ \times \Omega, \\ \mu \partial_t H = -\nabla \times E & \text{in } \mathbb{R}_+ \times \Omega, \\ E(0) = E^0, \quad H(0) = H^0 & \text{in } \Omega. \end{cases} \tag{2.32}$$

The solutions to this equation are the **electric field** $E \colon \Omega \to \mathbb{R}^3$ and the **magnetic field** $H \colon \Omega \to \mathbb{R}^3$. Further, $J \in L^2(\Omega)^3$ is the **current density**, $\sigma \in L^\infty(\Omega)^3$ is the **conductivity**, and $\varepsilon, \mu \in L^\infty(\Omega)^3$ are the **permittivity** and the **permeability**, respectively. We assume that ε and μ are uniformly positive and $\sigma \geq 0$ a.e. on Ω.

Formulation as a wave-type equation

We set

$$u = \begin{pmatrix} E \\ H \end{pmatrix}, \qquad u^0 = \begin{pmatrix} E^0 \\ H^0 \end{pmatrix}, \qquad g = \begin{pmatrix} J \\ 0 \end{pmatrix}$$

and

$$\widetilde{\mathscr{L}} = \begin{pmatrix} 0 & \nabla\times \\ -\nabla\times & 0 \end{pmatrix} - \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}, \qquad M = \begin{pmatrix} \varepsilon & 0 \\ 0 & \mu \end{pmatrix}$$

to rewrite (2.32) in the form of (2.22), i.e.,

$$\begin{cases} M\partial_t u = \widetilde{\mathscr{L}} u + g & \text{in } \mathbb{R}_+ \times \Omega, \\ u(0) = u^0 & \text{in } \Omega. \end{cases}$$

The Maxwell operator $\widetilde{\mathscr{L}}$ can be cast into the form of a Friedrichs' operator by defining the coefficients

$$L_0 = \begin{pmatrix} -\sigma & 0 \\ 0 & 0 \end{pmatrix}, \qquad L_i = \begin{pmatrix} 0 & \ell_i^T \\ \ell_i & 0 \end{pmatrix}, \qquad i = 1, 2, 3,$$

where $\ell_1, \ell_2, \ell_3 \in \mathbb{R}^{3\times 3}$ $\ell_1 = e_2 e_3^T - e_3 e_2^T$, $\ell_2 = e_3 e_1^T - e_1 e_3^T$ and $\ell_3 = e_1 e_2^T - e_2 e_1^T$.

As we have $\nabla\cdot\widetilde{\mathscr{L}} = 0 \in L^\infty(\Omega)^{6\times 6}$, the conditions of Definition 2.15 are fulfilled, showing that the operator is indeed a Friedrichs' operator. The graph space of $\widetilde{\mathscr{L}}$ is given by

$$\begin{aligned} H(\widetilde{\mathscr{L}}) &= \{(E, H) \in L^2(\Omega)^3 \times L^2(\Omega)^3 \mid \nabla\times E \in L^2(\Omega)^3,\ \nabla\times H \in L^2(\Omega)^3\} \\ &= H(\operatorname{curl}; \Omega) \times H(\operatorname{curl}; \Omega), \end{aligned}$$

where $H(\operatorname{curl}; \Omega)$ is the graph space of the curl operator, cf., [Monk, 2003, Section 3.5.3] or [Dautray and Lions, 1988, Section IX.1.2].

Associated boundary operator

Throughout the rest of this section, let

$$v = \begin{pmatrix} E \\ H \end{pmatrix} \qquad \text{and} \qquad w = \begin{pmatrix} \widetilde{E} \\ \widetilde{H} \end{pmatrix}.$$

By [Monk, 2003, Theorem 3.29] functions in $H(\operatorname{curl}; \Omega)$ in general posses tangential traces in $H^{-1/2}(\Gamma)$. Thus, we are not able to represent the boundary operator in the fashion of the corresponding operator for the wave or advection equation. Because of this we use Definition 2.19 to obtain the representation

$$\langle \widetilde{\mathscr{L}}_\partial v \,|\, w \rangle = \left(\nabla\times H \,\middle|\, \widetilde{E} \right)_\Omega - \left(\nabla\times E \,\middle|\, \widetilde{H} \right)_\Omega + \left(E \,\middle|\, \nabla\times \widetilde{H} \right)_\Omega - \left(H \,\middle|\, \nabla\times \widetilde{E} \right)_\Omega$$

for all $v, w \in H(\widetilde{\mathscr{L}})$.

Perfectly conducting boundary conditions

We consider **perfectly conducting boundary conditions**, which amount to vanishing tangential traces of the electric field. They can be modeled by defining the boundary operator

$$\langle \widetilde{\mathscr{L}}_\Gamma v \,|\, w \rangle = \left(\nabla\times H \,\middle|\, \widetilde{E} \right)_\Omega + \left(\nabla\times E \,\middle|\, \widetilde{H} \right)_\Omega - \left(E \,\middle|\, \nabla\times \widetilde{H} \right)_\Omega - \left(H \,\middle|\, \nabla\times \widetilde{E} \right)_\Omega \tag{2.33}$$

for all $v, w \in H(\widetilde{\mathcal{L}})$. This operator is skew-symmetric as we have

$$\begin{aligned}\langle \widetilde{\mathcal{L}}_\Gamma w \,|\, v\rangle &= \big(\nabla\times\widetilde{H} \,\big|\, E\big)_\Omega + \big(\nabla\times\widetilde{E} \,\big|\, H\big)_\Omega - \big(\widetilde{E} \,\big|\, \nabla\times H\big)_\Omega - \big(\widetilde{H} \,\big|\, \nabla\times E\big)_\Omega \\ &= -\langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w\rangle\end{aligned}$$

for all $v, w \in H(\widetilde{\mathcal{L}})$. Hence, $\widetilde{\mathcal{L}}_\Gamma$ fulfills condition (B1) of Definition 2.21. To confirm condition (B2), let $v, w \in H(\widetilde{\mathcal{L}})$. We have

$$\langle (\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma) w \,|\, v\rangle = 2\,\big(E \,\big|\, \nabla\times\widetilde{H}\big)_\Omega - 2\,\big(\nabla\times E \,\big|\, \widetilde{H}\big)_\Omega$$

and

$$\langle (\widetilde{\mathcal{L}}_\partial + \widetilde{\mathcal{L}}_\Gamma) w \,|\, v\rangle = 2\,\big(\nabla\times H \,\big|\, \widetilde{E}\big)_\Omega - 2\,\big(H \,\big|\, \nabla\times\widetilde{E}\big)_\Omega.$$

This yields

$$v = \begin{pmatrix} E \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ H \end{pmatrix} \qquad \text{with} \qquad \begin{pmatrix} E \\ 0 \end{pmatrix} \in \ker(\widetilde{\mathcal{L}}_\partial + \widetilde{\mathcal{L}}_\Gamma) \quad \text{and} \quad \begin{pmatrix} 0 \\ H \end{pmatrix} \in \ker(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma)$$

for arbitrary $v \in H(\widetilde{\mathcal{L}})$, showing Definition 2.21 (B2).

Remark 2.31. Let $v, w \in H(\widetilde{\mathcal{L}})$ be sufficiently smooth. Then, by the usual integration by parts formula for the curl operator, the boundary operators can be represented by

$$\langle \widetilde{\mathcal{L}}_\partial v \,|\, w\rangle = \big(\mathfrak{n}\times H \,\big|\, \widetilde{E}\big)_\Gamma - \big(\mathfrak{n}\times E \,\big|\, \widetilde{H}\big)_\Gamma$$

and

$$\langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w\rangle = \big(\mathfrak{n}\times H \,\big|\, \widetilde{E}\big)_\Gamma + \big(\mathfrak{n}\times E \,\big|\, \widetilde{H}\big)_\Gamma.$$

Hence, for $v \in \ker(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma)$ smooth enough this amounts to

$$\langle (\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma) v \,|\, w\rangle = -2\,\big(\mathfrak{n}\times E \,\big|\, \widetilde{H}\big)_\Gamma = 0 \qquad \text{for all } w \in C^\infty(\overline{\Omega})^3 \times C^\infty(\overline{\Omega})^3,$$

since $C^\infty(\overline{\Omega})^3 \times C^\infty(\overline{\Omega})^3 \subset H(\widetilde{\mathcal{L}}) = H(\mathrm{curl};\Omega) \times H(\mathrm{curl};\Omega)$. This implies

$$\mathfrak{n}\times E = 0 \qquad \text{a.e. on } \Gamma,$$

recovering the usual representation for perfectly conducting boundary conditions.

In fact, it can be shown (cf., [Di Pietro and Ern, 2012, Lemma 3.5 (ii)]) that we have

$$\ker(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma) = \{(E, H) \in H(\widetilde{\mathcal{L}}) \mid (\mathfrak{n}\times E)|_\Gamma = 0\}.$$

This is the domain usually associated with the Maxwell operator under perfectly conducting boundary conditions, see e.g., [Hochbruck et al., 2015a, Section 3.2]. ◇

Conclusion

Condition (2.20) is fulfilled, since we have $\nabla\cdot\widetilde{\mathcal{L}} = 0$ and $L_0 \leq 0$ since $\sigma \geq 0$. Thus, by Definition 2.23, the restriction of the Maxwell operator $\widetilde{\mathcal{L}}$ to $\ker(\widetilde{\mathcal{L}}_\partial - \widetilde{\mathcal{L}}_\Gamma)$ is a dissipative Friedrichs' operator. Consequently, (2.32) fits our framework.

3 | Spatial discretization

In this section, we discuss the spatial discretization of the wave-type problem (2.24). In particular, we employ the discontinuous Galerkin (dG) method to discretize the spatial differential operators, cf., [Di Pietro and Ern, 2012,Hesthaven and Warburton, 2008]. Using these discrete operators, we then state and analyze the spatially semidiscrete version of (2.24) in a suitable subspace of $L^2(\Omega)^m$.

The chapter is organized as follows. Sections 3.1–3.3 are devoted to establish the discrete setting needed for the formulation and analysis of the spatially semidiscrete problem. In Section 3.4 we briefly discuss some concepts related to Friedrichs' operators in this discrete setting. Having done this, we introduce the central fluxes dG discretization of a general Friedrichs' operator in Section 3.5. Consequently, in the same section, we investigate some properties of such discrete operators crucial for the analysis of the spatial discretization. We then use the discrete Friedrichs' operators in Section 3.6 to formulate the spatially discrete version of the wave-type problem (2.24). Finally, we investigate the error made by approximating the exact solution of (2.24) by the spatially semidiscrete solution in Section 3.7.

3.1 Meshes

We begin by posing the following assumption to avoid unnecessary technicalities.

Assumption 3.1. *We assume that the domain Ω is a polyhedron in $\mathbb{R}^d$.*

In particular, this means that we can discretize Ω exactly by using a polyhedral mesh. We are mostly interested in tensorial meshes. However, the simpler case of a simplicial mesh is needed to derive important properties of such meshes.

Definition 3.2. *Let $\{x_0, \dots, x_d\}$ be a set of $d+1$ points in $\mathbb{R}^d$ such that the vectors $x_1 - x_0, \dots, x_d - x_0$ are linearly independent. We call the interior of the convex hull of $\{x_0, \dots, x_d\}$ a **non-degenerate simplex** in $\mathbb{R}^d$.*

Definition 3.3. *A finite set $\mathcal{T} = \{K\}$ is called a **simplicial mesh** of the domain Ω if it satisfies the following.*

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

(ii) *The set $\mathcal{T}$ forms a partition of Ω, i.e., $\overline{\Omega} = \bigcup_{K\in\mathcal{T}} \overline{K}$ and $K \cap \widehat{K} = \emptyset$ for all $K, \widehat{K} \in \mathcal{T}$ with $K \neq \widehat{K}$.*

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

As stated before, we are interested in tensorial meshes, which need a more general notion of a mesh.

Definition 3.4. *A finite set $\mathcal{T} = \{K\}$ of polyhedra K is called a **general mesh** of the domain Ω if it satisfies Definition 3.3 (ii). Each $K \in \mathcal{T}$ is called a **mesh element**.*

Since everything in the following can be done on general meshes without additional effort, we will only restrict ourselves to the tensorial case, when we need it. That will be in Chapter 6, when we discuss the efficiency of the ADI method.

In the next definition we introduce some notation concerning the geometrical properties of a mesh and its elements.

Definition 3.5. *Let $\mathcal{T}$ be a mesh of Ω. For all $K \in \mathcal{T}$ we denote the **diameter** of K by h_K and define the piecewise constant function $h \in L^\infty(\Omega)$ by $h|_K \equiv h_K$ for all $K \in \mathcal{T}_\hbar$. Furthermore, we define the **radius of the largest ball inscribed in** K by r_K and the **meshsize** $\hbar$ as the maximal diameter of all mesh elements of $\mathcal{T}$, i.e.,*

$$\hbar = \max_{K \in \mathcal{T}} h_K.$$

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

Remark 3.6. The definition of the piecewise function h enables us to write mesh-dependent norms used in [Sturm, 2017, Section 3.5] more concisely, since for $v \in L^2(\Omega)^m$ and $p \in \mathbb{Z}$ we have

$$\|h^p v\|_\Omega^2 = \sum_{K \in \mathcal{T}_\hbar} \|h^p v\|_K^2 = \sum_{K \in \mathcal{T}_\hbar} h_K^{2p} \|v\|_K^2.$$

Note that the mapping $\|h^p \cdot\|_\Omega$ indeed defines a norm. Hence, we will refer to this as mesh-dependent norms throughout the thesis. Further, we will use a similar notation for the broken Sobolev norms introduced later.

For the rest of this thesis, keep in mind that on $K \in \mathcal{T}_\hbar$ we have

$$\|hv\|_K = h_K \|v\|_K.$$

We will use the notation on the left hand side throughout the thesis, since it better reflects the notation for the mesh-dependent norms. ◇

Next, we define the faces of a mesh.

Definition 3.7. *Let $\mathcal{T}_\hbar$ be a general mesh of Ω. We say that a closed subset F of $\overline{\Omega}$ is a **mesh face** if F has positive $(d-1)$-dimensional Hausdorff measure and if either one of the following conditions is satisfied.*

(i) *There are distinct mesh elements $K_1, K_2 \in \mathcal{T}_\hbar$ such that $F = \partial K_1 \cap \partial K_2$. In this case, we call F an **interface**.*

(ii) *There is a mesh element $K \in \mathcal{T}_\hbar$ such that $F = \partial K \cap \partial\Omega$. In this case, we call F a **boundary face**.*

We denote the set of all interfaces by $\mathcal{F}_\hbar^{\text{int}}$ and the set of all boundary faces by $\mathcal{F}_\hbar^{\text{bnd}}$. The set of all faces is denoted by

$$\mathcal{F}_\hbar = \mathcal{F}_\hbar^{\text{int}} \cup \mathcal{F}_\hbar^{\text{bnd}}.$$

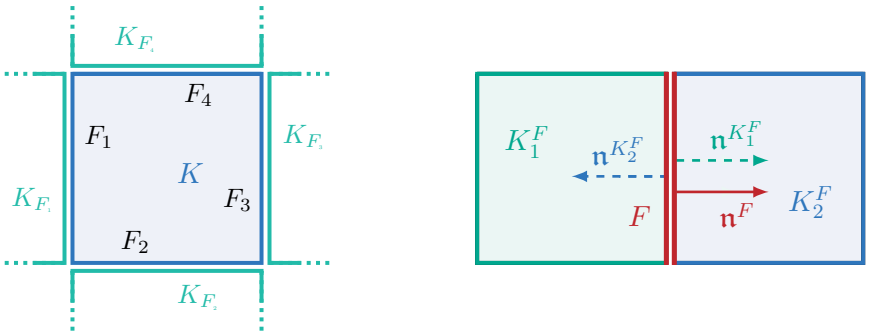


Figure 3.1: Convention for the element neighbors and normal vectors corresponding to Definitions 3.7 and 3.8.

For given $K \in \mathcal{T}_h$, we denote the set of all interfaces and the set of all boundary faces composing the boundary of K by

$$\mathcal{F}_h^{K,\mathrm{int}} = \{F \in \mathcal{F}_h^{\mathrm{int}} \mid F \subset \partial K\} \qquad \text{and} \qquad \mathcal{F}_h^{K,\mathrm{bnd}} = \{F \in \mathcal{F}_h^{\mathrm{bnd}} \mid F \subset \partial K\},$$

respectively. Consequently, the set of all faces composing the boundary of K is denoted by

$$\mathcal{F}_h^K = \mathcal{F}_h^{K,\mathrm{int}} \cup \mathcal{F}_h^{K,\mathrm{bnd}},$$

and we denote the ***maximum number of faces per element*** *in $\mathcal{T}_h$ by*

$$N_\partial = \max_{K \in \mathcal{T}_h} |\mathcal{F}_h^K|.$$

Lastly, given an interface $F \in \mathcal{F}_h^{K,\mathrm{int}}$ of the element K we denote by K_F the element with $F = \partial K \cap \partial K_F$. We call K_F the ***neighbor of K w.r.t. the interface*** *F.*

Definition 3.8. *Let $\mathcal{T}_h$ be a general mesh of Ω. For all $K \in \mathcal{T}_h$ we define the* ***element normal vector*** *$\mathfrak{n}^K$ a.e. on ∂K as the outward unit normal vector to K.*

For all interfaces $F \in \mathcal{F}_h^{\mathrm{int}}$ we arbitrarily denote the two neighboring elements, whose boundaries contain F, as K_1^F and K_2^F. We fix this choice and define the ***face normal vector*** *$\mathfrak{n}^F$ a.e. on F as the outward unit normal vector to K_1^F. For all boundary faces $F \in \mathcal{F}_h^{\mathrm{bnd}}$, we define $\mathfrak{n}^F$ a.e. on F as the outward unit normal vector to $\partial\Omega$.*

Function spaces considered in the dG method consist of functions that are only piecewise smooth, i.e., smooth on each mesh element K but not necessarily on the whole domain Ω. Such functions are smooth enough to admit traces on the faces of the mesh, but can be discontinuous across these faces. Hence, it is possible that such functions possess two-valued traces on each face $F \in \mathcal{F}_h^{\mathrm{int}}$. This motivates the following definition of the jump and the average of a function across an interface.

Definition 3.9. *Let $v \colon \Omega \to \mathbb{R}$ be a function such that for all mesh elements $K \in \mathcal{T}_h$ the restriction $v|_K$ admits a trace a.e. on ∂K. Then, for all $K \in \mathcal{T}_h$, we denote with v^K the extension of $v|_K$ to $\overline{K}$. We omit this superscript if there is no ambiguity (e.g., on boundary faces $F \in \mathcal{F}_h^{\mathrm{bnd}}$, where the trace is one-valued).*

With this, we define the ***average of*** *v across an interior face $F \in \mathcal{F}_h^{\mathrm{int}}$ as*

$$\{\!\{v\}\!\}_F = \frac{v^{K_1^F}|_F + v^{K_2^F}|_F}{2}$$

and the ***jump of*** v *across* F *as*

$$[\![v]\!]_F = v^{K_1^F}|_F - v^{K_2^F}|_F.$$

For vector or matrix valued fields, these operations act componentwise.

To investigate the error made by discretization through the dG method, we consider a sequence of meshes

$$\mathcal{T}_\mathcal{H} = (\mathcal{T}_\hbar)_{\hbar\in\mathcal{H}}$$

discretizing the spatial domain Ω. Here and in the following, $\mathcal{H}$ denotes a countable collection of positive numbers with 0 as only accumulation point. As we are only interested in the behavior for $\hbar \to 0$, we assume $\hbar < 1$ for all $\hbar \in \mathcal{H}$.

To derive error bounds involving constants that are independent of the meshsize $\hbar$, we assume the mesh sequence $\mathcal{T}_\mathcal{H}$ to possess certain qualities. One of them is shape- and contact-regularity. Before we can define this property, we need some auxiliary definitions.

Definition 3.10. *We call* $\mathcal{T}_\hbar$ *a* ***matching simplicial mesh*** *if it is a simplicial mesh, and if for every* $K \in \mathcal{T}_\hbar$ *with vertices* $\{x_0, \dots, x_d\}$*, the set* $\partial K \cup \partial\widehat{K}$*,* $\widehat{K} \in \mathcal{T}_\hbar$*, is the convex hull of a (possibly empty) subset of* $\{x_0, \dots, x_d\}$*.*

Definition 3.11. *Let* $\mathcal{T}_\hbar$ *be a general mesh. We call* $\mathcal{T}'_\hbar$ *a* ***matching simplicial submesh*** *if the following holds.*

(i) *The mesh* $\mathcal{T}'_\hbar$ *is a matching simplicial mesh with set of all faces denoted by* $\mathcal{F}'_\hbar$*.*

(ii) *For all* $K' \in \mathcal{T}'_\hbar$ *there is exactly one* $K \in \mathcal{T}_\hbar$ *such that* $K' \subset K$*.*

(iii) *For all* $F' \in \mathcal{F}'_\hbar$ *there is at most one* $F \in \mathcal{F}_\hbar$ *such that* $F' \subset F$*.*

With this we can define shape- and contact-regular mesh sequences.

Definition 3.12. *Let* $\mathcal{T}_\mathcal{H} = (\mathcal{T}_\hbar)_{\hbar\in\mathcal{H}}$ *be a mesh sequence, which admits a matching simplicial submesh* $\mathcal{T}'_\hbar$ *for all* $\hbar \in \mathcal{H}$*.*

(i) $\mathcal{T}_\mathcal{H}$ *is* ***shape-regular*** *if there is* $\rho_1 > 0$ *independent of* $\hbar$*, such that for all* $K' \in \mathcal{T}'_\hbar$ *we have*

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

(ii) $\mathcal{T}_\mathcal{H}$ *is* ***contact-regular*** *if there is* $\rho_2 > 0$ *independent of* $\hbar$*, such that for all* $K \in \mathcal{T}_\hbar$ *and all* $K' \in \mathcal{T}'_\hbar$ *with* $K' \in K$ *we have*

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

We denote the product of the mesh parameters ρ_1 *and* ρ_2 *by*

$$\rho = \rho_1\rho_2.$$

One important property of shape- and contact-regular mesh sequences is that the maximum number of faces per element N_∂ is bounded independently of $\hbar \in \mathcal{H}$, cf., [Di Pietro and Ern, 2012, Lemma 1.41]. The next lemma further shows that the diameters of neighboring elements of meshes belonging to shape- and contact-regular mesh sequences can be compared using a constant independent of the discretization parameter $\hbar$. The proof can be found in [Di Pietro and Ern, 2012, Lemma 1.43].

Lemma 3.13. *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence. Then, for all $h \in \mathcal{H}$ and each interface $F \in \mathcal{F}_h^{\text{int}}$ we have*

$$\max\{h_{K_1^F}, h_{K_2^F}\} \leq \rho \min\{h_{K_1^F}, h_{K_2^F}\}.$$

3.2 Broken polynomial spaces

To approximate functions in space, we use piecewise polynomials of degree at most k in each variable. These functions are gathered in the **broken polynomial space**

$$\mathbb{Q}_d^k(\mathcal{T}_h) = \{\, \boldsymbol{v} \in L^2(\Omega) \mid \boldsymbol{v}|_K \in \mathbb{Q}_d^k(K) \text{ for all } K \in \mathcal{T}_h \,\}, \tag{3.1}$$

where $\mathbb{Q}_d^k(K)$ denotes the set of polynomials on $\mathbb{R}^d$ of degree at most k in each variable on K. The space $\mathbb{Q}_d^k(\mathcal{T}_h)$ consists of functions that are such polynomials on each mesh element K but need not to be continuous across mesh faces.

Since we need to approximate $\mathbb{R}^m$-vector fields, we further introduce the **approximation space** (or **dG space**)

$$V_h = (\mathbb{Q}_d^k(\mathcal{T}_h))^m. \tag{3.2}$$

Remark 3.14. 1. Note that the dG method is flexible enough to admit varying polynomial degrees on each element $K \in \mathcal{T}_h$ without much effort. But for the sake of readability, we only present the case, where we have the same maximal polynomial degree on all elements. However, since all derived error bounds are given in an elementwise manner, the generalization to varying polynomial degrees is straightforward.

2. Other broken polynomial spaces can be considered. Most prominently the space of broken polynomials of total degree at most k given by

$$\mathbb{P}_d^k(\mathcal{T}_h) = \{\, \boldsymbol{v} \in L^2(\Omega) \mid \boldsymbol{v}|_K \in \mathbb{P}_d^k(K) \text{ for all } K \in \mathcal{T}_h \,\},$$

which is usually associated with simplicial elements. Here, $\mathbb{P}_d^k(K)$ denotes the set of polynomials on $\mathbb{R}^d$ of total degree at most k on K. We refer to [Di Pietro and Ern, 2012, Section 1.2.4.3] for more details. ◇

Throughout, we will frequently make use of the projection of an L^2-function onto the broken polynomial space $\mathbb{Q}_d^k(\mathcal{T}_h)$. We introduce it in the next definition.

Definition 3.15. *We define the L^2-**orthogonal projection** $\pi_h \colon L^2(\Omega) \to \mathbb{Q}_d^k(\mathcal{T}_h)$ onto $\mathbb{Q}_d^k(\mathcal{T}_h)$, such that*

$$\big(v - \pi_h v \,\big|\, \boldsymbol{\varphi}\big)_\Omega = 0 \qquad \text{for all } \boldsymbol{\varphi} \in \mathbb{Q}_d^k(\mathcal{T}_h).$$

*Using this, we define the **projection error of** v a function $v \in L^2(\Omega)$ as*

$$e_\pi^v = v - \pi_h v.$$

For vector fields the projection π_h and the projection error are defined componentwise.

We often need the boundedness of the L^2-projection. We give the result for the scalar case as the vector field case follows by the componentwise definition of the projection.

Lemma 3.16. *For $v \in L^2(\Omega)$ we have*

$$\|\pi_h v\|_\Omega \leq \|v\|_\Omega.$$

Proof. The bound follows by

$$\|\pi_h v\|_\Omega = \sup_{\substack{\boldsymbol{\varphi} \in \mathbb{Q}_d^k(\mathcal{T}_\hbar) \\ \|\boldsymbol{\varphi}\|_\Omega = 1}} \left(\pi_h v \,\middle|\, \boldsymbol{\varphi}\right)_\Omega = \sup_{\substack{\boldsymbol{\varphi} \in \mathbb{Q}_d^k(\mathcal{T}_\hbar) \\ \|\boldsymbol{\varphi}\|_\Omega = 1}} \left(v \,\middle|\, \boldsymbol{\varphi}\right)_\Omega \leq \sup_{\substack{\boldsymbol{\varphi} \in \mathbb{Q}_d^k(\mathcal{T}_\hbar) \\ \|\boldsymbol{\varphi}\|_\Omega = 1}} \|v\|_\Omega \|\boldsymbol{\varphi}\|_\Omega = \|v\|_\Omega,$$

where we have used the definition of the L^2-projection in the second and the Cauchy–Schwarz inequality in the third step. □

By the elementwise nature of the broken polynomial spaces we obtain that the L^2-projection can also be carried out elementwise. Again, we state the scalar case, whereby the vector field case readily follows.

Lemma 3.17. *Let $v \in L^2(\Omega)$. Then, for all $K \in \mathcal{T}_\hbar$, we have*

$$\left(v - \pi_h v \,\middle|\, \boldsymbol{\varphi}\right)_K = 0 \qquad \text{for all } \boldsymbol{\varphi} \in \mathbb{Q}_d^k(K).$$

Proof. We have

$$\left(v - \pi_h v \,\middle|\, \boldsymbol{\varphi}\right)_K = \left(v - \pi_h v \,\middle|\, \mathbb{1}_K \boldsymbol{\varphi}\right)_\Omega \qquad \text{for all } \boldsymbol{\varphi} \in \mathbb{Q}_d^k(K).$$

The assertion now follows by $\mathbb{1}_K \boldsymbol{\varphi} \in \mathbb{Q}_d^k(\mathcal{T}_\hbar)$. □

3.2.1 Inverse and trace inequality

Next, we investigate properties of the piecewise polynomial spaces that are important for the derivation of error bounds.

Lemma 3.18. *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence. Then, for all $\hbar \in \mathcal{H}$, all $\boldsymbol{v} \in \mathbb{Q}_d^k(\mathcal{T}_\hbar)$ and for all $K \in \mathcal{T}_\hbar$ we have*

$$\|\nabla \boldsymbol{v}\|_K \leq C'_{\mathrm{inv}} \|h^{-1} \boldsymbol{v}\|_K,$$

where C'_{inv} only depends on the dimension d, the polynomial degree k and the mesh regularity parameters ρ_1 and ρ_2.

Proof. The proof is analogous to the proof of [Di Pietro and Ern, 2012, Lemma 1.44], where the space $\mathbb{P}_d^k(\mathcal{T}_\hbar)$ is considered. The only part, where this plays a role is the application of [Ern and Guermond, 2004, Lemma 1.138], which can also be applied in the case of $\mathbb{Q}_d^k(\mathcal{T}_\hbar)$. □

As this gives a bound on the individual partial derivatives, we can easily derive a similar bound for Friedrichs' operators.

Lemma 3.19. *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence and let $\mathcal{F}$ be a Friedrichs' operator with coefficients $\left(F_i\right)_{i=0}^d$. Then, for all $\hbar \in \mathcal{H}$, all $\boldsymbol{v} \in V_\hbar$ and for all $K \in \mathcal{T}_\hbar$ we have*

$$\|\mathcal{F} \boldsymbol{v}\|_K \leq C_\mathcal{F} C_{\mathrm{inv}} \|h^{-1} \boldsymbol{v}\|_K$$

with $C_\mathcal{F} = \max_{i=0,\dots,d} \|F_i\|_{\infty,\Omega}$ and $C_{\mathrm{inv}} = \sqrt{d}\, C'_{\mathrm{inv}} + 1$.

Proof. The boundedness of the coefficients of $\mathcal{F}$ yields

$$\|\mathcal{F}\boldsymbol{v}\|_K \leq C_{\mathcal{F}}\big(\sum_{i=1}^{d}\|\partial_i\boldsymbol{v}\|_K + \|\boldsymbol{v}\|_K\big).$$

Using the equivalence of the 1- and 2-norm on $\mathbb{R}^d$ we obtain

$$\sum_{i=1}^{d}\|\partial_i\boldsymbol{v}\|_K \leq \sqrt{d}\,\|\nabla\boldsymbol{v}\|_K.$$

The proof is concluded by using Lemma 3.18 and $\|\boldsymbol{v}\|_K \leq \|h^{-1}\boldsymbol{v}\|_K$, since $h_K \leq \hbar < 1$ for all $K \in \mathcal{T}_\hbar$. □

We also need the discrete trace inequality.

Lemma 3.20. *Let $\mathcal{T}_\mathcal{H}$ be a shape- and contact-regular mesh sequence. Then, for all $\hbar \in \mathcal{H}$, all $\boldsymbol{v} \in \mathbb{Q}_d^k(\mathcal{T}_\hbar)$ and for all $K \in \mathcal{T}_\hbar$ and $F \in \mathcal{F}_\hbar^K$ we have*

$$\|\boldsymbol{v}\|_F \leq C_{\mathrm{tr}}\|h^{-1/2}\boldsymbol{v}\|_K,$$

where C_{tr} only depends on the dimension d, the polynomial degree k and the mesh regularity parameters ρ_1 and ρ_2.

Proof. This is proven analogously to [Di Pietro and Ern, 2012, Lemma 1.46], where the space $\mathbb{P}_d^k(\mathcal{T}_\hbar)$ is considered. The only difference is that we use the compactness of the unit sphere in $\mathbb{Q}_d^k(\widehat{K})$ w.r.t. the L^2-norm instead of the respective property of $\mathbb{P}_d^k(\widehat{K})$ with $\widehat{K}$ being the unit simplex. □

3.2.2 Optimal polynomial approximation

In the dG method we approximate functions by discrete functions contained in the space $\mathbb{Q}_d^k(\mathcal{T}_\hbar)$. Consequently, we are interested in investigating the error made by this approximation. It turns out that this depends on the mesh sequence under consideration. In this thesis we focus on mesh sequences that allow polynomial approximation properties in the following sense (cf., [Di Pietro and Ern, 2012, Definition 1.55]).

Definition 3.21. *We say that the mesh sequence $\mathcal{T}_\mathcal{H}$ has **optimal polynomial approximation properties** if for all $\hbar \in \mathcal{H}$, all $K \in \mathcal{T}_\hbar$ and all k, there is a linear interpolation operator $\mathcal{I}_K^k \colon L^2(K) \to \mathbb{Q}_d^k(K)$ such that for all $s \in \{0,\dots,k\}$ and all $v \in H^s(K)$, there holds*

$$|v - \mathcal{I}_K^k|_{q,K} \leq C'_{\mathrm{app}}h_K^{s-q}|v|_{s,K} \qquad \text{for all } q \in \{0,\dots,s\},$$

where C'_{app} is independent of both K and $\hbar$.

Mesh sequences that fulfill both shape- and contact regularity as well as having optimal polynomial approximation properties are called admissible mesh sequences.

Definition 3.22. *A shape- and contact-regular mesh sequence $\mathcal{T}_\mathcal{H}$ with optimal polynomial approximation properties is called an **admissible mesh sequence**.*

As we need this property to derive error bounds, we assume that $\mathcal{T}_\mathcal{H}$ possesses it.

Assumption 3.23. *We assume that $\mathcal{T}_\mathcal{H}$ is an **admissible mesh sequence**.*

With this, we obtain that the L^2-projection yields an optimal approximation in $\mathbb{Q}_d^k(\mathcal{T}_\hbar)$ to a given function if this function is smooth enough. As the proofs to [Di Pietro and Ern, 2012, Lemmas 1.58, 1.59] only need the inverse and discrete trace inequalities from Lemmas 3.18 and 3.20, respectively, they can be adjusted to the space $\mathbb{Q}_d^k(\mathcal{T}_\hbar)$ in a straightforward manner. This yields the following lemma.

Lemma 3.24. *Let $v \in H^{q+1}(K)$ for $0 \leq q \leq k$. Then, for all $\hbar \in \mathcal{H}$ and for all $K \in \mathcal{T}_\hbar$ we have*

$$\|e_\pi^v\|_K \leq C_{\text{app}} |h^{q+1} v|_{q+1,K},$$

and for all $F \in \mathcal{F}_\hbar^K$ we have

$$\|e_\pi^v\|_F \leq C_{\text{app},\partial} |h^{q+1/2} v|_{q+1,K},$$

where C_{app} and $C_{\text{app},\partial}$ are independent of both K and $\hbar$.

3.3 Broken Sobolev spaces

For the analysis of the discrete operators we need broken Sobolev spaces, which we introduce in the next definition.

Definition 3.25. *For $q \in \mathbb{N}_0$ we define the **broken Sobolev space of order** q as*

$$H^q(\mathcal{T}_\hbar) = \{ v \in L^2(\Omega) \mid v|_K \in H^q(K) \text{ for all } K \in \mathcal{T}_\hbar \}.$$

*On $H^q(\mathcal{T}_\hbar)$ we define the **broken Sobolev seminorm** and **norm** for $v \in H^q(\mathcal{T}_\hbar)$ by*

$$|v|_{q,\mathcal{T}_\hbar}^2 = \sum_{K \in \mathcal{T}_\hbar} |v|_{q,K}^2 \qquad \text{and} \qquad \|v\|_{q,\mathcal{T}_\hbar}^2 = \sum_{j=0}^{q} |v|_{j,\mathcal{T}_\hbar}^2,$$

respectively.

One important property of functions in these spaces is that they can be approximated optimally in $\mathbb{Q}_d^k(\mathcal{T}_\hbar)$ by Lemma 3.24. Further, for $v \in H^1(\mathcal{T}_\hbar)$ and $K \in \mathcal{T}_\hbar$ we have $v|_K \in H^1(K)$, and thus $v|_K$ has a well-defined trace in $L^2(\partial K)$.

By Definition 3.25 the usual Sobolev spaces are subspaces of their broken counterparts, i.e., for all $q \in \mathbb{N}_0$ we have $H^q(\Omega) \subset H^q(\mathcal{T}_\hbar)$. However, the converse is not true in general, as functions in $H^q(\mathcal{T}_\hbar)$ might have non-zero jumps across interfaces of the mesh. In contrast, jumps of functions in $H^q(\Omega)$ across interfaces vanish. In fact, together with being a subset of $H^q(\mathcal{T}_\hbar)$ this is a defining property of the usual Sobolev spaces, as the next lemma shows. The proof can be found in [Di Pietro and Ern, 2012, Lemma 1.23].

Lemma 3.26. *Let $v \in H^1(\mathcal{T}_\hbar)$. Then we have $v \in H^1(\Omega)$ if and only if*

$$[\![v]\!]_F = 0 \qquad \text{a.e. on } F \text{ for all } F \in \mathcal{F}_\hbar^{\text{int}}.$$

3.4 Friedrichs' operators in the discrete setting

The aim of this section is to introduce the dG discretization of a dissipative Friedrichs' operator. Hence, let $\mathcal{F}\colon D(\mathcal{F}) \to L^2(\Omega)^m$ be a dissipative Friedrichs' operator with coefficients $(F_i)_{i=0}^d$ and dissipative boundary condition $\mathcal{F}_\Gamma$.

Before introducing the discrete operator, we make additional assumptions on the coefficients of $\mathcal{F}$ and the boundary condition $\mathcal{F}_\Gamma$. We further investigate the intersection of the spaces $H(\mathcal{F})$ and $D(\mathcal{F})$ with the broken Sobolev spaces introduced in the last section. As we will see later, these results are needed for the analysis of the discrete operator.

3.4.1 Trace operators

For the definition of the discrete operators we assume more regularity of the coefficients $F_1, \dots, F_d$.

Assumption 3.27. *We assume that the coefficients fulfill* $F_1, \dots, F_d \in W^{1,\infty}(\Omega)^{m\times m}$.

This implies the following integration by parts formula.

Lemma 3.28. *Let* $K \subset \Omega$. *Then for* $v, w \in H^1(K)^m$ *we have*

$$\big(\mathcal{F}v \,\big|\, w\big)_K - \big(v \,\big|\, \mathcal{F}^{\circledast} w\big)_K = \big(\sum_{i=1}^{d} \mathfrak{n}_i^K F_i v \,\big|\, w\big)_{\partial K}, \tag{3.3}$$

where $\mathfrak{n}^K$ *denotes the outward unit normal vector to* K.

Proof. Using Definition 2.15, we obtain

$$\begin{aligned}\big(\mathcal{F}v \,\big|\, w\big)_K &= \big(\sum_{i=1}^{d} F_i \partial_i v + F_0 v \,\big|\, w\big)_K \\ &= \sum_{i=1}^{d} \big(\partial_i v \,\big|\, F_i w\big)_K + \big(v \,\big|\, F_0^* w\big)_K.\end{aligned}$$

Because of Assumption 3.27, for $i = 1, \dots, d$, we have

$$\partial_i\big(F_i w\big) = \partial_i F_i w + F_i \partial_i w \in L^2(\Omega)^m,$$

and we can therefore use the usual integration by parts formula, yielding

$$\begin{aligned}\big(\mathcal{F}v \,\big|\, w\big)_K &= \sum_{i=1}^{d} \Big(\big(v \,\big|\, -\partial_i(F_i w)\big)_K + \big(\mathfrak{n}_i^K v \,\big|\, F_i w\big)_{\partial K}\Big) + \big(v \,\big|\, F_0^* w\big)_K \\ &= \big(v \,\big|\, -\sum_{i=1}^{d} \partial_i(F_i w) + F_0^* w\big)_K + \big(\sum_{i=1}^{d} \mathfrak{n}_i^K F_i v \,\big|\, w\big)_{\partial K}.\end{aligned}$$

Using Definition 2.18 of the formal adjoint $\mathcal{F}^{\circledast}$ of $\mathcal{F}$ concludes the proof. □

In particular, choosing $K = \Omega$, this means that for $v, w \in H^1(\Omega)^m$, the operator $\mathcal{F}_\partial$ from Definition 2.19 can be represented as

$$\big\langle \mathcal{F}_\partial v \,\big|\, w \big\rangle = \big(\sum_{i=1}^{d} \mathfrak{n}_i F_i v \,\big|\, w\big)_\Gamma, \tag{3.4}$$

where $\mathfrak{n}$ denotes the outward unit normal vector to Γ.

Interfaces

We now introduce for each $K \in \mathfrak{T}_\hbar$ the boundary field $\mathcal{F}_\partial^K \colon \partial K \to \mathbb{R}^m$ associated with the elements of the mesh defined by

$$\mathcal{F}_\partial^K = \sum_{i=1}^{d} \mathfrak{n}_i^K F_i$$

and, for all $F \in \mathfrak{F}_\hbar$, the boundary field $\mathcal{F}_\partial^F \colon F \to \mathbb{R}^m$ associated with the faces of the mesh defined by

$$\mathcal{F}_\partial^F = \sum_{i=1}^{d} \mathfrak{n}_i^F F_i.$$

Because of Assumption 3.27, the coefficients of $\mathcal{F}$ are continuous over the faces of the mesh, and hence, $\mathcal{F}_\partial^F$ is well-defined on each face $F \in \mathfrak{F}_\hbar$.

Boundary faces

To obtain a representation of the boundary operator $\mathcal{F}_\Gamma$ similar to (3.4), we further assume the following.

Assumption 3.29. *We assume that the boundary operator $\mathcal{F}_\Gamma$ is associated with a matrix-valued boundary field $\widetilde{\mathcal{F}}_\Gamma \colon \Gamma \to \mathbb{R}^{m\times m}$ such that for $v, w \in H^1(\mathfrak{T}_\hbar)$ we have*

$$\big\langle \mathcal{F}_\Gamma v \,\big|\, w \big\rangle = \big(\widetilde{\mathcal{F}}_\Gamma v \,\big|\, w\big)_\Gamma.$$

For the sake of presentation, in the rest of the thesis we slightly abuse notation and identify the operator $\mathcal{F}_\Gamma$ with its associated matrix-valued field $\widetilde{\mathcal{F}}_\Gamma$ and use the symbol $\mathcal{F}_\Gamma$ instead of $\widetilde{\mathcal{F}}_\Gamma$ for this field.

3.4.2 The spaces $H(\mathcal{F}) \cap H^1(\mathfrak{T}_\hbar)^m$ and $D(\mathcal{F}) \cap H^1(\mathfrak{T}_\hbar)^m$

The goal of this section is to derive a defining property similar to Lemma 3.26 for the spaces $H(\mathcal{F})$ and $D(\mathcal{F})$. However, in general, functions in these spaces do not admit traces in L^2. Thus, we assume slightly more regularity and investigate $H(\mathcal{F}) \cap H^1(\mathfrak{T}_\hbar)^m$ and $D(\mathcal{F}) \cap H^1(\mathfrak{T}_\hbar)^m$. In particular, functions in these spaces do have a well-defined L^2-trace on each face $F \in \mathfrak{F}_\hbar$. Note that this is not really an additional restriction as we will need it later to show convergence. We start with the following auxiliary result.

Lemma 3.30. *Let $v, w \in H^1(\mathfrak{T}_\hbar)^m$. Then we have*

$$\begin{aligned}\sum_{K\in\mathfrak{T}_\hbar} \big(\mathcal{F}_\partial^K v \,\big|\, w\big)_{\partial K} = \sum_{F\in\mathfrak{F}_\hbar^{\mathrm{int}}} \Big(\big(\mathcal{F}_\partial^F \{\!\{v\}\!\}_F \,\big|\, [\![w]\!]_F\big)_F + \big(\mathcal{F}_\partial^F [\![v]\!]_F \,\big|\, \{\!\{w\}\!\}_F\big)_F\Big) \\ + \sum_{F\in\mathfrak{F}_\hbar^{\mathrm{bnd}}} \big(\mathcal{F}_\partial^F v \,\big|\, w\big)_F.\end{aligned}$$

Proof. By the definition of $\mathcal{F}_\partial^K$, $\mathcal{F}_\partial^F$ and the jump $[\![\cdot]\!]_F$, we have

$$\begin{aligned}\sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F}_\partial^K v \,\middle|\, w\right)_{\partial K} &= \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \left(\left(\mathcal{F}_\partial^F v^{K_1^F} \,\middle|\, w^{K_1^F}\right)_F - \left(\mathcal{F}_\partial^F v^{K_2^F} \,\middle|\, w^{K_2^F}\right)_F\right) \\ &\quad + \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left(\mathcal{F}_\partial^F v \,\middle|\, w\right)_F \\ &= \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \left([\![(\mathcal{F}_\partial^F v)\cdot w]\!]_F \,\middle|\, 1\right)_F + \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left(\mathcal{F}_\partial^F v \,\middle|\, w\right)_F.\end{aligned}$$

A straightforward calculation yields $[\![v\cdot w]\!]_F = \{\!\{v\}\!\}_F \cdot [\![w]\!]_F + [\![v]\!]_F \cdot \{\!\{w\}\!\}_F$ for all $v, w\colon \Omega \to \mathbb{R}^m$, concluding the proof. □

Now, we are able to characterize functions in $H(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$ and $D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$.

Lemma 3.31. *Let $v \in H^1(\mathcal{T}_\hbar)^m$. Then we have $v \in H(\mathcal{F})$ if and only if*

$$\mathcal{F}_\partial^F [\![v]\!]_F = 0 \qquad \text{a.e. on } F \text{ for all } F \in \mathcal{F}_\hbar^{\mathrm{int}}. \tag{3.5}$$

Additionally, for $v \in D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$, we have

$$(\mathcal{F}_\partial^F - \mathcal{F}_\Gamma) v = 0 \qquad \text{a.e. on } F \text{ for all } F \in \mathcal{F}_\hbar^{\mathrm{bnd}}. \tag{3.6}$$

Proof. Let $v \in H^1(\mathcal{T}_\hbar)^m$. We first show that (3.5) implies $v \in H(\mathcal{F})$. By definition (2.17) of the graph space, we have $v \in H(\mathcal{F})$ if

$$\sum_{i=1}^d F_i \partial_i v \in L^2(\Omega)^m.$$

This is equivalent to the mapping

$$C_c^\infty(\Omega)^m \to \mathbb{R}, \quad \varphi \mapsto \left(v \,\middle|\, \mathcal{F}^\circledast \varphi\right)_\Omega \tag{3.7}$$

being bounded in $L^2(\Omega)^m$, since $F_0^* \in L^\infty(\Omega)^{m\times m}$. Hence, let $\varphi \in C_c^\infty(\Omega)^m$. Then, using the integration by parts formula (3.3) on each element and the symmetry of $\mathcal{F}_\partial^K$, we have

$$\left(v \,\middle|\, \mathcal{F}^\circledast \varphi\right)_\Omega = \sum_{K\in\mathcal{T}_\hbar} \left(v \,\middle|\, \mathcal{F}^\circledast \varphi\right)_K = \sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F} v \,\middle|\, \varphi\right)_K + \sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F}_\partial^K v \,\middle|\, \varphi\right)_{\partial K}.$$

Since $[\![\varphi]\!]_F = 0$ and $\{\!\{\varphi\}\!\}_F = \varphi|_F$ for all $F \in \mathcal{F}_\hbar^{\mathrm{int}}$ and $\varphi|_F = 0$ for all $F \in \mathcal{F}_\hbar^{\mathrm{bnd}}$, Lemma 3.30 yields

$$\left(v \,\middle|\, \mathcal{F}^\circledast \varphi\right)_\Omega = \sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F} v \,\middle|\, \varphi\right)_K + \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \left(\mathcal{F}_\partial^K [\![v]\!]_F \,\middle|\, \varphi\right)_F = \sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F} v \,\middle|\, \varphi\right)_K, \tag{3.8}$$

where we have used that (3.5) holds true in the last step. Using the Cauchy–Schwarz inequality and $v \in H^1(\mathcal{T}_\hbar)^m$, this yields the boundedness of (3.7) and thus the assertion.

Now assume we have $v \in H(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$. Because of Assumption 3.27, [Jensen, 2004, Theorem 1.2] implies that $H(\mathcal{F}) \cap C^\infty(\Omega)^m$ is dense in $H(\mathcal{F})$. Hence, we can choose a sequence $(v_n)_{n\in\mathbb{N}}$ in $H(\mathcal{F}) \cap C^\infty(\Omega)^m$ with

$$v_n \to v, \qquad \mathcal{F} v_n \to \mathcal{F} v \qquad \text{in } L^2(\Omega)^m.$$

Let $\varphi \in C_c^\infty(\Omega)^m$ and denote by $\mathfrak{n}$ the outward unit normal vector to Γ. Then, by Lemma 3.28, we have

$$\begin{aligned}\big(\mathcal{F}v \,\big|\, \varphi\big)_\Omega &= \lim_{n\to\infty} \big(\mathcal{F}v_n \,\big|\, \varphi\big)_\Omega \\ &= \lim_{n\to\infty} \Big(\big(v_n \,\big|\, \mathcal{F}^\circledast\varphi\big)_\Omega + \big(\sum_{i=1}^d \mathfrak{n}_i F_i v_n \,\big|\, \varphi\big)_\Gamma\Big) \\ &= \big(v \,\big|\, \mathcal{F}^\circledast\varphi\big)_\Omega,\end{aligned}$$

where we have used $\varphi|_\Gamma = 0$ in the last step. Using the first equality in (3.8), this implies

$$\begin{aligned}\big(\mathcal{F}v \,\big|\, \varphi\big)_\Omega &= \sum_{K\in\mathcal{T}_\hbar} \big(\mathcal{F}v \,\big|\, \varphi\big)_K + \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \big(\mathcal{F}_\partial^K [\![v]\!]_F \,\big|\, \varphi\big)_F \\ &= \big(\mathcal{F}v \,\big|\, \varphi\big)_\Omega + \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \big(\mathcal{F}_\partial^K [\![v]\!]_F \,\big|\, \varphi\big)_F\end{aligned}$$

and thus

$$\sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \big(\mathcal{F}_\partial^F [\![v]\!]_F \,\big|\, \varphi\big)_F = 0.$$

In particular, this holds for arbitrary $\varphi \in C_c^\infty(\Omega)^m$ with $\operatorname{supp}\varphi$ intersecting only a single interface, implying (3.5).

Lastly, assume $v \in D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$ and let $F \in \mathcal{F}_\hbar^{\mathrm{bnd}}$. Then we have

$$\big((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)v \,\big|\, \varphi\big)_F = 0 \qquad \text{for all } \varphi \in C^\infty(\overline{\Omega})^m,$$

since v is smooth enough for the left hand side to make sense. This shows (3.6). □

3.5 Discretization of a Friedrichs' operator

We are now ready to define the dG discretization of the dissipative Friedrichs' operator $\mathcal{F}$. As we seek an approximation to the solution of the wave-type problem (2.24) in the discrete space $V_\hbar$, we would naturally define the discrete operator on this space. However, for our error analysis we extend this definition to the space $D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$. Hence, we combine both spaces and denote the **discrete operator domain** associated with $\mathcal{F}$ by

$$V_\hbar^{\mathcal{F}} = V_\hbar + (D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m).$$

3.5.1 Definition of a discrete Friedrichs' operator

Having introduced the domain of the discrete operator, we define the operator itself.

Definition 3.32. *Let $\mathcal{F}\colon D(\mathcal{F}) \to L^2(\Omega)^m$ be a dissipative Friedrichs' operator with coefficients $(F_i)_{i=0}^d$ and dissipative boundary condition $\mathcal{F}_\Gamma$. We define the **central flux dG discretization of** $\mathcal{F}$ as the operator $\boldsymbol{\mathcal{F}}\colon V_\hbar^{\mathcal{F}} \to V_\hbar$ such that*

$$\begin{aligned}\big(\boldsymbol{\mathcal{F}}v \,\big|\, \boldsymbol{\varphi}\big)_\Omega = &\sum_{K\in\mathcal{T}_\hbar} \big(\mathcal{F}v \,\big|\, \boldsymbol{\varphi}\big)_K - \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \big(\mathcal{F}_\partial^F [\![v]\!]_F \,\big|\, \{\!\{\boldsymbol{\varphi}\}\!\}_F\big)_F \\ &- \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \big((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)v \,\big|\, \boldsymbol{\varphi}\big)_F\end{aligned} \tag{3.9}$$

for all $\boldsymbol{\varphi} \in V_\hbar$.

Remark 3.33. 1. We assume in this thesis that the inner products used in Definition 3.32 are evaluated exactly. If the coefficients of the Friedrichs' operator are piecewise polynomials, this can be done by using quadrature formulas of sufficiently high order.

However, if this is not the case, additional quadrature errors have to be taken into account. As a consequence, the coefficients need to be smooth enough to still obtain the convergence rates proven in this thesis. One way to analyze this additional error is to use the unified error analysis in [Hipp et al., 2018].

2. It is possible to define the discrete operator in a more general way by using a weighted average instead of the standard average in (3.9). This weighted average can be defined as

$$\{\!\{v\}\!\}_F^\Lambda = \{\!\{\Lambda\}\!\}_F^{-1}\{\!\{\Lambda v\}\!\}_F,$$

where $\Lambda \in L^\infty(\Omega)^{m\times m}$ is symmetric and uniformly positive a.e. on Ω. The following results can then be proven analogously, however using different constants involving the weights.

If chosen in a suitable way, the weight Λ can lessen some constants ocurring in the analysis. This is, e.g., used in [Sturm, 2017] for isotropic Maxwell's equations. ◇

3.5.2 Properties of discrete Friedrichs' operators

Throughout the rest of this section, let $\boldsymbol{\mathcal{F}}$ be the central flux dG discretization of $\mathcal{F}$. We now prove some important properties of $\boldsymbol{\mathcal{F}}$ that will be used for the spatial, as well as the fully discrete error analysis.

Consistency

The first property is a consistency property. Namely, if we apply the discrete operator $\boldsymbol{\mathcal{F}}$ to a function belonging to $D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$, we obtain the L^2-projection onto $V_\hbar$ of the continuous operator $\mathcal{F}$ applied to this function. This is stated in the next lemma.

Proposition 3.34. *The discrete operator $\boldsymbol{\mathcal{F}}$ is **consistent** in the following sense. For all $v \in D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$ we have*

$$\boldsymbol{\mathcal{F}} v = \pi_h \mathcal{F} v.$$

Proof. First, note that by $v \in D(\mathcal{F})$ and Lemma 3.31 we have $\mathcal{F}_\partial^F [\![v]\!]_F = 0$ for all $F \in \mathcal{F}_\hbar^{\text{int}}$ and $(\mathcal{F}_\partial - \mathcal{F}_\Gamma)v = 0$ a.e. on Γ. Therefore, the interface as well as the boundary terms in (3.9) vanish.

Hence, for $v \in D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$ we have

$$\left(\boldsymbol{\mathcal{F}} v \,\middle|\, \boldsymbol{\varphi}\right)_\Omega = \sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F} v \,\middle|\, \boldsymbol{\varphi}\right)_K = \left(\mathcal{F} v \,\middle|\, \boldsymbol{\varphi}\right)_\Omega \qquad \text{for all } \boldsymbol{\varphi} \in V_\hbar$$

or, equivalently,

$$\left(\boldsymbol{\mathcal{F}} v - \mathcal{F} v \,\middle|\, \boldsymbol{\varphi}\right)_\Omega = 0 \qquad \text{for all } \boldsymbol{\varphi} \in V_\hbar,$$

proving $\boldsymbol{\mathcal{F}} v = \pi_h \mathcal{F} v$. □

Dissipativity

Next, we show that the discrete operator $\boldsymbol{\mathcal{F}}$ inherits the dissipativity of the continuous operator on the approximation space $V_\hbar$. To do so, we employ a similar approach as in the continuous case in Section 2.2. Hence, we first derive the adjoint operator of $\boldsymbol{\mathcal{F}}$.

Lemma 3.35. *Let $\boldsymbol{\mathcal{F}}^\circledast \colon V_\hbar \to V_\hbar$ be defined such that*

$$\begin{aligned}\left(\boldsymbol{\mathcal{F}}^\circledast \boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Omega = \sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F}^\circledast \boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_K + \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \left(\mathcal{F}_\partial^F [\![\boldsymbol{v}]\!]_F \,\middle|\, \{\!\{\boldsymbol{\varphi}\}\!\}_F\right)_F \\ + \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left((\mathcal{F}_\partial^F + \mathcal{F}_\Gamma^*)\boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_F\end{aligned}$$

for all $\boldsymbol{\varphi} \in V_\hbar$. Then $\boldsymbol{\mathcal{F}}^\circledast$ is the adjoint operator of $\boldsymbol{\mathcal{F}}$ on $V_\hbar$, i.e., $\boldsymbol{\mathcal{F}}^\circledast = (\boldsymbol{\mathcal{F}}|_{V_\hbar})^*$.

Proof. Let $\boldsymbol{v}, \boldsymbol{w} \in V_\hbar$. We use integration by parts (3.3) on each element and subsequently Lemma 3.30 to obtain

$$\begin{aligned}\left(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\middle|\, \boldsymbol{w}\right)_\Omega &= \sum_{K\in\mathcal{T}_\hbar} \left(\boldsymbol{v} \,\middle|\, \mathcal{F}^\circledast \boldsymbol{w}\right)_K + \sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F}_\partial^K \boldsymbol{v} \,\middle|\, \boldsymbol{w}\right)_{\partial K} - \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \left(\mathcal{F}_\partial^F [\![\boldsymbol{v}]\!]_F \,\middle|\, \{\!\{\boldsymbol{w}\}\!\}_F\right)_F \\ &\quad - \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)\boldsymbol{v} \,\middle|\, \boldsymbol{w}\right)_F \\ &= \sum_{K\in\mathcal{T}_\hbar} \left(\boldsymbol{v} \,\middle|\, \mathcal{F}\boldsymbol{w}\right)_K + \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \left(\mathcal{F}_\partial^F \{\!\{\boldsymbol{v}\}\!\}_F \,\middle|\, [\![\boldsymbol{w}]\!]_F\right)_F \\ &\quad + \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left((\mathcal{F}_\partial^F + \mathcal{F}_\Gamma)\boldsymbol{v} \,\middle|\, \boldsymbol{w}\right)_F.\end{aligned}$$

Hence, by the symmetry of $\mathcal{F}_\partial^F$, we have $\left(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\middle|\, \boldsymbol{w}\right)_\Omega = \left(\boldsymbol{v} \,\middle|\, \boldsymbol{\mathcal{F}}^\circledast \boldsymbol{w}\right)_\Omega$, proving the claim. □

Remark 3.36. In fact, the discrete adjoint operator $\boldsymbol{\mathcal{F}}^\circledast$ can be extended to functions in $D(\mathcal{F}^\circledast) \cap H^1(\mathcal{T}_\hbar)^m$ with $D(\mathcal{F}^\circledast) = \ker(\mathcal{F}_\partial + \mathcal{F}_\Gamma^*)$. This yields the central flux dG discretization of the adjoint operator of $\mathcal{F}$ introduced in Remark 2.24. ◇

We now show that $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{F}}^\circledast$ fulfill an analogous relation to their continuous counterparts. Namely, the discrete version of (2.18).

Lemma 3.37. *Let $\boldsymbol{v} \in V_\hbar$. Then we have*

$$\left(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Omega + \left(\boldsymbol{\mathcal{F}}^\circledast \boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Omega = \left((F_0 + F_0^* - \nabla\cdot\mathcal{F})\boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Omega + \tfrac{1}{2}\left((\mathcal{F}_\Gamma + \mathcal{F}_\Gamma^*)\boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Gamma$$

for all $\boldsymbol{\varphi} \in V_\hbar$.

Proof. Since the interface terms in $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{F}}^\circledast$ are identical but occur with opposite signs, we have

$$\left(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Omega + \left(\boldsymbol{\mathcal{F}}^\circledast \boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Omega = \left((\mathcal{F} + \mathcal{F}^\circledast)\boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Omega + \tfrac{1}{2}\left((\mathcal{F}_\Gamma + \mathcal{F}_\Gamma^*)\boldsymbol{v} \,\middle|\, \boldsymbol{\varphi}\right)_\Gamma.$$

Using (2.18) concludes the proof. □

The dissipativity of the discrete operator can now be proven similar to the continuous case, i.e., Theorem 2.22.

Proposition 3.38. *Let $F_0 + F_0^* - \nabla\cdot\mathcal{F} \leq 0$. Then the restriction of the discrete operator $\boldsymbol{\mathcal{F}}$ to $V_\hbar$ is **dissipative w.r.t.** $\left(\cdot \,\middle|\, \cdot\right)_\Omega$.*

Proof. Let $\boldsymbol{v} \in V_\hbar$. By the adjointness of $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{F}}^{\circledast}$ on $V_\hbar$ and Lemma 3.37 we have

$$\begin{aligned} 2\big(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\big|\, \boldsymbol{v}\big)_\Omega &= \big(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\big|\, \boldsymbol{v}\big)_\Omega + \big(\boldsymbol{\mathcal{F}}^{\circledast}\boldsymbol{v} \,\big|\, \boldsymbol{v}\big)_\Omega \\ &= \big((F_0 + F_0^* - \nabla\cdot\mathcal{F})\boldsymbol{v} \,\big|\, \boldsymbol{v}\big)_\Omega + \tfrac{1}{2}\big((\mathcal{F}_\Gamma + \mathcal{F}_\Gamma^*)\boldsymbol{v} \,\big|\, \boldsymbol{v}\big)_\Gamma \\ &\leq \big(\mathcal{F}_\Gamma\boldsymbol{v} \,\big|\, \boldsymbol{v}\big)_\Gamma \\ &\leq 0, \end{aligned}$$

where we have used the dissipativity of $\mathcal{F}$ and $\mathcal{F}_\Gamma$ in the sense of Definitions 2.23 and 2.21, respectively. □

Inverse inequality

We now show that the discrete operator $\boldsymbol{\mathcal{F}}$ fulfills an inverse inequality on the approximation space $V_\hbar$. In fact, the inverse inequality from Section 3.2.1 fulfilled by $\mathcal{F}$ (along with the trace inequality) is crucial to obtain this result.

As we apply this result to concatenations of discrete operators, we need to show a slightly more general result. Namely, we show that the inequality can also be applied in the mesh-dependent norms addressed in Remark 3.6. This is necessary to treat locally refined meshes which are not quasi-uniform (cf., e.g., [Ern and Guermond, 2004, Definition 1.140 & Corollary 1.141]).

To derive the aforementioned inverse inequality, we will use an element-based approach. Thus, the first thing we show is a representation of the discrete operator $\boldsymbol{\mathcal{F}}$ on a single element.

Lemma 3.39. *Let $K \in \mathcal{T}_\hbar$ and $\boldsymbol{v} \in V_\hbar$. Then we have*

$$\begin{aligned} \big(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K = \big(\mathcal{F}\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K &- \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}} \big(\mathcal{F}_\partial^F[\![\boldsymbol{v}]\!]_F \,\big|\, \boldsymbol{\varphi}^K\big)_F \\ &- \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{bnd}}} \big((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_F \end{aligned}$$

for all $\boldsymbol{\varphi} \in V_\hbar$.

Proof. Using the definition of the L^2-inner product yields

$$\big(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K = \big(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\big|\, \mathbb{1}_K\boldsymbol{\varphi}\big)_\Omega \qquad \text{for all } \boldsymbol{\varphi} \in V_\hbar. \tag{3.10}$$

Note that for $F \in \mathcal{F}_\hbar^{\mathrm{int}}$ we have

$$\{\!\{\mathbb{1}_K\boldsymbol{\varphi}\}\!\}_F \equiv \begin{cases} \tfrac{1}{2}\boldsymbol{\varphi}^K|_F & \text{for } F \in \mathcal{F}_\hbar^{K,\mathrm{int}}, \\ 0 & \text{else.} \end{cases}$$

Hence, by using the definition of $\boldsymbol{\mathcal{F}}$, we obtain

$$\begin{aligned}\big(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\big|\, \mathbb{1}_K\boldsymbol{\varphi}\big)_\Omega &= \sum_{\widehat{K}\in\mathcal{T}_\hbar} \big(\mathcal{F}\boldsymbol{v} \,\big|\, \mathbb{1}_K\boldsymbol{\varphi}\big)_{\widehat{K}} - \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \big(\mathcal{F}_\partial^F[\![\boldsymbol{v}]\!]_F \,\big|\, \{\!\{\mathbb{1}_K\boldsymbol{\varphi}\}\!\}_F\big)_F \\ &\quad - \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \big((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)\boldsymbol{v} \,\big|\, \mathbb{1}_K\boldsymbol{\varphi}\big)_F \\ &= \big(\mathcal{F}\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K - \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}} \big(\mathcal{F}_\partial^F[\![\boldsymbol{v}]\!]_F \,\big|\, \boldsymbol{\varphi}^K\big)_F \\ &\quad - \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{bnd}}} \big((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_F.\end{aligned}$$

Using (3.10) concludes the proof. □

Having derived the elementwise representation of $\boldsymbol{\mathcal{F}}$, we can now derive elementwise bounds.

Lemma 3.40. *Let $K \in \mathcal{T}_\hbar$ and $\boldsymbol{v} \in V_\hbar$. Then we have*

$$\big(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K \le \Big(C_{\mathcal{F},\mathrm{el},1}\|h^{-1}\boldsymbol{v}\|_K + C_{\mathcal{F},\mathrm{el},2}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}}\|h^{-1}\boldsymbol{v}\|_{K_F}\Big)\|\boldsymbol{\varphi}\|_K$$

for all $\boldsymbol{\varphi} \in V_\hbar$, where

$$C_{\mathcal{F},\mathrm{el},1} = C_{\mathcal{F}}C_{\mathrm{inv}} + \tfrac{1}{2}C_{\mathrm{tr}}^2(C_{\Gamma,\mathcal{F}} + N_\partial C_{\mathcal{F}}) \quad \text{and} \quad C_{\mathcal{F},\mathrm{el},2} = \tfrac{1}{2}\rho^{1/2}C_{\mathcal{F}}C_{\mathrm{tr}}^2$$

with $C_{\Gamma,\mathcal{F}} = \max_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}}\|\mathcal{F}_\partial^F - \mathcal{F}_\Gamma\|_{\infty,F}$.

Proof. Lemma 3.39 yields for all $\boldsymbol{\varphi} \in V_\hbar$ that

$$\begin{aligned}\big(\boldsymbol{\mathcal{F}}\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K &= \big(\mathcal{F}\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K - \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}} \varepsilon_{K,F}\Big(\big(\mathcal{F}_\partial^F\boldsymbol{v}^K \,\big|\, \boldsymbol{\varphi}^K\big)_F - \big(\mathcal{F}_\partial^F\boldsymbol{v}^{K_F} \,\big|\, \boldsymbol{\varphi}^K\big)_F\Big) \\ &\quad - \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{bnd}}} \big((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_F,\end{aligned}$$

where $\varepsilon_{K,F} = \mathbf{n}^K \cdot \mathbf{n}^F$. We bound the three terms individually. Note that $\varepsilon_{K,F} = \pm 1$, so it does not affect the norm of the terms.

To bound the element term we use the Cauchy–Schwarz inequality and the inverse inequality from Lemma 3.19 to obtain

$$\big(\mathcal{F}\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K \le \|\mathcal{F}\boldsymbol{v}\|_K\|\boldsymbol{\varphi}\|_K \le C_{\mathcal{F}}C_{\mathrm{inv}}\|h^{-1}\boldsymbol{v}\|_K\|\boldsymbol{\varphi}\|_K.$$

Next, for the boundary term, we again use the Cauchy–Schwarz inequality and then the boundedness of $\mathcal{F}_\partial^F$ and $\mathcal{F}_\Gamma$ and the trace inequality, yielding

$$\begin{aligned}\big((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)\boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_F &\le C_{\Gamma,\mathcal{F}}\|\boldsymbol{v}\|_F\|\boldsymbol{\varphi}\|_F \\ &\le C_{\Gamma,\mathcal{F}}C_{\mathrm{tr}}\|h^{-1/2}\boldsymbol{v}\|_K C_{\mathrm{tr}}\|h^{-1/2}\boldsymbol{\varphi}\|_K \\ &= C_{\Gamma,\mathcal{F}}C_{\mathrm{tr}}^2 h_K^{-1/2}\|h^{-1/2}\boldsymbol{v}\|_K\|\boldsymbol{\varphi}\|_K \\ &= C_{\Gamma,\mathcal{F}}C_{\mathrm{tr}}^2\|h^{-1}\boldsymbol{v}\|_K\|\boldsymbol{\varphi}\|_K.\end{aligned}$$

The first interface term can be bounded analogously to the boundary term by

$$\big(\mathcal{F}_\partial^F \boldsymbol{v}^K \,\big|\, \boldsymbol{\varphi}^K\big)_F \leq C_\mathcal{F} C_{\mathrm{tr}}^2 \|h^{-1}\boldsymbol{v}\|_K \|\boldsymbol{\varphi}\|_K.$$

For the second interface term we additionally use Lemma 3.13 in the third inequality to obtain

$$\begin{aligned}
\big(\mathcal{F}_\partial^F \boldsymbol{v}^{K_F} \,\big|\, \boldsymbol{\varphi}^K\big)_F &\leq \|\mathcal{F}_\partial^F \boldsymbol{v}^{K_F}\|_F \|\boldsymbol{\varphi}\|_F \\
&\leq C_\mathcal{F} C_{\mathrm{tr}} \|h^{-1/2}\boldsymbol{v}\|_{K_F} C_{\mathrm{tr}} \|h^{-1/2}\boldsymbol{\varphi}\|_K \\
&= C_\mathcal{F} C_{\mathrm{tr}}^2 h_K^{-1/2} \|h^{-1/2}\boldsymbol{v}\|_{K_F} \|\boldsymbol{\varphi}\|_K \\
&\leq \rho^{1/2} C_\mathcal{F} C_{\mathrm{tr}}^2 h_{K_F}^{-1/2} \|h^{-1/2}\boldsymbol{v}\|_{K_F} \|\boldsymbol{\varphi}\|_K \\
&= \rho^{1/2} C_\mathcal{F} C_{\mathrm{tr}}^2 \|h^{-1}\boldsymbol{v}\|_{K_F} \|\boldsymbol{\varphi}\|_K.
\end{aligned}$$

Combining these bounds and using the fact that each element has at most N_∂ interfaces and at most one boundary face concludes the proof. □

Now, we only need to put the elementwise bounds together to obtain global bounds on the whole of Ω.

Proposition 3.41. *Let $\boldsymbol{v} \in V_\hbar$. Then, for all $p \in \mathbb{Z}$, the discrete operator $\boldsymbol{\mathcal{F}}$ fulfills the following* ***inverse inequality***

$$\|h^p \boldsymbol{\mathcal{F}} \boldsymbol{v}\|_\Omega \leq C_{\mathrm{inv},\mathcal{F},p} \|h^{p-1}\boldsymbol{v}\|_\Omega$$

and in particular

$$\|\boldsymbol{\mathcal{F}}\boldsymbol{v}\|_\Omega \leq C_{\mathrm{inv},\mathcal{F}} \|h^{-1}\boldsymbol{v}\|_\Omega$$

with $C_{\mathrm{inv},\mathcal{F},p} = C_{\mathcal{F},\mathrm{el},1} + \rho^p N_\partial C_{\mathcal{F},\mathrm{el},2}$ and $C_{\mathrm{inv},\mathcal{F}} = C_{\mathrm{inv},\mathcal{F},0}$.

Proof. We show

$$\big(h^p \boldsymbol{\mathcal{F}} \boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_\Omega \leq C_{\mathrm{inv},\mathcal{F},p} \|h^{p-1}\boldsymbol{v}\|_\Omega \|\boldsymbol{\varphi}\|_\Omega.$$

The claim then follows because of

$$\|h^p \boldsymbol{\mathcal{F}} \boldsymbol{v}\|_\Omega = \sup_{\substack{\|\boldsymbol{\varphi}\|_\Omega = 1 \\ \boldsymbol{\varphi} \in V_\hbar}} \big(h^p \boldsymbol{\mathcal{F}} \boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_\Omega.$$

First, note that

$$\big(h^p \boldsymbol{\mathcal{F}} \boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_\Omega = \sum_{K \in \mathcal{T}_\hbar} \big(h^p \boldsymbol{\mathcal{F}} \boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K = \sum_{K \in \mathcal{T}_\hbar} h_K^p \big(\boldsymbol{\mathcal{F}} \boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_K.$$

Hence, by Lemmas 3.40 and 3.13 we have

$$\begin{aligned}
\big(h^p \boldsymbol{\mathcal{F}} \boldsymbol{v} \,\big|\, \boldsymbol{\varphi}\big)_\Omega &\leq \sum_{K \in \mathcal{T}_\hbar} h_K^p \Big(C_{\mathcal{F},\mathrm{el},1} \|h^{-1}\boldsymbol{v}\|_K + C_{\mathcal{F},\mathrm{el},2} \sum_{F \in \mathcal{F}_\hbar^{K,\mathrm{int}}} \|h^{-1}\boldsymbol{v}\|_{K_F} \Big) \|\boldsymbol{\varphi}\|_K \\
&= \sum_{K \in \mathcal{T}_\hbar} C_{\mathcal{F},\mathrm{el},1} \|h^{p-1}\boldsymbol{v}\|_K \|\boldsymbol{\varphi}\|_K + C_{\mathcal{F},\mathrm{el},2} \sum_{K \in \mathcal{T}_\hbar} \sum_{F \in \mathcal{F}_\hbar^{K,\mathrm{int}}} h_K^p \|h^{-1}\boldsymbol{v}\|_{K_F} \|\boldsymbol{\varphi}\|_K \\
&\leq C_{\mathcal{F},\mathrm{el},1} \sum_{K \in \mathcal{T}_\hbar} \|h^{p-1}\boldsymbol{v}\|_K \|\boldsymbol{\varphi}\|_K + C_{\mathcal{F},\mathrm{el},2}\, \rho^p \sum_{K \in \mathcal{T}_\hbar} \sum_{F \in \mathcal{F}_\hbar^{K,\mathrm{int}}} \|h^{p-1}\boldsymbol{v}\|_{K_F} \|\boldsymbol{\varphi}\|_K.
\end{aligned}$$

To bound the first term we use the Cauchy–Schwarz inequality in $\mathbb{R}^{|\mathcal{T}_\hbar|}$ to obtain

$$\sum_{K\in\mathcal{T}_\hbar} \|h^{p-1}\boldsymbol{v}\|_K\|\boldsymbol{\varphi}\|_K \le \Big(\sum_{K\in\mathcal{T}_\hbar} \|h^{p-1}\boldsymbol{v}\|_K^2\Big)^{1/2}\Big(\sum_{K\in\mathcal{T}_\hbar} \|\boldsymbol{\varphi}\|_K^2\Big)^{1/2} = \|h^{p-1}\boldsymbol{v}\|_\Omega\|\boldsymbol{\varphi}\|_\Omega.$$

The same argument for the second term yields

$$\sum_{K\in\mathcal{T}_\hbar}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}} \|h^{p-1}\boldsymbol{v}\|_{K_F}\|\boldsymbol{\varphi}\|_K \le \Big(\sum_{K\in\mathcal{T}_\hbar}\Big(\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}} \|h^{p-1}\boldsymbol{v}\|_{K_F}\Big)^2\Big)^{1/2}\Big(\sum_{K\in\mathcal{T}_\hbar} \|\boldsymbol{\varphi}\|_K^2\Big)^{1/2}.$$

By the equivalence of the 1- and 2-norm on $\mathbb{R}^{|\mathcal{F}_\hbar^K|}$ and $|\mathcal{F}_\hbar^K| \le N_\partial$ we have

$$\sum_{K\in\mathcal{T}_\hbar}\Big(\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}} \|h^{p-1}\boldsymbol{v}\|_{K_F}\Big)^2 \le N_\partial \sum_{K\in\mathcal{T}_\hbar}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}} \|h^{p-1}\boldsymbol{v}\|_{K_F}^2 \le N_\partial^2 \sum_{K\in\mathcal{T}_\hbar} \|h^{p-1}\boldsymbol{v}\|_K^2$$

and thus

$$\sum_{K\in\mathcal{T}_\hbar}\sum_{F\in\mathcal{F}_\hbar^{K,\mathrm{int}}} \|h^{p-1}\boldsymbol{v}\|_{K_F}\|\boldsymbol{\varphi}\|_K \le N_\partial\|h^{p-1}\boldsymbol{v}\|_\Omega\|\boldsymbol{\varphi}\|_\Omega,$$

concluding the proof. □

Approximation properties

It remains to study the approximation properties of the discrete operator. The next result gives a bound on the application of $\boldsymbol{\mathcal{F}}$ to the projection error of functions in $D(\mathcal{F}) \cap H^{q+1}(\mathcal{T}_\hbar)^m$. This can be seen as a measure of how well the discrete operator approximates its continuous counterpart, as for $v \in D(\mathcal{F}) \cap H^1(\mathcal{T}_\hbar)^m$ we have

$$\boldsymbol{\mathcal{F}} e_\pi^v = \boldsymbol{\mathcal{F}} v - \boldsymbol{\mathcal{F}}\pi_h v = (\pi_h\mathcal{F} - \boldsymbol{\mathcal{F}}\pi_h)v$$

by the consistency of $\boldsymbol{\mathcal{F}}$. We prove the next result using the same strategy we used to obtain Proposition 3.41.

Lemma 3.42. *Let $v \in D(\mathcal{F}) \cap H^{q+1}(\mathcal{T}_\hbar)^m$ for $0 \le q \le k$. Then, for all $p \in \mathbb{Z}$, we have*

$$\|h^p\boldsymbol{\mathcal{F}} e_\pi^v\|_\Omega \le C_{\pi,\mathcal{F},p}|h^{p+q}v|_{q+1,\mathcal{T}_\hbar}$$

and in particular

$$\|\boldsymbol{\mathcal{F}} e_\pi^v\|_\Omega \le C_{\pi,\mathcal{F}}|h^q v|_{q+1,\mathcal{T}_\hbar},$$

where $C_{\pi,\mathcal{F},p} = C_{\mathcal{F}}C_{\mathrm{app}}C_{\mathrm{inv}} + \frac{1}{2}N_\partial C_{\mathrm{tr}}C_{\mathrm{app},\partial}(C_{\Gamma,\mathcal{F}} + C_{\mathcal{F}} + \rho^{p+1/2}C_{\mathcal{F}})$ and $C_{\pi,\mathcal{F}} = C_{\pi,\mathcal{F},0}$.

Proof. As in the proof of Lemma 3.40 we first derive a bound on one element $K \in \mathcal{T}_\hbar$. To do so, we use the elementwise representation of $\boldsymbol{\mathcal{F}}$ in Lemma 3.39 and consequently

integration by parts (3.3) on the element K to obtain

$$\begin{aligned}\big(\boldsymbol{\mathcal{F}} e_\pi^v \,\big|\, \boldsymbol{\varphi}\big)_K &= \big(\mathcal{F} e_\pi^v \,\big|\, \boldsymbol{\varphi}\big)_K - \tfrac{1}{2} \sum_{F \in \mathcal{F}_\hbar^{K,\mathrm{int}}} \big(\mathcal{F}_\partial^F [\![e_\pi^v]\!]_F \,\big|\, \boldsymbol{\varphi}^K\big)_F \\ &\quad - \tfrac{1}{2} \sum_{F \in \mathcal{F}_\hbar^{K,\mathrm{bnd}}} \big((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma) e_\pi^v \,\big|\, \boldsymbol{\varphi}\big)_F \\ &= \big(e_\pi^v \,\big|\, \mathcal{F}^{\circledast} \boldsymbol{\varphi}\big)_K + \tfrac{1}{2} \sum_{F \in \mathcal{F}_\hbar^{K,\mathrm{int}}} \big(\mathcal{F}_\partial^F \{\!\{ e_\pi^v \}\!\}_F \,\big|\, \boldsymbol{\varphi}^K\big)_F \\ &\quad + \tfrac{1}{2} \sum_{F \in \mathcal{F}_\hbar^{K,\mathrm{bnd}}} \big((\mathcal{F}_\partial^F + \mathcal{F}_\Gamma) e_\pi^v \,\big|\, \boldsymbol{\varphi}\big)_F\end{aligned}$$

for all $\boldsymbol{\varphi} \in V_\hbar$.

For the first term we use the Cauchy–Schwarz inequality, the approximation properties from Lemma 3.24 and the inverse inequality from Lemma 3.19 to obtain

$$\begin{aligned}\big(e_\pi^v \,\big|\, \mathcal{F}^{\circledast} \boldsymbol{\varphi}\big)_K &\le \|e_\pi^v\|_K \|\mathcal{F}^{\circledast} \boldsymbol{\varphi}\|_K \\ &\le C_{\mathcal{F}} C_{\mathrm{app}} C_{\mathrm{inv}} |h^{q+1} v|_{q+1,K} \|h^{-1} \boldsymbol{\varphi}\|_K \\ &= C_{\mathcal{F}} C_{\mathrm{app}} C_{\mathrm{inv}} |h^q v|_{q+1,K} \|\boldsymbol{\varphi}\|_K.\end{aligned}$$

Hence, by similar arguments as in the proof of Lemma 3.40, albeit replacing the trace inequality by the bounds in Lemma 3.24, we have

$$\big(\boldsymbol{\mathcal{F}} e_\pi^v \,\big|\, \boldsymbol{\varphi}\big)_K \le \Big(\widetilde{C}_{\mathcal{F},\mathrm{el},1} |h^q v|_{q+1,K} + \widetilde{C}_{\mathcal{F},\mathrm{el},2} \sum_{F \in \mathcal{F}_\hbar^{K,\mathrm{int}}} |h^q v|_{q+1,K_F}\Big) \|\boldsymbol{\varphi}\|_K$$

with $\widetilde{C}_{\mathcal{F},\mathrm{el},1} = C_{\mathcal{F}} C_{\mathrm{app}} C_{\mathrm{inv}} + \frac{1}{2} N_\partial C_{\mathrm{tr}} C_{\mathrm{app},\partial} (C_{\Gamma,\mathcal{F}} + C_{\mathcal{F}})$ and $\widetilde{C}_{\mathcal{F},\mathrm{el},2} = \frac{1}{2} \rho^{1/2} C_{\mathcal{F}} C_{\mathrm{tr}} C_{\mathrm{app},\partial}$. Proceeding completely analogously to the proof of Proposition 3.41 concludes the proof. □

For the analysis of the Peaceman–Rachford scheme (or other perturbations of the Crank–Nicolson scheme), we need a similar approximation result for concatenations of more than one discrete operator. We show this in the next lemma, which can be seen as a generalization of [Pažur, 2013, Theorem 6.3]. There, a similar result was proven for arbitrary powers of one discrete operator under the assumption of quasi-uniform mesh sequences.

As mentioned before, we circumvent the quasi-uniformity assumption by using the more general versions of the inverse inequality from Proposition 3.41 and the approximation property from the last lemma. Without these more general results, we would need to take the inverse of the maximal diameter $\hbar$ out of the norm. This needs a uniform bound on h_K^{-1} for all $K \in \mathcal{T}_\hbar$, requiring the assumption of quasi-uniformity.

We will use this result only for $q = 2$ in Lemmas 5.12 and 5.13 below and will have to adjust it a bit, as we have to take material parameters into account. Thus, we do not state the explicit constants as in the results before, but rather their dependencies. Further, we assume the coefficients of the Friedrichs' operators to be constant to avoid technicalities.

Lemma 3.43. *Let $\mathcal{F}_1, \dots, \mathcal{F}_r$ be dissipative Friedrichs' operators with constant coefficients and $\boldsymbol{\mathcal{F}}_1, \dots, \boldsymbol{\mathcal{F}}_r$ their respective central flux dG discretizations. Further, let $v \in D(\mathcal{F}_r \dots \mathcal{F}_1) \cap H^{q+1}(\mathcal{T}_\hbar)^m$ for $0 \le q \le k$. Then, for all $p \in \mathbb{Z}$ and all $r \in \{1, \dots, q+1\}$ we have*

$$\|h^p (\boldsymbol{\mathcal{F}}_r \dots \boldsymbol{\mathcal{F}}_2 \boldsymbol{\mathcal{F}}_1 \pi_h - \pi_h \mathcal{F}_r \dots \mathcal{F}_1) v\|_\Omega \le C \|h^{p+(q+1)-r} v\|_{q+1,\mathcal{T}_\hbar}$$

and in particular

$$\|(\boldsymbol{\mathcal{F}}_r \dots \boldsymbol{\mathcal{F}}_2 \boldsymbol{\mathcal{F}}_1 \pi_h - \pi_h \mathcal{F}_r \dots \mathcal{F}_1) v\|_\Omega \leq C \|h^{(q+1)-r} v\|_{q+1,\mathcal{T}_h},$$

where the constants only depend on r and the dependencies of $C_{\mathrm{inv},\mathcal{F}_i,p}$ and $C_{\pi,\mathcal{F}_i,p}$ for $i = 1, \dots, r$.

Proof. We prove the assertion by induction over r. Lemma 3.42 provides the initial step, since for $r = 1$ the left hand side equals

$$\|h^p(\boldsymbol{\mathcal{F}}_1 \pi_h - \pi_h \mathcal{F}_1) v\|_\Omega = \|h^p(\boldsymbol{\mathcal{F}}_1 \pi_h - \boldsymbol{\mathcal{F}}_1) v\|_\Omega = \|h^p \boldsymbol{\mathcal{F}}_1 e_\pi^v\|_\Omega.$$

Assume the assertion holds for $r < q+1$. By the consistency of $\boldsymbol{\mathcal{F}}_{r+1}$, i.e., Proposition 3.34, we have

$$\begin{aligned}
&\|h^p(\boldsymbol{\mathcal{F}}_{r+1} \dots \boldsymbol{\mathcal{F}}_1 \pi_h - \pi_h \mathcal{F}_{r+1} \dots \mathcal{F}_1) v\|_\Omega \\
&\quad = \|h^p(\boldsymbol{\mathcal{F}}_{r+1} \dots \boldsymbol{\mathcal{F}}_1 \pi_h - \boldsymbol{\mathcal{F}}_{r+1} \mathcal{F}_r \dots \mathcal{F}_1) v\|_\Omega \\
&\quad = \|h^p \boldsymbol{\mathcal{F}}_{r+1}(\boldsymbol{\mathcal{F}}_r \dots \boldsymbol{\mathcal{F}}_1 \pi_h - \mathcal{F}_r \dots \mathcal{F}_1) v\|_\Omega.
\end{aligned}$$

We use Proposition 3.41 and Lemma 3.42 to obtain

$$\begin{aligned}
&\|h^p \boldsymbol{\mathcal{F}}_{r+1}(\boldsymbol{\mathcal{F}}_r \dots \boldsymbol{\mathcal{F}}_1 \pi_h - \mathcal{F}_r \dots \mathcal{F}_1) v\|_\Omega \\
&= \|h^p \big(\boldsymbol{\mathcal{F}}_{r+1}(\boldsymbol{\mathcal{F}}_r \dots \boldsymbol{\mathcal{F}}_1 \pi_h - \pi_h \mathcal{F}_r \dots \mathcal{F}_1) v - \boldsymbol{\mathcal{F}}_{r+1}(\mathcal{F}_r \dots \mathcal{F}_1 - \pi_h \mathcal{F}_r \dots \mathcal{F}_1) v\big)\|_\Omega \\
&\leq \|h^p \boldsymbol{\mathcal{F}}_{r+1}(\boldsymbol{\mathcal{F}}_r \dots \boldsymbol{\mathcal{F}}_1 \pi_h - \pi_h \mathcal{F}_r \dots \mathcal{F}_1) v\|_\Omega + \|h^p \boldsymbol{\mathcal{F}}_{r+1} e_\pi^{\mathcal{F}_r \dots \mathcal{F}_1 v}\|_\Omega \\
&\leq C_{\mathrm{inv},\mathcal{F}_{r+1},p} \|h^{p-1}(\boldsymbol{\mathcal{F}}_r \dots \boldsymbol{\mathcal{F}}_1 \pi_h - \pi_h \mathcal{F}_r \dots \mathcal{F}_1) v\|_\Omega \\
&\quad + C_{\pi,\mathcal{F}_{r+1},p} |h^{p+q-r} \mathcal{F}_r \dots \mathcal{F}_1 v|_{(q+1)-r,\mathcal{T}_h},
\end{aligned}$$

since by Lemma 2.17, for $v \in H^{q+1}(\mathcal{T}_h)^m \cap D(\mathcal{F}_r \dots \mathcal{F}_1)$, we have $\mathcal{F}_r \dots \mathcal{F}_1 v \in H^{(q+1)-r}(\mathcal{T}_h)^m$. To bound the first term, we use the induction hypothesis with $p-1$ instead of p. For the second term we use r times Lemma 2.17, yielding

$$|h^{p+q-r} \mathcal{F}_r \dots \mathcal{F}_1 v|_{(q+1)-r,\mathcal{T}_h} \leq (d+1)^{r/2} C_{\mathcal{F}_r} \dots C_{\mathcal{F}_1} \|h^{p+(q+1)-(r+1)} v\|_{q+1,\mathcal{T}_h}.$$

Combining these bounds concludes the proof. □

3.6 Spatial discretization of the wave-type problem

We can now formulate the spatially semidiscrete version of the wave-type problem (2.24). To do so, we replace the spatial differential operator $\mathcal{L}$ in (2.24a) by its discrete counterpart and seek an approximation to the solution in the approximation space V_h. Similar to the continuous case, we equip V_h with the weighted inner product $\big(\cdot \,\big|\, \cdot\big)_M$.

In light of Section 3.4, we have to make some more assumptions on the Friedrichs' operator $\widetilde{\mathcal{L}}$ and the corresponding split operators $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ to ensure that we can apply the results established in the last section. Namely, we assume that $\widetilde{\mathcal{L}}$,$\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ fulfill Assumptions 3.27 and 3.29.

3.6.1 Formulation of the semidiscrete wave-type problem

First, we construct the discrete counterpart of the spatial operator $\mathcal{L}$. In what follows, let $\widetilde{\mathcal{L}}\colon V_h^{\widetilde{\mathcal{L}}} \to V_h$ be the central flux dG discretization of the dissipative Friedrichs' operator $\widetilde{\mathcal{L}}$ and let $V_h^{\mathcal{L}} = V_h^{\widetilde{\mathcal{L}}}$. Analogously to the continuous case, we define the discrete operator $\boldsymbol{\mathcal{L}}\colon V_h^{\mathcal{L}} \to V_h$ such that we have

$$\boldsymbol{\mathcal{L}} = M^{-1}\widetilde{\boldsymbol{\mathcal{L}}}.$$

Before we state the semidiscrete problem using this operator, we will pose another assumption on the material tensor M to avoid technicalities.

Assumption 3.44. *We assume that the material tensor M is **piecewise constant**. We further assume that for all $h \in \mathcal{H}$ the mesh $\mathcal{T}_h$ is **matched** to the material, i.e., for all $K \in \mathcal{T}_h$ we have $M|_K \equiv M_K$ with constant $M_K \in \mathbb{R}^{m\times m}$.*

Due to Assumption 3.44, the L^2-projection can also be calculated by using $\left(\cdot \,\middle|\, \cdot\right)_M$, and we obtain an analogous result to Lemma 3.16.

Lemma 3.45. *For $v \in L^2(\Omega)^m$ we have*

$$\left(v - \pi_h v \,\middle|\, \boldsymbol{\varphi}\right)_M = 0 \qquad \text{for all } \boldsymbol{\varphi} \in V_h \tag{3.11}$$

and

$$\|\pi_h v\|_M \leq \|v\|_M. \tag{3.12}$$

Proof. By Lemma 3.17, for all $K \in \mathcal{T}_h$, we have

$$\left(v - \pi_h v \,\middle|\, \boldsymbol{\varphi}\right)_K = 0 \qquad \text{for all } \boldsymbol{\varphi} \in \mathbb{Q}_d^k(K)^m.$$

This implies

$$\begin{aligned}\left(v - \pi_h v \,\middle|\, \boldsymbol{\varphi}\right)_M &= \left(M(v - \pi_h v) \,\middle|\, \boldsymbol{\varphi}\right)_\Omega \\ &= \sum_{K\in\mathcal{T}_h} \left(M_K(v - \pi_h v) \,\middle|\, \boldsymbol{\varphi}\right)_K \\ &= \sum_{K\in\mathcal{T}_h} \left(v - \pi_h v \,\middle|\, M_K\boldsymbol{\varphi}\right)_K \\ &= 0\end{aligned}$$

for all $\boldsymbol{\varphi} \in V_h$, since $M_K\boldsymbol{\varphi}|_K \in \mathbb{Q}_d^k(K)^m$.

The second assertion now follows with the same strategy of proof as for Lemma 3.16. This yields

$$\|\pi_h v\|_M = \sup_{\substack{\boldsymbol{\varphi}\in V_h \\ \|\boldsymbol{\varphi}\|_M=1}} \left(\pi_h v \,\middle|\, \boldsymbol{\varphi}\right)_M = \sup_{\substack{\boldsymbol{\varphi}\in V_h \\ \|\boldsymbol{\varphi}\|_M=1}} \left(v \,\middle|\, \boldsymbol{\varphi}\right)_M \leq \sup_{\substack{\boldsymbol{\varphi}\in V_h \\ \|\boldsymbol{\varphi}\|_M=1}} \|v\|_M\|\boldsymbol{\varphi}\|_M = \|v\|_M,$$

where we used (3.11) in the second step. □

This leads to the operator $\boldsymbol{\mathcal{L}}$ exhibiting the same consistency property as $\widetilde{\mathcal{L}}$. Namely, for all $v \in D(\mathcal{L}) \cap H^1(\mathcal{T}_h)^m$ we have

$$\boldsymbol{\mathcal{L}}v = \pi_h \mathcal{L}v.$$

We are now able to pose the **central flux dG discretization of the wave-type problem** (2.24) as follows. Seek $\boldsymbol{u}\colon \mathbb{R}_+ \to V_h$ such that

$$\begin{cases} \mathrm{d}_t \boldsymbol{u}(t) = \boldsymbol{\mathcal{L}}\boldsymbol{u}(t) + \boldsymbol{f}_\pi(t), & t \in \mathbb{R}_+, \quad (3.13a) \\ \boldsymbol{u}(0) = \boldsymbol{u}_\pi^0, & (3.13b) \end{cases}$$

where $\boldsymbol{u}_\pi^0 = \pi_h u^0$ and $\boldsymbol{f}_\pi(t) = \pi_h f(t)$.

3.6.2 Wellposedness of the semidiscrete wave-type problem

Owing to the results from Section 3.5.2 we can immediately show wellposedness of (3.13). This is due to the fact that as a consequence of Proposition 3.38 the discrete operator $\boldsymbol{\mathcal{L}}$ is dissipative on V_h. Since V_h is a finite-dimensional space, this implies that $\boldsymbol{\mathcal{L}}$ is maximal dissipative, and hence, $\boldsymbol{\mathcal{L}}$ generates a contraction semigroup on V_h. We slightly abuse notation by denoting this semigroup as $\left(\mathrm{e}^{t\boldsymbol{\mathcal{L}}}\right)_{t\geq 0}$ instead of introducing a new symbol for the restriction of $\boldsymbol{\mathcal{L}}$ to V_h. This yields the following result.

Corollary 3.46. *The restriction of $\boldsymbol{\mathcal{L}}$ to V_h generates the contraction semigroup $\left(\mathrm{e}^{t\boldsymbol{\mathcal{L}}}\right)_{t\geq 0}$ on V_h, and hence, there exists a unique solution $\boldsymbol{u} \in C^1(\mathbb{R}_+; V_h)$ of (3.13) given by*

$$\boldsymbol{u}(t) = \mathrm{e}^{t\boldsymbol{\mathcal{L}}}\,\boldsymbol{u}_\pi^0 + \int_0^t \mathrm{e}^{(t-s)\boldsymbol{\mathcal{L}}}\,\boldsymbol{f}_\pi(s)\,\mathrm{d}s.$$

Proof. The discrete operator $\boldsymbol{\mathcal{L}}$ is dissipative on V_h as we have

$$\left(\boldsymbol{\mathcal{L}}\boldsymbol{v} \,\middle|\, \boldsymbol{v}\right)_M = \left(M\boldsymbol{\mathcal{L}}\boldsymbol{v} \,\middle|\, \boldsymbol{v}\right)_\Omega = \left(\widetilde{\boldsymbol{\mathcal{L}}}\boldsymbol{v} \,\middle|\, \boldsymbol{v}\right)_\Omega \leq 0 \qquad \text{for all } \boldsymbol{v} \in V_h$$

by Proposition 3.38. Since V_h is finite-dimensional, this implies the maximal dissipativity of $\boldsymbol{\mathcal{L}}$. The claim now follows by the Lumer–Phillips Theorem 2.12. □

3.7 Error analysis of the spatially semidiscrete problem

Using the results obtained in this chapter, we are able to analyze the error of the semidiscrete approximation given by (3.13). Throughout this section, let u be the solution of the continuous problem (2.24) and let $\boldsymbol{u}$ be the approximation given by the spatially semidiscrete problem (3.13).

For all $t \in \mathbb{R}_+$, we denote the **spatially semidiscrete error** by

$$e(t) = u(t) - \boldsymbol{u}(t).$$

We further introduce the **projection error** $e_\pi(t) = u(t) - \pi_h u(t)$ and the **space discretization error** $\boldsymbol{e}(t) = \pi_h u(t) - \boldsymbol{u}(t)$ to obtain the error splitting

$$e(t) = e_\pi(t) + \boldsymbol{e}(t).$$

As a direct consequence of Lemma 3.24 and the equivalence of the standard and the weighted L^2-norm (2.23) we can bound the $\|\cdot\|_M$-norm of the projection error by

$$\|e_\pi(t)\|_M \leq C_{\mathrm{app},M} |h^{k+1} u(t)|_{k+1,\mathcal{T}_h} \qquad (3.14)$$

with $C_{\mathrm{app},M} = \|M\|_{\infty,\Omega}^{1/2} C_{\mathrm{app}}$ if the exact solution is smooth enough. Hence, it remains to bound the space discretization error.

3.7.1 Error recursion

We proceed in the usual way and show that the space discretization error $\boldsymbol{e}$ satisfies a perturbed version of the semidiscrete problem (3.13). We already know that the discrete solution satisfies the semidiscrete problem. Hence, we start by investigating the defect caused by inserting the projected exact solution into the semidiscrete problem (3.13).

Lemma 3.47. *Assume $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}) \cap H^1(\mathcal{T}_\hbar)^m)$. Then the projected exact solution $\pi_h u$ satisfies*

$$\mathrm{d}_t \pi_h u(t) = \boldsymbol{\mathcal{L}} \pi_h u(t) + \boldsymbol{f}_\pi(t) + \boldsymbol{d}_\pi(t), \qquad t \in \mathbb{R}_+, \tag{3.15}$$

*where the defect $\boldsymbol{d}_\pi \colon \mathbb{R}_+ \to \mathbb{R}^m$ is called the **space truncation error** and is given by*

$$\boldsymbol{d}_\pi(t) = \boldsymbol{\mathcal{L}} e_\pi(t). \tag{3.16}$$

Proof. Using the fact that that d_t and the L^2-projection commute, and that u satisfies the continuous problem (2.24), we obtain

$$\mathrm{d}_t \pi_h u = \pi_h \mathrm{d}_t u = \pi_h(\mathcal{L} u + f) = \boldsymbol{\mathcal{L}} u + \boldsymbol{f}_\pi,$$

where we have used the consistency of $\boldsymbol{\mathcal{L}}$ in the sense of Proposition 3.34 in the last equation. Using

$$\boldsymbol{\mathcal{L}} u = \boldsymbol{\mathcal{L}} \pi_h u + \boldsymbol{\mathcal{L}}(u - \pi_h u) = \boldsymbol{\mathcal{L}} \pi_h u + \boldsymbol{\mathcal{L}} e_\pi$$

concludes the proof. □

This readily yields the aforementioned error recursion.

Corollary 3.48. *Assume $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}) \cap H^1(\mathcal{T}_\hbar)^m)$. Then the space discretization error $\boldsymbol{e}(t) = \pi_h u(t) - \boldsymbol{u}(t)$ satisfies*

$$\begin{cases} \mathrm{d}_t \boldsymbol{e}(t) = \boldsymbol{\mathcal{L}} \boldsymbol{e}(t) + \boldsymbol{d}_\pi(t), & t \in \mathbb{R}_+, \\ \boldsymbol{e}(0) = 0. \end{cases} \tag{3.17}$$

Proof. We have $\boldsymbol{e}(0) = \pi_h u^0 - \boldsymbol{u}_\pi^0 = 0$, since we use $\boldsymbol{u}_\pi^0 = \pi_h u^0$ as initial value of the semidiscrete problem. The remaining assertion follows by subtracting the semidiscrete recursion (3.13a) from (3.15). □

3.7.2 Spatial convergence result

We can now state the **convergence result for the central fluxes dG discretization** of the wave-type problem (2.24).

Theorem 3.49. *Assume that the exact solution u of the wave-type problem (2.24) satisfies $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}) \cap H^{k+1}(\mathcal{T}_\hbar)^m)$. Then, for $t \in \mathbb{R}_+$, the spatially semidiscrete error satisfies*

$$\begin{aligned} \|u(t) - \boldsymbol{u}(t)\|_M &\leq C_{\mathrm{app},M} |h^{k+1} u(t)|_{k+1,\mathcal{T}_\hbar} + C_{\pi,\widetilde{\mathcal{L}},M} \int_0^t |h^k u(s)|_{k+1,\mathcal{T}_\hbar} \,\mathrm{d}s \\ &\leq C \hbar^k, \end{aligned}$$

where $C_{\pi,\widetilde{\mathcal{L}},M} = \|M^{-1}\|_{\infty,\Omega}^{1/2} C_{\pi,\widetilde{\mathcal{L}}}$ and C only depends on $C_{\mathrm{app},M}$, $C_{\pi,\widetilde{\mathcal{L}},M}$ and $|u(s)|_{k+1,\mathcal{T}_\hbar}$, $s \in [0,t]$.

Proof. As mentioned before, the projection error satisfies the bound (3.14). This yields the first term.

To bound the space discretization error we use Corollary 3.46 to solve (3.17), yielding

$$\boldsymbol{e}(t) = \int_0^t \mathrm{e}^{(t-s)\boldsymbol{\mathcal{L}}}\, \boldsymbol{d}_\pi(s)\,\mathrm{d}s,$$

since $\boldsymbol{e}(0) = 0$. By the contractivity of $\mathrm{e}^{(t-s)\boldsymbol{\mathcal{L}}}$ and (3.16), i.e., the definition of the space truncation error $\boldsymbol{d}_\pi$, this yields

$$\|\boldsymbol{e}(t)\|_M \leq \int_0^t \|\boldsymbol{\mathcal{L}} e_\pi(s)\|_M \,\mathrm{d}s.$$

Hence, it remains to bound $\|\boldsymbol{\mathcal{L}} e_\pi(s)\|_M$. We now use Lemma 3.42 to obtain

$$\begin{aligned} \|\boldsymbol{\mathcal{L}} e_\pi(s)\|_M &\leq \|M^{-1/2}\widetilde{\boldsymbol{\mathcal{L}}} e_\pi(s)\|_\Omega \\ &\leq \|M^{-1}\|_{\infty,\Omega}^{1/2}\, \|\widetilde{\boldsymbol{\mathcal{L}}} e_\pi(s)\|_\Omega \\ &\leq \|M^{-1}\|_{\infty,\Omega}^{1/2}\, C_{\pi,\widetilde{\mathcal{L}}}\, |h^k u(s)|_{k+1,\mathcal{T}_h}, \end{aligned} \tag{3.18}$$

which proves the claim. □

Remark 3.50. Note that we can derive a similar result if we replace $H^{k+1}(\mathcal{T}_h)^m$ by $H^{q+1}(\mathcal{T}_h)^m$ with $q < k$. However, in that case, we would also obtain convergence rates determined by q instead of k. ◇

4 | Temporal discretization

In this chapter, we present and analyze the two temporal discretization schemes we consider in this thesis. Namely, the Crank–Nicolson and the Peaceman–Rachford scheme. As stated before, our main focus lies on the Peaceman–Rachford scheme, however, as our analysis of the scheme is based upon interpreting this scheme as a perturbation of the Crank–Nicolson scheme, we as well consider the latter.

Throughout, let $\tau > 0$ be the timestep size, $t_q = q\tau$, $q \in \mathbb{Q}$ and $f^{n+1} = f(t_{n+1})$, $n \in \mathbb{N}$. Further, assume that we have $f \in C(\mathbb{R}_+; L^2(\Omega)^m)$. Note that this is not really a restriction as this assumption is a necessary condition for the wellposedness of the continuous problem (2.24).

The chapter is structured as follows. We begin with a short derivation of the Crank–Nicolson scheme in Section 4.1. Subsequently, wellposedness and the stability of the method applied to the wave-type problem (2.24) are discussed in the same section. We proceed accordingly in Section 4.2 for the Peaceman–Rachford scheme. Finally, in Section 4.3 we investigate the approximation errors of both schemes. In particular, we show that both schemes converge with order 2 to the exact solution, given this solution fulfills appropriate regularity assumptions.

4.1 The Crank–Nicolson scheme

The **Crank–Nicolson scheme** or implicit trapezoidal rule is a well-known implicit time integration scheme of classical order two, cf., e.g., [Hairer et al., 2006, Section II.1.1] and [Hairer and Wanner, 1996, Section IV.3]. The analysis presented in this section is based upon the work in [Sturm, 2017, Chapter 4], where the Crank–Nicolson scheme applied to a dG discretization of Maxwell's equations is analyzed. We transfer this analysis to the abstract wave-type equation (2.24).

Before stating the scheme itself, we motivate it by giving a short derivation. We will see in Section 4.2 that the Peaceman–Rachford scheme can be derived similarly, underlining the connection between both schemes.

Example 4.1. Consider again the inhomogeneous initial value problem from Example 2.1, i.e.,

$$\begin{cases} \mathrm{d}_t u(t) = Au(t) + f(t), & t \in \mathbb{R}_+, \\ u(0) = u^0. \end{cases} \tag{4.1}$$

Given an approximation u_τ^n to the solution u at time t_n, we are interested in computing an approximation u_τ^{n+1} to u at time t_{n+1}. Starting with the given initial value u^0 at time t_0, we are then able to approximate the solution at all times t_n, $n \in \mathbb{N}$.

To derive this approximation, we apply the fundamental theorem of calculus to the exact solution and use (4.1), yielding

$$\begin{aligned} u(t_{n+1}) &= u(t_n) + \int_{t_n}^{t_{n+1}} \mathrm{d}_t u(s)\,\mathrm{d}s \\ &= u(t_n) + \int_{t_n}^{t_{n+1}} Au(s) + f(s)\,\mathrm{d}s. \end{aligned}$$

To obtain the iteration for the Crank–Nicolson scheme we approximate the integral on the right hand side by the trapezoidal rule. Doing so, we end up with

$$u(t_{n+1}) \approx u(t_n) + \tfrac{\tau}{2}\Big(\big(Au(t_n) + f^n\big) + \big(Au(t_{n+1}) + f^{n+1}\big)\Big).$$

Replacing $u(t_{n+1})$ and $u(t_n)$ by the approximations u_τ^{n+1} and u_τ^n and rearranging yields the **Crank–Nicolson scheme**

$$\begin{cases} (I - \tfrac{\tau}{2}A)u_\tau^{n+1} = (I + \tfrac{\tau}{2}A)u_\tau^n + \tfrac{\tau}{2}\big(f^{n+1} + f^n\big), & n \in \mathbb{N}_0, \\ u_\tau^0 = u^0. \end{cases}$$

Note that the scheme is implicit, i.e., we have to solve a linear system with coefficient matrix $I - \frac{\tau}{2}A$ in each step. ◇

Though the solution of a linear system in each step may be feasible for ordinary differential equations with a moderate number of unknowns, it can strongly affect the performance of the scheme if applied to spatially discretized partial differential equations, in particular in higher dimensions. This is due to the fact that the number of degrees of freedom, and hence, the dimension of the linear system grows under mesh refinement. On the other hand, the implicitness comes with an advantage, namely the unconditional stability of the scheme.

The Crank–Nicolson method applied to the wave-type problem (2.24) is given by

$$\begin{cases} \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{L}\big)u_\tau^{n+1} = \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{L}\big)u_\tau^n + \tfrac{\tau}{2}(f^{n+1} + f^n), & n \in \mathbb{N}_0, \qquad (4.2a) \\ u_\tau^0 = u^0. \qquad (4.2b) \end{cases}$$

We start by investigating the conditions on the initial value u^0 and the inhomogeneity f under which this scheme is wellposed.

4.1.1 Wellposedness

To investigate the wellposedness of the Crank–Nicolson scheme (4.2), we rewrite the scheme in a compact form. This is done by introducing the operators $\mathcal{R}_{\text{CN}} \colon L^2(\Omega)^m \to D(\mathcal{L})$ and $\mathcal{S}_{\text{CN}} \colon D(\mathcal{L}) \to D(\mathcal{L})$ defined as

$$\mathcal{R}_{\text{CN}} = \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{L}\big)^{-1}$$

and

$$\mathcal{S}_{\text{CN}} = \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{L}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{L}\big).$$

Since $\big(\mathcal{I} - \frac{\tau}{2}\mathcal{L}\big)^{-1}$ is the resolvent of a maximal dissipative operator, these operators are well-defined by Lemma 2.10 (i).

Given $u_\tau^n \in D(\mathcal{L})$ for all $n \in \mathbb{N}_0$, the Crank–Nicolson scheme (4.2) is equivalent to

$$\begin{cases} u_\tau^{n+1} = \mathcal{S}_{\text{CN}} u_\tau^n + \frac{\tau}{2}\mathcal{R}_{\text{CN}}(f^{n+1} + f^n), & n \in \mathbb{N}_0, \\ u_\tau^0 = u^0. \end{cases} \tag{4.3}$$

From this, we can already see that $u_\tau^{n+1} \in D(\mathcal{L})$ if $u_\tau^n \in D(\mathcal{L})$ for all $n \in \mathbb{N}_0$. Hence, if $u^0 \in D(\mathcal{L})$, the scheme is wellposed. We state this and the discrete version of the variation-of-constants formula in the next theorem.

Theorem 4.2. *Let $u^0 \in D(\mathcal{L})$. Then, for all $n \in \mathbb{N}_0$ and all $\tau > 0$, there exists a unique $u_\tau^{n+1} \in D(\mathcal{L})$ fulfilling the Crank–Nicolson scheme (4.2) given by the discrete variation-of-constants formula*

$$u_\tau^{n+1} = \mathcal{S}_{\text{CN}}^{n+1} u^0 + \frac{\tau}{2} \sum_{j=0}^{n} \mathcal{S}_{\text{CN}}^{n-j} \mathcal{R}_{\text{CN}}(f^{j+1} + f^j). \tag{4.4}$$

Proof. Existence and uniqueness of $u_\tau^{n+1} \in D(\mathcal{L})$ for all $n \in \mathbb{N}_0$ can be directly seen in (4.3). The discrete variation-of-constants formula can be verified by a straightforward induction argument. □

Remark 4.3. The operator $\mathcal{S}_{\text{CN}}$ corresponds to the stability function of the Crank–Nicolson scheme given by

$$S(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}, \qquad z \in \mathbb{C}.$$

Thus, at least formally, we have $\mathcal{S}_{\text{CN}} = S(\tau\mathcal{L})$. Further, $\mathcal{S}_{\text{CN}}$ can be seen as an approximation of the semigroup $\left(\mathrm{e}^{t\mathcal{L}}\right)_{t\geq 0}$ evaluated at $t = \tau$. Heuristically, this can be recognized by noting that the stability function S is the (1,1)-Padé-approximation to the exponential function (cf., [Hairer and Wanner, 1996, Section IV.3] for more details on the stability function). Having this in mind, the discrete variation-of-constants formula (4.4) can be seen as a discrete version of the usual variation-of-constants formula (2.25), justifying the name. ◇

4.1.2 Stability

We now show that the operator $\mathcal{S}_{\text{CN}}$ is contractive. This reflects the corresponding property of the semigroup $\left(\mathrm{e}^{t\mathcal{L}}\right)_{t\geq 0}$.

Lemma 4.4. *Let $v \in D(\mathcal{L})$. Then, for all $\tau > 0$, we have*

$$\|\mathcal{S}_{\text{CN}} v\|_M \leq \|v\|_M,$$

i.e., the operator $\mathcal{S}_{\text{CN}}$ is a contraction.

Proof. Note that $\mathcal{L}$ and its resolvent commute on $D(\mathcal{L})$, cf., e.g., [Schnaubelt, 2015, Theorem 1.13]. Consequently, $\mathcal{S}_{\text{CN}} v$ is the transform encountered in Lemma 2.10 (ii) for $\lambda = \frac{\tau}{2}$ applied to $v \in D(\mathcal{L})$. Further, recall that by Theorem 2.26, $\mathcal{L}$ is maximal dissipative w.r.t. $\left(\cdot \,\middle|\, \cdot\right)_M$ and $\left(\mathcal{I} - \frac{\tau}{2}\mathcal{L}\right) \colon D(\mathcal{L}) \to L^2(\Omega)^m$ is an isomorphism.

Thus, Lemma 2.10 (ii) yields

$$\|\mathcal{S}_{\text{CN}} v\|_M = \|\left(\mathcal{I} - \tfrac{\tau}{2}\mathcal{L}\right)^{-1}\left(\mathcal{I} + \tfrac{\tau}{2}\mathcal{L}\right) v\|_M = \|\left(\mathcal{I} + \tfrac{\tau}{2}\mathcal{L}\right)\left(\mathcal{I} - \tfrac{\tau}{2}\mathcal{L}\right)^{-1} v\|_M \leq \|v\|_M,$$

showing the contractivity of $\mathcal{S}_{\text{CN}}$. □

As a consequence, arbitrary powers of $\mathcal{S}_{\text{CN}}$ applied to elements in $D(\mathcal{L})$ are bounded by the weighted L^2-norm of that element.

Corollary 4.5. *Let $q \in \mathbb{N}$ and $\tau > 0$. Then, for all $v \in D(\mathcal{L})$ we have*

$$\|\mathcal{S}_{\text{CN}}^q v\|_M \leq \|v\|_M,$$

and for all $v \in L^2(\Omega)^m$ we have

$$\|\mathcal{S}_{\text{CN}}^q \mathcal{R}_{\text{CN}} v\|_M \leq \|v\|_M.$$

Proof. Let $v \in D(\mathcal{L})$. Since we have $\mathcal{S}_{\text{CN}}^q v \in D(\mathcal{L})$ for $q \in \mathbb{N}$, we can apply Lemma 4.4 to obtain

$$\|\mathcal{S}_{\text{CN}}^q v\|_M = \|\mathcal{S}_{\text{CN}} \mathcal{S}_{\text{CN}}^{q-1} v\|_M \leq \|\mathcal{S}_{\text{CN}}^{q-1} v\|_M.$$

Repeating this argument $q-1$ times yields the asserted inequality.

For the second assertion, recall that for $v \in L^2(\Omega)^m$ we have $\mathcal{R}_{\text{CN}} v \in D(\mathcal{L})$. Hence, this is a direct consequence of the first assertion and Lemma 2.10 (i). □

Using this, we can state the **stability result for the Crank–Nicolson scheme**.

Corollary 4.6. *Let $u^0 \in D(\mathcal{L})$. Then, for all $n \in \mathbb{N}_0$ and all $\tau > 0$, the approximation u_τ^{n+1} given by the Crank–Nicolson scheme (4.2) satisfies*

$$\|u_\tau^{n+1}\|_M \leq \|u^0\|_M + \tfrac{\tau}{2} \sum_{j=0}^{n} \|f^{j+1} + f^j\|_M.$$

Proof. We use the discrete variation-of-constants formula (4.4) and the triangle inequality to obtain

$$\|u_\tau^{n+1}\|_M = \|\mathcal{S}_{\text{CN}}^{n+1} u^0\|_M + \tfrac{\tau}{2} \sum_{j=0}^{n} \|\mathcal{S}_{\text{CN}}^{n-j} \mathcal{R}_{\text{CN}} (f^{j+1} + f^j)\|_M.$$

The claim now follows by Corollary 4.5, since we have $u^0 \in D(\mathcal{L})$ and $f(t) \in L^2(\Omega)^m$ for all $t \in \mathbb{R}_+$. □

Remark 4.7. 1. As we have already pointed out in Remark 4.3, the operator $\mathcal{S}_{\text{CN}}$ approximates the semigroup $\left(\mathrm{e}^{t\mathcal{L}}\right)_{t\geq 0}$ at $t = \tau$. Thinking along the same lines, the contractivity of $\mathcal{S}_{\text{CN}}$ shown in this section reflects the corresponding property of the semigroup.

In particular, given a skew-adjoint operator, the corresponding semigroup is norm-conserving. This behavior is preserved by $\mathcal{S}_{\text{CN}}$, cf., e.g., [Sturm, 2017, Section 4.2.1] for the discretized undamped Maxwell's equations.

2. In light of Remark 2.28 on the necessity of dissipativity we want to point out that in the shift-dissipative case, the operator $\mathcal{S}_{\text{CN}}$ is not contractive. However, in that case it reflects the corresponding property of the semigroup generated by the shift-dissipative operator. Namely, we have exponential growth over time, respectively number of steps. ◇

4.2 The Peaceman–Rachford scheme

The **Peaceman–Rachford scheme** proposed in [Peaceman and Rachford, 1955] is a time integration method designed for problems that can be split into two distinct subproblems. It was originally used in the context of dimension splitting. However, other splittings have been considered, e.g., into a linear and a nonlinear part, cf., [Hansen and Henningsson, 2013].

Combining the Peaceman–Rachford method with the approach of dimensional splitting leads to an alternating direction implicit (ADI) method. The main feat of this approach is the fact that it is unconditionally stable and formally of order two in time but can be performed much more efficiently than the Crank–Nicolson scheme if applied to problems possessing a certain structure. Namely, it can be performed roughly at the cost of an explicit scheme. In fact, we determine a class of wave-type problems for which this is the case in Chapter 6.

It is well-known that the Peaceman–Rachford scheme can be interpreted as a perturbation of the Crank–Nicolson scheme. The analysis performed in this section relies on this fact, as it adapts the fully discrete analysis of the Crank–Nicolson scheme introduced in [Sturm, 2017] to the Peaceman–Rachford scheme.

Similar results have been shown in [Hochbruck et al., 2015a], where an ADI scheme applied to the homogeneous Maxwell's equations was analyzed in the setting of abstract evolution equations. These results were recently generalized to the inhomogeneous case including a damping term in the two papers [Eilinghoff and Schnaubelt, 2018, Eilinghoff and Schnaubelt, 2017] and the dissertation [Eilinghoff, 2017].

The analyses in these publications differ in two main points from the analysis in this thesis. First, techniques from [Hansen and Ostermann, 2008] are used to obtain the convergence results as opposed to the approach in this thesis to interpret the scheme as a perturbed Crank–Nicolson scheme. We chose the latter as it enables us to obtain full discretization results by transferring the results in [Sturm, 2017].

Second, the regularity assumptions in these publications are formulated in terms of regularity of the data, i.e., the initial value and the inhomogeneity. As this heavily relies on the particular structure of Maxwell's equations and we study a broader class of equations, the regularity assumptions in this thesis are given in terms of the regularity of the exact solution.

As for the Crank–Nicolson scheme we give a short derivation of the scheme itself.

Example 4.8. To derive the Peaceman–Rachford scheme we proceed similar to the Crank–Nicolson scheme in Example 4.1. This time, consider the homogeneous initial value problem from Example 2.1 given by

$$\begin{cases} \mathrm{d}_t u(t) = Au(t), & t \in \mathbb{R}_+, \\ u(0) = u^0. \end{cases}$$

Assume that we can split the matrix A into

$$A = A_1 + A_2,$$

such that we can solve linear systems with coefficient matrix $I - \frac{\tau}{2}A_1$ or $I - \frac{\tau}{2}A_2$ more efficiently than systems with coefficient matrix $I - \frac{\tau}{2}A$.

We proceed as in Example 4.1 to get

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} A_1 u(s)\,\mathrm{d}s + \int_{t_n}^{t_{n+1}} A_2 u(s)\,\mathrm{d}s.$$

However, instead of using the trapezoidal rule to approximate the integrals, we now apply the midpoint rule to the first and the trapezoidal rule to the second integral. This yields

$$u(t_{n+1}) \approx u(t_n) + \tau A_1 u(t_{n+1/2}) + \tfrac{\tau}{2} A_2\big(u(t_n) + u(t_{n+1})\big).$$

If we want to use this to compute an approximation to $u(t_{n+1})$ from $u(t_n)$ we additionally need an approximation to $u(t_{n+1/2})$. To establish this approximation we proceed analogously to obtain

$$u(t_{n+1/2}) = u(t_n) + \int_{t_n}^{t_{n+1/2}} A_1 u(s)\,\mathrm{d}s + \int_{t_n}^{t_{n+1/2}} A_2 u(s)\,\mathrm{d}s.$$

This time, we use the right rectangular rule for the first and the left rectangular rule for the second integral, yielding

$$u(t_{n+1/2}) \approx u(t_n) + \tfrac{\tau}{2} A_1 u(t_{n+1/2}) + \tfrac{\tau}{2} A_2 u(t_n).$$

Again, we replace the exact values $u(t_n)$, $u(t_{n+1/2})$ and $u(t_{n+1})$ by the approximations u_τ^n, $u_\tau^{n+1/2}$ and u_τ^{n+1} to obtain

$$\begin{aligned} u_\tau^{n+1/2} &= u_\tau^n + \tfrac{\tau}{2} A_1 u_\tau^{n+1/2} + \tfrac{\tau}{2} A_2 u_\tau^n, \\ u_\tau^{n+1} &= u_\tau^n + \tau A_1 u_\tau^{n+1/2} + \tfrac{\tau}{2} A_2\big(u_\tau^n + u_\tau^{n+1}\big). \end{aligned}$$

Subtracting the first equation from the second to eliminate the terms involving u_τ^n and rearranging the terms yields the **Peaceman–Rachford scheme**

$$\begin{cases} (I - \tfrac{\tau}{2} A_1) u_\tau^{n+1/2} = (I + \tfrac{\tau}{2} A_2) u_\tau^n, \\ (I - \tfrac{\tau}{2} A_2) u_\tau^{n+1} = (I + \tfrac{\tau}{2} A_1) u_\tau^{n+1/2}, \qquad n \in \mathbb{N}_0, \\ u_\tau^0 = u^0. \end{cases}$$

Note that, like the Crank–Nicolson scheme, the Peaceman–Rachford scheme is implicit. Namely, in each step we have to solve two linear systems with coefficient matrices given by either $I - \frac{\tau}{2} A_1$ or $I - \frac{\tau}{2} A_2$. However, if these are easier to solve than systems with coefficient matrix $I - \frac{\tau}{2} A$, as we assumed in the beginning of the example, this can be favorable over the Crank–Nicolson scheme.

In this thesis we consider inhomogeneous problems. Hence, take again the linear system of inhomogeneous ordinary differential equations

$$\begin{cases} \mathrm{d}_t u(t) = A u(t) + f(t), \qquad t \in \mathbb{R}_+, \\ u(0) = u^0. \end{cases}$$

The Peaceman–Rachford scheme applied to the inhomogeneous equation we consider in this thesis is given by

$$\begin{cases} (I - \tfrac{\tau}{2} A_1) u_\tau^{n+1/2} = (I + \tfrac{\tau}{2} A_2) u_\tau^n, \\ (I - \tfrac{\tau}{2} A_2) u_\tau^{n+1} = (I + \tfrac{\tau}{2} A_1)\big(u_\tau^{n+1/2} + \tfrac{\tau}{2}\big(f^{n+1} + f^n\big)\big), \qquad n \in \mathbb{N}_0, \\ u_\tau^0 = u^0. \end{cases}$$

This treatment of the right-hand side is taken from [Eilinghoff and Schnaubelt, 2018]. It is chosen such that only two linear systems have to be solved in each full step. However, other treatments are possible, cf., Section 4.4. ◇

Recall that we have split $\mathcal{L} = \mathcal{A} + \mathcal{B}$ in (2.26). Hence, together with this splitting, the Peaceman–Rachford method applied to the wave-type problem (2.24) is given by

$$\begin{cases} \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)u_\tau^{n+1/2} = \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big)u_\tau^n, & \text{(4.5a)} \\ \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)u_\tau^{n+1} = \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(u_\tau^{n+1/2} + \tfrac{\tau}{2}(f^{n+1} + f^n)\big), \qquad n \in \mathbb{N}_0, & \text{(4.5b)} \\ u_\tau^0 = u^0. & \text{(4.5c)} \end{cases}$$

Again, we start by investigating the conditions on the initial value u^0 and the inhomogeneity f under which this scheme is wellposed.

4.2.1 Wellposedness

We rewrite (4.5a) and (4.5b) equivalently as

$$u_\tau^{n+1/2} = \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big)u_\tau^n, \tag{4.6a}$$

$$u_\tau^{n+1} = \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(u_\tau^{n+1/2} + \tfrac{\tau}{2}(f^{n+1} + f^n)\big). \tag{4.6b}$$

From this, we can already see that $u_\tau^{n+1/2} \in D(\mathcal{A})$ and thus also $u_\tau^{n+1} \in D(\mathcal{B})$ if $u_\tau^n \in D(\mathcal{B})$ and $f \in C(\mathbb{R}_+; D(\mathcal{A}))$.

To obtain a discrete variation-of-constants formula for the Peaceman–Rachford scheme, we rewrite the scheme similar to the compact form of the Crank–Nicolson scheme (4.3). To this end we define the operator $\mathcal{S}_{\text{PR}} \colon D(\mathcal{B}) \to D(\mathcal{B})$ as

$$\mathcal{S}_{\text{PR}} = \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big).$$

As we will need it for the analysis of the scheme, we also define the operator $\mathcal{R}_{\text{PR}} \colon L^2(\Omega)^m \to D(\mathcal{B})$ as

$$\mathcal{R}_{\text{PR}} = \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)^{-1}.$$

Note that as for the Crank–Nicolson scheme both operators are well-defined, since $\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}$ and $\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)^{-1}$ are the resolvents of maximal dissipative operators.

By inserting (4.6a) into (4.6b) we can rewrite the Peaceman–Rachford scheme (4.5) equivalently as

$$\begin{cases} u_\tau^{n+1} = \mathcal{S}_{\text{PR}}u_\tau^n + \tfrac{\tau}{2}\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(f^{n+1} + f^n\big), \qquad n \in \mathbb{N}_0, & \text{(4.7a)} \\ u_\tau^0 = u^0. & \text{(4.7b)} \end{cases}$$

We can now state the wellposedness result and give the aforementioned discrete variation-of-constants formula for u_τ^{n+1} in the next theorem.

Theorem 4.9. *Let $u^0 \in D(\mathcal{B})$ and $f \in C(\mathbb{R}_+; D(\mathcal{A}))$. Then, for all $n \in \mathbb{N}_0$ and all $\tau > 0$, there exist unique $u_\tau^{n+1} \in D(\mathcal{B})$ and $u_\tau^{n+1/2} \in D(\mathcal{A})$ fulfilling the Peaceman–Rachford scheme (4.5). Further, u_τ^{n+1} is given for all $n \in \mathbb{N}_0$ by the discrete variation-of-constants formula*

$$u_\tau^{n+1} = \mathcal{S}_{\text{PR}}^{n+1}u^0 + \tfrac{\tau}{2}\sum_{j=0}^{n} \mathcal{S}_{\text{PR}}^{n-j}\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(f^{j+1} + f^j\big). \tag{4.8}$$

Proof. Existence and uniqueness of $u_\tau^{n+1} \in D(\mathcal{B})$ and $u_\tau^{n+1/2} \in D(\mathcal{A})$ for all $n \in \mathbb{N}_0$ can be seen in (4.7) by acknowledging $\big(\mathcal{I} + \frac{\tau}{2}\mathcal{A}\big) f(t) \in L^2(\Omega)^m$ for all $t \in \mathbb{R}_+$. The discrete variation-of-constants formula can again be verified by a straightforward induction argument. □

Remark 4.10. 1. The approach used in this section (including the following part on stability) is based on [Hochbruck et al., 2015a, Section 2.2], where the unconditional stability of the scheme was shown in a similar way. Further, a similar proof for matrices can be found in [Faragó et al., 2005] and alternative strategies of proof can be found in the literature, e.g., in [Lee and Fornberg, 2004].

2. The discrete variation-of-constants formula (4.8) yields an explicit formula for u_τ^{n+1} for all $n \in \mathbb{N}_0$. Using this and (4.5a), we can explicitly express the intermediate solution $u_\tau^{n+1/2}$ for all $n \in \mathbb{N}_0$. ◇

4.2.2 Stability

We now show that the operator $\mathcal{S}_{\text{PR}}$ fulfills a bound similar to (2.5). However, unlike the corresponding bound for the semigroup and the Crank–Nicolson operator, we can only bound its application in the graph norm of $\mathcal{B}$. For the analysis we further need two additional bounds.

Lemma 4.11. *Let $q \in \mathbb{N}$ and $\tau > 0$. Then, for $v \in D(\mathcal{B})$ we have*

$$\|\mathcal{S}_{\text{PR}}^q v\|_M \leq \|(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B})v\|_M,$$

and for $v \in L^2(\Omega)^m$ we have

$$\|\mathcal{S}_{\text{PR}}^q \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1} v\|_M \leq \|v\|_M$$

and

$$\|\mathcal{S}_{\text{PR}}^q \mathcal{R}_{\text{PR}} v\|_M \leq \|v\|_M.$$

Proof. The first inequality can easily be seen by noting that concatenating $\mathcal{S}_{\text{PR}}$ multiple times creates concatenations of the transforms encountered in Lemma 2.10 (ii) corresponding to the operators $\mathcal{A}$ and $\mathcal{B}$. Abbreviating

$$\mathcal{C} = \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big)\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}$$

yields

$$\begin{aligned} \mathcal{S}_{\text{PR}}^q v &= \big(\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big)\big)^q v \\ &= \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\mathcal{C}^{q-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big)v. \end{aligned} \tag{4.9}$$

Since $\mathcal{A}$ and $\mathcal{B}$ are maximal dissipative, we have $\|\big(\mathcal{I} - \frac{\tau}{2}\mathcal{B}\big)^{-1} w\|_M \leq \|w\|_M$, $\|\mathcal{C}w\|_M \leq \|w\|_M$ and $\|\big(\mathcal{I} + \frac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} - \frac{\tau}{2}\mathcal{A}\big)^{-1} w\|_M \leq \|w\|_M$ for all $w \in L^2(\Omega)^m$ by Lemma 2.10. This yields the first assertion.

Using (4.9), we further have

$$\mathcal{S}_{\text{PR}}^q \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1} v = \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\mathcal{C}^q v$$

and

$$\mathcal{S}_{\mathrm{PR}}^{q}\mathcal{R}_{\mathrm{PR}}v = \left(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\right)^{-1}\mathcal{C}^{q}\left(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\right)^{-1}v$$

for $v \in L^2(\Omega)^m$. Again using the contractivity of $\left(\mathcal{I} - \frac{\tau}{2}\mathcal{A}\right)^{-1}$, $\left(\mathcal{I} - \frac{\tau}{2}\mathcal{B}\right)^{-1}$ and $\mathcal{C}$ proves the remaining assertions. □

Note that the statements in the second paragraph of Remark 4.7 also hold for these bounds, as they rely on the contractivity of the resolvents and the transforms from Lemma 2.10 (ii). In the case of shift dissipative operators $\mathcal{A}$ and $\mathcal{B}$, we again obtain bounds growing exponentially in time.

We conclude this section by the **stability result for the Peaceman–Rachford scheme**.

Corollary 4.12. *Let $u^0 \in D(\mathcal{B})$ and $f \in C(\mathbb{R}_+; D(\mathcal{A}))$. Then, for all $n \in \mathbb{N}_0$ and all $\tau > 0$, the approximation u_τ^{n+1} given by the Peaceman–Rachford scheme (4.5) satisfies*

$$\|u_\tau^{n+1}\|_M \leq \left\|\left(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\right)u^0\right\|_M + \tfrac{\tau}{2}\sum_{j=0}^{n}\left\|\left(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\right)\left(f^{j+1} + f^j\right)\right\|_M. \tag{4.10}$$

Proof. The proof can be performed completely analogously to the one for the Crank–Nicolson scheme, i.e., the proof of Corollary 4.6. The only difference is that we use Lemma 4.11 instead of Corollary 4.5. □

4.3 Error analysis of the temporal semidiscretization

Having established wellposedness and stability of both time integration schemes, we are now able to perform the error analysis of these schemes. For this, let u be the solution of the continuous problem (2.24) and u_τ^n be the approximation given by either the Crank–Nicolson scheme (4.2) or the Peaceman–Rachford scheme (4.5) after n steps. We denote the **temporally semidiscrete error after n steps** by

$$e^n = u(t_n) - u_\tau^n.$$

We follow a strategy similar to the spatial error analysis in Section 3.7. In other words, we show that the error satisfies a perturbed version of the scheme under consideration. However, we use the temporally semidiscrete schemes (4.2) and (4.5) instead of the spatially semidiscrete scheme (3.13).

4.3.1 Error recursions

As in Section 3.7 we show that the errors satisfy perturbed versions of the respective scheme.

Crank–Nicolson method

We start by inserting the solution of the continuous problem (2.24) into the Crank–Nicolson scheme (4.2). The defect caused by this is investigated in the next lemma. Note that the regularity assumptions on the solution coincide with those of the wellposedness result Corollary 2.27. However, we will need more regularity to prove convergence of the scheme.

Lemma 4.13. *Assume $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}))$. Then, for all $n \in \mathbb{N}_0$ and all $\tau > 0$, the exact solution u satisfies*

$$\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{L}\big)u(t_{n+1}) = \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{L}\big)u(t_n) + \tfrac{\tau}{2}(f^{n+1} + f^n) + \delta^n_{\mathrm{CN}}, \tag{4.11}$$

*where the **Crank–Nicolson defect** δ^n_{CN} is given by*

$$\delta^n_{\mathrm{CN}} = \int_{t_n}^{t_{n+1}} \mathrm{d}_t u(s)\,\mathrm{d}s - \tfrac{\tau}{2}\big(\mathrm{d}_t u(t_n) + \mathrm{d}_t u(t_{n+1})\big).$$

Proof. Equation (4.11) is obtained by replacing u_τ^{n+1} and u_τ^n in (4.2a) with the exact solution at times t_{n+1} and t_n. Since the exact solution does not satisfy the scheme, we obtain a defect, which we define as δ^n_{CN}. Rearranging the terms in (4.11) yields

$$\begin{aligned}\delta^n_{\mathrm{CN}} &= u(t_{n+1}) - u(t_n) - \tfrac{\tau}{2}\Big(\big(\mathcal{L}u(t_n) + f^n\big) + \big(\mathcal{L}u(t_{n+1}) + f^{n+1}\big)\Big)\\ &= u(t_{n+1}) - u(t_n) - \tfrac{\tau}{2}\big(\mathrm{d}_t u(t_n) + \mathrm{d}_t u(t_{n+1})\big),\end{aligned}$$

where we have used that u fulfills the continuous problem (2.24) in the second step. Using the fundamental theorem of calculus concludes the proof. □

We can now derive the desired error recursion.

Corollary 4.14. *Assume $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}))$. Then, for all $\tau > 0$, the error $e^n = u(t_n) - u_\tau^n$ of the Crank–Nicolson scheme satisfies*

$$\begin{cases} \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{L}\big)e^{n+1} = \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{L}\big)e^n + \delta^n_{\mathrm{CN}}, & n \in \mathbb{N}_0,\\ e^0 = 0. \end{cases}$$

Proof. We have $e^0 = 0$ since we chose the exact initial value for the Crank–Nicolson scheme. The remaining assertion follows by subtracting (4.2a) from (4.11). □

Peaceman–Rachford method

Similar to the Crank–Nicolson scheme we start by investigating the defect caused by inserting the exact solution into the Peaceman–Rachford scheme (4.7). However, to derive this defect, we require a bit more regularity of the exact solution and the inhomogeneity.

Lemma 4.15. *Assume $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{A}) \cap D(\mathcal{AB}))$ and that we have $f \in C(\mathbb{R}_+; D(\mathcal{A}^2))$. Then, for all $n \in \mathbb{N}_0$ and all $\tau > 0$, the exact solution u satisfies*

$$u(t_{n+1}) = \mathcal{S}_{\mathrm{PR}}u(t_n) + \tfrac{\tau}{2}\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(f^{n+1} + f^n\big) + \mathcal{R}_{\mathrm{PR}}\big(\delta^n_{\mathrm{CN}} + \delta^n_{\mathrm{PR}}\big), \tag{4.12}$$

*where δ^n_{CN} is the Crank–Nicolson defect from Lemma 4.13, and the (Peaceman–Rachford) **perturbation defect** δ^n_{PR} is given by*

$$\delta^n_{\mathrm{PR}} = \delta^n_{\mathrm{PR},u} + \delta^n_{\mathrm{PR},f}$$

with

$$\delta^n_{\mathrm{PR},u} = \frac{\tau^2}{4}\mathcal{AB}\big(u(t_{n+1}) - u(t_n)\big) \qquad \text{and} \qquad \delta^n_{\mathrm{PR},f} = \frac{\tau^3}{8}\mathcal{A}^2(f^{n+1} + f^n). \tag{4.13}$$

Proof. Equation (4.12) is obtained by inserting the exact solution into (4.7a) and defining the defect made by this as $\mathcal{R}_{\mathrm{PR}}(\delta_{\mathrm{CN}}^n + \delta_{\mathrm{PR}}^n)$. Since $u(t) \in D(\mathcal{AB})$ for all $t \geq 0$, we can apply $\mathcal{R}_{\mathrm{PR}}^{-1}$ to (4.12). Because of

$$\begin{aligned}\mathcal{R}_{\mathrm{PR}}^{-1}\mathcal{S}_{\mathrm{PR}}u(t_n) &= \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big)u(t_n)\\ &= \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big)u(t_n)\end{aligned}$$

and

$$\mathcal{R}_{\mathrm{PR}}^{-1}\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)^{-1}\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(f^{n+1} + f^n\big) = \big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(f^{n+1} + f^n\big),$$

this yields

$$\begin{aligned}\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\big)u(t_{n+1}) &= \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{B}\big)u(t_n)\\ &\quad + \tfrac{\tau}{2}\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\big)\big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\big)\big(f^{n+1} + f^n\big) + \delta_{\mathrm{CN}}^n + \delta_{\mathrm{PR}}^n.\end{aligned}$$

Expanding the first three terms and using $\mathcal{L} = \mathcal{A} + \mathcal{B}$ we obtain

$$\begin{aligned}\big(\mathcal{I} - \tfrac{\tau}{2}\mathcal{L} + \tfrac{\tau^2}{4}\mathcal{AB}\big)u(t_{n+1}) &= \big(\mathcal{I} + \tfrac{\tau}{2}\mathcal{L} + \tfrac{\tau^2}{4}\mathcal{AB}\big)u(t_n)\\ &\quad + \tfrac{\tau}{2}\big(\mathcal{I} - \tfrac{\tau^2}{8}\mathcal{A}^2\big)\big(f^{n+1} + f^n\big) + \delta_{\mathrm{CN}}^n + \delta_{\mathrm{PR}}^n.\end{aligned}$$

Rearranging the terms yields

$$\begin{aligned}\delta_{\mathrm{CN}}^n + \delta_{\mathrm{PR}}^n &= u(t_{n+1}) - u(t_n) - \tfrac{\tau}{2}\Big(\big(\mathcal{L}u(t_n) + f^n\big) + \big(\mathcal{L}u(t_{n+1}) + f^{n+1}\big)\Big)\\ &\quad + \tfrac{\tau^2}{4}\mathcal{AB}\big(u(t_{n+1}) - u(t_n)\big) + \tfrac{\tau^3}{8}\mathcal{A}^2\big(f^{n+1} + f^n\big).\end{aligned} \tag{4.14}$$

By the proof of Lemma 4.13, the first few terms coincide with the Crank–Nicolson defect δ_{CN}^n. This concludes the proof. □

From this we can derive the error recursion.

Corollary 4.16. *Assume $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{A}) \cap D(\mathcal{AB}))$ and that we have $f \in C(\mathbb{R}_+; D(\mathcal{A}^2))$. Then, for all $\tau > 0$, the error $e^n = u(t_n) - u_\tau^n$ of the Peaceman–Rachford scheme satisfies*

$$\begin{cases} e^{n+1} = \mathcal{S}_{\mathrm{PR}}e^n + \mathcal{R}_{\mathrm{PR}}(\delta_{\mathrm{CN}}^n + \delta_{\mathrm{PR}}^n), & n \in \mathbb{N}_0,\\ e^0 = 0. \end{cases} \tag{4.15}$$

Proof. We have $e^0 = 0$ since we chose the exact initial value for the Peaceman–Rachford scheme. The remaining assertion follows by subtracting (4.7a) from (4.12). □

4.3.2 Bounds on the defects

In the last section we have seen that the errors of the Crank–Nicolson and the Peaceman–Rachford scheme satisfy a perturbed version of the respective scheme if the exact solution is smooth enough. Hence, by Theorems 4.2 and 4.9, we can adapt the respective discrete variation-of-constants to solve them. Note that we have $e^0 = 0$, as we use the exact initial values for both schemes. This yields

$$e^{n+1} = \sum_{j=0}^{n} \mathcal{S}_{\mathrm{CN}}^{n-j}\mathcal{R}_{\mathrm{CN}}\delta_{\mathrm{CN}}^j \tag{4.16}$$

for the Crank–Nicolson scheme and

$$e^{n+1} = \sum_{j=0}^{n} \mathcal{S}_{\mathrm{PR}}^{n-j} \mathcal{R}_{\mathrm{PR}} (\delta_{\mathrm{CN}}^{j} + \delta_{\mathrm{PR}}^{j}) \tag{4.17}$$

for the Peaceman–Rachford scheme. Owing to the stability of the schemes—or more precisely Corollary 4.5 and Lemma 4.11—we already have suitable bounds on the operators $\mathcal{S}_{\mathrm{CN}}^{n-j}\mathcal{R}_{\mathrm{CN}}$ and $\mathcal{S}_{\mathrm{PR}}^{n-j}\mathcal{R}_{\mathrm{PR}}$. Hence, it remains to bound the defects δ_{CN}^{j} and δ_{PR}^{j}.

Crank–Nicolson defect

We start with the Crank–Nicolson defect δ_{CN}^{n}. In fact, δ_{CN}^{n} is the quadrature error of the trapezoidal rule applied to $\mathrm{d}_t u$. This enables us to derive a different representation if the exact solution is sufficiently smooth, yielding a suitable bound on the defect.

Corollary 4.17. *Assume $u \in C^3(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}))$. Then, for all $\tau > 0$, we have*

$$\delta_{\mathrm{CN}}^{n} = \tau^2 \int_{t_n}^{t_{n+1}} \frac{(s-t_n)(s-t_{n+1})}{2\tau^2} \, \mathrm{d}_t^3 u(s) \, \mathrm{d}s \tag{4.18}$$

and thus

$$\|\delta_{\mathrm{CN}}^{n}\|_M \leq \frac{\tau^2}{8} \int_{t_n}^{t_{n+1}} \|\mathrm{d}_t^3 u(s)\|_M \, \mathrm{d}s. \tag{4.19}$$

Proof. As mentioned before, δ_{CN}^{n} is the quadrature error of the trapezoidal rule applied to $\mathrm{d}_t u$. It is a well-known fact that this quadrature error can be expressed via the Peano kernel $s(s-1)/2$ if u is three times continuously differentiable, cf., e.g., [Hochbruck, 2015, Theorem 1.10]. This yields (4.18). The bound (4.19) is a straightforward consequence of this representation. □

Perturbation defect

It remains to bound the perturbation defect δ_{PR}^{n} of the Peaceman–Rachford scheme. The defect $\delta_{\mathrm{PR},f}^{n}$ can be bounded in a straightforward manner as we will see later. Hence, we only derive a bound on $\delta_{\mathrm{PR},u}^{n}$ in this section. Again, we require slightly more regularity of the exact solution to acquire a suitable bound. We state this in the next lemma.

Corollary 4.18. *Assume $u \in C^1(\mathbb{R}_+; D(\mathcal{AB}))$. Then, for all $\tau > 0$, we have*

$$\delta_{\mathrm{PR},u}^{n} = \frac{\tau^2}{4} \int_{t_n}^{t_{n+1}} \mathcal{AB}\mathrm{d}_t u(s) \, \mathrm{d}s \tag{4.20}$$

and thus

$$\|\delta_{\mathrm{PR},u}^{n}\|_M \leq \frac{\tau^2}{4} \int_{t_n}^{t_{n+1}} \|\mathcal{AB}\mathrm{d}_t u(s)\|_M \, \mathrm{d}s. \tag{4.21}$$

Proof. Since $u \in C^1(\mathbb{R}_+; D(\mathcal{AB}))$, we can use the fundamental theorem of calculus to obtain

$$\delta_{\mathrm{PR},u}^{n} = \frac{\tau^2}{4} \mathcal{AB} \int_{t_n}^{t_{n+1}} \mathrm{d}_t u(s) \, \mathrm{d}s = \frac{\tau^2}{4} \int_{t_n}^{t_{n+1}} \mathcal{AB}\mathrm{d}_t u(s) \, \mathrm{d}s.$$

This shows (4.20). The bound (4.21) follows by the triangle inequality for integrals. □

4.3.3 Temporal convergence results

Since we have established bounds on the defects of both the Crank–Nicolson and the Peaceman–Rachford scheme, we are now able to give the convergence results. We show that both schemes are of order two in time, provided the exact solution is smooth enough.

Crank–Nicolson method

The next theorem shows the **convergence result for the Crank–Nicolson discretization** of the wave-type problem (2.24).

Theorem 4.19. *Assume that the exact solution of the wave-type problem (2.24) satisfies $u \in C^3(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}))$. Then, for all $n \in \mathbb{N}_0$ and all $\tau > 0$, the error of the Crank–Nicolson scheme satisfies*

$$\begin{aligned}\|u(t_{n+1}) - u_\tau^{n+1}\|_M &\leq \frac{\tau^2}{8} \int_0^{t_{n+1}} \|\mathrm{d}_t^3 u(s)\|_M \,\mathrm{d}s \\ &\leq C\tau^2,\end{aligned}$$

where C only depends on t_{n+1}, $\|\mathrm{d}_t^3 u(s)\|_M$, $s \in [0, t_{n+1}]$.

Proof. First, by Corollary 4.14, the error e^n satisfies the Crank–Nicolson scheme (4.2) with initial value 0 and $\frac{\tau}{2}(f^{n+1} + f^n)$ replaced by the Crank–Nicolson defect δ_{CN}^n. Further, for all $n \in \mathbb{N}_0$, we have $e^n \in D(\mathcal{L})$, since $u(t_n) \in D(\mathcal{L})$ and $u_\tau^n \in D(\mathcal{L})$.

Hence, we can apply Theorem 4.2 and the discrete variation-of-constants formula (4.4) to obtain (4.16). By using the triangle inequality and subsequently Corollary 4.5, we obtain

$$\begin{aligned}\|e^{n+1}\|_M &\leq \sum_{j=0}^{n} \|\mathcal{S}_{\mathrm{CN}}^{n-j} \mathcal{R}_{\mathrm{CN}} \delta_{\mathrm{CN}}^j\|_M \\ &\leq \sum_{j=0}^{n} \|\delta_{\mathrm{CN}}^j\|_M.\end{aligned}$$

Using (4.19) to bound the defect concludes the proof. □

Peaceman–Rachford method

Now, we state the **convergence result for the Peaceman–Rachford discretization** of the wave-type problem (2.24). Note that we have to assume additional regularity of the exact solution and the inhomogeneity, as we also need to bound the perturbation defect δ_{PR}^n.

Theorem 4.20. *Assume that the exact solution of the wave-type problem (2.24) satisfies $u \in C^3(\mathbb{R}_+; L^2(\Omega)^m) \cap C^1(\mathbb{R}_+; D(\mathcal{A}\mathcal{B})) \cap C(\mathbb{R}_+; D(\mathcal{A}))$ and that $f \in C(\mathbb{R}_+; D(\mathcal{A}^2))$. Then, for all $n \in \mathbb{N}_0$ and all $\tau > 0$, the error of the Peaceman–Rachford scheme satisfies*

$$\begin{aligned}\|u(t_{n+1}) - u_\tau^{n+1}\|_M \leq \frac{\tau^2}{4}\Big(&\int_0^{t_{n+1}} \tfrac{1}{2}\|\mathrm{d}_t^3 u(s)\|_M + \|\mathcal{A}\mathcal{B}\mathrm{d}_t u(s)\|_M \,\mathrm{d}s \\ &+ \tfrac{\tau}{2} \sum_{j=0}^{n} \|\mathcal{A}^2\big(f^j + f^{j+1}\big)\|_M\Big) \\ \leq C\tau^2,\end{aligned}$$

where C only depends on t_{n+1}, $\|\mathrm{d}_t^3 u(s)\|_M$, $\|\mathcal{A}\mathcal{B}\mathrm{d}_t u(s)\|_M$ and $\|\mathcal{A}^2 f(s)\|_M$, $s \in [0, t_{n+1}]$.

Proof. By Corollary 4.16, the error e^n satisfies the perturbed Peaceman–Rachford scheme (4.15). Since $u(t_n) \in D(\mathcal{B})$ and $u_\tau^n \in D(\mathcal{B})$, we further have $e^n \in D(\mathcal{B})$ for all $n \in \mathbb{N}_0$.

Hence, we can apply the discrete variation-of-constants formula (4.8) to obtain (4.17). Taking the $\|\cdot\|_M$-norm and using the triangle inequality and subsequently Lemma 4.11 yields

$$\begin{aligned}\|e^{n+1}\|_M &\leq \sum_{j=0}^{n} \|\mathcal{S}_{\mathrm{PR}}^{n-j}\mathcal{R}_{\mathrm{PR}}(\delta_{\mathrm{CN}}^j + \delta_{\mathrm{PR}}^j)\|_M \\ &\leq \sum_{j=0}^{n} \|\delta_{\mathrm{CN}}^j\|_M + \sum_{j=0}^{n} \|\delta_{\mathrm{PR},u}^j\|_M + \sum_{j=0}^{n} \|\delta_{\mathrm{PR},f}^j\|_M.\end{aligned}$$

The first two terms can be treated by using the bounds (4.19) and (4.21). The bound on the third term is a straightforward consequence of the definition of $\delta_{\mathrm{PR},f}^n$, see (4.13). □

4.4 Concluding remarks

We conclude this chapter by some remarks.

4.4.1 Regularity assumptions

If we demand less regularity of the solution, we can still show convergence. However, the convergence order decreases accordingly.

In particular, by only demanding the exact solution to lie in $C^2(\mathbb{R}_+; L^2(\Omega)^m)$ instead of $C^3(\mathbb{R}_+; L^2(\Omega)^m)$, we can still bound the Crank–Nicolson defect with order one in τ. This can be achieved by using Peano kernels of order one.

Similarly, if we only demand the regularity of the exact solution from Lemma 4.15, we can still bound the Peaceman–Rachford defect. Again, we lose one order in τ, as we can not apply the fundamental theorem of calculus.

4.4.2 Variants of the schemes

Different treatments of the inhomogeneity in both schemes are possible using the same strategy of proof. We shortly remark upon two of them.

The implicit midpoint scheme

We can replace $\frac{\tau}{2}(f^{n+1} + f^n)$ in the Crank–Nicolson scheme (4.2a) by $\tau f(t_{n+1/2})$ to obtain the implicit midpoint scheme. We refer to [Sturm, 2017, Section 4.4] for the fully discrete analysis of this method for Maxwell's equations.

Variant of the Peaceman–Rachford scheme

Similarly, we can replace

$$\left(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\right)(f^{n+1} + f^n)$$

by

$$\left(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\right)^{-1}(f^{n+1} + f^n)$$

in the Peaceman–Rachford scheme (4.5b). Using this, we don't need the additional regularity assumption $f \in C(\mathbb{R}_+; D(\mathcal{A}^2))$ on the inhomogeneity. Further, the corresponding error term in Theorem 4.20 vanishes.

This is due to the fact that the different treatment of the inhomogeneity gives rise to the term

$$\mathcal{R}_{\mathrm{PR}}(f^{n+1} + f^n)$$

instead of

$$\left(\mathcal{I} - \tfrac{\tau}{2}\mathcal{B}\right)^{-1}\left(\mathcal{I} + \tfrac{\tau}{2}\mathcal{A}\right)(f^{n+1} + f^n)$$

in (4.7a). Because of this, the last term in (4.14), and hence, the defect $\delta^n_{\mathrm{PR},f}$ does not occur. Further, the stability bound (4.10) would change accordingly.

However, this comes at the cost of having to solve an additional linear system in each step to compute $\left(\mathcal{I} - \tfrac{\tau}{2}\mathcal{A}\right)^{-1}(f^{n+1} + f^n)$. We refer to Chapter 6 for more details.

5 | Full discretization

This chapter is devoted to the fully discrete numerical scheme obtained by combining the dG method in space with the Crank–Nicolson and Peaceman–Rachford scheme in time, respectively. These schemes are derived via a method of lines approach, meaning we first discretize in space and subsequently in time. To analyze the schemes, we proceed in a similar way as for the temporal discretization, however, we have to deal with additional defects stemming from the spatial discretization.

Full discretization results for the Peaceman–Rachford scheme in literature are usually performed for non-stiff ordinary differential equations, see e.g., [Hundsdorfer and Verwer, 1989, Pearcy, 1962]. This leads to error bounds in which the constants depend on the norm of the matrices resulting from the discretization of the spatial operators. As these matrices approximate unbounded operators, their norm tends to infinity under refinement of the spatial mesh. Consequently, such techniques can only provide reliable results if a fixed spatial mesh is considered.

The only rigorous convergence result for a full discretization obtained by using the Peaceman–Rachford scheme in time known to the author is given in [Hansen and Henningsson, 2016] based on the techniques from [Hansen and Ostermann, 2008]. Therein, a general class of space discretization techniques fulfilling certain conditions is considered for the discretization of the spatial differential operators. However, to obtain these results, assumptions are posed on the norm of certain concatenations of the discrete spatial operators, which are, to the author's understanding, far from being trivial to verify in applications.

We circumvent such assumptions by using and extending techniques from [Sturm, 2017] used to analyze a dG discretization combined with the Crank–Nicolson scheme applied to Maxwell's equations. In particular, the key step for the analysis of the dG-Peaceman–Rachford scheme is to use a result similar to Lemma 3.43. This allows us to use the regularity of the exact solution and approximation properties of the discrete operators to bound additional defects emerging from the spatial discretization.

The chapter is organized similarly to Chapter 4. That is, in Section 5.1 we begin by applying the Crank–Nicolson scheme to the spatially semidiscrete problem (3.13) and subsequently investigate wellposedness and stability of the resulting scheme. We proceed in the same way in Section 5.2 for the Peaceman–Rachford scheme. In Section 5.3 we then investigate the error made by approximating the exact solution of the wave-type problem (2.24) using these schemes. In particular, we show that both schemes converge to the exact solution with order k in space and 2 in time, given the exact solution fulfills appropriate regularity assumptions.

5.1 The dG-Crank–Nicolson scheme

Applying the Crank–Nicolson scheme (4.2) to the semidiscrete problem (3.13) yields the **fully discrete dG-Crank–Nicolson (dG-CN) scheme**

$$\begin{cases} \left(\boldsymbol{\mathcal{I}} - \frac{\tau}{2}\boldsymbol{\mathcal{L}}\right)\boldsymbol{u}_\tau^{n+1} = \left(\boldsymbol{\mathcal{I}} + \frac{\tau}{2}\boldsymbol{\mathcal{L}}\right)\boldsymbol{u}_\tau^{n} + \frac{\tau}{2}(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^{n}), \qquad n \in \mathbb{N}_0, & \text{(5.1a)} \\ \boldsymbol{u}_\tau^0 = \boldsymbol{u}_\pi^0. & \text{(5.1b)} \end{cases}$$

As in the semidiscrete case, we start by investigating wellposedness.

5.1.1 Wellposedness

We have already seen in Section 3.6.2 that $\boldsymbol{\mathcal{L}}$ is maximal dissipative on $V_\hbar$. Consequently, $\left(\boldsymbol{\mathcal{I}} - \frac{\tau}{2}\boldsymbol{\mathcal{L}}\right)\colon V_\hbar \to V_\hbar$ is an isomorphism. Similar to the semidiscrete case, we can therefore define the operators $\boldsymbol{\mathcal{R}}_{\text{CN}}\colon V_\hbar \to V_\hbar$ and $\boldsymbol{\mathcal{S}}_{\text{CN}}\colon V_\hbar \to V_\hbar$ as

$$\boldsymbol{\mathcal{R}}_{\text{CN}} = \left(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\right)^{-1}$$

and

$$\boldsymbol{\mathcal{S}}_{\text{CN}} = \left(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\right)^{-1}\left(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\right).$$

The dG-CN scheme (5.1) is thus equivalent to

$$\begin{cases} \boldsymbol{u}_\tau^{n+1} = \boldsymbol{\mathcal{S}}_{\text{CN}}\boldsymbol{u}_\tau^{n} + \frac{\tau}{2}\boldsymbol{\mathcal{R}}_{\text{CN}}(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^{n}), \qquad n \in \mathbb{N}_0, \\ \boldsymbol{u}_\tau^0 = \boldsymbol{u}_\pi^0. \end{cases} \tag{5.2}$$

Using this, we can state the wellposedness of the scheme and the fully discrete version of the variation-of-constants formula.

Theorem 5.1. *For all $n \in \mathbb{N}_0$, all $\hbar \in \mathcal{H}$ and all $\tau > 0$ there exists a unique $\boldsymbol{u}_\tau^{n+1} \in V_\hbar$ fulfilling the dG-CN scheme (5.1) given by the discrete variation-of-constants formula*

$$\boldsymbol{u}_\tau^{n+1} = \boldsymbol{\mathcal{S}}_{\text{CN}}^{n+1}\boldsymbol{u}_\pi^0 + \frac{\tau}{2}\sum_{j=0}^{n}\boldsymbol{\mathcal{S}}_{\text{CN}}^{n-j}\boldsymbol{\mathcal{R}}_{\text{CN}}(\boldsymbol{f}_\pi^{j+1} + \boldsymbol{f}_\pi^{j}). \tag{5.3}$$

Proof. Existence and uniqueness of $\boldsymbol{u}_\tau^{n+1}$ for all $n \in \mathbb{N}_0$ can be seen by (5.2) and the definition of the operators occurring therein. The variation-of-constants can again be shown by a straightforward induction argument. □

5.1.2 Stability

To show stability of the dG-CN scheme, we proceed as in the semidiscrete case. We start by showing contractivity of $\boldsymbol{\mathcal{S}}_{\text{CN}}$.

Lemma 5.2. *Let $\hbar \in \mathcal{H}$ and $\tau > 0$. Then, for all $\boldsymbol{v} \in V_\hbar$ we have*

$$\|\boldsymbol{\mathcal{S}}_{\text{CN}}\boldsymbol{v}\|_M \leq \|\boldsymbol{v}\|_M,$$

i.e., the operator $\boldsymbol{\mathcal{S}}_{\text{CN}}$ is a contraction.

Proof. This is a direct consequence of the fact that $\boldsymbol{\mathcal{S}}_{\text{CN}}$ is the transform encountered in Lemma 2.10 (ii) corresponding to the maximal dissipative operator $\boldsymbol{\mathcal{L}}|_{V_\hbar}$. □

Similar to the semidiscrete case, the contractivity of $\boldsymbol{\mathcal{S}}_{\text{CN}}$ yields that arbitrary powers of $\boldsymbol{\mathcal{S}}_{\text{CN}}$ applied to elements of $V_\hbar$ can be bounded. We state this in the next corollary.

Corollary 5.3. *Let $\hbar \in \mathcal{H}$ and $\tau > 0$. Then, for all $q \in \mathbb{N}$ and all $\boldsymbol{v} \in V_\hbar$ we have*

$$\|\boldsymbol{\mathcal{S}}_{\text{CN}}^q \boldsymbol{v}\|_M \leq \|\boldsymbol{v}\|_M$$

and

$$\|\boldsymbol{\mathcal{S}}_{\text{CN}}^q \boldsymbol{\mathcal{R}}_{\text{CN}} \boldsymbol{v}\|_M \leq \|\boldsymbol{v}\|_M.$$

Proof. The first inequality is a direct consequence of applying Lemma 5.2 q times. The second one additionally uses Lemma 2.10 (i) together with the fact that $\boldsymbol{\mathcal{R}}_{\text{CN}}$ is the resolvent of the maximal dissipative operator $\boldsymbol{\mathcal{L}}|_{V_\hbar}$. □

Using this, we obtain the **stability result for the dG-Crank–Nicolson scheme**.

Corollary 5.4. *Let $\hbar \in \mathcal{H}$, $\tau > 0$ and $u^0 \in D(\mathcal{L})$. Then, for all $n \in \mathbb{N}_0$, the approximation $\boldsymbol{u}_\tau^{n+1}$ given by the dG-Crank–Nicolson scheme (5.1) satisfies*

$$\|\boldsymbol{u}_\tau^{n+1}\|_M \leq \|u^0\|_M + \tfrac{\tau}{2} \sum_{j=1}^{n} \|f^{j+1} + f^j\|_M.$$

Proof. The claim can be proven analogously to the semidiscrete case, i.e., Corollary 4.6. The only difference is that we use the discrete variation-of-constants given in (5.3) instead of (4.4), Corollary 5.3 instead of Corollary 4.5, and that we additionally use the boundedness of the L^2-projection (3.12). □

5.2 The dG-Peaceman–Rachford scheme

Before we can apply the Peaceman–Rachford scheme to the semidiscrete problem (3.13), we need to define the discrete versions of the split operators from Section 2.4. Thus, let $\widetilde{\boldsymbol{\mathcal{A}}}\colon V_\hbar^{\widetilde{\mathcal{A}}} \to V_\hbar$ and $\widetilde{\boldsymbol{\mathcal{B}}}\colon V_\hbar^{\widetilde{\mathcal{B}}} \to V_\hbar$ be the central flux discretization of the dissipative Friedrichs' operators $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$, respectively. Further, let $V_\hbar^{\mathcal{A}} = V_\hbar^{\widetilde{\mathcal{A}}}$ and $V_\hbar^{\mathcal{B}} = V_\hbar^{\widetilde{\mathcal{B}}}$ and define the discrete operators $\boldsymbol{\mathcal{A}}\colon V_\hbar^{\mathcal{A}} \to V_\hbar$ and $\boldsymbol{\mathcal{B}}\colon V_\hbar^{\mathcal{B}} \to V_\hbar$ as

$$\boldsymbol{\mathcal{A}} = M^{-1}\widetilde{\boldsymbol{\mathcal{A}}} \qquad \text{and} \qquad \boldsymbol{\mathcal{B}} = M^{-1}\widetilde{\boldsymbol{\mathcal{B}}}.$$

Note that by the same argumentation as in Section 3.6.1 for $\boldsymbol{\mathcal{L}}$, the operators $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ are consistent in the sense of Proposition 3.34.

Applying the Peaceman–Rachford scheme (4.5) to the semidiscrete problem (3.13) yields the **fully discrete dG-Peaceman–Rachford (dG-PR) scheme**

$$\begin{cases} \left(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\right)\boldsymbol{u}_\tau^{n+1/2} = \left(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\right)\boldsymbol{u}_\tau^{n}, & \text{(5.4a)} \\ \left(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\right)\boldsymbol{u}_\tau^{n+1} = \left(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\right)\left(\boldsymbol{u}_\tau^{n+1/2} + \tfrac{\tau}{2}(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^{n})\right), \qquad n \in \mathbb{N}_0, & \text{(5.4b)} \\ \boldsymbol{u}_\tau^0 = \boldsymbol{u}_\pi^0. & \text{(5.4c)} \end{cases}$$

Again, we begin by showing wellposedness of the scheme.

5.2.1 Wellposedness

By transferring the argumentation from Section 3.6.2 to the split operators $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$, we see that both operators are maximal dissipative on $V_\hbar$. Again, this yields that $\big(\boldsymbol{\mathcal{I}}-\frac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\colon V_\hbar \to V_\hbar$ and $\big(\boldsymbol{\mathcal{I}}-\frac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\colon V_\hbar \to V_\hbar$ are isomorphisms. We can thus rewrite (5.4a) and (5.4b) for all $n \in \mathbb{N}_0$ equivalently as

$$\boldsymbol{u}_\tau^{n+1/2} = \big(\boldsymbol{\mathcal{I}}-\tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)^{-1}\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\boldsymbol{u}_\tau^n, \tag{5.5a}$$

$$\boldsymbol{u}_\tau^{n+1} = \big(\boldsymbol{\mathcal{I}}-\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)^{-1}\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{u}_\tau^{n+1/2}+\tfrac{\tau}{2}(\boldsymbol{f}_\pi^{n+1}+\boldsymbol{f}_\pi^n)\big). \tag{5.5b}$$

We proceed analogously to the semidiscrete case in Section 4.2.1. Namely, we rewrite the scheme in the same form as the dG-CN scheme (5.2). For this, we define the operators $\boldsymbol{\mathcal{R}}_{\mathrm{PR}}\colon V_\hbar \to V_\hbar$ and $\boldsymbol{\mathcal{S}}_{\mathrm{PR}}\colon V_\hbar \to V_\hbar$ as

$$\boldsymbol{\mathcal{R}}_{\mathrm{PR}} = \big(\boldsymbol{\mathcal{I}}-\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)^{-1}\big(\boldsymbol{\mathcal{I}}-\tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)^{-1}$$

and

$$\boldsymbol{\mathcal{S}}_{\mathrm{PR}} = \big(\boldsymbol{\mathcal{I}}-\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)^{-1}\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{\mathcal{I}}-\tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)^{-1}\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big).$$

Inserting (5.5a) into (5.5b) and making use of $\boldsymbol{\mathcal{S}}_{\mathrm{PR}}$, we can rewrite the scheme (5.4) equivalently as

$$\begin{cases} \boldsymbol{u}_\tau^{n+1} = \boldsymbol{\mathcal{S}}_{\mathrm{PR}}\boldsymbol{u}_\tau^n + \frac{\tau}{2}\big(\boldsymbol{\mathcal{I}}-\frac{\tau}{2}\boldsymbol{\mathcal{B}}\big)^{-1}\big(\boldsymbol{\mathcal{I}}+\frac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{f}_\pi^{n+1}+\boldsymbol{f}_\pi^n\big), \qquad n \in \mathbb{N}_0, & \text{(5.6a)} \\ \boldsymbol{u}_\tau^0 = \boldsymbol{u}_\pi^0. & \text{(5.6b)} \end{cases}$$

From this, we can derive the wellposedness of the dG-PR scheme.

Theorem 5.5. *For all $n \in \mathbb{N}_0$, all $\hbar \in \mathcal{H}$ and all $\tau > 0$, there exists a unique $\boldsymbol{u}_\tau^{n+1} \in V_\hbar$ fulfilling the dG-PR scheme (5.4) given by the discrete variation-of-constants formula*

$$\boldsymbol{u}_\tau^{n+1} = \boldsymbol{\mathcal{S}}_{\mathrm{PR}}^{n+1}\boldsymbol{u}_\pi^0 + \frac{\tau}{2}\sum_{j=0}^{n}\boldsymbol{\mathcal{S}}_{\mathrm{PR}}^{n-j}\big(\boldsymbol{\mathcal{I}}-\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)^{-1}\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{f}_\pi^{j+1}+\boldsymbol{f}_\pi^j\big). \tag{5.7}$$

Proof. Uniqueness and existence follow by (5.6) and the definition of the operators occurring therein. The variation-of-constants formula can again be verified by a straightforward induction argument. □

5.2.2 Stability

Stability is handled as in the semidiscrete case. We begin by showing analogous bounds to the ones in Lemma 4.11.

Lemma 5.6. *Let $\hbar \in \mathcal{H}$ and $\tau > 0$. Then, for all $q \in \mathbb{N}$ and all $\boldsymbol{v} \in V_\hbar$ we have*

$$\|\boldsymbol{\mathcal{S}}_{\mathrm{PR}}^q\boldsymbol{v}\|_M \le \|\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\boldsymbol{v}\|_M.$$

Further, we have

$$\|\boldsymbol{\mathcal{S}}_{\mathrm{PR}}^q\big(\boldsymbol{\mathcal{I}}-\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)^{-1}\boldsymbol{v}\|_M \le \|\boldsymbol{v}\|_M$$

and

$$\|\boldsymbol{\mathcal{S}}_{\mathrm{PR}}^q\boldsymbol{\mathcal{R}}_{\mathrm{PR}}\boldsymbol{v}\|_M \le \|\boldsymbol{v}\|_M.$$

Proof. The proof of these results is completely analogous to their semidiscrete counterparts in Lemma 4.11. This can be seen as the proofs only rely on the fact that the operators $\mathcal{A}$ and $\mathcal{B}$ are maximally dissipative. As stated in the beginning of Section 5.2.1, the discrete versions $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ inherit this property on $V_\hbar$. □

Using this, we obtain the **stability result for the dG-Peaceman–Rachford scheme**. Recall that the semidiscrete version of this result, i.e., Corollary 4.12, involves the norms of $\big(\mathcal{I}+\frac{\tau}{2}\mathcal{B}\big)v$ and $\big(\mathcal{I}+\frac{\tau}{2}\mathcal{A}\big)\big(f^{j+1}+f^j\big)$. As a consequence, we have to deal with additional terms stemming from the discretization of $\mathcal{A}$ and $\mathcal{B}$.

Corollary 5.7. *Let $\hbar \in \mathcal{H}$, $\tau > 0$, $u^0 \in D(\mathcal{B}) \cap H^1(\mathcal{T}_\hbar)$ and $f \in C(\mathbb{R}_+; D(\mathcal{A}) \cap H^1(\mathcal{T}_\hbar))$. Then, for all $n \in \mathbb{N}_0$, the approximation $\boldsymbol{u}_\tau^{n+1}$ given by the dG-PR scheme (5.4) satisfies*

$$\begin{aligned}\|\boldsymbol{u}_\tau^{n+1}\|_M \le{}& \|\big(\mathcal{I}+\tfrac{\tau}{2}\mathcal{B}\big)u^0\|_M + \tfrac{\tau}{2}C_{\pi,\widetilde{\mathcal{B}},M}|u^0|_{1,\mathcal{T}_\hbar} \\ &+ \tfrac{\tau}{2}\sum_{j=1}^{n}\Big(\|\big(\mathcal{I}+\tfrac{\tau}{2}\mathcal{A}\big)\big(f^{j+1}+f^j\big)\|_M + \tfrac{\tau}{2}C_{\pi,\widetilde{\mathcal{A}},M}|f^{j+1}+f^j|_{1,\mathcal{T}_\hbar}\Big)\end{aligned}$$

with $C_{\pi,\widetilde{\mathcal{A}},M} = \|M^{-1}\|_{\infty,\Omega}^{1/2}\, C_{\pi,\widetilde{\mathcal{A}}}$ and $C_{\pi,\widetilde{\mathcal{B}},M} = \|M^{-1}\|_{\infty,\Omega}^{1/2}\, C_{\pi,\widetilde{\mathcal{B}}}$.

Proof. We proceed as in the semidiscrete case and use the discrete variation-of-constants formula (5.7) and Lemma 5.6 to obtain

$$\|\boldsymbol{u}_\tau^{n+1}\|_M \le \|\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\boldsymbol{u}_\pi^0\|_M + \tfrac{\tau}{2}\sum_{j=1}^{n}\|\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\boldsymbol{f}_\pi^{j+1}+\boldsymbol{f}_\pi^j\|_M.$$

To bound the first term, we use the consistency of $\boldsymbol{\mathcal{B}}$ in the sense of Proposition 3.34 to obtain

$$\begin{aligned}\|\big(\boldsymbol{\mathcal{I}}+\tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\pi_h u^0\|_M &= \|\pi_h u^0 + \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}u^0 + \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}(\pi_h u^0 - u^0)\|_M \\ &\le \|\pi_h u^0 + \tfrac{\tau}{2}\pi_h\mathcal{B}u^0\|_M + \tfrac{\tau}{2}\|\boldsymbol{\mathcal{B}}e_\pi^{u^0}\|_M \\ &\le \|\pi_h\big(\mathcal{I}+\tfrac{\tau}{2}\mathcal{B}\big)u^0\|_M + \tfrac{\tau}{2}C_{\pi,\widetilde{\mathcal{B}},M}|u^0|_{1,\mathcal{T}_\hbar},\end{aligned}$$

where we used an analogous argument to the one in (3.18) to bound $\|\boldsymbol{\mathcal{B}}e_\pi^{u^0}\|_M$. Treating the second term in the same way concludes the proof. □

5.3 Error analysis of the full discretization

As we have established wellposedness and stability of both fully discrete schemes, we now conduct the error analysis. This is done by combining the ideas used in both the spatially as well as the temporally semidiscrete analysis. In principle, we follow the approach used in Section 4.3. However, additional difficulties stemming from the spatial discretization have to be dealt with.

Let u be the solution of the continuous problem (2.24) and $\boldsymbol{u}_\tau^n$ be the solution of either the dG-CN scheme (5.1) or the dG-PR scheme (5.4) after n steps. To conduct the error analysis, for all $n \in \mathbb{N}$, we denote the **fully discrete error after n steps** by

$$e^n = u(t_n) - \boldsymbol{u}_\tau^n.$$

As in the analysis of the spatially semidiscrete problem, we perform an error splitting. We use the **projection error** $e_\pi^n = u(t_n) - \pi_h u(t_n)$ at time t_n and introduce the **full discretization error** $\boldsymbol{e}^n = \pi_h u(t_n) - \boldsymbol{u}_\tau^n$ to obtain

$$e^n = e_\pi^n + \boldsymbol{e}^n.$$

Note that the full discretization error $\boldsymbol{e}^n$ includes the error made by both the space and the time discretization. Further, recall that the projection error e_π^n satisfies the bound (3.14). Hence, it only remains to bound $\boldsymbol{e}^n$.

5.3.1 Error recursion

Similar to the semidiscrete cases, we begin by deriving an error recursion. As the full discretization error $\boldsymbol{e}^n$ is obtained by measuring the fully discrete approximation against the projected exact solution, we proceed as in Section 3.7.1. That is to say, we insert the projected exact solution into the fully discrete schemes.

Crank–Nicolson method

We start by investigating the defect caused by inserting the projected exact solution into the dG-CN scheme (5.1).

Lemma 5.8. *Let $h \in \mathcal{H}$ and $\tau > 0$. Further, assume that we have $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}) \cap H^1(\mathcal{T}_h)^m)$. Then, for all $n \in \mathbb{N}_0$, the projected exact solution $\pi_h u$ satisfies*

$$\big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\big)\pi_h u(t_{n+1}) = \big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\big)\pi_h u(t_n) + \tfrac{\tau}{2}(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^n) + \boldsymbol{d}_{\mathrm{CN}}^n, \tag{5.8}$$

*where the **fully discrete Crank–Nicolson defect** $\boldsymbol{d}_{\mathrm{CN}}^n$ is given by*

$$\boldsymbol{d}_{\mathrm{CN}}^n = \pi_h \delta_{\mathrm{CN}}^n + \tfrac{\tau}{2}\big(\boldsymbol{d}_\pi(t_{n+1}) + \boldsymbol{d}_\pi(t_n)\big). \tag{5.9}$$

Proof. Equation (5.8) is obtained by inserting the projected exact solution into the dG-CN iteration (5.1a) and defining the defect caused by this as $\boldsymbol{d}_{\mathrm{CN}}^n$. We solve (5.8) for $\boldsymbol{d}_{\mathrm{CN}}^n$, yielding

$$\begin{aligned} \boldsymbol{d}_{\mathrm{CN}}^n &= \pi_h u(t_{n+1}) - \pi_h u(t_n) - \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\big(\pi_h u(t_{n+1}) + \pi_h u(t_n)\big) - \tfrac{\tau}{2}(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^n) \\ &= \pi_h\big(u(t_{n+1}) - u(t_n)\big) - \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\pi_h\big(u(t_{n+1}) + u(t_n)\big) - \tfrac{\tau}{2}\pi_h(f^{n+1} + f^n). \end{aligned} \tag{5.10}$$

To obtain the projected Crank–Nicolson defect $\pi_h \delta_{\mathrm{CN}}^n$, we rewrite the second term. By the definition of the projection error and Proposition 3.34, i.e., the consistency of $\boldsymbol{\mathcal{L}}$, we have

$$\begin{aligned} \boldsymbol{\mathcal{L}}\pi_h\big(u(t_{n+1}) + u(t_n)\big) &= \boldsymbol{\mathcal{L}}\big(u(t_{n+1}) + u(t_n)\big) - \boldsymbol{\mathcal{L}}\big(e_\pi^{n+1} + e_\pi^n\big) \\ &= \pi_h \mathcal{L}\big(u(t_{n+1}) + u(t_n)\big) - \boldsymbol{\mathcal{L}}\big(e_\pi^{n+1} + e_\pi^n\big). \end{aligned}$$

Inserting this into (5.10) and rearranging the terms, we get

$$\begin{aligned} \boldsymbol{d}_{\mathrm{CN}}^n = \pi_h\Big(u(t_{n+1}) - u(t_n) - \tfrac{\tau}{2}\Big(\big(\mathcal{L}u(t_n) + f^n\big) + \big(\mathcal{L}u(t_{n+1}) + f^{n+1}\big)\Big)\Big) \\ + \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\big(e_\pi^{n+1} + e_\pi^n\big). \end{aligned}$$

Taking into account the definition of $\boldsymbol{d}_\pi$ in Lemma 3.47 and the proof of Lemma 4.13 proves the claim. □

We state the error recursion in the next corollary.

Corollary 5.9. *Let $\hbar \in \mathcal{H}$ and $\tau > 0$. Further, assume that we have $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}) \cap H^1(\mathcal{T}_\hbar)^m)$. Then, for all $n \in \mathbb{N}_0$, the full discretization error $\boldsymbol{e}^{n+1} = \pi_h u(t_{n+1}) - \boldsymbol{u}_\tau^{n+1}$ of the dG-CN scheme satisfies*

$$\begin{cases} \big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\big)\boldsymbol{e}^{n+1} = \big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{L}}\big)\boldsymbol{e}^n + \boldsymbol{d}_{\mathrm{CN}}^n, & n \in \mathbb{N}_0, \\ \boldsymbol{e}^0 = 0. \end{cases} \tag{5.11}$$

Proof. We have $\boldsymbol{e}^0 = 0$, since we use the projected initial value for the dG-CN scheme. The remaining assertion follows by subtracting (5.1a) from (5.8). □

Peaceman–Rachford method

We proceed in the same way for the dG-PR scheme. However, we have to deal with a new defect stemming from the perturbation defect δ_{PR}^n.

In the next lemma we derive the full defect, comprising the fully discrete Crank–Nicolson defect and the semidiscrete as well as a discrete perturbation defect. As before, we combine the regularity assumptions of the semidiscrete cases.

Lemma 5.10. *Let $\hbar \in \mathcal{H}$ and $\tau > 0$. Further, assume that we have $u \in C^1(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{A}) \cap D(\mathcal{AB}) \cap H^1(\mathcal{T}_\hbar)^m)$ and $f \in C(\mathbb{R}_+; D(\mathcal{A}^2) \cap H^1(\mathcal{T}_\hbar)^m)$. Then, for all $n \in \mathbb{N}_0$, the projected exact solution $\pi_h u$ satisfies*

$$\begin{aligned} \pi_h u(t_{n+1}) = \boldsymbol{\mathcal{S}}_{\mathrm{PR}} \pi_h u(t_n) + \tfrac{\tau}{2}\big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)^{-1}\big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^n\big) \\ + \boldsymbol{\mathcal{R}}_{\mathrm{PR}}\big(\boldsymbol{d}_{\mathrm{CN}}^n + \boldsymbol{d}_{\mathrm{PR}}^n\big), \end{aligned} \tag{5.12}$$

*where the **fully discrete (Peaceman–Rachford) perturbation defect** $\boldsymbol{d}_{\mathrm{PR}}^n$ is given by*

$$\boldsymbol{d}_{\mathrm{PR}}^n = \pi_h \delta_{\mathrm{PR}}^n + \boldsymbol{d}_{\mathrm{PR},u}^n + \boldsymbol{d}_{\mathrm{PR},f}^n$$

with

$$\boldsymbol{d}_{\mathrm{PR},u}^n = \frac{\tau^2}{4}\big(\boldsymbol{\mathcal{AB}}\pi_h - \pi_h \mathcal{AB}\big)\big(u(t_{n+1}) - u(t_n)\big)$$

and

$$\boldsymbol{d}_{\mathrm{PR},f}^n = \frac{\tau^3}{8}\big(\boldsymbol{\mathcal{A}}^2\pi_h - \pi_h \mathcal{A}^2\big)\big(f^{n+1} + f^n\big).$$

Proof. Equation (5.12) is obtained by inserting the projected solution into (5.6a) and defining the resulting defect as $\boldsymbol{\mathcal{R}}_{\mathrm{PR}}(\boldsymbol{d}_{\mathrm{CN}}^n + \boldsymbol{d}_{\mathrm{PR}}^n)$. We apply $\boldsymbol{\mathcal{R}}_{\mathrm{PR}}^{-1} = \big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)$ to (5.12) and solve for $\boldsymbol{d}_{\mathrm{CN}}^n + \boldsymbol{d}_{\mathrm{PR}}^n$ to obtain

$$\begin{aligned} \boldsymbol{d}_{\mathrm{PR}}^n + \boldsymbol{d}_{\mathrm{CN}}^n = \big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\pi_h u(t_{n+1}) - \big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\pi_h u(t_n) \\ - \tfrac{\tau}{2}\big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^n\big), \end{aligned}$$

where we have used $\boldsymbol{\mathcal{R}}_{\mathrm{PR}}^{-1}\boldsymbol{\mathcal{S}}_{\mathrm{PR}} = \big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)$. Expanding all three terms on the right hand side and using $\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}}$ yields

$$\begin{aligned} \boldsymbol{d}_{\mathrm{PR}}^n + \boldsymbol{d}_{\mathrm{CN}}^n &= \big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{L}} + \tfrac{\tau^2}{4}\boldsymbol{\mathcal{AB}}\big)\pi_h u(t_{n+1}) - \big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{L}} + \tfrac{\tau^2}{4}\boldsymbol{\mathcal{AB}}\big)\pi_h u(t_n) \\ &\quad - \tfrac{\tau}{2}\big(\boldsymbol{\mathcal{I}} - \tfrac{\tau^2}{4}\boldsymbol{\mathcal{A}}^2\big)\big(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^n\big) \\ &= \boldsymbol{d}_{\mathrm{CN}}^n + \tfrac{\tau^2}{4}\boldsymbol{\mathcal{AB}}\pi_h\big(u(t_{n+1}) - u(t_n)\big) + \tfrac{\tau^3}{8}\boldsymbol{\mathcal{A}}^2\pi_h\big(f^{n+1} + f^n\big) \end{aligned}$$

with $\boldsymbol{d}_{\mathrm{CN}}^n$ defined in (5.9). For the remaining two terms, we have

$$\begin{aligned}\mathcal{AB}\pi_h\big(u(t_{n+1})-u(t_n)\big) &= \big(\mathcal{AB}\pi_h-\pi_h\mathcal{AB}\big)\big(u(t_{n+1})-u(t_n)\big)\\ &\quad+\pi_h\mathcal{AB}\big(u(t_{n+1})-u(t_n)\big)\end{aligned}$$

and

$$\mathcal{A}^2\pi_h\big(f^{n+1}+f^n\big)=\big(\mathcal{A}^2\pi_h-\pi_h\mathcal{A}^2\big)\big(f^{n+1}+f^n\big)+\pi_h\mathcal{A}^2\big(f^{n+1}+f^n\big).$$

The sum of the respective last terms is the projected Peaceman–Rachford defect δ_{PR}^n by its definition (4.13). This concludes the proof. □

This readily implies the error recursion for the dG-PR scheme.

Corollary 5.11. *Let $\hbar\in\mathcal{H}$ and $\tau>0$. Further, assume that we have $u\in C^1(\mathbb{R}_+;L^2(\Omega)^m)\cap C(\mathbb{R}_+;D(\mathcal{A})\cap D(\mathcal{AB})\cap H^1(\mathcal{T}_\hbar)^m)$ and $f\in C(\mathbb{R}_+;D(\mathcal{A}^2)\cap H^1(\mathcal{T}_\hbar)^m)$. Then, for all $n\in\mathbb{N}_0$, the full discretization error $\boldsymbol{e}^{n+1}=\pi_h u(t_{n+1})-\boldsymbol{u}_\tau^{n+1}$ of the dG-PR scheme satisfies*

$$\begin{cases}\boldsymbol{e}^{n+1}=\boldsymbol{\mathcal{S}}_{\mathrm{PR}}\boldsymbol{e}^n+\boldsymbol{\mathcal{R}}_{\mathrm{PR}}(\boldsymbol{d}_{\mathrm{CN}}^n+\boldsymbol{d}_{\mathrm{PR}}^n), & n\in\mathbb{N}_0,\\ \boldsymbol{e}^0=0.\end{cases}$$

Proof. We have $\boldsymbol{e}^0=0$, since we use the projected initial value for the dG-PR scheme. The remaining assertion follows by subtracting (5.6a) from (5.12). □

5.3.2 Bounds on the defects

We follow the same strategy of proof as in the temporally semidiscrete case, cf., Section 4.3.2. Hence, we need bounds on all occurring defects.

We have already dealt with the defects $\boldsymbol{d}_\pi^n$, δ_{CN}^n and δ_{PR}^n, as these occurred in the analysis of the semidiscrete schemes. However, we still need to bound the projection errors $\boldsymbol{d}_{\mathrm{PR},u}^n$ and $\boldsymbol{d}_{\mathrm{PR},f}^n$. We begin with the former.

Lemma 5.12. *Let $\hbar\in\mathcal{H}$, $\tau>0$ and $k\geq 1$. Further, assume that we have $u\in C^1(\mathbb{R}_+;D(\mathcal{AB})\cap H^2(\mathcal{T}_\hbar))$ and $B_0\in W^{1,\infty}(K)^{m\times m}$ for all $K\in\mathcal{T}_\hbar$. Then we have*

$$\|\boldsymbol{d}_{\mathrm{PR},u}^n\|_M\leq C_{\mathrm{PR},u}\frac{\tau^2}{4}\int_{t_n}^{t_{n+1}}\|\mathrm{d}_t u(s)\|_{2,\mathcal{T}_\hbar}\,\mathrm{d}s,$$

where $C_{\mathrm{PR},u}=\|M^{-1}\|_{\infty,\Omega}^{3/2}(C_{\mathrm{inv},\widetilde{\mathcal{A}}}C_{\pi,\widetilde{\mathcal{B}},-1}+C_{\pi,\widetilde{\mathcal{A}}}C_{1,\widetilde{\mathcal{B}}})$ with $C_{1,\widetilde{\mathcal{B}}}=\max\limits_{K\in\mathcal{T}_\hbar}C_{1,K,\widetilde{\mathcal{B}}}$.

Proof. By the fundamental theorem of calculus and since the spatial operators commute with the temporal integration we have

$$\boldsymbol{d}_{\mathrm{PR},u}^n=\frac{\tau^2}{4}\int_{t_n}^{t_{n+1}}\big(\mathcal{AB}\pi_h-\pi_h\mathcal{AB}\big)\mathrm{d}_t u(s)\,\mathrm{d}s$$

and thus

$$\|\boldsymbol{d}_{\mathrm{PR},u}^n\|_M\leq\frac{\tau^2}{4}\int_{t_n}^{t_{n+1}}\|\big(\mathcal{AB}\pi_h-\pi_h\mathcal{AB}\big)\mathrm{d}_t u(s)\|_M\,\mathrm{d}s.$$

We follow the proof of Lemma 3.43 to bound the integrand. However, we have to make some adaptations owing to the material parameters.

For what follows, recall that we have $\boldsymbol{\mathcal{A}} = M^{-1}\widetilde{\boldsymbol{\mathcal{A}}}$, $\boldsymbol{\mathcal{B}} = M^{-1}\widetilde{\boldsymbol{\mathcal{B}}}$, $\mathcal{A} = M^{-1}\widetilde{\mathcal{A}}$ and $\mathcal{B} = M^{-1}\widetilde{\mathcal{B}}$, and that $\|M^{-1}\|_{\infty,\Omega}$ is the supremum of the spectral norm of M^{-1} on Ω. Further, by Lemma 3.45, the discrete operators $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ are consistent in the sense of Proposition 3.34. We will make use of this several times in what follows.

In particular, this yields

$$\begin{aligned}\|\big(\boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{B}}\pi_h - \pi_h\mathcal{A}\mathcal{B}\big)\mathrm{d}_t u(s)\|_M &= \|\boldsymbol{\mathcal{A}}\big(\boldsymbol{\mathcal{B}}\pi_h - \mathcal{B}\big)\mathrm{d}_t u(s)\|_M \\ &= \|M^{1/2}\boldsymbol{\mathcal{A}}\big(\boldsymbol{\mathcal{B}}\pi_h - \mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega \\ &= \|M^{-1/2}\widetilde{\boldsymbol{\mathcal{A}}}\big(\boldsymbol{\mathcal{B}}\pi_h - \mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega \\ &\leq \|M^{-1}\|_{\infty,\Omega}^{1/2}\|\widetilde{\boldsymbol{\mathcal{A}}}\big(\boldsymbol{\mathcal{B}}\pi_h - \mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega.\end{aligned}$$

By adding and subtracting $\widetilde{\boldsymbol{\mathcal{A}}}\pi_h\mathcal{B}\mathrm{d}_t u(s)$ we get

$$\|\widetilde{\boldsymbol{\mathcal{A}}}\big(\boldsymbol{\mathcal{B}}\pi_h - \mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega \leq \|\widetilde{\boldsymbol{\mathcal{A}}}\big(\boldsymbol{\mathcal{B}}\pi_h - \pi_h\mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega + \|\widetilde{\boldsymbol{\mathcal{A}}}\big(\pi_h\mathcal{B} - \mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega. \tag{5.13}$$

For the first term in (5.13) we use Proposition 3.41 to obtain

$$\begin{aligned}\|\widetilde{\boldsymbol{\mathcal{A}}}\big(\boldsymbol{\mathcal{B}}\pi_h - \pi_h\mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega &\leq C_{\mathrm{inv},\widetilde{\mathcal{A}}}\|h^{-1}\big(\boldsymbol{\mathcal{B}}\pi_h - \boldsymbol{\mathcal{B}}\big)\mathrm{d}_t u(s)\|_\Omega \\ &= C_{\mathrm{inv},\widetilde{\mathcal{A}}}\|h^{-1}M^{-1}\big(\widetilde{\boldsymbol{\mathcal{B}}}\pi_h - \widetilde{\boldsymbol{\mathcal{B}}}\big)\mathrm{d}_t u(s)\|_\Omega \\ &\leq C_{\mathrm{inv},\widetilde{\mathcal{A}}}\|M^{-1}\|_{\infty,\Omega}\|h^{-1}\widetilde{\boldsymbol{\mathcal{B}}}\big(\pi_h - \mathcal{I}\big)\mathrm{d}_t u(s)\|_\Omega \\ &= C_{\mathrm{inv},\widetilde{\mathcal{A}}}\|M^{-1}\|_{\infty,\Omega}\|h^{-1}\widetilde{\boldsymbol{\mathcal{B}}}e_\pi^{\mathrm{d}_t u(s)}\|_\Omega \\ &\leq C_{\mathrm{inv},\widetilde{\mathcal{A}}}\|M^{-1}\|_{\infty,\Omega}C_{\pi,\widetilde{\mathcal{B}},-1}|\mathrm{d}_t u(s)|_{2,\mathcal{T}_h},\end{aligned}$$

where we have used Lemma 3.42 with $p = -1$ and $q = 1$ in the last step.

For the second term in (5.13) observe that

$$\|\widetilde{\boldsymbol{\mathcal{A}}}\big(\pi_h\mathcal{B} - \mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega = \|\widetilde{\boldsymbol{\mathcal{A}}}\big(\pi_h - \mathcal{I}\big)\mathcal{B}\mathrm{d}_t u(s)\|_\Omega = \|\widetilde{\boldsymbol{\mathcal{A}}}e_\pi^{\mathcal{B}\mathrm{d}_t u(s)}\|_\Omega.$$

Hence, we can use Lemma 3.42 with $p = 0$ and $q = 0$ to obtain

$$\|\widetilde{\boldsymbol{\mathcal{A}}}\big(\pi_h\mathcal{B} - \mathcal{B}\big)\mathrm{d}_t u(s)\|_\Omega \leq C_{\pi,\widetilde{\mathcal{A}}}|\mathcal{B}\mathrm{d}_t u(s)|_{1,\mathcal{T}_h}.$$

By the definition of the $H^1(\mathcal{T}_h)$-seminorm in Definition 3.25 and since M is elementwise constant by Assumption 3.44, we have

$$\begin{aligned}|\mathcal{B}\mathrm{d}_t u(s)|_{1,\mathcal{T}_h}^2 &= \sum_{K\in\mathcal{T}_h}|\mathcal{B}\mathrm{d}_t u(s)|_{1,K}^2 \\ &= \sum_{K\in\mathcal{T}_h}|M_K^{-1}\widetilde{\mathcal{B}}\mathrm{d}_t u(s)|_{1,K}^2 \\ &\leq \sum_{K\in\mathcal{T}_h}\|M_K^{-1}\|^2|\widetilde{\mathcal{B}}\mathrm{d}_t u(s)|_{1,K}^2 \\ &\leq \|M^{-1}\|_{\infty,\Omega}^2\sum_{K\in\mathcal{T}_h}|\widetilde{\mathcal{B}}\mathrm{d}_t u(s)|_{1,K}^2.\end{aligned}$$

Thus, using Lemma 2.16 yields

$$|\mathcal{B}\mathrm{d}_t u(s)|_{1,\mathcal{T}_h}^2 \leq \|M^{-1}\|_{\infty,\Omega}^2\sum_{K\in\mathcal{T}_h}C_{1,K,\widetilde{\mathcal{B}}}^2\,\|\mathrm{d}_t u(s)\|_{2,K}^2$$

$$\begin{aligned}&\leq \|M^{-1}\|_{\infty,\Omega}^2 C_{1,\widetilde{\mathcal{B}}}^2 \sum_{K\in\mathcal{T}_\hbar} \|\mathrm{d}_t u(s)\|_{2,K}^2\\ &= \|M^{-1}\|_{\infty,\Omega}^2 C_{1,\widetilde{\mathcal{B}}}^2 \|\mathrm{d}_t u(s)\|_{2,\mathcal{T}_\hbar}^2,\end{aligned}$$

concluding the proof. □

The bound on the second defect $\boldsymbol{d}_{\mathrm{PR},f}^n$ can be proven in the same way.

Lemma 5.13. *Let $\hbar \in \mathcal{H}$, $\tau > 0$ and $k \geq 1$. Further, assume that we have $f \in C(\mathbb{R}_+; D(\mathcal{A}^2) \cap H^2(\mathcal{T}_\hbar))$ and $A_0 \in W^{1,\infty}(K)^{m\times m}$ for all $K \in \mathcal{T}_\hbar$. Then we have*

$$\|\boldsymbol{d}_{\mathrm{PR},f}^n\|_M \leq C_{\mathrm{PR},f} \frac{\tau^3}{8} \|f^{n+1} + f^n\|_{2,\mathcal{T}_\hbar},$$

where $C_{\mathrm{PR},f} = \|M^{-1}\|_{\infty,\Omega}^{3/2}(C_{\mathrm{inv},\widetilde{\mathcal{A}}} C_{\pi,\widetilde{\mathcal{A}},-1} + C_{\pi,\widetilde{\mathcal{A}}} C_{1,\widetilde{\mathcal{A}}})$ with $C_{1,\widetilde{\mathcal{A}}} = \max_{K\in\mathcal{T}_\hbar} C_{1,\widetilde{\mathcal{A}}}$.

Proof. This can be proven analogously to the second part of the proof of Lemma 5.12. □

5.3.3 Fully discrete convergence results

We are now able to give the fully discrete convergence results. The proofs follow the same strategy as the proofs of the temporally semidiscrete counterparts Theorems 4.19 and 4.20. However, we also have to deal with the defects introduced by the spatial discretization.

Crank–Nicolson method

We state the **convergence result for the fully discrete dG-Crank–Nicolson discretization** of the wave-type problem (2.24) in the next theorem.

Theorem 5.14. *Let $\hbar \in \mathcal{H}$ and $\tau > 0$. Further, assume that the exact solution of the wave-type problem (2.24) satisfies $u \in C^3(\mathbb{R}_+; L^2(\Omega)^m) \cap C(\mathbb{R}_+; D(\mathcal{L}) \cap H^{k+1}(\mathcal{T}_\hbar)^m)$. Then, for all $n \in \mathbb{N}_0$, the dG-Crank–Nicolson error satisfies*

$$\begin{aligned}\|u(t_{n+1}) - \boldsymbol{u}_\tau^{n+1}\|_M &\leq C_{\mathrm{app},M} |h^{k+1} u(t_{n+1})|_{k+1,\mathcal{T}_\hbar}\\ &\quad + C_{\pi,\widetilde{\mathcal{L}},M} \tfrac{\tau}{2} \sum_{j=0}^{n} |h^k \big(u(t_{j+1}) + u(t_j)\big)|_{k+1,\mathcal{T}_\hbar}\\ &\quad + \frac{\tau^2}{8} \int_0^{t_{n+1}} \|\mathrm{d}_t^3 u(s)\|_M \,\mathrm{d}s\\ &\leq C\big(\hbar^k + \tau^2\big),\end{aligned} \tag{5.14}$$

where C only depends on t_{n+1}, $C_{\mathrm{app},M}$, $C_{\pi,\widetilde{\mathcal{L}},M}$, $|u(s)|_{k+1,\mathcal{T}_\hbar}$ and $\|\mathrm{d}_t^3 u(s)\|_M$, $s \in [0, t_{n+1}]$.

Proof. The projection error $e_\pi^{n+1} = u(t_{n+1}) - \pi_h u(t_{n+1})$ can be bounded by (3.14). This yields the first term in (5.14).

It remains to bound the discretization error $\boldsymbol{e}^{n+1} = \pi_h u(t_{n+1}) - \boldsymbol{u}_\tau^{n+1}$. To do so, we proceed analogously to the proof of the semidiscrete case, i.e., Theorem 4.19. More precisely, we use Theorem 5.1 to solve the error recursion (5.11). Because of $\boldsymbol{e}^0 = 0$ and Corollary 5.3,

this leads to

$$\begin{aligned}\|\boldsymbol{e}^{n+1}\|_M &\le \sum_{j=0}^{n} \|\boldsymbol{\mathcal{S}}_{\mathrm{CN}}^{n-j}\boldsymbol{\mathcal{R}}_{\mathrm{CN}}\boldsymbol{d}_{\mathrm{CN}}^{j}\|_M \\ &\le \sum_{j=0}^{n} \|\boldsymbol{d}_{\mathrm{CN}}^{j}\|_M \\ &\le \sum_{j=0}^{n} \|\pi_h\delta_{\mathrm{CN}}^{j}\|_M + \sum_{j=0}^{n} \tfrac{\tau}{2}\|\boldsymbol{d}_\pi(t_{j+1}) + \boldsymbol{d}_\pi(t_j)\|_M.\end{aligned} \tag{5.15}$$

The first term is bounded by using (4.19) and the boundedness of the L^2-projection (3.12). This yields the third term in (5.14).

For the second term recall that

$$\begin{aligned}\|\boldsymbol{d}_\pi(t_{j+1}) + \boldsymbol{d}_\pi(t_j)\|_M &= \|\boldsymbol{\mathcal{L}}\big(e_\pi^{j+1} + e_\pi^j\big)\|_M \\ &= \|M^{-1/2}\widetilde{\boldsymbol{\mathcal{L}}}\big(e_\pi^{j+1} + e_\pi^j\big)\|_\Omega \\ &\le \|M^{-1}\|_{\infty,\Omega}^{1/2}\|\widetilde{\boldsymbol{\mathcal{L}}}\big(e_\pi^{j+1} + e_\pi^j\big)\|_\Omega.\end{aligned}$$

Now, note that we have

$$e_\pi^{j+1} + e_\pi^j = \big(u(t_{j+1}) + u(t_j)\big) - \pi_h\big(u(t_{j+1}) + u(t_j)\big).$$

This is the projection error of $u(t_{j+1}) + u(t_j)$. Hence, we can use Lemma 3.42 to conclude the proof. □

Peaceman–Rachford method

As we can bound the projection error $\boldsymbol{d}_{\mathrm{PR}}^n$ by Lemmas 5.12 and 5.13, we also get the **convergence result for the fully discrete dG-Peaceman–Rachford discretization** of the wave-type problem (2.24). We state it in the next theorem.

Theorem 5.15. *Let $\hbar \in \mathcal{H}$, $\tau > 0$ and $k \ge 1$. Further, assume that the exact solution of the wave-type problem (2.24) satisfies*

$$u \in C^3(\mathbb{R}_+; L^2(\Omega)^m) \cap C^1(\mathbb{R}_+; D(\mathcal{A}\mathcal{B}) \cap H^2(\mathcal{T}_\hbar)) \cap C(\mathbb{R}_+; D(\mathcal{A}) \cap H^{k+1}(\mathcal{T}_\hbar)),$$

that we have $A_0, B_0 \in W^{1,\infty}(K)^{m\times m}$ for all $K \in \mathcal{T}_\hbar$ and that the inhomogeneity fulfills $f \in C(\mathbb{R}_+; D(\mathcal{A}^2) \cap H^2(\mathcal{T}_\hbar))$. Then, for all $n \in \mathbb{N}_0$, the dG-Peaceman–Rachford error satisfies

$$\begin{aligned}\|u(t_{n+1}) - \boldsymbol{u}_\tau^{n+1}\|_M &\le C_{\mathrm{app},M}|h^{k+1}u(t_{n+1})|_{k+1,\mathcal{T}_\hbar} \\ &\quad + C_{\pi,\widetilde{\mathcal{L}},M}\tfrac{\tau}{2}\sum_{j=0}^{n}|h^k\big(u(t_{j+1}) + u(t_j)\big)|_{k+1,\mathcal{T}_\hbar} \\ &\quad + \frac{\tau^2}{4}\Big(\int_0^{t_{n+1}} \tfrac{1}{2}\|\mathrm{d}_t^3 u(s)\|_M \,\mathrm{d}s \\ &\qquad + \int_0^{t_{n+1}} \|\mathcal{A}\mathcal{B}\mathrm{d}_t u(s)\|_M + C_{\mathrm{PR},u}\|\mathrm{d}_t u(s)\|_{2,\mathcal{T}_\hbar} \,\mathrm{d}s \\ &\qquad + \tfrac{\tau}{2}\sum_{j=0}^{n}\big(\|\mathcal{A}^2\big(f^{j+1} + f^j\big)\|_M + C_{\mathrm{PR},f}\|f^{j+1} + f^j\|_{2,\mathcal{T}_\hbar}\big)\Big) \\ &\le C\big(\hbar^k + \tau^2\big),\end{aligned}$$

where C only depends on t_{n+1}, $C_{\mathrm{app},M}$, $C_{\pi,\widetilde{\mathcal{L}},M}$, $C_{\mathrm{PR},u}$, $C_{\mathrm{PR},f}$, $\|\mathrm{d}_t^3 u(s)\|_M$, $\|\mathcal{A}\mathcal{B}\mathrm{d}_t u(s)\|_M$, $|u(s)|_{k+1,\mathcal{T}_\hbar}$, $\|\mathrm{d}_t u(s)\|_{2,\mathcal{T}_\hbar}$, $\|\mathcal{A}^2 f(s)\|_M$ and $\|f(s)\|_{2,\mathcal{T}_\hbar}$, $s \in [0, t_{n+1}]$.

Proof. The proof is analogous to the one of Theorem 5.14. The only difference being the additional defects $\pi_h \delta^n_{\mathrm{PR}}$ and $\boldsymbol{d}^n_{\mathrm{PR}}$ emerging in (5.15). The first can be treated by the boundedness of the L^2-projection (3.12) and Corollary 4.18, the latter by Lemmas 5.12 and 5.13. □

6 Efficient implementation of a dG-Peaceman–Rachford ADI method

This chapter is devoted to the implementation of the dG-Peaceman–Rachford method. It is an extension of [Hochbruck and Köhler, 2019]. In particular, we identify a class of wave-type problems, for which the dG-PR scheme is of linear complexity in each step w.r.t. the number of elements in the utilized mesh (and thus of the same complexity as an explicit scheme).

This class of problems is characterized by a special structure of the operator $\mathscr{L}$. This structure ensures that we can split the operator, resulting in two distinct subproblems, for which the flows of the spatial derivatives completely decouple. We call the operators resulting from such a splitting Friedrichs' operators having decoupled partial derivatives, cf., Definition 6.4 below.

In fact, the combination of such a splitting with the Peaceman–Rachford method can be seen as a generalization of the classical ADI methods. This will become apparent in Section 6.5, where we revisit the examples given in Section 2.5. We will see that the splitting for the acoustic wave equation turns out to be similar to the splitting in the original ADI method proposed in [Peaceman and Rachford, 1955]. Further, the splitting for Maxwell's equations we consider is the one proposed in [Namiki, 1999, Zhen et al., 2000] (or rather the extension of this scheme proposed in [Eilinghoff and Schnaubelt, 2018], as we include damping and currents).

We begin this chapter by briefly looking at the implementation of the general dG-PR scheme (5.4) in Section 6.1. In Section 6.2 we introduce the aforementioned class of Friedrichs' operators and consequently investigate their structure in 6.3. These results lead to a class of wave-type problems for which one step of the dG-PR scheme can be performed in linear complexity w.r.t. the total number of elements in the chosen mesh. We show this in Section 6.4 and give instructions on the specific implementation. Lastly, as stated before, we revisit the examples from Section 2.5 in Section 6.5.

The aforementioned favorable runtime behavior of the method for suitable problems comes with one main drawback. Namely, it can only be achieved if Ω is a union of tensorial domains, e.g., rectangles or cuboids. Whence the following assumption.

Assumption 6.1. *We assume that the domain Ω is a union of a finite number of tensorial sets. Further, we assume that $\mathcal{T}_h$ consists of tensorial elements without hanging nodes.*

6.1 Implementation of the dG-Peaceman–Rachford scheme

Recall that the dG-Peaceman–Rachford scheme (5.4) reads

$$\begin{cases} \big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\boldsymbol{u}_\tau^{n+1/2} = \big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\boldsymbol{u}_\tau^{n}, & \text{(6.1a)} \\ \big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\boldsymbol{u}_\tau^{n+1} = \big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{u}_\tau^{n+1/2} + \tfrac{\tau}{2}(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^{n})\big), \qquad n \in \mathbb{N}_0, & \text{(6.1b)} \\ \boldsymbol{u}_\tau^0 = \boldsymbol{u}_\pi^0 & \text{(6.1c)} \end{cases}$$

and is posed on the finite-dimensional Hilbert space $\big(V_\hbar, \big(\,\cdot \,\big|\, \cdot\,\big)_M\big)$. Hence, to implement the scheme, we begin by constructing a basis of the approximation space $V_\hbar$.

6.1.1 Construction of a basis of $V_\hbar$

Recall that, by (3.2), the approximation space $V_\hbar$ is the space of $\mathbb{R}^m$-valued functions, whose components lie in the broken polynomial space $\mathbb{Q}_d^k(\mathcal{T}_\hbar)$ defined in (3.1). Functions in the latter space are polynomials that are defined on each element $K \in \mathcal{T}_\hbar$ independently without any coupling between the elements. This allows us to choose a basis $\mathcal{Q}_\hbar$ of $\mathbb{Q}_d^k(\mathcal{T}_\hbar)$ of the form

$$\mathcal{Q}_\hbar = \bigcup_{K\in\mathcal{T}_\hbar} \{\, \boldsymbol{\phi}_1^K, \dots, \boldsymbol{\phi}_{N_k}^K \,\}, \qquad N_k = (k+1)^d,$$

where, for all $K \in \mathcal{T}_\hbar$, the set $\{\boldsymbol{\phi}_1^K|_K, \dots, \boldsymbol{\phi}_{N_k}^K|_K\}$ is a basis of $\mathbb{Q}_d^k(K)$ and

$$\operatorname{supp}(\boldsymbol{\phi}_i^K) \subset \overline{K} \qquad \text{for all } i = 1, \dots, N_k. \tag{6.2}$$

The exact nature of the basis of $\mathbb{Q}_d^k(K)$ is not important for what follows, but this can, e.g., be a standard nodal or modal basis.

This leads to a basis $\mathcal{V}_\hbar$ of $V_\hbar$ in a straightforward manner. Namely, we consider the m-fold Cartesian product of $\mathcal{Q}_\hbar$, i.e.,

$$\mathcal{V}_\hbar = \mathcal{Q}_\hbar^m$$

as the basis of $V_\hbar$. This implies that the number of elements in $\mathcal{V}_\hbar$, and thus the dimension of $V_\hbar$, is given by

$$N_\hbar = m\, N_k\, |\mathcal{T}_\hbar|.$$

Hence, we can enumerate the basis functions in $\mathcal{V}_\hbar$ to obtain

$$\mathcal{V}_\hbar = \{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{N_\hbar}\}.$$

6.1.2 Representation of the scheme in $V_\hbar$

Using the basis $\mathcal{V}_\hbar$, for all $n \in \mathbb{N}_0$, equations (6.1a) and (6.1b) are equivalent to

$$\begin{aligned} \big(\big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\boldsymbol{u}_\tau^{n+1/2} \,\big|\, \boldsymbol{\psi}_\ell\big)_M &= \big(\big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\boldsymbol{u}_\tau^{n} \,\big|\, \boldsymbol{\psi}_\ell\big)_M, \\ \big(\big(\boldsymbol{\mathcal{I}} - \tfrac{\tau}{2}\boldsymbol{\mathcal{B}}\big)\boldsymbol{u}_\tau^{n+1} \,\big|\, \boldsymbol{\psi}_\ell\big)_M &= \big(\big(\boldsymbol{\mathcal{I}} + \tfrac{\tau}{2}\boldsymbol{\mathcal{A}}\big)\big(\boldsymbol{u}_\tau^{n+1/2} + \tfrac{\tau}{2}(\boldsymbol{f}_\pi^{n+1} + \boldsymbol{f}_\pi^{n})\big) \,\big|\, \boldsymbol{\psi}_\ell\big)_M \end{aligned} \tag{6.3}$$

for all $\ell \in \{1, \dots, N_\hbar\}$. As (6.3) is an equation on an $N_\hbar$-dimensional space, we are able to rewrite it as as a linear system on $\mathbb{R}^{N_\hbar}$. This is done by representing $\boldsymbol{u}_\tau^n, \boldsymbol{u}_\tau^{n+1/2}, \boldsymbol{f}_\pi^n \in V_\hbar$ in the basis $\mathcal{V}_\hbar$ as

$$\boldsymbol{u}_\tau^n = \sum_{i=1}^{N_\hbar} \mathrm{u}_i^n \boldsymbol{\psi}_i, \qquad \boldsymbol{u}_\tau^{n+1/2} = \sum_{i=1}^{N_\hbar} \mathrm{u}_i^{n+1/2} \boldsymbol{\psi}_i, \qquad \boldsymbol{f}_\pi^n = \sum_{i=1}^{N_\hbar} \mathrm{f}_i^n \boldsymbol{\psi}_i,$$

and defining the corresponding **coefficient vectors** $\mathrm{u}^n, \mathrm{u}^{n+1/2}, \mathrm{f}^n \in \mathbb{R}^{N_\hbar}$ by

$$\mathrm{u}^n = \begin{pmatrix} \mathrm{u}_1^n \\ \vdots \\ \mathrm{u}_{N_\hbar}^n \end{pmatrix}, \qquad \mathrm{u}^{n+1/2} = \begin{pmatrix} \mathrm{u}_1^{n+1/2} \\ \vdots \\ \mathrm{u}_{N_\hbar}^{n+1/2} \end{pmatrix}, \qquad \mathrm{f}^n = \begin{pmatrix} \mathrm{f}_1^n \\ \vdots \\ \mathrm{f}_{N_\hbar}^n \end{pmatrix}.$$

Further, we define the **mass matrix** $\mathbb{M} \in \mathbb{R}^{N_\hbar \times N_\hbar}$ by

$$\mathbb{M} = \Big(\big(\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i \big)_M \Big)_{i,j=1}^{N_\hbar} \tag{6.4}$$

and the **stiffness matrices** $\mathbb{A}, \mathbb{B} \in \mathbb{R}^{N_\hbar \times N_\hbar}$ corresponding to $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$, respectively, by

$$\mathbb{A} = \Big(\big(\boldsymbol{\mathcal{A}} \boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i \big)_M \Big)_{i,j=1}^{N_\hbar} \quad \text{and} \quad \mathbb{B} = \Big(\big(\boldsymbol{\mathcal{B}} \boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i \big)_M \Big)_{i,j=1}^{N_\hbar}. \tag{6.5}$$

Consequently, we can rewrite (6.3) equivalently as the linear system

$$\big(\mathbb{M} - \tfrac{\tau}{2} \mathbb{A} \big) \mathrm{u}^{n+1/2} = \big(\mathbb{M} + \tfrac{\tau}{2} \mathbb{B} \big) \mathrm{u}^n, \tag{6.6a}$$

$$\big(\mathbb{M} - \tfrac{\tau}{2} \mathbb{B} \big) \mathrm{u}^{n+1} = \big(\mathbb{M} + \tfrac{\tau}{2} \mathbb{A} \big) \big(\mathrm{u}^{n+1/2} + (\mathrm{f}^{n+1} + \mathrm{f}^n) \big) \tag{6.6b}$$

on $\mathbb{R}^{N_\hbar}$.

In this chapter we are interested in the computational complexity of the scheme. In order to solve (6.6) for u^{n+1}, we have to perform matrix-vector multiplications, vector-vector and matrix-matrix additions and solve two linear systems, one in each half-step. As it is well-known for the dG method, the mass and stiffness matrices are sparse. Thus, we can perform the matrix-vector multiplications and the matrix-matrix additions in linear time w.r.t. the total number of degrees of freedom $N_\hbar$. The same holds for vector-vector additions. However, solving the linear systems in (6.6) is in general more costly if we have $d > 1$.

6.2 Friedrichs' operators with decoupled partial derivatives

In this section we introduce Friedrichs' operators having decoupled partial derivatives. Such operators lead to stiffness matrices for which the linear systems occurring in (6.6) can be solved in linear complexity w.r.t. the total number of elements in $\mathcal{T}_\hbar$, since the flows associated with such operators effectively decouple into one-dimensional flows. We start by introducing the following concept of decoupled block diagonal matrices.

Definition 6.2. *Let $M_1, \dots, M_d \in \mathbb{R}^{m \times m}$ be symmetric matrices and denote by*

$$\mathcal{I}_i = \{ j \in \{1, \dots, m\} \mid M_i e_j \neq 0 \}$$

*the set of indices of non-zero columns (or rows) in M_i, $i = 1, \dots, d$. Then we call $M_1, \dots, M_d \in \mathbb{R}^{m \times m}$ **decoupled block-diagonal** if*

$$\mathcal{I}_i \cap \mathcal{I}_j = \emptyset \qquad \text{for all } i \neq j.$$

By Definition 6.2, symmetric and decoupled block-diagonal matrices have pairwise disjoint non-zero rows and columns. The name is motivated by the following property.

Theorem 6.3. *Let $M_1, \dots, M_d \in \mathbb{R}^{m\times m}$ be symmetric and decoupled block-diagonal. Then there is a permutation matrix $P \in \mathbb{R}^{m\times m}$ such that for all $i = 1, \dots, d$, the matrix $P^T M_i P$ is block-diagonal with at most one non-zero diagonal block, which vanishes in all other matrices $P^T M_j P$, $j \neq i$.*

Proof. The assertion follows from the symmetry of the matrices M_i if we reorder the rows and columns by the indices in $\mathcal{I}_1$, then $\mathcal{I}_2, \dots, \mathcal{I}_d$, and last the indices of those columns which vanish in all matrices. □

Using the notion of decoupled block-diagonal matrices, we are able to characterize Friedrichs' operators whose partial derivatives completely decouple. As we will see later, the structure of the coefficient F_0 and the boundary condition $\mathcal{F}_\Gamma$ also play a role in the representation of the discrete operator. Hence, they have to be admissible as well, leading to the additional conditions on these objects.

Definition 6.4. *Let $\mathcal{F}$ be a dissipative Friedrichs' operator with coefficients $(F_i)_{i=0}^d$ and boundary condition $\mathcal{F}_\Gamma$. We say that $\mathcal{F}$ has **decoupled partial derivatives** if the following holds.*

(i) *The matrices $\sup_{x\in\Omega} |F_1(x)|, \dots, \sup_{x\in\Omega} |F_d(x)|$ are decoupled block-diagonal. We denote by $P_\mathcal{F}$ the permutation matrix from Theorem 6.3 corresponding to these matrices.*

(ii) *The matrix $P_\mathcal{F}^T F_0 P_\mathcal{F}$ is block-diagonal a.e. on Ω with diagonal blocks of the same size as $P_\mathcal{F}^T \big(\sup_{x\in\Omega} |F_i(x)|\big) P_\mathcal{F}$ for $i = 1, \dots, d$.*

(iii) *The matrix $P_\mathcal{F}^T \mathcal{F}_\Gamma P_\mathcal{F}$ is block-diagonal a.e. on Γ with the only non-zero blocks being those occurring in $P_\mathcal{F}^T \big(\sup_{x\in\Omega} |F_i(x)|\big) P_\mathcal{F}$ for $i = 1, \dots, d$.*

In the following, given a Friedrichs' operator $\mathcal{F}$ having decoupled partial derivatives, we denote the sets of indices from Definition 6.2 corresponding to $\mathcal{F}$ by $\mathcal{I}_1^\mathcal{F}, \dots, \mathcal{I}_d^\mathcal{F}$ and the set of remaining indices in $\{1, \dots, m\}$ by $\mathcal{I}_0^\mathcal{F}$. Further, as in Definition 6.4, we denote by $P_\mathcal{F}$ the corresponding permutation matrix from Theorem 6.3.

As a consequence of Definition 6.4, the coefficients $F_1, \dots, F_d$ are decoupled block-diagonal on Ω. Moreover, the sets of indices from Definition 6.2 corresponding to these coefficients are subsets of the $\mathcal{I}_1^\mathcal{F}, \dots, \mathcal{I}_d^\mathcal{F}$. We show this in the next lemma.

Lemma 6.5. *Let $\mathcal{F}$ be a dissipative Friedrichs' operator with coefficients $(F_i)_{i=0}^d$ and boundary condition $\mathcal{F}_\Gamma$ having decoupled partial derivatives. Then the fields $F_1, \dots, F_d$ are decoupled block-diagonal on Ω. Further, for all $x \in \Omega$ and all $i = 1, \dots, d$, the set of indices from Definition 6.2 corresponding to $F_i(x)$ is a subset of $\mathcal{I}_i^\mathcal{F}$.*

Proof. For all $i = 1, \dots, d$, each entry of $\sup_{x\in\Omega} |F_i(x)|$ is the L^∞-norm of the corresponding entry of F_i. Hence, if an entry of $\sup_{x\in\Omega} |F_i(x)|$ equals 0, the corresponding element of F_i is 0 on Ω, since by Assumption 3.27, F_i is continuous. This proves the claim. □

Clearly, since the coefficients $F_1, \dots, F_d$ are continuous by Assumption 3.27, this especially holds on the faces of $\mathcal{T}_h$.

6.3 Structure of a discrete Friedrichs' operator with decoupled partial derivatives

Having defined Friedrichs' operators with decoupled partial derivatives, we investigate the structure of their central flux dG discretization in this section. Hence, let $\mathcal{F}$ be a dissipative Friedrichs' operator with coefficients $(F_i)_{i=0}^d$ and boundary condition $\mathcal{F}_\Gamma$ having decoupled partial derivatives. Further, let $\boldsymbol{\mathcal{F}}$ be its central flux dG discretization, cf., Definition 3.32.

Analogously to Section 6.1.2, the dG discretization $\boldsymbol{\mathcal{F}}$ can be represented by the stiffness matrix $\mathbb{F} \in \mathbb{R}^{N_\hbar \times N_\hbar}$ given by

$$\mathbb{F} = \Big(\big(\boldsymbol{\mathcal{F}} \boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i \big)_M \Big)_{i,j=1}^{N_\hbar}. \tag{6.7}$$

In the rest of this section we investigate the structure of $\mathbb{F}$. In fact, we show that we can bring $\mathbb{F}$ into a block-tridiagonal form with block sizes independent of the number of elements in $\mathcal{T}_\hbar$. This can be achieved by reordering the basis functions in $\mathcal{V}_\hbar$ in a suitable way.

One crucial ingredient for this to work is the tensorial structure of the mesh $\mathcal{T}_\hbar$, cf., Assumption 6.1. More precisely, this is due to the fact that the normal vectors of the interfaces are parallel to the canonical unit vectors. This leads to the fact that coupling between elements of the mesh can only occur in one direction for each spatial derivative.

6.3.1 Decomposition of $\mathcal{V}_\hbar$ and $\mathcal{F}_\hbar^{\text{int}}$

To exploit the fact that $\mathcal{F}$ has decoupled partial derivatives, we decompose the basis $\mathcal{V}_\hbar$ into subsets corresponding to the index sets $\mathcal{I}_0^{\mathcal{F}}, \mathcal{I}_1^{\mathcal{F}}, \dots, \mathcal{I}_d^{\mathcal{F}}$. More precisely, we define the sets

$$\mathcal{V}_{\hbar,i}^{\mathcal{F}} = \bigcup_{j \in \mathcal{I}_i^{\mathcal{F}}} \{\, \boldsymbol{\phi}\, e_j \in \mathcal{V}_\hbar \mid \boldsymbol{\phi} \in \mathcal{Q}_\hbar \,\}, \qquad i = 0, \dots, d \tag{6.8}$$

with $e_j \in \mathbb{R}^m$, yielding

$$\mathcal{V}_\hbar = \bigcup_{i=0}^{d} \mathcal{V}_{\hbar,i}^{\mathcal{F}}.$$

We proceed similar for the interfaces $\mathcal{F}_\hbar^{\text{int}}$ of the mesh $\mathcal{T}_\hbar$. Due to the tensorial structure of $\mathcal{T}_\hbar$, normal vectors to the faces in $\mathcal{F}_\hbar$ are $\pm e_j \in \mathbb{R}^d$ for some $j \in \{1, \dots, d\}$. In the following, for $F \in \mathcal{F}_\hbar^{\text{int}}$, we fix the face normal vectors $\mathfrak{n}^F$ defined in Definition 3.8 to be the vector pointing in positive coordinate direction. This enables us to define the sets

$$\mathcal{F}_{\hbar,i} = \{\, F \in \mathcal{F}_\hbar^{\text{int}} \mid \mathfrak{n}^F = e_i \,\} \tag{6.9}$$

and decompose the set of interfaces into

$$\mathcal{F}_\hbar^{\text{int}} = \bigcup_{i=1}^{d} \mathcal{F}_{\hbar,i}.$$

6.3.2 Ordering of the basis functions

By (6.7) the entries of $\mathbb{F}$ are determined by inserting the basis functions into the definition of $\boldsymbol{\mathcal{F}}$, i.e., (3.9). In this section we investigate the non-zero pattern of $\mathbb{F}$ and determine an ordering of the basis functions in $\mathcal{V}_\hbar$ for which $\mathbb{F}$ is block-tridiagonal.

For the rest of this section, let $\boldsymbol{\vartheta}_1 \in \mathcal{V}_{\hbar,i}^{\mathcal{F}}$ and $\boldsymbol{\vartheta}_2 \in \mathcal{V}_{\hbar,j}^{\mathcal{F}}$ for $i,j \in \{0,\dots,d\}$. Inserting $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ into (3.9) yields

$$\begin{aligned}
\left(\boldsymbol{\mathcal{F}}\boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_\Omega &= \sum_{K\in\mathcal{T}_\hbar} \left(\mathcal{F}\boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_K - \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \left(\mathcal{F}_\partial^F [\![\boldsymbol{\vartheta}_1]\!]_F \,\middle|\, \{\!\{\boldsymbol{\vartheta}_2\}\!\}_F\right)_F \\
&\quad - \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left((\mathcal{F}_\partial^F - \mathcal{F}_\Gamma)\boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_F \\
&= \sum_{K\in\mathcal{T}_\hbar} \sum_{r=1}^{d} \left(F_r \partial_r \boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_K + \sum_{K\in\mathcal{T}_\hbar} \left(F_0 \boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_K \\
&\quad - \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \sum_{r=1}^{d} \left(F_r \mathfrak{n}_r^F [\![\boldsymbol{\vartheta}_1]\!]_F \,\middle|\, \{\!\{\boldsymbol{\vartheta}_2\}\!\}_F\right)_F \\
&\quad - \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \sum_{r=1}^{d} \left(F_r \mathfrak{n}_r^F \boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_F + \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left(\mathcal{F}_\Gamma \boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_F.
\end{aligned}$$

By (6.8) there exist $\ell_1 \in \mathcal{I}_i^{\mathcal{F}}$, $\ell_2 \in \mathcal{I}_j^{\mathcal{F}}$ and $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathcal{Q}_\hbar$ such that we have $\boldsymbol{\vartheta}_1 = \boldsymbol{\phi}_1 e_{\ell_1}$ and $\boldsymbol{\vartheta}_2 = \boldsymbol{\phi}_2 e_{\ell_2}$. This yields

$$\begin{aligned}
\left(\boldsymbol{\mathcal{F}}\boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_\Omega &= \sum_{K\in\mathcal{T}_\hbar} \sum_{r=1}^{d} \left(e_{\ell_2}^T F_r e_{\ell_1} \partial_r \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_K + \sum_{K\in\mathcal{T}_\hbar} \left(e_{\ell_2}^T F_0 e_{\ell_1} \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_K \\
&\quad - \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \sum_{r=1}^{d} \left(e_{\ell_2}^T F_r e_{\ell_1} \mathfrak{n}_r^F [\![\boldsymbol{\phi}_1]\!]_F \,\middle|\, \{\!\{\boldsymbol{\phi}_2\}\!\}_F\right)_F \\
&\quad - \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \sum_{r=1}^{d} \left(e_{\ell_2}^T F_r e_{\ell_1} \mathfrak{n}_r^F \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_F + \tfrac{1}{2} \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left(e_{\ell_2}^T \mathcal{F}_\Gamma e_{\ell_1} \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_F.
\end{aligned}$$

We now use the fact that $\mathcal{F}$ has decoupled partial derivatives. In particular, by Definition 6.4 and with δ_{ij} denoting the Kronecker delta, this means that for $i,j \neq 0$ we have

$$e_{\ell_2}^T F_0 e_{\ell_1} = \delta_{ij} e_{\ell_2}^T F_0 e_{\ell_1} \qquad \text{a.e. on } \Omega$$

and

$$e_{\ell_2}^T \mathcal{F}_\Gamma e_{\ell_1} = \delta_{ij} e_{\ell_2}^T \mathcal{F}_\Gamma e_{\ell_1} \qquad \text{a.e. on } \Gamma.$$

Further, as a consequence of Lemma 6.5 we have

$$\sum_{r=1}^{d} e_{\ell_2}^T F_r e_{\ell_1} = \delta_{ij} e_{\ell_2}^T F_i e_{\ell_1} \qquad \text{on } \Omega.$$

Using this, the sums over r vanish and we obtain

$$\begin{aligned}\left(\boldsymbol{\mathcal{F}}\boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_\Omega = \delta_{ij}\Big(&\sum_{K\in\mathcal{T}_\hbar} \left(e_{\ell_2}^T F_i e_{\ell_1} \partial_i \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_K + \sum_{K\in\mathcal{T}_\hbar} \left(e_{\ell_2}^T F_0 e_{\ell_1} \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_K \\ &- \sum_{F\in\mathcal{F}_\hbar^{\mathrm{int}}} \left(e_{\ell_2}^T F_i e_{\ell_1} \mathfrak{n}_i^F [\![\boldsymbol{\phi}_1]\!]_F \,\middle|\, \{\!\{\boldsymbol{\phi}_2\}\!\}_F\right)_F \\ &- \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left(e_{\ell_2}^T F_i e_{\ell_1} \mathfrak{n}_i^F \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_F + \tfrac{1}{2}\sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left(e_{\ell_2}^T \mathcal{F}_\Gamma e_{\ell_1} \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_F\Big)\end{aligned} \tag{6.10}$$

if $i, j \neq 0$.

Further, if $i = 0$ or $j = 0$, by the same reasoning we have

$$\left(\boldsymbol{\mathcal{F}}\boldsymbol{\vartheta}_1 \,\middle|\, \boldsymbol{\vartheta}_2\right)_\Omega = \delta_{ij}\Big(\sum_{K\in\mathcal{T}_\hbar} \left(e_{\ell_2}^T F_0 e_{\ell_1} \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_K\Big), \tag{6.11}$$

as basis functions in $\mathcal{V}_{\hbar,0}^{\mathcal{F}}$ correspond to the set of indices whose columns vanish in all F_r, $r = 1, \dots, d$ and $\mathcal{F}_\Gamma$ by Definition 6.4.

Hence, two basis functions can only generate a non-zero entry if they belong to the same set $\mathcal{V}_{\hbar,i}^{\mathcal{F}}$, $i = 0, \dots, d$. This means that if we order the basis functions according to these sets, the resulting matrix is block-diagonal. However, the block sizes still depend on the number of elements in $\mathcal{T}_\hbar$, since each set $\mathcal{V}_{\hbar,i}^{\mathcal{F}}$, $i = 0, \dots, d$, contains basis functions belonging to all elements of the mesh.

To see that these blocks, and therefore the stiffness matrix $\mathbb{F}$, can be brought into a block-tridiagonal structure, we next investigate the terms occurring in (6.10) and (6.11). In the following, assume $i = j$, i.e., $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2 \in \mathcal{V}_{\hbar,i}^{\mathcal{F}}$ and thus $\ell_1, \ell_2 \in \mathcal{I}_i^{\mathcal{F}}$.

Volume terms

We start by investigating the first two terms in (6.10) and the term in (6.11). Since $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathcal{Q}_\hbar$, by (6.2) we have $\operatorname{supp}(\boldsymbol{\phi}_1) \subset \overline{K}_1$ and $\operatorname{supp}(\boldsymbol{\phi}_2) \subset \overline{K}_2$ for some $K_1, K_2 \in \mathcal{T}_\hbar$. Hence, if $K_1 \neq K_2$, we have

$$\sum_{K\in\mathcal{T}_\hbar} \left(e_{\ell_2}^T F_i e_{\ell_1} \partial_i \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_K = \sum_{K\in\mathcal{T}_\hbar} \left(e_{\ell_2}^T F_0 e_{\ell_1} \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_K = 0.$$

Therefore, the only non-zero entries generated by these terms are due to basis functions that are associated with the same mesh element. By ordering the functions in the sets $\mathcal{V}_{\hbar,i}^{\mathcal{F}}$, $i = 0, \dots, d$ elementwise, these terms only contribute to diagonal blocks. These blocks have block size $|\mathcal{I}_i^{\mathcal{F}}|N_k$, since per element we have N_k basis functions for each index in $\mathcal{I}_i^{\mathcal{F}}$, $i = 0, \dots, d$.

Boundary terms

Next, we investigate the terms in (6.10) involving boundary faces. These can be treated analogously to the volume terms, since each boundary face belongs to only one element. Hence, for $K_1 \neq K_2$ we have

$$\sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left(e_{\ell_2}^T F_i e_{\ell_1} \mathfrak{n}_i^F \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_F = \sum_{F\in\mathcal{F}_\hbar^{\mathrm{bnd}}} \left(e_{\ell_2}^T \mathcal{F}_\Gamma e_{\ell_1} \boldsymbol{\phi}_1 \,\middle|\, \boldsymbol{\phi}_2\right)_F = 0.$$

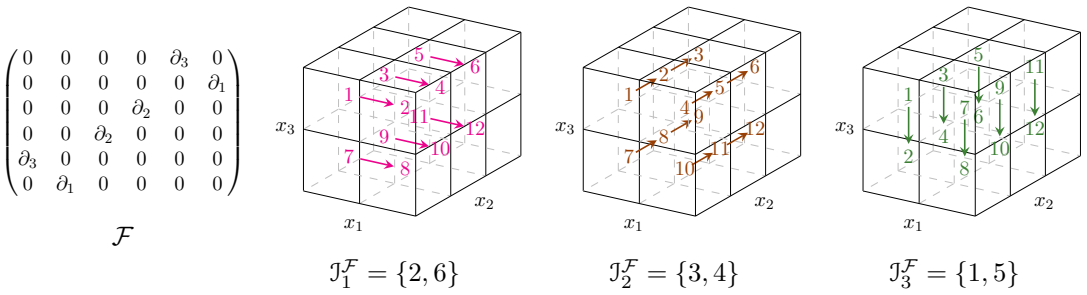


Figure 6.1: Suitable ordering of the elements for a given Friedrichs' operator $\mathcal{F}$ with decoupled partial derivatives.

With the elementwise ordering proposed in the last paragraph, these terms therefore contribute to the same blockdiagonal as the volume terms. In particular, the only non-zero entries stem from basis functions belonging to boundary elements.

Interface terms

Lastly, we investigate the third term in (6.10), which corresponds to the interfaces of the mesh. As this term involves averages and jumps across interfaces, we obtain contributions outside of the blockdiagonal.

However, by the definition of $\mathcal{F}_{\hbar,i}$ in (6.9) we have $\mathfrak{n}_i^F = 1$ and $\mathfrak{n}_\ell^F = 0$ for $\ell \neq i$ and thus

$$\sum_{F \in \mathcal{F}_\hbar^{\text{int}}} \left(e_{\ell_2}^T F_i e_{\ell_1} \mathfrak{n}_i^F [\![\boldsymbol{\phi}_1]\!]_F \,\middle|\, \{\!\{\boldsymbol{\phi}_2\}\!\}_F\right)_F = \sum_{F \in \mathcal{F}_{\hbar,i}} \left(e_{\ell_2}^T F_i e_{\ell_1} [\![\boldsymbol{\phi}_1]\!]_F \,\middle|\, \{\!\{\boldsymbol{\phi}_2\}\!\}_F\right)_F.$$

Further, if $F \not\subset \partial K_\ell$, we have

$$\{\!\{\boldsymbol{\phi}_\ell\}\!\}_F = [\![\boldsymbol{\phi}_\ell]\!]_F = 0 \qquad \text{for } \ell = 1, 2.$$

Hence, if $\partial K_1 \cap \partial K_2 \notin \mathcal{F}_{\hbar,i}$, i.e., K_1 and K_2 do not share a common face with normal in the ith coordinate direction, this yields

$$\sum_{F \in \mathcal{F}_\hbar^{\text{int}}} \left(e_{\ell_2}^T F_i e_{\ell_1} \mathfrak{n}_i^F [\![\boldsymbol{\phi}_1]\!]_F \,\middle|\, \{\!\{\boldsymbol{\phi}_2\}\!\}_F\right)_F = 0.$$

Thus, this term only contributes to off-blockdiagonal entries if $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ have their support on elements sharing a face in $\mathcal{F}_{\hbar,i}$. As we do not admit the mesh to have hanging nodes by Assumption 6.1, each element has at most two neighbors w.r.t. faces in $\mathcal{F}_{\hbar,i}$, one in positive and one in negative ith direction. Consequently, if we additionally order the elements of the mesh along the normal vectors in the ith direction, the only additional entries appear in the first sub- and super-blockdiagonals.

Summary

Let us summarize the discussion of the last sections. We have seen that we can represent the central flux dG discretization of a Friedrichs' operator $\mathcal{F}$ having decoupled partial derivatives by a block-tridiagonal stiffness matrix. This is achieved by choosing a suitably ordered basis $\mathcal{V}_\hbar$ of the approximation space $V_\hbar$.

This basis is constructed as explained in Section 6.1.1. We can then achieve the aforementioned block-tridiagonal structure by ordering the basis functions as follows.

1. Order the basis functions in $\mathcal{V}_\hbar$ according to the index sets $\mathcal{I}_0^{\mathcal{F}}, \dots, \mathcal{I}_d^{\mathcal{F}}$ corresponding to the coefficients of $\mathcal{F}$. This yields $d+1$ subbases $\mathcal{V}_{\hbar,0}^{\mathcal{F}}, \dots, \mathcal{V}_{\hbar,d}^{\mathcal{F}}$.
2. For each subbasis $\mathcal{V}_{\hbar,1}^{\mathcal{F}}, \dots, \mathcal{V}_{\hbar,d}^{\mathcal{F}}$, order the elements such that elements sharing a face in $\mathcal{F}_{\hbar,i}$ are consecutive. Consequently, order the basis functions in each $\mathcal{V}_{\hbar,i}^{\mathcal{F}}$ elementwise according to this ordering of the elements. Lastly, order the basis functions in $\mathcal{V}_{\hbar,0}^{\mathcal{F}}$ elementwise for some arbitrary ordering of the elements.

In particular, this leads to a block-tridiagonal structure with block sizes that only depend on the polynomial degree k, the spatial dimension d and the number of indices belonging to each $\mathcal{I}_i^{\mathcal{F}}$, $i = 1, \dots, d$. Hence, the bandwidth is independent of the total number of elements.

Further, off-diagonal blocks can only appear if the corresponding basis functions belong to elements sharing a face in $\mathcal{F}_{\hbar,i}$, $i = 1, \dots, d$. Thus, the resulting matrix has diagonal blocks corresponding to rows of elements along the ith direction. These diagonal blocks completely decouple.

Given a Friedrichs' operator $\mathcal{F}$ having decoupled partial derivatives, we denote the basis functions ordered according to this procedure as

$$\boldsymbol{\psi}_1^{\mathcal{F}}, \dots, \boldsymbol{\psi}_{N_\hbar}^{\mathcal{F}}.$$

As it does not matter in which order we enumerate the subbases $\mathcal{V}_{\hbar,i}^{\mathcal{F}}$ or the rows of elements in the ith direction, $i = 1, \dots, d$, the ordering gained by this procedure is in general not unique. Hence, we arbitrarily fix one suitable ordering.

Remark 6.6. It is possible to admit the mesh $\mathcal{T}_\hbar$ to have hanging nodes and still obtain a banded structure of $\mathbb{F}$ with bandwidth independent of the total number of elements in $\mathcal{T}_\hbar$. To do so, one possibility is to start with a uniform mesh and refine the individual elements (preferably preserving the tensorial structure within the element) as needed. In that case $\mathbb{F}$ does not end up being block-tridiagonal, but having a higher number of sub- and super-blockdiagonals. This is due to the fact that elements may have more than two neighbors w.r.t. faces in $\mathcal{F}_{\hbar,i}$. The number of additional diagonals depends on the grade of refinement along one row of unrefined elements sharing faces in the ith direction.

The strategy given in this section is still applicable, albeit with some modifications. Namely, cluster the elements created by refinement of one element together and proceed as in Section 6.3.2 for the unrefined elements. Lastly, order the clustered fine elements corresponding to the ith coordinate of their center of mass. ◇

6.4 Efficiency of Peaceman–Rachford ADI schemes

In this section we apply the results from Section 6.3 to the dG-PR discretization of the wave-type problem (2.24). To this end, we start with assumptions on the type of problem we are considering.

Assumption 6.7. *We assume that the following holds.*

(i) *The split operators $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ defined in Section 2.4 have decoupled partial derivatives with permutation matrices $P_{\widetilde{\mathcal{A}}}$ and $P_{\widetilde{\mathcal{B}}}$, respectively.*

(ii) *The field* $P_{\widetilde{\mathcal{A}}}^T M P_{\widetilde{\mathcal{A}}}$ *is block-diagonal a.e. on* Ω *with diagonal blocks of the same size as* $P_{\widetilde{\mathcal{A}}}^T\big(\sup_{x\in\Omega}|A_i(x)|\big)P_{\widetilde{\mathcal{A}}}$ *for* $i = 1,\dots,d$.

(iii) *The field* $P_{\widetilde{\mathcal{B}}}^T M P_{\widetilde{\mathcal{B}}}$ *is block-diagonal a.e. on* Ω *with diagonal blocks of the same size as* $P_{\widetilde{\mathcal{B}}}^T\big(\sup_{x\in\Omega}|B_i(x)|\big)P_{\widetilde{\mathcal{B}}}$ *for* $i = 1,\dots,d$.

Assumption 6.7 poses conditions on the wave-type problem (2.24). Hence, it depends on the original problem if such a splitting is even possible.

We call the combination of splitting a problem fulfilling Assumption 6.7 with using the Peaceman–Rachford scheme for temporal discretization an **alternating direction implicit** (ADI) scheme. The reason is that this splitting corresponds to a dimension splitting of the operator $\mathcal{L}$ of the wave-type problem (2.24). We will see in Section 6.5 that applying this scheme to the wave equation in $\mathbb{R}^2$ or to the isotropic Maxwell's equations in $\mathbb{R}^3$ indeed results in (variants of) the classical ADI schemes proposed in [Peaceman and Rachford, 1955] and [Namiki, 1999, Zhen et al., 2000], respectively.

In the course of this chapter we show that under Assumption 6.7 one step of the dG-PR scheme can be performed in linear complexity w.r.t. the total number of elements in $\mathcal{T}_\hbar$. To this end, we make use of the special structure of $\widetilde{\mathcal{A}}$, $\widetilde{\mathcal{B}}$ and M and two different orderings of $\mathcal{V}_\hbar$ given by

$$\mathcal{V}_\hbar = \{\boldsymbol{\psi}_1^{\widetilde{\mathcal{A}}},\dots,\boldsymbol{\psi}_{N_\hbar}^{\widetilde{\mathcal{A}}}\} = \{\boldsymbol{\psi}_1^{\widetilde{\mathcal{B}}},\dots,\boldsymbol{\psi}_{N_\hbar}^{\widetilde{\mathcal{B}}}\}. \tag{6.12}$$

In fact, we show that (6.6a), i.e.,

$$\big(\mathbb{M} - \tfrac{\tau}{2}\mathbb{A}\big)\mathrm{u}^{n+1/2} = \big(\mathbb{M} + \tfrac{\tau}{2}\mathbb{B}\big)\mathrm{u}^n$$

can be solved efficiently by using the ordering w.r.t. $\widetilde{\mathcal{A}}$. Analogously, we show that (6.6b), i.e.,

$$\big(\mathbb{M} - \tfrac{\tau}{2}\mathbb{B}\big)\mathrm{u}^{n+1} = \big(\mathbb{M} + \tfrac{\tau}{2}\mathbb{A}\big)\big(\mathrm{u}^{n+1/2} + (\mathrm{f}^{n+1} + \mathrm{f}^n)\big)$$

can be solved in an efficient manner by using the ordering w.r.t. $\widetilde{\mathcal{B}}$.

6.4.1 Structure of the matrices

We start by investigating the structure of the stiffness matrices $\mathbb{A}$ and $\mathbb{B}$. By their definition in (6.5), we have

$$\mathbb{A} = \Big(\big(\boldsymbol{\mathcal{A}}\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i\big)_M\Big)_{i,j=1}^{N_\hbar} = \Big(\big(M\boldsymbol{\mathcal{A}}\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i\big)_\Omega\Big)_{i,j=1}^{N_\hbar} = \Big(\big(\widetilde{\boldsymbol{\mathcal{A}}}\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i\big)_\Omega\Big)_{i,j=1}^{N_\hbar}$$

and

$$\mathbb{B} = \Big(\big(\boldsymbol{\mathcal{B}}\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i\big)_M\Big)_{i,j=1}^{N_\hbar} = \Big(\big(M\boldsymbol{\mathcal{B}}\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i\big)_\Omega\Big)_{i,j=1}^{N_\hbar} = \Big(\big(\widetilde{\boldsymbol{\mathcal{B}}}\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i\big)_\Omega\Big)_{i,j=1}^{N_\hbar}.$$

Thus, by Assumption 6.7 both matrices are the stiffness matrices corresponding to Friedrichs' operators having decoupled partial derivatives. By the discussion in Section 6.3 this means that $\mathbb{A}$ and $\mathbb{B}$ can be reordered such that they are block-tridiagonal by using the orderings $\boldsymbol{\psi}_1^{\widetilde{\mathcal{A}}},\dots,\boldsymbol{\psi}_{N_\hbar}^{\widetilde{\mathcal{A}}}$ and $\boldsymbol{\psi}_1^{\widetilde{\mathcal{B}}},\dots,\boldsymbol{\psi}_{N_\hbar}^{\widetilde{\mathcal{B}}}$, respectively.

It remains to investigate the structure of the mass matrix $\mathbb{M}$ under the two orderings of $V_\hbar$ given in (6.12). By the definition of the mass matrix (6.4), we have

$$\mathbb{M} = \Big(\big(\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i\big)_M\Big)_{i,j=1}^{N_\hbar} = \Big(\big(M\boldsymbol{\psi}_j \,\big|\, \boldsymbol{\psi}_i\big)_\Omega\Big)_{i,j=1}^{N_\hbar}.$$

Comparing Assumption 6.7 (ii) and (iii) to Definition 6.4 (ii) implies that M has the same block-diagonal structure as A_0 and B_0. Hence, by the same reasoning as in Section 6.3.2 the mass matrix can only take non-zero values on the diagonal blocks of the tridiagonal structure of $\mathbb{A}$ and $\mathbb{B}$.

6.4.2 Implementation

The discussion in Section 6.4.1 yields that both $\left(\mathbb{M}-\frac{\tau}{2}\mathbb{A}\right)$ and $\left(\mathbb{M}-\frac{\tau}{2}\mathbb{B}\right)$ can be reordered to have block-tridiagonal structure. Hence, we apply the following strategy to perform one step of the dG Peaceman–Rachford scheme.

- Determine the ordering of basis functions given by
$$\boldsymbol{\psi}_1^{\widetilde{\mathcal{A}}},\dots,\boldsymbol{\psi}_{N_\hbar}^{\widetilde{\mathcal{A}}} \qquad \text{and} \qquad \boldsymbol{\psi}_1^{\widetilde{\mathcal{B}}},\dots,\boldsymbol{\psi}_{N_\hbar}^{\widetilde{\mathcal{B}}}.$$
- Solve the linear system
$$\left(\mathbb{M}-\tfrac{\tau}{2}\mathbb{A}\right)\mathrm{u}^{n+1/2}=\left(\mathbb{M}+\tfrac{\tau}{2}\mathbb{B}\right)\mathrm{u}^{n}$$
using the ordering w.r.t. $\widetilde{\mathcal{A}}$.
- Solve the linear system
$$\left(\mathbb{M}-\tfrac{\tau}{2}\mathbb{B}\right)\mathrm{u}^{n+1}=\left(\mathbb{M}+\tfrac{\tau}{2}\mathbb{A}\right)\left(\mathrm{u}^{n+1/2}+(\mathrm{f}^{n+1}+\mathrm{f}^{n})\right)$$
using the ordering w.r.t. $\widetilde{\mathcal{B}}$.

Using the suggested ordering, the matrices $\left(\mathbb{M}-\frac{\tau}{2}\mathbb{A}\right)$ and $\left(\mathbb{M}-\frac{\tau}{2}\mathbb{B}\right)$ are banded with bandwidth independent of the total number of elements in $\mathcal{T}_\hbar$. Therefore, this yields an algorithm that can be performed in linear complexity w.r.t. this number.

Further, as already stated in Section 6.3.2, off-diagonal blocks in $\mathbb{A}$ and $\mathbb{B}$ can only appear if the corresponding basis functions belong to elements sharing a face in $\mathcal{F}_{\hbar,i}$, $i=1,\dots,d$. This means that the linear systems one has to solve decouple into smaller linear systems corresponding to rows of elements along one direction. Hence, we can easily parallelize one half-step of the scheme by solving these linear systems simultaneously.

Remark 6.8. We can further speed up the method if the coefficients of $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ and the material tensor have a product structure. In that case, by choosing a tensorial basis for $\mathbb{Q}_d^{k+1}(K)$ on each $K\in\mathcal{T}_\hbar$, the inner products comprising the stiffness and mass matrices reduce to a product of one-dimensional integrals. This leads to a Kronecker product structure of the resulting stiffness and mass matrices, which can be exploited to solve the occurring linear systems more efficiently. ◇

Remark 6.9. As stated in Section 4.4.2 we can vary the Peaceman–Rachford scheme (4.5) by treating the inhomogeneity differently and still obtain a second order scheme. The associated linear system on $\mathbb{R}^{N_\hbar}$ then reads
$$\begin{aligned}\left(\mathbb{M}-\tfrac{\tau}{2}\mathbb{A}\right)\mathrm{u}^{n+1/2}&=\left(\mathbb{M}+\tfrac{\tau}{2}\mathbb{B}\right)\mathrm{u}^{n},\\ \left(\mathbb{M}-\tfrac{\tau}{2}\mathbb{B}\right)\mathrm{u}^{n+1}&=\left(\mathbb{M}+\tfrac{\tau}{2}\mathbb{A}\right)\mathrm{u}^{n+1/2}+\left(\mathbb{M}-\tfrac{\tau}{2}\mathbb{A}\right)^{-1}\left(\mathrm{f}^{n+1}+\mathrm{f}^{n}\right).\end{aligned}$$
As this requires us to compute $\left(\mathbb{M}-\frac{\tau}{2}\mathbb{A}\right)^{-1}\left(\mathrm{f}^{n+1}+\mathrm{f}^{n}\right)$, we need to solve an additional linear system in each step. Hence, this variant is slightly more costly, but can still be evaluated with the same linear complexity. However, this additional cost comes at the advantage of less error terms occurring in the discretization error. ◇

6.5 Examples

Lastly, let us revisit the examples given in Section 2.5. As we need Assumption 6.7 to be fulfilled, this restricts the generality of the examples considered in this section. Namely, to fulfill this assumption, we restrict ourselves to the two-dimensional advection and wave equation. Further, for Maxwell's equations we restrict ourselves to the isotropic case.

In the course of this section we show that under the aforementioned circumstances, our examples admit a splitting such that the assumptions in Section 2.4 on the split operators and Assumption 6.7 are fulfilled. To be more precise, we show that the split operators are dissipative Friedrichs' operators fulfilling the splitting properties (2.26) and (2.27). Further, we show that these operators have decoupled partial derivatives, and that M fulfills Assumption 6.7 (ii) and (iii).

6.5.1 The two-dimensional advection equation

Consider the advection equation (2.28) in $\Omega \subset \mathbb{R}^2$, i.e.,

$$\begin{cases} \partial_t u = \alpha \cdot \nabla u + g & \text{in } \mathbb{R}_+ \times \Omega, \\ u(0) = u^0 & \text{in } \Omega. \end{cases}$$

Split operators

First, note that since we consider $d = 2$, we have

$$\widetilde{\mathcal{L}} = \alpha \cdot \nabla = \alpha_1 \partial_1 + \alpha_2 \partial_2.$$

We define the split operators by

$$\widetilde{\mathcal{A}} = \alpha_1 \partial_1 \qquad \text{and} \qquad \widetilde{\mathcal{B}} = \alpha_2 \partial_2.$$

Both $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ fulfill the conditions of Definition 2.15, and are therefore Friedrichs' operators with coefficients

$$\begin{aligned} A_0 &= 0, & A_1 &= \alpha_1, & A_2 &= 0, \\ B_0 &= 0, & B_1 &= 0, & B_2 &= \alpha_2 \end{aligned} \tag{6.13}$$

and graph spaces

$$\begin{aligned} H(\widetilde{\mathcal{A}}) &= \{v \in L^2(\Omega) \mid \alpha_1 \partial_1 v \in L^2(\Omega)\}, \\ H(\widetilde{\mathcal{B}}) &= \{v \in L^2(\Omega) \mid \alpha_2 \partial_2 v \in L^2(\Omega)\}. \end{aligned}$$

Hence, we have

$$\widetilde{\mathcal{L}} v = \widetilde{\mathcal{A}} v + \widetilde{\mathcal{B}} v \qquad \text{for all } v \in H(\widetilde{\mathcal{A}}) \cap H(\widetilde{\mathcal{B}}),$$

confirming (2.26). Further, by (6.13), both the coefficients of $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ fulfill Definition 6.4 (i) and (ii), meaning they have decoupled partial derivatives if the boundary conditions are of a certain structure.

We additionally assume $\partial_1 \alpha_1 > 0$ and $\partial_2 \alpha_2 > 0$ in order to fulfill condition (2.20) assuring that the split operators are dissipative. However, we want to point out that this is only a restriction of our analytical framework as we excluded shift-dissipative operators. We do not need this for the efficient implementation.

Splitting of the boundary conditions

Next, we split the boundary operator $\widetilde{\mathcal{L}}_\Gamma$ defined in (2.29). Recall that, for $v, w \in H(\widetilde{\mathcal{L}})$, we have

$$\langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w \rangle = -\big(|\alpha \cdot \mathfrak{n}| v \,|\, w\big)_\Gamma = -\big(|\alpha_1 \mathfrak{n}_1 + \alpha_2 \mathfrak{n}_2| v \,|\, w\big)_\Gamma.$$

By Assumption 6.1 the domain Ω is a finite union of bounded paraxial tensorial sets. Hence, we can decompose the boundary Γ into two separate sets

$$\Gamma = \Gamma_1 \cup \Gamma_2$$

with

$$\Gamma_i = \{\, x \in \Gamma \mid \mathfrak{n}(x) = \pm e_i \,\}, \qquad i = 1, 2.$$

This yields

$$\langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w \rangle = -\big(|\alpha_1| v \,|\, w\big)_{\Gamma_1} - \big(|\alpha_2| v \,|\, w\big)_{\Gamma_2}.$$

We now define the boundary operators $\widetilde{\mathcal{A}}_\Gamma \colon H(\widetilde{\mathcal{A}}) \to H(\widetilde{\mathcal{A}})'$ and $\widetilde{\mathcal{B}}_\Gamma \colon H(\widetilde{\mathcal{B}}) \to H(\widetilde{\mathcal{B}})'$ by

$$\langle \widetilde{\mathcal{A}}_\Gamma v \,|\, w \rangle = -\big(|\alpha_1| v \,|\, w\big)_{\Gamma_1} \qquad \text{for all } w \in H(\widetilde{\mathcal{A}})$$

and

$$\langle \widetilde{\mathcal{B}}_\Gamma v \,|\, w \rangle = -\big(|\alpha_2| v \,|\, w\big)_{\Gamma_2} \qquad \text{for all } w \in H(\widetilde{\mathcal{B}}).$$

Hence, we have

$$\widetilde{\mathcal{L}}_\Gamma v = \widetilde{\mathcal{A}}_\Gamma v + \widetilde{\mathcal{B}}_\Gamma v \qquad \text{for all } v \in H(\widetilde{\mathcal{A}}) \cap H(\widetilde{\mathcal{B}}),$$

confirming (2.27). Further, since we have $m = 1$, Definition 6.4 (iii) is trivially fulfilled.

It remains to show that $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$ are dissipative boundary conditions for $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$, respectively. However, note that we can retrieve the split operators $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{A}}_\Gamma$ from $\widetilde{\mathcal{L}}$ and $\widetilde{\mathcal{L}}_\Gamma$, respectively, by setting $\alpha_2 = 0$. The same holds for $\widetilde{\mathcal{B}}$ and $\widetilde{\mathcal{B}}_\Gamma$ if we set $\alpha_1 = 0$. Since we have shown that $\widetilde{\mathcal{L}}_\Gamma$ is a dissipative boundary condition for $\widetilde{\mathcal{L}}$ for more general α, the claim readily follows.

Structure of the material tensor

Recall that the material tensor M for the advection equation is given by $M = 1$. This trivially yields Assumptions 6.7 (ii) and (iii).

6.5.2 The two-dimensional acoustic wave equation

Next, we revisit the acoustic wave equation (2.30), i.e.,

$$\begin{cases} \rho \partial_t p = \nabla \cdot q + \tilde{g} & \text{in } \mathbb{R}_+ \times \Omega, \\ \partial_t q = \nabla p & \text{in } \mathbb{R}_+ \times \Omega, \\ p(0) = p^0, \quad q(0) = q^0 & \text{in } \Omega. \end{cases}$$

Again, we restrict ourselves to two spatial dimensions, i.e., $\Omega \subset \mathbb{R}^2$ and thus $m = 3$.

Throughout this section, let

$$v = \begin{pmatrix} p \\ q \end{pmatrix} \quad \text{with} \quad q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} \qquad \text{and} \qquad w = \begin{pmatrix} \widetilde{p} \\ \widetilde{q} \end{pmatrix} \quad \text{with} \quad \widetilde{q} = \begin{pmatrix} \widetilde{q}_1 \\ \widetilde{q}_2 \end{pmatrix}.$$

Split operators

Because of $d = 2$, we have

$$\widetilde{\mathcal{L}} = \begin{pmatrix} 0 & \nabla\cdot \\ \nabla & 0 \end{pmatrix} = \begin{pmatrix} 0 & \partial_1 & \partial_2 \\ \partial_1 & 0 & 0 \\ \partial_2 & 0 & 0 \end{pmatrix}.$$

We define the split operators by

$$\widetilde{\mathcal{A}} = \begin{pmatrix} 0 & \partial_1 & 0 \\ \partial_1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad \widetilde{\mathcal{B}} = \begin{pmatrix} 0 & 0 & \partial_2 \\ 0 & 0 & 0 \\ \partial_2 & 0 & 0 \end{pmatrix}.$$

Hence, $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ are Friedrichs' operators with coefficients

$$\begin{aligned} &A_0 = 0, && A_1 = \begin{pmatrix} 0 & e_1^T \\ e_1 & 0 \end{pmatrix}, && A_2 = 0, \\ &B_0 = 0, && B_1 = 0, && B_2 = \begin{pmatrix} 0 & e_2^T \\ e_2 & 0 \end{pmatrix} \end{aligned} \tag{6.14}$$

and graph spaces

$$\begin{aligned} H(\widetilde{\mathcal{A}}) &= \{(p,q) \in L^2(\Omega) \times L^2(\Omega)^2 \mid \partial_1 p \in L^2(\Omega),\ \partial_1 q_1 \in L^2(\Omega)\}, \\ H(\widetilde{\mathcal{B}}) &= \{(p,q) \in L^2(\Omega) \times L^2(\Omega)^2 \mid \partial_2 p \in L^2(\Omega),\ \partial_2 q_2 \in L^2(\Omega)\}. \end{aligned}$$

This yields

$$\widetilde{\mathcal{L}} v = \widetilde{\mathcal{A}} v + \widetilde{\mathcal{B}} v \qquad \text{for all } v \in H(\widetilde{\mathcal{A}}) \cap H(\widetilde{\mathcal{B}}),$$

confirming (2.26). Further, by (6.14) the coefficients of $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ fulfill Definition 6.4 (i) and (ii), meaning they have decoupled partial derivatives if equipped with dissipative boundary conditions that possess a certain structure.

Splitting of the boundary conditions

We now split the boundary operator $\widetilde{\mathcal{L}}_\Gamma$ given in (2.31). First, recognize that, for $v, w \in H(\widetilde{\mathcal{L}})$, integration by parts yields

$$\begin{aligned} \langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w \rangle &= \langle \mathbf{n} \cdot q \,|\, \widetilde{p} \rangle - \langle \mathbf{n} \cdot \widetilde{q} \,|\, p \rangle \\ &= \big(\nabla\cdot q \,\big|\, \widetilde{p} \big)_\Omega + \big(q \,\big|\, \nabla \widetilde{p} \big)_\Omega - \big(\nabla p \,\big|\, \widetilde{q} \big)_\Omega - \big(p \,\big|\, \nabla\cdot \widetilde{q} \big)_\Omega \\ &= \big(\partial_1 q_1 \,\big|\, \widetilde{p} \big)_\Omega + \big(q_1 \,\big|\, \partial_1 \widetilde{p} \big)_\Omega - \big(\partial_1 p \,\big|\, \widetilde{q}_1 \big)_\Omega - \big(p \,\big|\, \partial_1 \widetilde{q}_1 \big)_\Omega \\ &\quad + \big(\partial_2 q_2 \,\big|\, \widetilde{p} \big)_\Omega + \big(q_2 \,\big|\, \partial_2 \widetilde{p} \big)_\Omega - \big(\partial_2 p \,\big|\, \widetilde{q}_2 \big)_\Omega - \big(p \,\big|\, \partial_2 \widetilde{q}_2 \big)_\Omega. \end{aligned}$$

Now, we define $\widetilde{\mathcal{A}}_\Gamma \colon H(\widetilde{\mathcal{A}}) \to H(\widetilde{\mathcal{A}})'$ and $\widetilde{\mathcal{B}}_\Gamma \colon H(\widetilde{\mathcal{B}}) \to H(\widetilde{\mathcal{B}})'$ by

$$\langle \widetilde{\mathcal{A}}_\Gamma v \,|\, w \rangle = \big(\partial_1 q_1 \,\big|\, \widetilde{p} \big)_\Omega - \big(\partial_1 p \,\big|\, \widetilde{q}_1 \big)_\Omega + \big(q_1 \,\big|\, \partial_1 \widetilde{p} \big)_\Omega - \big(p \,\big|\, \partial_1 \widetilde{q}_1 \big)_\Omega \tag{6.15}$$

for all $w \in H(\widetilde{\mathcal{A}})$ and

$$\langle \widetilde{\mathcal{B}}_\Gamma v \,|\, w \rangle = \big(\partial_2 q_2 \,\big|\, \widetilde{p} \big)_\Omega - \big(\partial_2 p \,\big|\, \widetilde{q}_2 \big)_\Omega + \big(q_2 \,\big|\, \partial_2 \widetilde{p} \big)_\Omega - \big(p \,\big|\, \partial_2 \widetilde{q}_2 \big)_\Omega$$

for all $w \in H(\widetilde{\mathcal{B}})$. This yields

$$\widetilde{\mathcal{L}}_\Gamma v = \widetilde{\mathcal{A}}_\Gamma v + \widetilde{\mathcal{B}}_\Gamma v \qquad \text{for all } v \in H(\widetilde{\mathcal{A}}) \cap H(\widetilde{\mathcal{B}}),$$

confirming (2.27).

We next show that Definition 6.4 (iii) is fulfilled. Thus, we need to determine the matrix fields associated with $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$. For v, w sufficiently smooth, integration by parts yields

$$\langle \widetilde{\mathcal{A}}_\Gamma v \,|\, w \rangle = \left(\mathfrak{n}_1 q_1 \,|\, \widetilde{p}\right)_\Gamma - \left(\mathfrak{n}_1 p \,|\, \widetilde{q}_1\right)_\Gamma$$

and

$$\langle \widetilde{\mathcal{B}}_\Gamma v \,|\, w \rangle = \left(\mathfrak{n}_2 q_2 \,|\, \widetilde{p}\right)_\Gamma - \left(\mathfrak{n}_2 p \,|\, \widetilde{q}_2\right)_\Gamma.$$

Hence, the matrix fields associated with $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$ are given by

$$\mathcal{N}_{\widetilde{\mathcal{A}}} = \begin{pmatrix} 0 & \mathfrak{n}_1 & 0 \\ -\mathfrak{n}_1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathcal{N}_{\widetilde{\mathcal{B}}} = \begin{pmatrix} 0 & 0 & \mathfrak{n}_2 \\ 0 & 0 & 0 \\ -\mathfrak{n}_2 & 0 & 0 \end{pmatrix},$$

showing Definition 6.4 (iii).

It remains to show that $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$ are dissipative boundary conditions for $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$, respectively. Both operators fulfill Definition 2.21 (B1) as they are skew-symmetric. To show Definition 2.21 (B2), let $v, w \in H(\widetilde{\mathcal{A}})$. By Definition 2.19 the boundary operator $\widetilde{\mathcal{A}}_\partial$ is given by

$$\langle \widetilde{\mathcal{A}}_\partial v \,|\, w \rangle = \left(\partial_1 q_1 \,|\, \widetilde{p}\right)_\Omega + \left(\partial_1 p \,|\, \widetilde{q}_1\right)_\Omega + \left(q_1 \,|\, \partial_1 \widetilde{p}\right)_\Omega + \left(p \,|\, \partial_1 \widetilde{q}_1\right)_\Omega.$$

Hence, with (6.15) we have

$$\langle (\widetilde{\mathcal{A}}_\partial - \widetilde{\mathcal{A}}_\Gamma) v \,|\, w \rangle = 2\left(\partial_1 p \,|\, \widetilde{q}_1\right)_\Omega + 2\left(p \,|\, \partial_1 \widetilde{q}_1\right)_\Omega$$

and

$$\langle (\widetilde{\mathcal{A}}_\partial + \widetilde{\mathcal{A}}_\Gamma) v \,|\, w \rangle = 2\left(\partial_1 q_1 \,|\, \widetilde{p}\right)_\Omega + 2\left(q_1 \,|\, \partial_1 \widetilde{p}\right)_\Omega.$$

Therefore, we can decompose v into

$$v = \begin{pmatrix} p \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ q \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} p \\ 0 \end{pmatrix} \in \ker(\widetilde{\mathcal{A}}_\partial + \widetilde{\mathcal{A}}_\Gamma) \quad \text{and} \quad \begin{pmatrix} 0 \\ q \end{pmatrix} \in \ker(\widetilde{\mathcal{A}}_\partial - \widetilde{\mathcal{A}}_\Gamma),$$

confirming Definition 2.21 (B2) for $\widetilde{\mathcal{A}}_\Gamma$. Proceeding analogously shows the claim for $\widetilde{\mathcal{B}}_\Gamma$.

Structure of the material tensor

Recall that the material tensor M is given a.e. on Ω by

$$M = \begin{pmatrix} \rho & 0 \\ 0 & \mathcal{I} \end{pmatrix}$$

with ρ being scalar. Hence, M is diagonal a.e. on Ω, showing Assumptions 6.7 (ii) and (iii).

Remark 6.10. We want to point out that the splitting in this section is similar to the splitting in [Peaceman and Rachford, 1955], where the ADI method was originally proposed. The difference being that in this paper, the Laplace operator instead of the div-grad operator was split. However, the acoustic wave equation can be associated with the Laplace operator by changing into the second order formulation. Splitting the Laplace operator as in [Peaceman and Rachford, 1955] should lead to similar split problems as the ones in this section. ◇

6.5.3 Maxwell's equations

Lastly, we once again consider Maxwell's equations (2.32), i.e.,

$$\begin{cases} \varepsilon \partial_t E = \nabla\times H - \sigma E - J & \text{in } \mathbb{R}_+ \times \Omega, \\ \mu \partial_t H = -\nabla\times E & \text{in } \mathbb{R}_+ \times \Omega, \\ E(0) = E^0, \quad H(0) = H^0 & \text{in } \Omega. \end{cases}$$

To fulfill Assumption 6.7 we have to restrict ourselves to isotropic materials. Hence, assume that ε, μ, σ are diagonal.

Throughout this section, let

$$v = \begin{pmatrix} E \\ H \end{pmatrix} \qquad \text{with} \qquad E = \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} H_1 \\ H_2 \\ H_3 \end{pmatrix}$$

and

$$w = \begin{pmatrix} \widetilde{E} \\ \widetilde{H} \end{pmatrix} \qquad \text{with} \qquad \widetilde{E} = \begin{pmatrix} \widetilde{E}_1 \\ \widetilde{E}_2 \\ \widetilde{E}_3 \end{pmatrix} \quad \text{and} \quad \widetilde{H} = \begin{pmatrix} \widetilde{H}_1 \\ \widetilde{H}_2 \\ \widetilde{H}_3 \end{pmatrix}.$$

Split operators

Recall that the Maxwell operator is given by

$$\widetilde{\mathcal{L}} = \begin{pmatrix} 0 & \nabla\times \\ -\nabla\times & 0 \end{pmatrix} - \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

with

$$\nabla\times = \begin{pmatrix} 0 & -\partial_3 & \partial_2 \\ \partial_3 & 0 & -\partial_1 \\ -\partial_2 & \partial_1 & 0 \end{pmatrix}.$$

The idea to obtain suitable split operators is to first split the curl-operator into a positive and a negative part, namely

$$\nabla\times = \mathcal{C}_1 - \mathcal{C}_2 \qquad \text{with} \qquad \mathcal{C}_1 = \begin{pmatrix} 0 & 0 & \partial_2 \\ \partial_3 & 0 & 0 \\ 0 & \partial_1 & 0 \end{pmatrix} \quad \text{and} \quad \mathcal{C}_2 = \begin{pmatrix} 0 & \partial_3 & 0 \\ 0 & 0 & \partial_1 \\ \partial_2 & 0 & 0 \end{pmatrix}.$$

With this, we define the split operators

$$\widetilde{\mathcal{A}} = \begin{pmatrix} 0 & \mathcal{C}_1 \\ \mathcal{C}_2 & 0 \end{pmatrix} - \begin{pmatrix} \frac{\sigma}{2} & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad \widetilde{\mathcal{B}} = \begin{pmatrix} 0 & -\mathcal{C}_2 \\ -\mathcal{C}_1 & 0 \end{pmatrix} - \begin{pmatrix} \frac{\sigma}{2} & 0 \\ 0 & 0 \end{pmatrix}.$$

This yields that $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ are Friedrichs' operators with coefficients

$$A_0 = \begin{pmatrix} \frac{\sigma}{2} & 0 \\ 0 & 0 \end{pmatrix}, \qquad A_i = \begin{pmatrix} 0 & a_i^T \\ a_i & 0 \end{pmatrix}, \qquad i = 1,2,3,$$

$$B_0 = \begin{pmatrix} \frac{\sigma}{2} & 0 \\ 0 & 0 \end{pmatrix}, \qquad B_i = \begin{pmatrix} 0 & -b_i^T \\ -b_i & 0 \end{pmatrix}, \qquad i = 1,2,3,$$

where $a_1, a_2, a_3, b_1, b_2, b_3 \in \mathbb{R}^{3\times 3}$ with $a_1 = b_1^T = e_2 e_3^T$, $a_2 = b_2^T = e_3 e_1^T$ and $a_3 = b_3^T = e_1 e_2^T$. Hence, Definition 6.4 (i) and (ii) are fulfilled. Therefore, if combined with a boundary condition that has a suitable structure, $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ have decoupled partial derivatives. Further, the graph spaces of $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ are given by

$$H(\widetilde{\mathcal{A}}) = \{(E,H) \in L^2(\Omega)^3 \times L^2(\Omega)^3 \mid \mathcal{C}_2 E \in L^2(\Omega)^3,\ \mathcal{C}_1 H \in L^2(\Omega)^3\},$$
$$H(\widetilde{\mathcal{B}}) = \{(E,H) \in L^2(\Omega)^3 \times L^2(\Omega)^3 \mid \mathcal{C}_1 E \in L^2(\Omega)^3,\ \mathcal{C}_2 H \in L^2(\Omega)^3\}.$$

Altogether, this yields

$$\widetilde{\mathcal{L}} v = \widetilde{\mathcal{A}} v + \widetilde{\mathcal{B}} v \qquad \text{for all } v \in H(\widetilde{\mathcal{A}}) \cap H(\widetilde{\mathcal{B}}),$$

i.e., (2.26) is fulfilled.

Splitting of the boundary conditions

Next, we split the boundary operator $\widetilde{\mathcal{L}}_\Gamma$ defined in (2.33). For $v, w \in H(\widetilde{\mathcal{L}})$ we have

$$\begin{aligned}\langle \widetilde{\mathcal{L}}_\Gamma v \,|\, w\rangle &= \big(\nabla\times H \,\big|\, \widetilde{E}\big)_\Omega + \big(\nabla\times E \,\big|\, \widetilde{H}\big)_\Omega - \big(E \,\big|\, \nabla\times \widetilde{H}\big)_\Omega - \big(H \,\big|\, \nabla\times \widetilde{E}\big)_\Omega \\ &= \big(\mathcal{C}_1 H \,\big|\, \widetilde{E}\big)_\Omega - \big(\mathcal{C}_2 E \,\big|\, \widetilde{H}\big)_\Omega - \big(E \,\big|\, \mathcal{C}_1 \widetilde{H}\big)_\Omega + \big(H \,\big|\, \mathcal{C}_2 \widetilde{E}\big)_\Omega \\ &\quad - \big(\mathcal{C}_2 H \,\big|\, \widetilde{E}\big)_\Omega + \big(\mathcal{C}_1 E \,\big|\, \widetilde{H}\big)_\Omega + \big(E \,\big|\, \mathcal{C}_2 \widetilde{H}\big)_\Omega - \big(H \,\big|\, \mathcal{C}_1 \widetilde{E}\big)_\Omega.\end{aligned}$$

Hence, define the split operators $\widetilde{\mathcal{A}}_\Gamma \colon H(\widetilde{\mathcal{A}}) \to H(\widetilde{\mathcal{A}})'$ and $\widetilde{\mathcal{B}}_\Gamma \colon H(\widetilde{\mathcal{B}}) \to H(\widetilde{\mathcal{B}})'$ by

$$\langle \widetilde{\mathcal{A}}_\Gamma v \,|\, w\rangle = \big(\mathcal{C}_1 H \,\big|\, \widetilde{E}\big)_\Omega - \big(\mathcal{C}_2 E \,\big|\, \widetilde{H}\big)_\Omega - \big(E \,\big|\, \mathcal{C}_1 \widetilde{H}\big)_\Omega + \big(H \,\big|\, \mathcal{C}_2 \widetilde{E}\big)_\Omega$$

for all $w \in H(\widetilde{\mathcal{A}})$ and

$$\langle \widetilde{\mathcal{B}}_\Gamma v \,|\, w\rangle = -\big(\mathcal{C}_2 H \,\big|\, \widetilde{E}\big)_\Omega + \big(\mathcal{C}_1 E \,\big|\, \widetilde{H}\big)_\Omega + \big(E \,\big|\, \mathcal{C}_2 \widetilde{H}\big)_\Omega - \big(H \,\big|\, \mathcal{C}_1 \widetilde{E}\big)_\Omega$$

for all $w \in H(\widetilde{\mathcal{B}})$. By construction, we have

$$\widetilde{\mathcal{L}}_\Gamma v = \widetilde{\mathcal{A}}_\Gamma v + \widetilde{\mathcal{B}}_\Gamma v \qquad \text{for all } v \in H(\widetilde{\mathcal{A}}) \cap H(\widetilde{\mathcal{B}}),$$

yielding (2.27).

Next, we determine the matrix fields associated with $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$ to confirm Definition 6.4 (iii). Hence, let v, w be sufficiently smooth. Then, using integration by parts in each component we obtain

$$\langle \widetilde{\mathcal{A}}_\Gamma v \,|\, w\rangle = \big(\mathcal{N}_1 H \,\big|\, \widetilde{E}\big)_\Gamma - \big(\mathcal{N}_2 E \,\big|\, \widetilde{H}\big)_\Gamma$$

and

$$\langle \widetilde{\mathcal{B}}_\Gamma v \,|\, w\rangle = -\big(\mathcal{N}_2 H \,\big|\, \widetilde{E}\big)_\Gamma + \big(\mathcal{N}_1 E \,\big|\, \widetilde{H}\big)_\Gamma$$

with

$$\mathcal{N}_1 = \begin{pmatrix} 0 & 0 & \mathfrak{n}_2 \\ \mathfrak{n}_3 & 0 & 0 \\ 0 & \mathfrak{n}_1 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathcal{N}_2 = \begin{pmatrix} 0 & \mathfrak{n}_3 & 0 \\ 0 & 0 & \mathfrak{n}_1 \\ \mathfrak{n}_2 & 0 & 0 \end{pmatrix}.$$

Thus, the matrix fields associated with $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$ are given by

$$\mathcal{N}_{\widetilde{\mathcal{A}}} = \begin{pmatrix} 0 & \mathcal{N}_1 \\ -\mathcal{N}_2 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathcal{N}_{\widetilde{\mathcal{B}}} = \begin{pmatrix} 0 & -\mathcal{N}_2 \\ \mathcal{N}_1 & 0 \end{pmatrix},$$

which shows that Definition 6.4 (iii) is fulfilled.

Lastly, we show that $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$ are dissipative boundary conditions for $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$, respectively. Definition 2.21 (B1) is apparent, since both $\widetilde{\mathcal{A}}_\Gamma$ and $\widetilde{\mathcal{B}}_\Gamma$ are skew-symmetric. For Definition 2.21 (B2), let $v, w \in H(\widetilde{\mathcal{A}})$. By Definition 2.19 the boundary operator $\widetilde{\mathcal{A}}_\partial$ associated with $\widetilde{\mathcal{A}}$ is given by

$$\langle \widetilde{\mathcal{A}}_\partial v \,|\, w \rangle = \left(\mathcal{C}_1 H \,|\, \widetilde{E}\right)_\Omega + \left(\mathcal{C}_2 E \,|\, \widetilde{H}\right)_\Omega + \left(E \,|\, \mathcal{C}_1 \widetilde{H}\right)_\Omega + \left(H \,|\, \mathcal{C}_2 \widetilde{E}\right)_\Omega.$$

Hence, we have

$$\langle (\widetilde{\mathcal{A}}_\partial - \widetilde{\mathcal{A}}_\Gamma) v \,|\, w \rangle = 2\left(\mathcal{C}_2 E \,|\, \widetilde{H}\right)_\Omega + 2\left(E \,|\, \mathcal{C}_1 \widetilde{H}\right)_\Omega$$

and

$$\langle (\widetilde{\mathcal{A}}_\partial + \widetilde{\mathcal{A}}_\Gamma) v \,|\, w \rangle = 2\left(\mathcal{C}_1 H \,|\, \widetilde{E}\right)_\Omega + 2\left(H \,|\, \mathcal{C}_2 \widetilde{E}\right)_\Omega.$$

This yields

$$v = \begin{pmatrix} E \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ H \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} E \\ 0 \end{pmatrix} \in \ker(\widetilde{\mathcal{A}}_\partial + \widetilde{\mathcal{A}}_\Gamma) \quad \text{and} \quad \begin{pmatrix} 0 \\ H \end{pmatrix} \in \ker(\widetilde{\mathcal{A}}_\partial - \widetilde{\mathcal{A}}_\Gamma),$$

which shows Definition 2.21 (B2) for $\widetilde{\mathcal{A}}_\Gamma$. The claim is proven analogously for $\widetilde{\mathcal{B}}_\Gamma$.

Structure of the material tensor

The material tensor M is given a.e. on Ω by

$$M = \begin{pmatrix} \varepsilon & 0 \\ 0 & \mu \end{pmatrix}$$

and is therefore diagonal a.e. on Ω. This yields Assumptions 6.7 (ii) and (iii).

Remark 6.11. This splitting for Maxwell's equations was originally proposed in [Namiki, 1999, Zhen et al., 2000] for a finite difference discretization of undamped Maxwell's equations on the Yee grid. In this context it is known as the ADI-FDTD method. The way to split the damping term involving the conductivity σ is taken from [Eilinghoff and Schnaubelt, 2018], and the operator splitting framework used in this example is inspired by [Hochbruck et al., 2015a]. ◇

Conclusion

As we have seen, all three examples in this section admit a splitting for which Assumption 6.7 is fulfilled. In particular, this means that the error analysis of the dG-PR scheme given in Chapter 5 is valid for these problems. Further, the special structure of these problems allows us to implement the scheme by following the procedure described in Section 6.4.2. This yields an unconditionally stable method of temporal order two and spatial order k, of which one step in time can be performed with linear complexity w.r.t. the number of total elements in the spatial mesh.

7 | Numerical experiments

In this chapter we present some numerical experiments to illustrate the theoretical results we obtained over the course of the thesis. As a proof of concept we have implemented the dG-PR method for Maxwell's equations with splitting performed as in Section 6.5.3. In particular, this was done for isotropic Maxwell's equations without damping or external currents, i.e., $\rho = J = 0$.

7.1 Implementation

We have implemented the scheme with the help of the `C++` finite element library `deal.ii` [Bangerth et al., 2007], which was used to assemble the mass and the stiffness matrices stemming from the dG discretization. The implementation of the dG-PR scheme itself was done as described in Section 6.4.2.

In particular, we implemented a matrix class adapted to the dG-PR scheme. Solving is performed by calculating a block LU-decomposition of the block-tridiagonal matrices before starting the time stepping process. Then, in each step, the occurring linear systems are solved by forward and backward substitutions. This amounts to only matrix-vector multiplications in each step with matrices of the size of the tridiagonal blocks. Since the block banded structure of the matrix is preserved by the LU-decomposition, each step can therefore even be performed in linear complexity w.r.t. the full number of basis functions N_h (as opposed to only the number of elements of the mesh).

We further parallelized the code by exploiting that the linear systems corresponding to lines of elements along one direction completely decouple, as described at the end of Section 6.4.2. This parallelization was achieved by using the `C++11`-class `std::thread`.

7.2 Problem setup

All experiments were performed on the cuboidal domain $\Omega = [0, 2] \times [0, 1]^2$ equipped with a uniform tensorial mesh of various mesh widths. As a reference example, we used the exact solution given by

$$E(t, x) = \varepsilon^{-1} \begin{pmatrix} \widehat{E}_1 \cos(\kappa_1 x_1) \sin(\kappa_2 x_2) \sin(\kappa_3 x_3) \\ \widehat{E}_2 \sin(\kappa_1 x_1) \cos(\kappa_2 x_2) \sin(\kappa_3 x_3) \\ \widehat{E}_3 \sin(\kappa_1 x_1) \sin(\kappa_2 x_2) \cos(\kappa_3 x_3) \end{pmatrix} \cos(\Theta t),$$

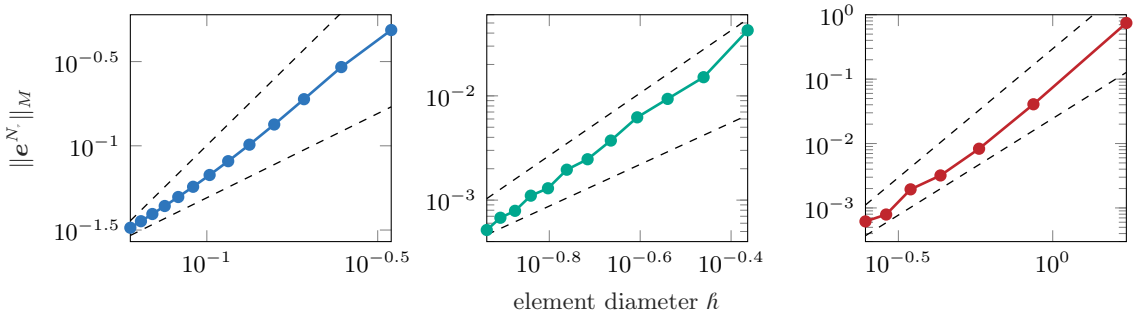


Figure 7.1: Convergence of the dG-PR method w.r.t. the mesh width $\hbar$. From left to right the polynomial degree is $k = 1$, $k = 2$ and $k = 3$. The dashed reference lines have slope k and $k + 1$. We have used $N_\tau = 20000$ steps for the time integration.

$$H(t,x) = c^2\Theta^{-1}\begin{pmatrix}(\widehat{E}_2\kappa_3 - \widehat{E}_3\kappa_2)\,\sin(\kappa_1 x_1)\,\cos(\kappa_2 x_2)\,\cos(\kappa_3 x_3)\\ (\widehat{E}_3\kappa_1 - \widehat{E}_1\kappa_3)\,\cos(\kappa_1 x_1)\,\sin(\kappa_2 x_2)\,\cos(\kappa_3 x_3)\\ (\widehat{E}_1\kappa_2 - \widehat{E}_2\kappa_1)\,\cos(\kappa_1 x_1)\,\cos(\kappa_2 x_2)\,\sin(\kappa_3 x_3)\end{pmatrix}\sin(\Theta t)$$

for $(t,x) \in \mathbb{R}_+ \times \Omega$. Here, $c = (\varepsilon\mu)^{-1/2}$ is the **speed of light**, $\kappa = (\kappa_1,\kappa_2,\kappa_3) \in \mathbb{R}^3_+$ is the **wave vector** and $\Theta = c\,\|\kappa\|$ is the **angular frequency**. Further, $\widehat{E}_1, \widehat{E}_2, \widehat{E}_3$ are preset amplitudes of the waves. For our experiments we chose

$$\varepsilon \equiv \mu \equiv 1, \qquad \kappa_1 = \kappa_2 = \kappa_3 = 2\pi \qquad \text{and} \qquad \widehat{E}_1 = -1, \quad \widehat{E}_2 = 0, \quad \widehat{E}_3 = 1.$$

This initial data also fulfills the perfectly conducting boundary conditions.

7.3 Convergence behavior

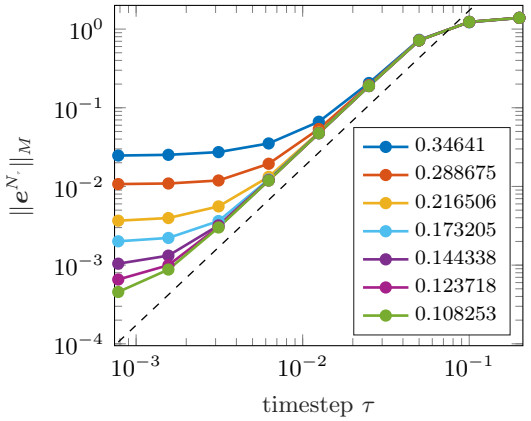
We start by confirming the bounds on the full discretization error obtained in Chapter 5. In other words, we investigate the spatial and temporal convergence rates of the dG-PR method applied to the aforementioned reference example.

Spatial convergence

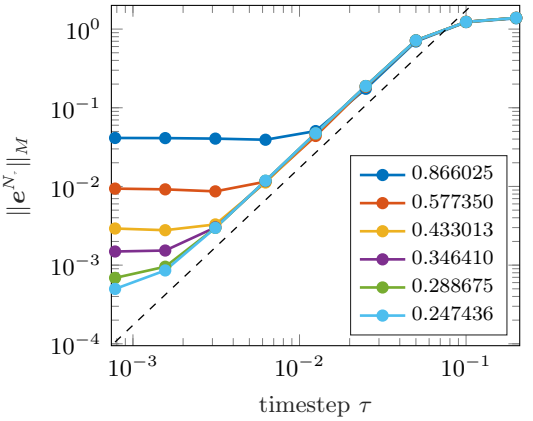
Spatial convergence, i.e., the behavior of the full discretization error under mesh refinement can be seen in Figure 7.1. The simulation interval for the experiments was $[0,2]$, and all errors were measured at the end time $T_{\text{end}} = 2$. These simulations were carried out using the timestep $\tau = 10^{-4}$ to ensure that the time integration error is negligible. We see that the convergence rates are between about one and half an order better than the ones predicted by Theorem 5.15. This is possibly due to the smooth exact solution and the regular structure of the meshes in use.

Temporal convergence

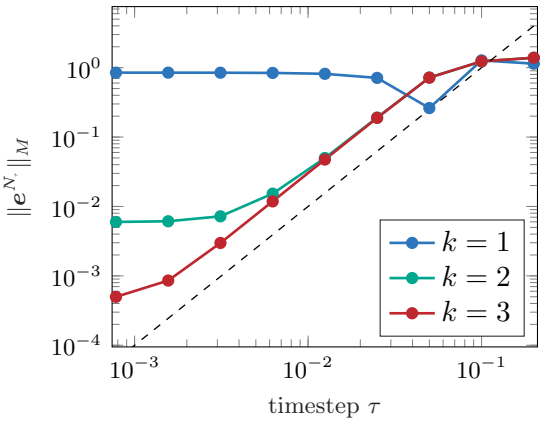
Next, we have a look at the temporal convergence order of the dG-PR method. Results of our numerical experiments are given in Figure 7.2. The simulation interval for these experiments was $[0,8]$, and all errors were measured at the end time $T_{\text{end}} = 8$. The temporal convergence order of two predicted by Theorem 5.15 is clearly validated by these results. Further, there is no sign of any stability constraints, again matching the theory.



(a) Error of the dG-PR scheme for polynomial degree $k = 2$ for mesh sizes given in the legend.



(b) Error of the dG-PR scheme for polynomial degree $k = 3$ for mesh sizes given in the legend.



(c) Error of the dG-PR scheme for fixed mesh size $h = 0.247436$ and polynomial degree k given in the legend.

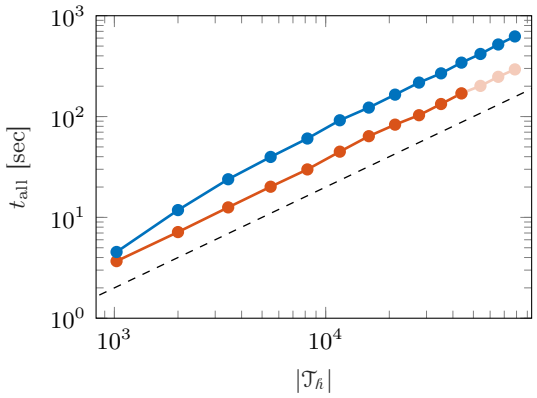
Figure 7.2: Convergence of the dG-PR method w.r.t. the timestep τ. We have used $N_\tau = \frac{T_{\text{end}}}{\tau}$ steps for the time integration. The dashed reference line in all three graphs have slope 2.

7.4 Runtime behavior

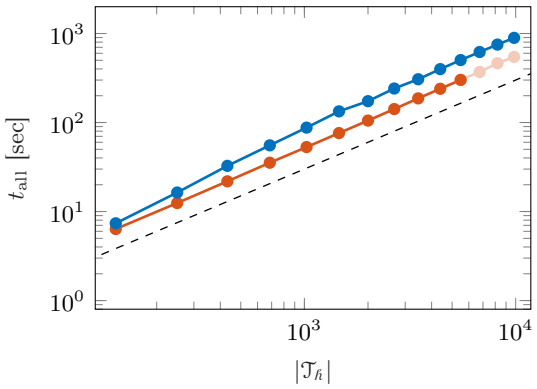
Lastly, we confirm the runtime behavior predicted in Chapter 6. All simulations in this section were carried out on the simulation interval $[0, 2]$ with timestep $\tau = 0.01$.

In Figure 7.3 we have plotted the total runtime t_{all} of the dG-PR method against the number of elements in the considered mesh $\mathcal{T}_h$. The resulting slopes confirm that the algorithm is indeed of linear complexity w.r.t. the number of mesh elements $|\mathcal{T}_h|$.

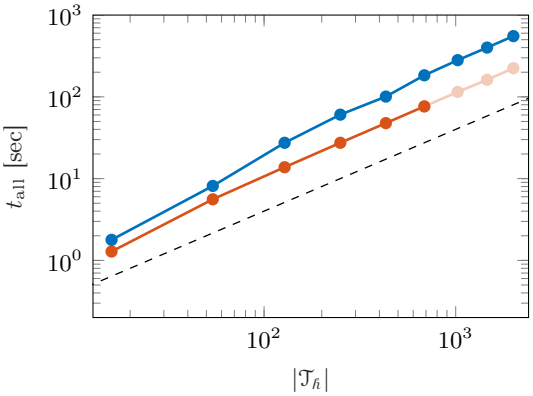
For comparison, we have also plotted the runtime of the fully explicit leapfrog or Verlet method from [Sturm, 2017, Section 4.2]. This scheme was combined with the same dG discretization of the spatial operators, leading to the dG-leapfrog scheme. The similarity in runtime behavior underlines the claim that the dG-PR method can be evaluated roughly at the cost of an explicit scheme. Further, we can see that the leapfrog scheme becomes unstable at some point, whereas the dG-PR scheme stays stable throughout all calculations.



(a) Polynomial degree $k = 1$. At the last data point the dG-PR scheme took ~ 2.1 times longer than the dG-leapfrog scheme.



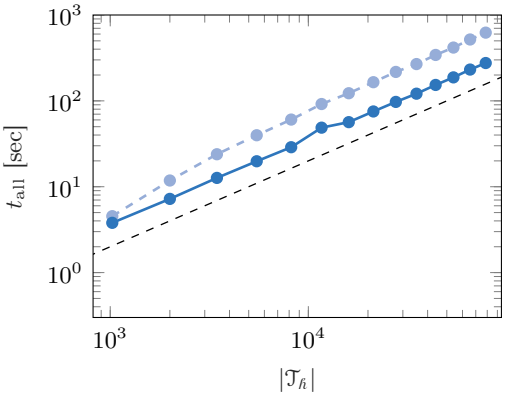
(b) Polynomial degree $k = 2$. At the last data point the dG-PR scheme took ~ 1.6 times longer than the dG-leapfrog scheme.



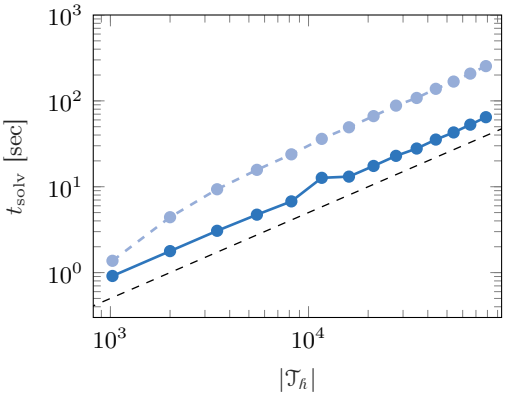
(c) Polynomial degree $k = 3$. At the last data point the dG-PR scheme took ~ 2.5 times longer than the dG-leapfrog scheme.

Figure 7.3: Total runtime t_{all} of the dG-PR and dG-leapfrog scheme for polynomial degrees $k = 1$, $k = 2$ and $k = 3$. The dashed reference lines in all three graphs have slope 1. The grayed out points indicate meshes for which the dG-leapfrog scheme was unstable.

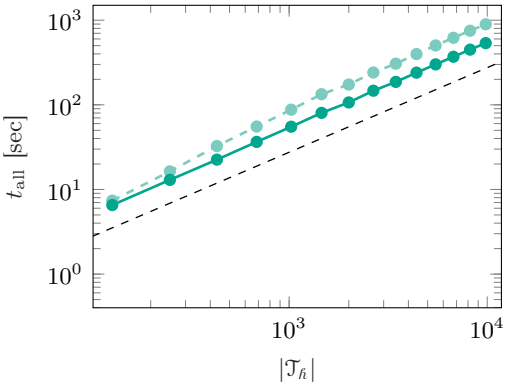
In Figure 7.4 we have further illustrated the benefit of the parallelization described at the end of Section 7.1. The first column shows the effect on the overall runtime t_{all} (including assembling, etc.). In the second column, the time t_{solv} needed to solve all occurring linear systems during runtime is plotted. This demonstrates that the straightforward parallelization described in Section 6.4.2 is almost optimal, at least if performed with shared memory.



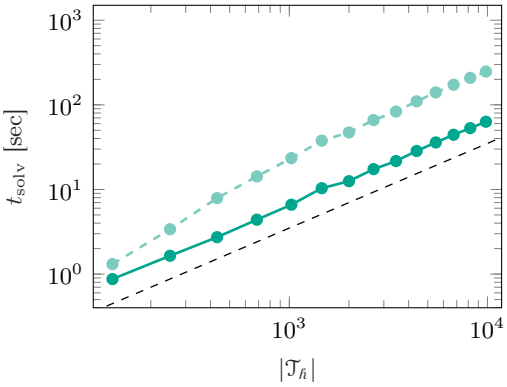
(a) Speed-up at the last data point: ~ 2.26.



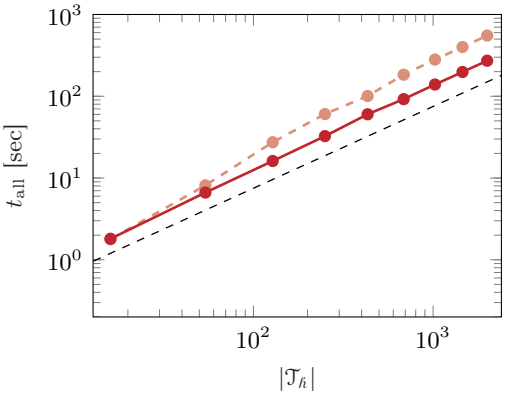
(b) Speed-up at the last data point: ~ 3.92.



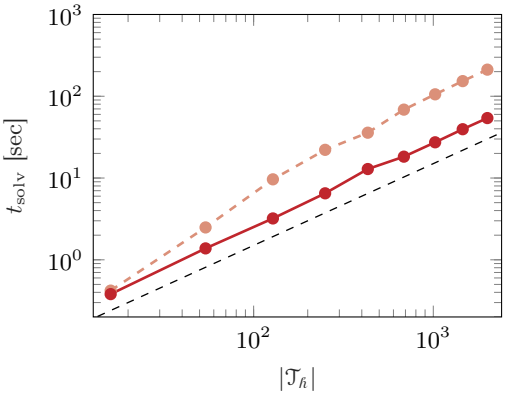
(c) Speed-up at the last data point: ~ 1.66.



(d) Speed-up at the last data point: ~ 3.93.



(e) Speed-up at the last data point: ~ 2.03.



(f) Speed-up at the last data point: ~ 3.91.

Figure 7.4: Total runtime t_{all} and time needed to solve all linear systems t_{solv} of the parallelized and serial dG-PR scheme for polynomial degrees $k = 1$, $k = 2$ and $k = 3$. The dashed reference lines in all six graphs have slope 1, and the grayed out, dashed lines correspond to the serial version. Parallelized results were generated by using four threads on a 4-core CPU with shared memory. The speed-up factors are given in the subcaptions.

Bibliography

[Bangerth et al., 2007]
W. Bangerth, R. Hartmann, G. Kanschat. deal.II—a general-purpose object-oriented finite element library. *ACM Trans. Math. Software*, 33(4), Art. 24, 2007. URL http://dx.doi.org/10.1145/1268776.1268779.

[Benzoni-Gavage and Serre, 2007]
S. Benzoni-Gavage, D. Serre. *Multidimensional hyperbolic partial differential equations. First-order systems and applications.* Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, Oxford, 2007. URL http://dx.doi.org/10.1093/acprof:oso/9780199211234.001.0001.

[Brian, 1961]
P. L. T. Brian. A finite-difference method of high-order accuracy for the solution of three-dimensional transient heat conduction problems. *AIChE Journal*, 7(3):pp. 367–370, 1961. URL http://dx.doi.org/10.1002/aic.690070305.

[Burazin and Erceg, 2016]
K. Burazin, M. Erceg. Non-stationary abstract Friedrichs systems. *Mediterr. J. Math.*, 13(6):pp. 3777–3796, 2016. URL http://dx.doi.org/10.1007/s00009-016-0714-8.

[Burman et al., 2010]
E. Burman, A. Ern, M. A. Fernández. Explicit Runge-Kutta schemes and finite elements with symmetric stabilization for first-order linear PDE systems. *SIAM J. Numer. Anal.*, 48(6):pp. 2019–2042, 2010. URL http://dx.doi.org/10.1137/090757940.

[Chen et al., 2008]
W. Chen, X. Li, D. Liang. Energy-conserved splitting FDTD methods for Maxwell's equations. *Numer. Math.*, 108(3):pp. 445–485, 2008. URL http://dx.doi.org/10.1007/s00211-007-0123-9.

[Chen et al., 2010]
W. Chen, X. Li, D. Liang. Energy-conserved splitting finite-difference time-domain methods for Maxwell's equations in three dimensions. *SIAM J. Numer. Anal.*, 48(4):pp. 1530–1554, 2010. URL http://dx.doi.org/10.1137/090765857.

[Courant et al., 1928]
R. Courant, K. Friedrichs, H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100(1):pp. 32–74, 1928. URL http://dx.doi.org/10.1007/BF01448839.

[Dautray and Lions, 1988]
R. DAUTRAY, J.-L. LIONS. *Mathematical analysis and numerical methods for science and technology. Vol. 2. Functional and variational methods.* Springer-Verlag, Berlin, 1988. URL http://dx.doi.org/10.1007/978-3-642-61566-5.

[Descombes et al., 2013]
S. DESCOMBES, S. LANTERI, L. MOYA. Locally implicit time integration strategies in a discontinuous Galerkin method for Maxwell's equations. *J. Sci. Comput.*, 56(1):pp. 190–218, 2013. URL http://dx.doi.org/10.1007/s10915-012-9669-5.

[Di Pietro and Ern, 2012]
D. A. DI PIETRO, A. ERN. *Mathematical aspects of discontinuous Galerkin methods*, vol. 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012. URL http://dx.doi.org/10.1007/978-3-642-22980-0.

[Diaz and Grote, 2009]
J. DIAZ, M. J. GROTE. Energy conserving explicit local time stepping for second-order wave equations. *SIAM J. Sci. Comput.*, 31(3):pp. 1985–2014, 2009. URL http://dx.doi.org/10.1137/070709414.

[Diehl et al., 2010]
R. DIEHL, K. BUSCH, J. NIEGEMANN. Comparison of Low-Storage Runge-Kutta Schemes for Discontinuous Galerkin Time-Domain Simulations of Maxwell's Equations. *J. Comput. Theor. Nanosci.*, 7(8):pp. 1572–1580, 2010. URL http://dx.doi.org/doi:10.1166/jctn.2010.1521.

[Douglas, 1962]
J. DOUGLAS, JR. Alternating direction methods for three space variables. *Numer. Math.*, 4:pp. 41–63, 1962. URL http://dx.doi.org/10.1007/BF01386295.

[Douglas and Gunn, 1964]
J. DOUGLAS, JR., J. E. GUNN. A general formulation of alternating direction methods. I. Parabolic and hyperbolic problems. *Numer. Math.*, 6:pp. 428–453, 1964. URL http://dx.doi.org/10.1007/BF01386093.

[Eilinghoff, 2017]
J. EILINGHOFF. *Error analysis of splitting methods for wave type equations.* Ph.D. thesis, Karlsruhe Institute of Technology, 2017. URL http://dx.doi.org/10.5445/IR/1000075070.

[Eilinghoff et al., 2018]
J. EILINGHOFF, T. JAHNKE, R. SCHNAUBELT. Error analysis of an energy preserving ADI splitting scheme for the Maxwell equation. CRC 1173-Preprint 2018/12, Karlsruhe Institute of Technology, 2018. URL http://www.waves.kit.edu/downloads/CRC1173_Preprint_2018-12.pdf.

[Eilinghoff and Schnaubelt, 2017]
J. EILINGHOFF, R. SCHNAUBELT. Error estimates in L^2 of an ADI splitting scheme for the inhomogeneous Maxwell equations. CRC 1173-Preprint 2017/32, Karlsruhe Institute of Technology, 2017. URL http://www.waves.kit.edu/downloads/CRC1173_Preprint_2017-32.pdf.

[Eilinghoff and Schnaubelt, 2018]
J. EILINGHOFF, R. SCHNAUBELT. Error analysis of an ADI splitting scheme for the inhomogeneous Maxwell equations. *Discrete Contin. Dyn. Syst. Ser. A*, 38(11):pp. 5685–5709, 2018. URL http://dx.doi.org/710.3934/dcds.2018248.

[Engel and Nagel, 2000]
K.-J. ENGEL, R. NAGEL. *One-parameter semigroups for linear evolution equations*, vol. 194 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2000. URL http://dx.doi.org/10.1007/b97696.

[Ern and Guermond, 2004]
A. ERN, J.-L. GUERMOND. *Theory and practice of finite elements*, vol. 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004. URL http://dx.doi.org/10.1007/978-1-4757-4355-5.

[Ern and Guermond, 2006a]
A. ERN, J.-L. GUERMOND. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2):pp. 753–778, 2006a. URL http://dx.doi.org/10.1137/050624133.

[Ern and Guermond, 2006b]
A. ERN, J.-L. GUERMOND. Discontinuous Galerkin methods for Friedrichs' systems. II. Second-order elliptic PDEs. *SIAM J. Numer. Anal.*, 44(6):pp. 2363–2388, 2006b. URL http://dx.doi.org/10.1137/05063831X.

[Ern and Guermond, 2008]
A. ERN, J.-L. GUERMOND. Discontinuous Galerkin methods for Friedrichs' systems. III. Multifield theories with partial coercivity. *SIAM J. Numer. Anal.*, 46(2):pp. 776–804, 2008. URL http://dx.doi.org/10.1137/060664045.

[Ern et al., 2007]
A. ERN, J.-L. GUERMOND, G. CAPLAIN. An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems. *Comm. Partial Differential Equations*, 32(1-3):pp. 317–341, 2007. URL http://dx.doi.org/10.1080/03605300600718545.

[Fahs, 2009]
H. FAHS. High-Order Leap-Frog Based Discontinuous Galerkin Method for the Time-Domain Maxwell Equations on Non-Conforming Simplicial Meshes. *Numerical mathematics*, 2(3):pp. 275 – 300, 2009. URL https://hal.archives-ouvertes.fr/hal-00600467.

[Faragó et al., 2005]
I. FARAGÓ, R. HORVÁTH, W. H. SCHILDERS. Investigation of numerical time-integrations of Maxwell's equations using the staggered grid spatial discretization. *Int. J. of Numer. Model. Electron. Netw. Dev. Fields*, 18(2):pp. 149–169, 2005. URL http://dx.doi.org/10.1002/jnm.570.

[Friedrichs, 1958]
K. O. FRIEDRICHS. Symmetric positive linear differential equations. *Comm. Pure Appl. Math.*, 11:pp. 333–418, 1958. URL http://dx.doi.org/10.1002/cpa.3160110306.

[Gao et al., 2013]
L. GAO, X. LI, W. CHEN. New energy identities and super convergence analysis of the energy conserved splitting FDTD methods for 3D Maxwell's equations. *Math. Methods Appl. Sci.*, 36(4):pp. 440–455, 2013. URL http://dx.doi.org/10.1002/mma.2605.

[Gao et al., 2007]
L. GAO, B. ZHANG, D. LIANG. The splitting finite-difference time-domain methods for Maxwell's equations in two dimensions. *J. Comput. Appl. Math.*, 205(1):pp. 207–230, 2007. URL http://dx.doi.org/10.1016/j.cam.2006.04.051.

[Grote and Mitkova, 2010]
M. J. GROTE, T. MITKOVA. Explicit local time-stepping methods for Maxwell's equations. *J. Comput. Appl. Math.*, 234(12):pp. 3283–3302, 2010. URL http://dx.doi.org/10.1016/j.cam.2010.04.028.

[Hairer et al., 2006]
E. HAIRER, C. LUBICH, G. WANNER. *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, vol. 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second ed., 2006. URL http://dx.doi.org/10.1007/3-540-30666-8.

[Hairer and Wanner, 1996]
E. HAIRER, G. WANNER. *Solving ordinary differential equations II: Stiff and differential-algebraic problems*, vol. 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second ed., 1996. URL http://dx.doi.org/10.1007/978-3-642-05221-7.

[Hansen and Henningsson, 2013]
E. HANSEN, E. HENNINGSSON. A convergence analysis of the Peaceman-Rachford scheme for semilinear evolution equations. *SIAM J. Numer. Anal.*, 51(4):pp. 1900–1910, 2013. URL http://dx.doi.org/10.1137/120890570.

[Hansen and Henningsson, 2016]
E. HANSEN, E. HENNINGSSON. A full space-time convergence order analysis of operator splittings for linear dissipative evolution equations. *Commun. Comput. Phys.*, 19(5):pp. 1302–1316, 2016. URL http://dx.doi.org/10.4208/cicp.scpde14.22s.

[Hansen and Ostermann, 2008]
E. HANSEN, A. OSTERMANN. Dimension splitting for evolution equations. *Numer. Math.*, 108(4):pp. 557–570, 2008. URL http://dx.doi.org/10.1007/s00211-007-0129-3.

[Hesthaven and Warburton, 2008]
J. S. HESTHAVEN, T. WARBURTON. *Nodal discontinuous Galerkin methods. Algorithms, analysis, and applications*, vol. 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. URL http://dx.doi.org/10.1007/978-0-387-72067-8.

[Hipp et al., 2018]
D. HIPP, M. HOCHBRUCK, C. STOHRER. Unified error analysis for nonconforming space discretizations of wave-type equations. *IMA J. Numer. Anal.*, 2018. URL http://dx.doi.org/10.1093/imanum/dry036.

[Hochbruck, 2015]
M. HOCHBRUCK. *Skriptum zur Vorlesung Numerik I, II und Numerik von Differentialgleichungen*. WS 2013/14 – WS 2014/15. Arbeitsgruppe Numerik, Institut für Angewandte und Numerische Mathematik, Karlsruher Institut für Technologie, 2013–2015. URL https://na.math.kit.edu/download/teaching/2014w/nummetdgl/skript/skript.pdf.

[Hochbruck et al., 2015a]
M. HOCHBRUCK, T. JAHNKE, R. SCHNAUBELT. Convergence of an ADI splitting for Maxwell's equations. *Numer. Math.*, 129(3):pp. 535–561, 2015a. URL http://dx.doi.org/10.1007/s00211-014-0642-0.

[Hochbruck and Köhler, 2019]
M. HOCHBRUCK, J. KÖHLER. On the efficiency of the Peaceman–Rachford ADI-dG method for wave-type problems. In F. RADU, K. KUMAR, I. BERRE, J. NORDBOTTEN, I. POP (editors), *Numerical Mathematics and Advanced Applications ENUMATH 2017*, vol. 126 of *Lecture Notes in Computational Science and Engineering*, pp. 135–144. Springer International Publishing, 2019. URL http://dx.doi.org/10.1007/978-3-319-96415-7.

[Hochbruck et al., 2015b]
M. HOCHBRUCK, T. PAŽUR, A. SCHULZ, E. THAWINAN, C. WIENERS. Efficient time integration for discontinuous Galerkin approximations of linear wave equations. *ZAMM*, 95(3):pp. 237–259, 2015b. URL http://dx.doi.org/10.1002/zamm.201300306.

[Hochbruck and Sturm, 2016]
M. HOCHBRUCK, A. STURM. Error analysis of a second-order locally implicit method for linear Maxwell's equations. *SIAM J. Numer. Anal.*, 54(5):pp. 3167–3191, 2016. URL http://dx.doi.org/10.1137/15M1038037.

[Hochbruck and Sturm, 2018]
M. HOCHBRUCK, A. STURM. Upwind discontinuous Galerkin space discretization and locally implicit time integration for linear Maxwell's equations. *Math. Comp.*, online first, 2018. URL http://dx.doi.org/10.1090/mcom/3365.

[Hundsdorfer and Verwer, 2003]
W. HUNDSDORFER, J. VERWER. *Numerical solution of time-dependent advection-diffusion-reaction equations*, vol. 33 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2003. URL http://dx.doi.org/10.1007/978-3-662-09017-6.

[Hundsdorfer and Verwer, 1989]
W. H. HUNDSDORFER, J. G. VERWER. Stability and convergence of the Peaceman-Rachford ADI method for initial-boundary value problems. *Math. Comp.*, 53(187):pp. 81–101, 1989. URL http://dx.doi.org/10.2307/2008350.

[Jacob and Zwart, 2012]
B. JACOB, H. J. ZWART. *Linear port-Hamiltonian systems on infinite-dimensional spaces*, vol. 223 of *Operator Theory: Advances and Applications*. Birkhäuser/Springer Basel AG, Basel, 2012. URL http://dx.doi.org/10.1007/978-3-0348-0399-1.

[Jensen, 2004]
M. JENSEN. *Discontinuous Galerkin Methods for Friedrichs Systems with Irregular Solutions*. Ph.D. thesis, University of Oxford, 2004. URL http://sro.sussex.ac.uk/45497/.

[Lee and Fornberg, 2003]
J. LEE, B. FORNBERG. A split step approach for the 3-D Maxwell's equations. *J. Comput. Appl. Math.*, 158(2):pp. 485–505, 2003. URL http://dx.doi.org/10.1016/S0377-0427(03)00484-9.

[Lee and Fornberg, 2004]
J. LEE, B. FORNBERG. Some unconditionally stable time stepping methods for the 3D Maxwell's equations. *J. Comput. Appl. Math.*, 166(2):pp. 497–523, 2004. URL http://dx.doi.org/10.1016/j.cam.2003.09.001.

[Monk, 2003]
P. MONK. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003. URL http://dx.doi.org/10.1093/acprof:oso/9780198508885.001.0001.

[Namiki, 1999]
T. NAMIKI. A new FDTD algorithm based on alternating-direction implicit method. *IEEE Trans. Microw. Theory Tech.*, 47(10):pp. 2003–2007, 1999. URL http://dx.doi.org/10.1109/22.795075.

[Pažur, 2013]
T. PAŽUR. *Error analysis of implicit and exponential time integration of linear Maxwell's equations*. Ph.D. thesis, Karlsruhe Institute of Technology, 2013. URL https://publikationen.bibliothek.kit.edu/1000038617.

[Pazy, 1983]
A. PAZY. *Semigroups of Linear Operators and Applications to Partial Differential Equations*, vol. 44 of *Applied Mathematical Sciences*. Springer, New York, 1983. URL http://dx.doi.org/10.1007/978-1-4612-5561-1.

[Peaceman and Rachford, 1955]
D. W. PEACEMAN, H. H. RACHFORD, JR. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Indust. Appl. Math.*, 3:pp. 28–41, 1955. URL http://dx.doi.org/10.1137/0103003.

[Pearcy, 1962]
C. PEARCY. On convergence of alternating direction procedures. *Numer. Math.*, 4:pp. 172–176, 1962. URL http://dx.doi.org/10.1007/BF01386310.

[Phillips, 1959]
R. S. PHILLIPS. Dissipative operators and hyperbolic systems of partial differential equations. *Trans. Amer. Math. Soc.*, 90:pp. 193–254, 1959. URL http://dx.doi.org/10.2307/1993202.

[Piperno, 2006]
S. PIPERNO. Symplectic local time-stepping in non-dissipative DGTD methods applied to wave propagation problems. *M2AN Math. Model. Numer. Anal.*, 40(5):pp. 815–841, 2006. URL http://dx.doi.org/10.1051/m2an:2006035.

[Schnaubelt, 2011]
R. SCHNAUBELT. *Lecture notes "Evolution Equations"*. WS 2010/11. Workgroup Functional Analysis, Institute for Analysis, Karlsruhe Institute of Technology, 2010–2011. URL http://www.math.kit.edu/iana3/~schnaubelt/media/evgl-skript.pdf.

[Schnaubelt, 2013]
R. SCHNAUBELT. *Lecture notes "Operator Semigroups and Dispersive Equations"*. WS 2012/13. Workgroup Functional Analysis, Institute for Analysis, Karlsruhe Institute of Technology, 2012–2013. URL http://www.math.kit.edu/iana3/~schnaubelt/media/isem16-skript.pdf.

[Schnaubelt, 2015]
R. SCHNAUBELT. *Lecture notes "Spectral Theory"*. SS 2015. Workgroup Functional Analysis, Institute for Analysis, Karlsruhe Institute of Technology, 2015. URL http://www.math.kit.edu/iana3/~schnaubelt/media/st-skript15.pdf.

[Sturm, 2017]
A. STURM. *Locally Implicit Time Integration for Linear Maxwell's Equations*. Ph.D. thesis, Karlsruhe Institute of Technology, 2017. URL http://dx.doi.org/10.5445/IR/1000069341.

[Verwer, 2011]
J. G. VERWER. Component splitting for semi-discrete Maxwell equations. *BIT*, 51(2):pp. 427–445, 2011. URL http://dx.doi.org/10.1007/s10543-010-0296-y.

[Zhen et al., 2000]
F. ZHEN, Z. CHEN, J. ZHANG. Toward the development of a three-dimensional unconditionally stable finite-difference time-domain method. *IEEE Trans. Microw. Theory Tech.*, 48(9):pp. 1550–1558, 2000. URL http://dx.doi.org/10.1109/22.869007.

List of Constants

Let

- $\Omega \subset \mathbb{R}^d$ be a bounded, open and connected Lipschitz domain with boundary $\Gamma = \partial\Omega$,
- $K \subset \Omega$ be an open subset of Ω,
- $p \in \mathbb{Z}$,
- $\mathcal{F}$ be a Friedrichs' operator with coefficients $(F_i)_{i=0}^d$ and dissipative boundary condition $\mathcal{F}_\Gamma$,
- $\mathfrak{T}_\mathcal{H}$ be a shape- and contact-regular sequence of meshes discretizing Ω,
- $\mathfrak{T}_h \in \mathfrak{T}_\mathcal{H}$ be a general mesh of Ω,
- $\widetilde{\mathscr{A}}$ and $\widetilde{\mathscr{B}}$ be the split operators used in Theorem 5.15.

Then the constants used throughout this thesis are given as follows.

Constant	Definition	Page
$\widetilde{C}_{1,K,\mathcal{F}}$	$\max\limits_{i=0,\dots,d}\{\max\limits_{j=1,\dots,d}\|\partial_j F_i\|_{\infty,K}, \|F_i\|_{\infty,K}\}$	16
$C_{1,K,\mathcal{F}}$	$\sqrt{2}\,(d+1)\widetilde{C}_{1,K,\mathcal{F}}$	16, 76, 77
$C_{\mathcal{F},K}$	$\max\limits_{i=0,\dots,d}\|F_i\|_{\infty,K}$	16, 17
N_∂	$\max_{\widehat{K}\in\mathfrak{T}_h}\|\mathcal{F}_h^{\widehat{K}}\|$	31, 32, 44, 45, 46, 47, 114
ρ_1	Shape-regularity parameter of $\mathfrak{T}_\mathcal{H}$	32, 34, 35
ρ_2	Contact-regularity parameter of $\mathfrak{T}_\mathcal{H}$	32, 34, 35
ρ	$\rho_1\rho_2$	32, 33, 44, 45, 46, 47, 114
C'_{inv}	Inverse inequality constant of ∇	34
C_{inv}	$\sqrt{d}\,C'_{\mathrm{inv}} + 1$	34, 44, 46, 47, 114

Constant	Definition	Page
$C_{\mathcal{F}}$	$\max_{i=0,\dots,d} \|F_i\|_{\infty,\Omega}$	34, 35, 44, 45, 46, 47, 48, 114
C_{tr}	Discrete trace inequality constant	35, 44, 45, 46, 47, 114
C'_{app}	Optimal polynomial approximation constant (Interpolation)	35
C_{app}	Optimal polynomial approximation constant (Projection)	36, 46, 47, 50, 114
$C_{\mathrm{app},\partial}$	Optimal polynomial approximation constant on faces (Projection)	36, 46, 47, 114
$C_{\Gamma,\mathcal{F}}$	$\max_{F\in\mathcal{F}_h^{\mathrm{bnd}}} \|\mathcal{F}_\partial^F - \mathcal{F}_\Gamma\|_{\infty,F}$	44, 46, 47, 114
$C_{\mathcal{F},\mathrm{el},1}$	$C_{\mathcal{F}}C_{\mathrm{inv}} + \frac{1}{2}{C_{\mathrm{tr}}}^2(C_{\Gamma,\mathcal{F}} + N_\partial C_{\mathcal{F}})$	44, 45, 114
$C_{\mathcal{F},\mathrm{el},2}$	$\frac{1}{2}\rho^{1/2}C_{\mathcal{F}}{C_{\mathrm{tr}}}^2$	44, 45, 114
$C_{\mathrm{inv},\mathcal{F},p}$	$C_{\mathcal{F},\mathrm{el},1} + \rho^p N_\partial C_{\mathcal{F},\mathrm{el},2}$	45, 48
$C_{\mathrm{inv},\mathcal{F}}$	$C_{\mathrm{inv},\mathcal{F},0}$	45, 76, 77, 78, 114
$C_{\pi,\mathcal{F},p}$	$C_{\mathcal{F}}C_{\mathrm{app}}C_{\mathrm{inv}} + \frac{1}{2}N_\partial C_{\mathrm{tr}}C_{\mathrm{app},\partial}(C_{\Gamma,\mathcal{F}} + C_{\mathcal{F}} + \rho^{p+1/2}C_{\mathcal{F}})$	46, 48
$C_{\pi,\mathcal{F}}$	$C_{\pi,\mathcal{F},0}$	46, 51, 52, 73, 76, 77, 78, 114
$\widetilde{C}_{\mathcal{F},\mathrm{el},1}$	$C_{\mathcal{F}}C_{\mathrm{app}}C_{\mathrm{inv}} + \frac{1}{2}N_\partial C_{\mathrm{tr}}C_{\mathrm{app},\partial}(C_{\Gamma,\mathcal{F}} + C_{\mathcal{F}})$	47
$\widetilde{C}_{\mathcal{F},\mathrm{el},2}$	$\frac{1}{2}\rho^{1/2}C_{\mathcal{F}}C_{\mathrm{tr}}C_{\mathrm{app},\partial}$	47
$C_{\mathrm{app},M}$	$\|M\|_{\infty,\Omega}^{1/2}\, C_{\mathrm{app}}$	50, 51, 78, 79
$C_{\pi,\mathcal{F},M}$	$\|M^{-1}\|_{\infty,\Omega}^{1/2}\, C_{\pi,\mathcal{F}}$	51, 73, 78, 79
$C_{1,\mathcal{F}}$	$\max_{\widehat{K}\in\mathcal{T}_h} C_{1,\widehat{K},\mathcal{F}}$	76, 78, 114
$C_{\mathrm{PR},u}$	$\|M^{-1}\|_{\infty,\Omega}^{3/2}(C_{\mathrm{inv},\widetilde{\mathcal{A}}}C_{\pi,\widetilde{\mathcal{B}},-1} + C_{\pi,\widetilde{\mathcal{A}}}C_{1,\widetilde{\mathcal{B}}})$	76, 79
$C_{\mathrm{PR},f}$	$\|M^{-1}\|_{\infty,\Omega}^{3/2}(C_{\mathrm{inv},\widetilde{\mathcal{A}}}C_{\pi,\widetilde{\mathcal{A}},-1} + C_{\pi,\widetilde{\mathcal{A}}}C_{1,\widetilde{\mathcal{A}}})$	78, 79