A unified error analysis for the numerical solution of nonlinear wave-type equations with application to kinetic boundary conditions

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

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

DISSERTATION

von

Jan Leibold

Tag der mündlichen Prüfung: 24. Februar 2021

1. Referentin: Prof. Dr. Marlis Hochbruck
2. Referent: Prof. Dr. Roland Schnaubelt
3. Referent: Prof. Dr. Christian Lubich

Acknowledgement

I gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 258734477 – SFB 1173.

I continue in German.

Ich möchte mich im Folgenden bei all jenen bedanken, die zum erfolgreichen Gelingen meiner Promotion beigetragen haben.

Zuallererst gilt mein Dank meiner Betreuerin Prof. Dr. Marlis Hochbruck. Bereits während meines Bachelorstudiums haben ihre Vorlesung mit dazu beigetragen, in mir das Interesse für numerische Mathematik zu wecken. Während des Masterstudiums ermöglichte sie es mir dann, als Hiwi erste Einblicke in mathematische Forschung zu erhalten. Durch ihr Vertrauen in meine Fähigkeiten habe ich eine Promotion überhaupt erst in Erwägung gezogen. Ihre klare Art, Mathematik zu praktizieren und zu vermitteln, sowie ihre pragmatische Art, Entscheidungen zu treffen, haben mich nachhaltig geprägt und mir in vielen Phasen der Promotion geholfen.

Weiterhin danke ich meinem Zweitbetreuer Prof. Dr. Roland Schnaubelt, an den ich mich stets mit analytischen Fragen wenden konnte. Allein zu wissen, dass es mit ihm jemanden gibt, der im Falle eines analytischen Problems helfen kann, war sehr beruhigend.

Bei Prof. Dr. Christian Lubich bedanke ich mich herzlich für sein Gutachten und die gute Zusammenarbeit im Rahmen des Sonderforschungsbereichs.

Ich bedanke mich bei allen ehemaligen und aktuellen Mitgliedern der erweiterten AG Numerik für das gute Miteinander bei und abseits der Arbeit. Nicht missen möchte ich die gemeinsamen Kompaktseminare, die jedes Jahr ein Highlight für mich waren. Hervorzuheben sind auch Laurette, Christian und Mathias, an die ich mich jederzeit mit technischen oder organisatorischen Fragen wenden konnte und die immer alles dafür tun, dass die restlichen Mitglieder der Arbeitsgruppe möglichst reibungslos arbeiten können.

Ein besonderer Dank gilt auch David. Er hat mit seiner Forschung das Fundament für meine Arbeit gelegt und stand mir zu Beginn meiner Promotionszeit mit Rat und Tat beim Einstieg in die Forschung zur Seite.

Bei allen ehemaligen und aktuellen Teilnehmern der täglichen Kaffeerunde bedanke ich mich dafür, dass sie den Start eines jeden Arbeitstages etwas angenehmer gemacht haben. Insbesondere an langen Homeoffice-Tagen im letzten Jahr war die Kaffeerunde immer ein kleiner Lichtblick.

Ein letzter Dank geht an Benjamin, Bernhard und Marc für das Korrekturlesen meiner Dissertation. Mit vielen hilfreichen fachlichen sowie sprachlichen Anmerkungen haben sie unmittelbar zur Qualität dieser Arbeit beigetragen.

Abstract

In this thesis, a unified error analysis for discretizations of nonlinear first- and second-order wave-type equations is provided. For this, the wave equations as well as their space discretizations are considered as nonlinear evolution equations in Hilbert spaces. The space discretizations are supplemented with Runge–Kutta time discretizations. By employing stability properties of monotone operators, abstract error bounds for the space, time, and full discretizations are derived.

Further, for semilinear second-order wave-type equations, an implicit-explicit time integration scheme is presented. This scheme only requires the solution of a linear system of equations in each time step and it is stable under a step size restriction only depending on the nonlinearity. It is proven that the scheme converges with second order in time and in combination with the abstract space discretization of the unified error analysis, corresponding full discretization error bounds are derived.

The abstract results are used to derive convergence rates for an isoparametric finite element space discretization of a wave equation with kinetic boundary conditions and nonlinear forcing and damping terms. For the combination of the finite element discretization with Runge–Kutta methods or the implicit-explicit scheme, respectively, error bounds of the resulting fully discrete schemes are proven. The theoretical results are illustrated by numerical experiments.

Contents

CHAPTER 1

Introduction

Motivation

Wave equations are fundamental models in physics that describe the propagation of various types of waves. One example is the acoustic wave equation which models the propagation of sound waves or the vibration of a membrane. If the wave propagation is modeled in a bounded domain, further conditions have to be imposed on the boundary. Usually, these boundary conditions simply prescribe the value or derivatives of the solution at the boundary. In the case of a vibrating membrane, prescribing the value would simply model that the membrane is fixed at the boundary.

In contrast to standard boundary conditions, kinetic boundary conditions model the propagation of the wave on the boundary. For example, they can be derived by considering a vibrating membrane which is not fixed, but where its boundary carries a mass density and is subject to linear tension (cf. [Goldstein, 2006, Section 5]). This leads a wave equation in the interior domain coupled to a wave equation on the boundary. Kinetic boundary conditions also can serve as an effective model for the interaction of an acoustic wave with a thin boundary layer with distinctive elastic or damping properties, and where the wave length is large compared to the width of the boundary layer (cf. [Nicaise, 2017, Section 3.2]).

Analysis and numerics of wave equations with kinetic boundary conditions have developed significantly in recent years. The analysis of the continuous problem already includes problems with nonlinear damping and forcing (cf. Vitillaro [2013, 2017]). However, to our knowledge, there are only results for the numerical analysis of linear (cf. Hipp [2017]) and semilinear problems (cf. Hochbruck and Leibold [2020]), so far.

In this thesis, we aim at proving error bounds for suitable space and time discretizations of wave equations with kinetic boundary conditions including nonlinear forcing and damping terms.

Thereby, two main difficulties arise:

- **Nonlinearities:** The nonlinear forcing and damping terms appearing in the equations render the numerical discretization as well as the error analysis and the implementation much more involved: One has to use suitable space and time discretization schemes that preserve properties of the nonlinear operators which yield stability of the equations. Furthermore, the error analysis of space and time discretization schemes involves nonlinear error terms that have to be bounded. In addition, to run a numerical method, nonlinear systems of equations have to be solved, which makes the implementation more involved and is computationally expensive.
- **Non-conforming space discretizations:** Equations with dynamic boundary conditions are usually posed on domains with smooth and possibly curved boundaries. Hence, numerical schemes as, e.g., finite element schemes, have to approximate the boundary of the domain which renders the discretization non-conforming. This significantly complicates the error analysis of the spatial discretization.

The main goal of this thesis is to present tools and techniques to tackle these difficulties in a systematic way. In the following, we outline our main contributions.

Unified error analysis

To analyze non-conforming space discretizations in a systematic way, a unified error analysis for first- and second-order linear wave-type equations was introduced in Hipp [2017] and Hipp et al. [2019]. Thereby, the differential equations as well as their space discretizations are considered in a framework of abstract linear evolution equations in Hilbert spaces. The authors employ stability bounds from semigroup theory to derive abstract error bounds in terms of interpolation, data, and conformity errors of the method. These abstract error bounds can then be used to derive convergence rates for a large class of problems in a simple, systematic and modular way by plugging in approximation properties of the corresponding space discretization. Thus, the main advantage of this unified approach is that one does not have to perform the error proof from scratch for every single problem, but gains precise insights into which terms have to be bounded. In particular, this applies to errors caused by the non-conforming space discretizations. For instance, in Hipp [2017], the unified error analysis was used to prove error bounds for finite element discretizations of wave equations with various types of boundary conditions on smooth domains and for discontinuous Galerkin discretizations of Maxwell equations. Further, it was used in Hochbruck et al. [2019] to prove error bounds for a heterogeneous multiscale method for linear Maxwell equations.

A first step towards nonlinear problems was made in the Master's thesis Leibold [2017], where the unified error analysis was extended to semilinear problems. To consider more general nonlinear problems, we extend the framework of the unified error analysis to abstract space discretizations of nonlinear evolution equations with maximal (quasi-)monotone operators. Using analytical stability properties of such evolution equation, we derive abstract error bounds in this nonlinear setting. Thus, our analysis shows which properties must be preserved when discretizing the nonlinearities and which nonlinear error terms must be estimated to derive convergence rates for specific examples.

It is also possible to combine the abstract space discretizations with time discretization schemes and to derive corresponding abstract time and full discretization error bounds. In this thesis, we show this exemplary for algebraically stable and coercive Runge–Kutta methods based on an error analysis from Hansen [2006b], where time discretization errors were analyzed in a similar framework as the one we use here. The combination with the unified error analysis has again the advantage that these results apply to all equations that fit into the framework.

We are not aware of other results that allow one to analyze non-conforming space and full discretizations of wave-type equations that involve nonlinear damping terms. Nevertheless, we should mention the following works, which go in the same direction. In Emmrich et al. [2015], an abstract full discretization in a framework similar to the one used in this thesis was considered. However, only a conforming space discretization was analyzed and no error bounds but only weak convergence of the discretization was shown. A related framework for quasilinear equations was introduced in Hochbruck and Maier [2021], Maier [2020], covering both quasilinear wave and Maxwell equations. But the error analysis in this work relies on properties of quasilinear operators that cannot be used for nonlinear damping terms.

Efficient time integration via implicit-explicit (IMEX) schemes

The time integration of nonlinear problems suffers from the necessity that, in general, nonlinear systems of equations have to be solved in each time step. IMEX time integration schemes overcome this problem in the case of semilinear equations. By semilinear we mean that the equation can be splitted into an unbounded, stiff linear and a (locally) Lipschitz continuous, nonstiff nonlinear part. In this case, the idea of IMEX schemes is to integrate the stiff linear part implicitly while the non-stiff part is treated explicitly. If done properly, this leads to stable schemes, where the time step size is only restricted by the Lipschitz constant of the nonlinearity. In particular, they do not suffer from a CFL condition when they are applied to spatial discretizations of partial differential equations. Additionally, the schemes are efficient, since due to the explicit treatment of the nonlinear part, only linear systems of equations have to be solved.

IMEX schemes are widely used in applications, e.g., in structural dynamics and fluid-structure interaction (van Zuijlen and Bijl [2005]), hydrodynamics (Kadioglu et al. [2010]), sea-ice dynamics (Lemieux et al. [2014]), or atmospheric dynamics (Gardner et al. [2018]).

There is a rich literature on IMEX schemes for first-order equations, in particular, there is a well-developed theory for IMEX multistep schemes (Akrivis et al. [1999], Ascher et al. [1995], Frank et al. [1997], Hundsdorfer and Ruuth [2007]) or IMEX Runge–Kutta schemes (Ascher et al. [1997], Boscarino [2007]). In Boscarino [2007], Hundsdorfer and Ruuth [2007], an error analysis for ordinary differential equations is presented, while Akrivis et al. [1999] contains discretization errors for IMEX schemes applied to conformal space discretizations of quasilinear parabolic evolution equations.

In contrast, there are only few articles on IMEX schemes that take advantage of the special structure of second-order equations. We refer to Stern and Grinspun [2009], Zhang and Skeel [1997], where IMEX schemes for undamped second-order ordinary differential equations are considered.

In this thesis, we propose a very efficient IMEX scheme which is tailor-made for semilinear second-order differential equations including linear damping terms. It is a combination of the explicit leapfrog method and the implicit Crank–Nicolson scheme.

We show that our scheme is unconditionally stable in the sense that the time step size is only restricted by the Lipschitz constant of the explicitly treated nonlinearity. Further, we prove a second-order error bound. We then combine the scheme with the abstract space discretization of the unified error analysis and prove an error bound that is second-order in time and contains the abstract space-discretization errors of the unified error analysis. This result allows us to derive full discretization error bounds for specific equations and space discretizations. To our knowledge, such a general and rigorous full discretization error analysis for IMEX schemes has also not been considered in the literature so far.

We emphasize that, although there already exists a so-called Crank–Nicolson-leapfrog IMEX scheme which is obtained from a combination of the Crank–Nicolson and the leapfrog scheme for first-order equations (cf. Layton and Trenchea [2012], Layton et al. [2016], and references therein), this scheme is not equivalent to the scheme we present in this thesis. This is due to the fact that the leapfrog schemes for first- and second-order equations are not equivalent and indeed have completely different stability properties. More precisely, the Crank–Nicolson-leapfrog scheme is only stable if the explicitly treated part is linear and skew symmetric, which is not the case in our setting.

The results on the IMEX scheme contained in this dissertation were already published in the paper Hochbruck and Leibold [2021].

Numerical analysis of a wave equation with kinetic boundary conditions

As an application of our abstract theory we study wave equations with kinetic boundary conditions. In Hipp [2017], based on the bulk-surface finite element method from Elliott and Ranner [2013], a non-conforming isoparametric finite element space discretization was introduced for linear wave equations with kinetic boundary conditions. In this thesis, we extend the discretization to nonlinear forcing and damping terms. Then, by using the results of the unified error analysis, we derive novel convergence rates for the space discretization and full discretization error bounds for suitable Runge–Kutta methods.

Further, we apply our IMEX scheme to the wave equation in the semilinear case and combine it with the finite element discretization. This yields a very efficient fully discrete scheme for which we also obtain error bounds using the abstract results in the framework of the unified error analysis.

Finally, we illustrate our theoretical results with some numerical experiments.

Conclusion

The main contributions of this thesis are:

- A unified error analysis for space and time discretizations of nonlinear first- and second-order wave-type equations in a quite general framework of nonlinear evolution equations with maximal monotone operators.

- A novel, efficient IMEX time integration scheme for semilinear second-order wave equations and its stability and error analysis in the general framework of the unified error analysis.
- The numerical analysis of the wave equation with kinetic boundary conditions and nonlinear forcing and damping terms, consisting of an isoparametric finite element space discretization, Runge–Kutta or IMEX time discretization, and a full discretization error analysis.

Outline

This thesis is organized as follows. We introduce the unified error analysis for first- and second-order nonlinear wave-type equations in Chapter 2. More precisely, we define the analytical framework, introduce general non-conforming space discretizations, and perform the abstract space discretization error analysis.

In Chapter 3, we analyze Runge–Kutta methods applied to first- and second-order nonlinear wave-type equations. We prove wellposedness and time discretization error bounds.

Chapter 4 is devoted to an IMEX scheme for second-order semilinear evolution equations. We present the construction of the scheme, show stability, and prove a second-order error bound.

In Chapter 5, the time discretization error analyses of the Runge–Kutta methods and the IMEX scheme are combined with the unified space discretization error analysis to obtain abstract full discretization error bounds for first- and second-order wave-type equations.

Finally, we consider the numerical analysis of the wave equation with kinetic boundary conditions in Chapter 6. We introduce an isoparametric finite element space discretization and prove space, time and full discretization error bounds by employing the abstract theory from Chapters 2 to 5. Furthermore, we illustrate the theoretical results with some numerical experiments.

Preliminaries

Here, we introduce some conventions and notion that we use throughout this thesis.

Hilbert spaces Let X be a real Hilbert space with norm $\|\cdot\|_X$. If p is a scalar product on X we use the notation (X, p) for X equipped with the scalar product p, i.e., we then have

$$\big(\cdot,\cdot\big)_X = p\big(\cdot,\cdot\big).$$

We further use the notation $\|\cdot\|_p^2 := p\big(\cdot,\cdot\big)$ for the norm induced by p.

We denote the dual space of X by X^* and the corresponding dual pairing by

$$\langle \phi, x\rangle_{X^*\times X} := \phi(x), \qquad \phi \in X^*, x \in X.$$

Let Y be another real Hilbert space. For elements in the product space $X \times Y$ we use the notation

$$[x, y]^\intercal = \begin{bmatrix} x \\ y \end{bmatrix} \in X \times Y.$$

Linear operators We denote by $\mathcal{L}(X; Y)$ the space of all bounded linear operators from X to Y and endow $\mathcal{L}(X; Y)$ with the norm

$$\|B\|_{Y\leftarrow X} := \sup_{\|x\|_X\neq 0} \frac{\|Bx\|_Y}{\|x\|_X}, \qquad B \in \mathcal{L}(X; Y).$$

By $D(A) \subset X$, we denote the domain of an operator $A\colon D(A) \to X$ on X.

The identity operator on X is denoted by I. We refrain from adding an additional X to the notation since the space should be clear from the context in each case.

Hilbert space valued functions By $C([0, T]; X)$ we denote the space of all continuous functions and by $C^k([0, T]; X), k \in \mathbb{N}$, the spaces of all k-times (Fréchet) differentiable functions from a time interval $[0, T], T > 0$, to X.

The Hilbert space valued Lebesgue spaces $L^q([0, T]; X), q \in [1, \infty]$, consist of all measurable functions $f\colon [0, T] \to X$ with

$$\infty > \|f\|_{L^q([0,T];X)} := \begin{cases} \big(\int_0^T |f(t)|^q \,\mathrm{d}t\big)^{1/q}, & q < \infty, \\ \operatorname{ess\,sup}_{t\in[0,T]} |f(t)|, & q = \infty. \end{cases} \tag{1.1}$$

The spaces $L^q_{loc}([0, \infty); X)$ contain all measurable functions $f\colon [0, \infty) \to X$ which satisfy (1.1) for all $0 < T < \infty$.

We make also use of the Hilbert space valued Sobolev spaces

$$W^{1,q}([0, T]; X) := \big\{f\colon [0, T] \to X \mid f(t) = f(0) + \int_0^t g(s)\,\mathrm{d}s,\ g \in L^q([0, T]; X)\big\} \qquad \text{and}$$
$$W^{1,q}_{loc}([0, \infty); X) := \big\{f\colon [0, \infty) \to X \mid f\big|_{[0,\,T]} \in W^{1,q}([0, T]; X) \text{ for all } T > 0\big\}.$$

Differential equations Time dependent differential equations in this thesis are usually posed on the infinite time interval $[0, \infty)$. We consider solutions of differential equations on compact time intervals $[0, T]$ for some $T \in (0, t^*)$, where $t^* \in (0, \infty]$ denotes the maximal existence time of the solution.

Partial derivatives Derivatives in this thesis are always understood in the sense of distributions.

Let $\Omega \subset \mathbb{R}^d, d \in \mathbb{N}$, be an open and bounded domain and $T > 0$. For a function $u\colon [0, T] \times \Omega \to \mathbb{R}$, we denote the temporal derivative by u_t. By ∇u and Δu, we denote the gradient and the Laplacian of u w.r.t. the spatial variables.

For a function $f\colon \Omega \times \mathbb{R}$, we use the notation $\partial_2 f$ for the derivative of f w.r.t. the second variable, i.e., $\partial_2 f(\mathbf{x}, \xi) = \partial_\xi f(\mathbf{x}, \xi)$.

Lebesgue and Sobolev spaces Let Ω have a Lipschitz boundary $\Gamma = \partial\Omega$. For $q \in [1, \infty]$, we denote the usual Lebesgue spaces over Ω and Γ by $L^q(\Omega)$ and $L^q(\Gamma)$, respectively. The Lebesgue measures of Ω and Γ are denoted by $\sigma(\Omega)$ and $\sigma(\Gamma)$, respectively.

The usual Sobolev spaces of order $k \in \mathbb{N}$ over Ω are denoted by $H^k(\Omega) = W^{k,2}(\Omega)$.

We also make use of the corresponding boundary Sobolev spaces $H^k(\Gamma) = W^{k,2}(\Gamma)$ which can be defined if Γ is at least C^k regular (cf. [Grisvard, 2011, Section 1.3]).

Dirichlet trace We denote the usual Dirichlet trace operator by $\gamma\colon H^1(\Omega) \to L^2(\Gamma)$ and define

$$H^1_0(\Omega) \coloneqq \{v \in H^1(\Omega) \mid \gamma(v) = 0\}.$$

Normal vector and surface integrals We denote the unit normal vector of Γ by $\mathbf{n}\colon \Gamma \to \mathbb{R}^d$ and the surface integral of a function $\varphi \in L^1(\Gamma)$ by

$$\int_\Gamma \varphi \,\mathrm{d}s.$$

Surface differential operators Let Γ be C^1 regular. For the normal derivative of a function $v \in H^1(\Omega)$ we write $\partial_{\mathbf{n}} v \coloneqq \mathbf{n} \cdot \nabla v$.

Further, we define the surface gradient by

$$\nabla_\Gamma v \coloneqq (\partial_{i,\Gamma} v)_{i=1}^d \coloneqq (\mathrm{I} - \mathbf{n}\mathbf{n}^\intercal)\nabla v.$$

The differential operators $\partial_{i,\Gamma}$ and ∇_Γ can be generalized to operators $\partial_{i,\Gamma} \in \mathcal{L}(H^1(\Gamma); L^2(\Gamma))$ and $\nabla_\Gamma \in \mathcal{L}\big(H^1(\Gamma); (L^2(\Gamma))^d\big)$ by defining them in terms of local variables (cf. Kashiwabara et al. [2015] or Disser et al. [2015] for details). Now, let Γ be C^2 regular. For $v \in H^2(\Omega)$, the surface Laplacian (or Laplace–Beltrami operator) is defined via

$$\Delta_\Gamma v \coloneqq \sum_{i=1}^d \partial_{i,\Gamma}^2 v.$$

Constants In the whole thesis, C denotes a generic constant which may have different values at different occurrences.

All constants with a hat (e.g., $\widehat{c}$) appear in the context of spatial discretizations which are related to a discretization parameter h (e.g., the mesh width of a spatial grid). The hat above a constant then indicates that the constant is independent of h.

CHAPTER 2

Abstract space discretizations of first- and second-order evolution equations

In this chapter, we present our unified error analysis for first- and second-order wave-type equations. We introduce a framework to consider both the equations and the corresponding abstract non-conforming space discretizations as nonlinear evolution equations in Hilbert spaces. As main results of this chapter, we derive abstract error bounds that can be applied to all equations that fit into the framework, e.g., to wave equations with kinetic boundary conditions which we consider in Chapter 6. We are mainly interested in second-order wave-type equations with nonlinear damping. However, we start by considering first-order equations, since these are easier to analyze. We then transfer the results to the second-order case.

This work generalizes the results from Hipp et al. [2019] for the linear and from Hochbruck and Leibold [2020] for the semilinear to the nonlinear case. We closely stick to the framework used in these papers.

Outline In Section 2.1 we introduce our setting for first-order evolution equations and corresponding non-conforming space discretizations and derive abstract error bounds. Afterwards, in Section 2.2, we use these results to prove error bounds for abstract space discretizations of second-order wave-type equations with nonlinear damping. We illustrate the application of the abstract results with a basic example.

2.1 First-order evolution equation with monotone operators

2.1.1 Analytical setting

We consider the following abstract first-order nonlinear evolution equation in a Hilbert space (X, p):

$$x'(t) + \mathcal{S}(x(t)) = G(x(t)) + g(t), \qquad t \geq 0, \tag{2.1a}$$

$$x(0) = x^0 \in D(\mathcal{S}). \tag{2.1b}$$

In the following, we will suppress the t arguments in evolution equations.

Definition 2.1 (Wellposedness)**.**

a) For $T > 0$ a function $x \in W^{1,\infty}([0,T];X)$ is called ***strong solution*** *of (2.1) on $[0,T]$, if $x(t) \in D(\mathcal{S})$ for all $t \in [0,T]$, $x(0) = x^0$, and (2.1a) is satisfied for almost all $t \in [0,T]$.*

b) The evolution equation (2.1) is called ***locally wellposed****, if for every initial value there exists a maximal existence time $t^*(x^0) \in (0,\infty]$, s.t. (2.1) has a unique strong solution on $[0,T]$ for all $T < t^*(x^0)$.*

Remark 2.2. *Note that by [Showalter, 1997, Proposition III.1.1], we have that all $y \in W^{1,\infty}([0,T];X)$ are continuous and weakly differentiable with $y' \in L^\infty([0,T];X)$. Thus, all conditions in Definition 2.1 a) and especially point evaluations of strong solutions are well defined. We will use this frequently in this thesis.*

We make the following assumptions on $\mathcal{S}$, G, and g such that (2.1) is locally wellposed.

Assumption 2.3.

a) The nonlinear operator $\mathcal{S}\colon D(\mathcal{S}) \to X$ is quasi-monotone and maximal, i.e., there is a $c_{\mathrm{qm}} > 0$ s.t.

$$p\big(\mathcal{S}(y) - \mathcal{S}(z), y - z\big) \geq -c_{\mathrm{qm}} \|y - z\|_X^2 \qquad \text{for all } y, z \in D(\mathcal{S}),$$

and there exists some $\lambda > c_{\mathrm{qm}}$ s.t. $\operatorname{range}(\lambda + \mathcal{S}) = X$. Furthermore, $D(\mathcal{S})$ is dense in X.

b) The nonlinearity $G\colon X \to X$ is locally Lipschitz continuous, i.e., for all $\rho > 0$ there exists a constant L_ρ s.t. for all $y, z \in X$ with $\|y\|_X, \|z\|_X \leq \rho$ we have

$$\|G(y) - G(z)\|_X \leq L_\rho \|y - z\|_X.$$

c) The inhomogeneity satisfies $g \in W^{1,1}_{loc}([0,\infty);X)$.

Theorem 2.4. *Under Assumption 2.3 the evolution equation (2.1) is locally wellposed.*

Proof. The result is stated in [Chueshov et al., 2002, Theorem 7.2] and generalizes the classical result from [Showalter, 1997, Corollary IV.4.1], which only covers the case of global Lipschitz continuous nonlinearities

G. Since in [Chueshov et al., 2002, Theorem 7.2] the additional assumption $\mathcal{S}(0) = 0$ is made, we obtain the assertion by applying this theorem to the equivalent evolution equation

$$x' + \widetilde{\mathcal{S}}(x) = G(x) + \widetilde{g}$$

with

$$\widetilde{\mathcal{S}}(x) = \mathcal{S}(x) - \mathcal{S}(0), \qquad \widetilde{g}(t) = g(t) - \mathcal{S}(0), \quad t \geq 0.$$

□

Remark 2.5. *Note that in the semilinear setting presented in Hochbruck and Leibold [2020], the Lipschitz-continuous nonlinearity G is also allowed to depend on the time t. However, the wellposedness results we mention in the proof of Theorem 2.4 are only stated for time-independent nonlinearities. The generalization of these results to time-dependent nonlinearities is out of the scope of this thesis.*

Further, we have the following stability result which is essential for the latter error analysis.

Theorem 2.6. *Let Assumption 2.3 be satisfied and for a $T > 0$ and $i = 1, 2$ let $x_i \in W^{1,\infty}([0,T]; X)$ be strong solutions of*

$$\begin{aligned} x_i' + \mathcal{S}(x_i) &= G(x_i) + g_i, \qquad t \in [0,T], \\ x_i(0) &= x_i^0 \end{aligned}$$

with $g_i \in W^{1,1}([0,T]; X)$ and $\|x_i\|_{L^\infty(0,T;X)} \leq \rho$, for some $\rho > 0$. Then for all $t \in [0,T]$

$$\|x_1(t) - x_2(t)\|_X \leq \mathrm{e}^{(c_{\mathrm{qm}} + L_\rho)t} \left(\|x_1^0 - x_2^0\|_X + \int_0^t \|g_1(s) - g_2(s)\|_X \,\mathrm{d}s \right).$$

Proof. The result can be derived with energy estimates as done in [Showalter, 1997, Theorem IV.4.1A]. The difference $\Delta(t) = x_1(t) - x_2(t)$ is the strong solution of the evolution equation

$$\Delta' + \mathcal{S}(x_1) - \mathcal{S}(x_2) = G(x_1) - G(x_2) + g_1 - g_2, \qquad t \in [0,T], \tag{2.2a}$$

$$\Delta(0) = x_1^0 - x_2^0. \tag{2.2b}$$

The following calculations hold true almost everywhere on $[0,T]$ and derivatives are meant in the weak sense. We assume without lost of generality, that $\|\Delta(t)\|_X \neq 0$ for almost all t. Taking the inner product of (2.2a) with Δ and exploiting $p\big(\Delta', \Delta\big) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\|_X^2 = \|\Delta\|_X \frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\|_X$ yields

$$\|\Delta\|_X \frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\|_X = -p\big(\mathcal{S}(x_1) - \mathcal{S}(x_2), \Delta\big) + p\big(G(x_1) - G(x_2), \Delta\big) + p\big(g_1 - g_2, \Delta\big).$$

By using the Cauchy–Schwarz inequality and the properties from Assumption 2.3, we obtain

$$\|\Delta\|_X \frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\|_X \leq (c_{\mathrm{qm}} + L_\rho)\|\Delta\|_X^2 + \|g_1 - g_2\|_X \|\Delta\|_X.$$

We divide by $\|\Delta\|_X$ and integrate from 0 to t which yields together with (2.2b)

$$\|\Delta(t)\|_X \leq \|x_1^0 - x_2^0\|_X + (c_{\mathrm{qm}} + L_\rho)\int_0^t \|\Delta(s)\|_X \,\mathrm{d}s + \int_0^t \|g_1(s) - g_2(s)\|_X \,\mathrm{d}s.$$

Finally, the assertion follows from applying Grönwall's lemma stated in Lemma A.1a). □

2.1.2 Space discretization

We now introduce an abstract space discretization of the evolution equation (2.1). Let $(X_h, p_h)_h$ be a family of finite dimensional vector spaces related to a discretization parameter h, e.g., the maximal mesh width of a finite element discretization. In each $X_h \in (X_h)_h$ we want to obtain an approximation x_h to the solution x of (2.1). We assume that $\mathcal{S}_h, G_h$ and g_h are approximations of $\mathcal{S}$,G, and g, respectively, that satisfy similar properties.

Assumption 2.7.

a) The nonlinear operator $\mathcal{S}_h\colon X_h \to X_h$ is quasi-monotone, i.e., there is a $\widehat{c}_{\mathrm{qm}} > 0$ s.t.

$$p_h\big(\mathcal{S}_h(y_h) - \mathcal{S}_h(z_h), y_h - z_h\big) \geq -\widehat{c}_{\mathrm{qm}} \|y_h - z_h\|_{X_h}^2 \qquad \text{for all } y_h, z_h \in X_h. \tag{2.3}$$

b) The nonlinearity $G_h\colon X_h \to X_h$ is locally Lipschitz continuous, i.e., for all $\rho > 0$ there exists a constant $\widehat{L}_\rho$ s.t. for all $y_h, z_h \in X_h$ with $\|y_h\|_{X_h}, \|z_h\|_{X_h} \leq \rho$:

$$\|G_h(y_h) - G_h(z_h)\|_{X_h} \leq \widehat{L}_\rho \|y_h - z_h\|_{X_h}.$$

c) The inhomogeneity satisfies $g_h \in W^{1,1}_{loc}([0;\infty); X_h)$.

The constants $\widehat{c}_{\mathrm{qm}}$ and $\widehat{L}_\rho$ are independent of h.

The discretized version of (2.1) is then given by

$$x_h' + \mathcal{S}_h(x_h) = G_h(x_h) + g_h, \qquad t \geq 0, \tag{2.4a}$$

$$x_h(0) = x_h^0. \tag{2.4b}$$

Since the assumptions are similar to the continuous case, we obtain by Theorem 2.4 that (2.4) is locally wellposed.

In the following, we present a framework for the error analysis of the abstract space discretization that is similar to the linear case presented in Hipp et al. [2019]. We allow for non-conforming space discretizations, where $X_h \not\subseteq X$. To still be able to relate the discrete and the continuous solution, we make the following assumptions:

Assumption 2.8.

a) There exists a lift operator $\mathcal{L}_h \in \mathcal{L}(X_h, X)$ that satisfies for some constant $\widehat{C}_X > 0$

$$\|\mathcal{L}_h y_h\|_X \leq \widehat{C}_X \|y_h\|_{X_h} \qquad \text{for all } y_h \in X_h. \tag{2.5}$$

By $\mathcal{L}_h^* \in \mathcal{L}(X, X_h)$ we denote the adjoint of the lift operator which is defined via

$$p_h\big(\mathcal{L}_h^* y, y_h\big) = p\big(y, \mathcal{L}_h y_h\big), \qquad \text{for all } y \in X, y_h \in X_h.$$

b) Let $Z \hookrightarrow X$ be a densely embedded subspace of X on which a reference operator $J_h \in \mathcal{L}(Z; X_h)$ is defined which satisfies for some constant $\widehat{C}_{J_h} > 0$

$$\|J_h\|_{X_h \leftarrow Z} \leq \widehat{C}_{J_h}.$$

The constants $\widehat{C}_X$ and $\widehat{C}_{J_h}$ are independent of h.

The reference operator should satisfy $\mathcal{L}_h J_h z \approx z$ for all $z \in Z$ in a suitable sense and could, e.g., be an interpolation or a projection operator.

Definition 2.9 (Remainder terms)**.**

*a) The **remainder of the nonlinear monotone operator** is defined via*

$$R_h \colon D(\mathcal{S}) \cap Z \to X_h, \qquad R_h(z) \coloneqq \mathcal{L}_h^* \mathcal{S}(z) - \mathcal{S}_h\left(J_h z\right).$$

*b) The **remainder of the Lipschitz continuous nonlinearity** is given by*

$$r_h \colon Z \to X_h, \qquad r_h(z) \coloneqq \mathcal{L}_h^* G(z) - G_h(J_h z).$$

We are now able to prove the following abstract error bound.

Theorem 2.10. *Let Assumptions 2.3, 2.7 and 2.8 be satisfied and x be the strong solution of (2.1) on $[0,T]$ with $x, x' \in L^\infty([0,T];Z)$. Furthermore, let x_h be the solution of (2.4) on $[0,T]$, and*

$$\rho_h \coloneqq \max\left\{\widehat{C}_{J_h}\|x\|_{L^\infty([0,T];X)}, \|x_h\|_{L^\infty([0,T];X_h)}\right\}.$$

Then, for all $t \in [0,T]$ the lifted discrete solution satisfies the error bound

$$\|\mathcal{L}_h x_h(t) - x(t)\|_X \leq \widehat{C}_X \mathrm{e}^{\left(\widehat{L}_{T,\rho_h} + \widehat{c}_{\mathrm{qm}}\right)t} E_h(t) + \|(\mathrm{I} - \mathcal{L}_h J_h)x(t)\|_X \tag{2.6}$$

with

$$\begin{aligned} E_h(t) = {} & \left\|x_h^0 - J_h x^0\right\|_{X_h} + t\big\|(\mathcal{L}_h^* - J_h)x'\big\|_{L^\infty([0,t];X_h)} \\ & + t\|R_h(x)\|_{L^\infty([0,t];X_h)} + t\|r_h(x)\|_{L^\infty([0,t];X_h)} + t\|\mathcal{L}_h^* g - g_h\|_{L^\infty([0,t];X_h)}. \end{aligned} \tag{2.7}$$

Proof. We split the error via $\mathcal{L}_h x_h(t) - x(t) = \mathcal{L}_h e_h + (\mathcal{L}_h J_h - \mathrm{I})x(t)$, where

$$e_h(t) = x_h(t) - J_h x(t) \in V_h$$

is the discrete error. Thus, the error can be bounded by

$$\|\mathcal{L}_h x_h(t) - x(t)\|_X \leq \widehat{C}_X \|e_h\|_{X_h} + \|(\mathcal{L}_h J_h - \mathrm{I})x(t)\|_X \tag{2.8}$$

and we have to further investigate the discrete error.

Applying the adjoint lift to (2.1a) yields

$$\mathcal{L}_h^* x' + \mathcal{L}_h^* \mathcal{S}(x) = \mathcal{L}_h^* G(x) + \mathcal{L}_h^* g.$$

By adding $J_h x'$ on both sides we obtain

$$J_h x' + \mathcal{S}_h(J_h x) = G_h(J_h x) + g_h + \phi_h \tag{2.9}$$

with

$$\phi_h = \left(J_h - \mathcal{L}_h^*\right) x' + \mathcal{S}_h(J_h x) - \mathcal{L}_h^* \mathcal{S}(x) + \mathcal{L}_h^* G(x) - G_h(J_h x) + \mathcal{L}_h^* g - g_h. \tag{2.10}$$

The stability estimate from Theorem 2.6 holds also true in the discrete case with $\widehat{c}_{\mathrm{qm}}$ and $\widehat{L}_\rho$ instead of c_{qm} and L_ρ, respectively, since we made the same assumptions. Note that due to our assumptions we have

$$\|J_h x(t)\|_{X_h} \leq \widehat{C}_{J_h} \|x(t)\|_Z \leq \rho_h.$$

Hence, we obtain by Theorem 2.6 applied to (2.4) and (2.9) the following bound for the discrete error

$$\begin{aligned} \|e_h(t)\|_{X_h} &\leq \mathrm{e}^{\left(\widehat{c}_{\mathrm{qm}}+\widehat{L}_{\rho_h}\right)t} \left(\|x_h^0 - J_h x^0\|_{X_h} + \int_0^t \|\phi_h(s)\|_{X_h} \,\mathrm{d}s \right) \\ &\leq \mathrm{e}^{\left(\widehat{c}_{\mathrm{qm}}+\widehat{L}_{\rho_h}\right)t} \left(\|x_h^0 - J_h x^0\|_{X_h} + t\|\phi_h\|_{L^\infty([0,T];X_h)} \right) \\ &\leq \mathrm{e}^{\left(\widehat{c}_{\mathrm{qm}}+\widehat{L}_{\rho_h}\right)t} E_h(t). \end{aligned} \tag{2.11}$$

Here, we used (2.10) and the definitions of the remainder terms from Definition 2.9. Together with (2.8), we finally obtain (2.6). □

In Theorem 2.10, we assume the existence of the numerical approximation x_h on $[0, T]$. Under a suitable consistency assumption, it is possible to ensure the existence and boundedness:

Theorem 2.11. *Let Assumptions 2.3, 2.7 and 2.8 be satisfied and x be the strong solution of (2.1) on $[0, T]$ with $x, x' \in L^\infty([0, T]; Z)$. Further, we assume that $E_h(t) \to 0$ for $h \to 0$ and for all $t \in [0, T]$.*

Then, there exists $h^* > 0$ s.t. for all $h < h^*$ the strong solution x_h of (2.6) exists on $[0, T]$ and satisfies

$$\|x_h\|_{L^\infty([0,T];X_h)} \leq \rho := 2\widehat{C}_{J_h} \|x\|_{L^\infty([0,T];Z)}. \tag{2.12}$$

Additionally, the error bound (2.6) holds true with $\rho_h = \rho$. If furthermore

$$\|(\mathrm{I} - \mathcal{L}_h J_h)x(t)\|_X \to 0 \quad \textit{for } h \to 0 \textit{ and all } t \in [0, T]$$

holds, then the lifted numerical solution converges to the continuous solution, i.e.,

$$\lim_{h\to 0} \|\mathcal{L}_h x_h(t) - x(t)\|_X = 0, \qquad \textit{for all } t \in [0, T].$$

Proof. We only have to show the existence of x_h on $[0, T]$ and the bound (2.12). The other assertions follow directly by Theorem 2.10.

The proof works by a standard contradiction argument: Let

$$T_h := \sup\left\{ t \geq 0 \;\middle|\; \|x_h\|_{L^\infty([0,t];X_h)} \leq \rho \right\}$$

be the maximal time s.t. the discrete solution is bounded by ρ. Since (2.6) is locally wellposed we have $T_h > 0$. We now show that for sufficiently small h we have $T_h \geq T$.

Assume that $T_h < T$. By (2.12) we obtain for all $t \leq T_h$

$$\|x_h(t)\|_{X_h} \leq \|x_h(t) - J_h x(t)\|_{X_h} + \|J_h x(t)\|_{X_h} \leq \|x_h(t) - J_h x(t)\|_{X_h} + \frac{\rho}{2}.$$

The first summand is $\|e_h\|_{X_h}$ and we obtain with (2.11) for all $t \leq T_h$

$$\|x_h(t)\|_{X_h} \leq \mathrm{e}^{\left(\widehat{c}_{\mathrm{qm}}+\widehat{L}_{\rho_h}\right)t} E_h(t) + \frac{\rho}{2} \to \frac{\rho}{2}, \quad h \to 0.$$

Hence, there is a $h^* > 0$ s.t. for all $h < h^*$ and $t \leq T_h$

$$\|x_h(t)\|_{X_h} \leq \frac{3}{4}\rho.$$

This is a contradiction to the definition of T_h since, due to the continuity of x_h in time, there is some $\varepsilon > 0$ s.t. we have $\|x_h(t)\|_{X_h} \leq \rho$ for all $t \in [0, T_h + \varepsilon]$. Hence, $T_h \geq T$ holds true for all $h < h^*$. □

2.2 Second-order evolution equations with nonlinear damping

In this section, we present an abstract framework for second-order differential equations and corresponding space discretizations. We follow the structure of Section 2.1 and use Theorems 2.10 and 2.11, to prove abstract error bounds for the second-order case.

We also illustrate how to use the abstract results to prove an error estimate for a finite element discretization of a nonlinear damped wave equation.

2.2.1 Analytical setting

Second-order formulation Let V, H be Hilbert spaces and V be densely embedded in H. We consider the following second-order differential equation in variational form, as it is typical for weak formulations of second-order differential equations

$$\begin{aligned} m\big(u'', \varphi\big) + m\big(\mathcal{D}(u'), \varphi\big) + a\big(u, \varphi\big) &= m\big(F(u), \varphi\big) + m\big(f, \varphi\big), \qquad \text{for } t \geq 0 \text{ and all } \varphi \in V, \\ u(0) = u^0, \qquad u'(0) &= v^0, \end{aligned} \tag{2.13}$$

where we make the following assumptions:

Assumption 2.12.

a) The bilinear form $m \colon H \times H \to \mathbb{R}$ is a scalar product on H with induced norm $\|\cdot\|_m$. In the following, we equip H with m.

b) The bilinear form $a \colon V \times V \to \mathbb{R}$ is symmetric and there exists a constant $c_G \geq 0$ s.t.

$$\tilde{a} := a + c_G m$$

is a scalar product on V with induced norm $\|\cdot\|_{\tilde{a}}$. From now on, we equip V with $\tilde{a}$.

c) The nonlinearity $\mathcal{D} \colon V \to H$ is quasi-monotone, i.e., there is a constant $\beta_{\mathrm{qm}} \geq 0$ s.t.

$$m\big(\mathcal{D}(v) - \mathcal{D}(w), v - w\big) \geq -\beta_{\mathrm{qm}} \|v - w\|_m^2 \qquad \text{for all } v, w \in V.$$

Furthermore, we have that $\mathcal{D} \in C(V, V^*)$.

d) The nonlinearity $F \colon V \to H$ is locally Lipschitz continuous with Lipschitz constant L_ρ, i.e., for all $v, w \in V$ with $\|v\|_V, \|w\|_V \leq \rho$ we have

$$\|F(v) - F(w)\|_m \leq L_\rho \|v - w\|_{\tilde{a}}.$$

e) The inhomogeneity satisfies $f \in W^{1,1}_{loc}([0,\infty);H)$.

By $C_{H,V}$ we denote the embedding constant of V into H, i.e.,

$$\|v\|_m \le C_{H,V}\|v\|_{\tilde{a}} \qquad \text{for all } v \in V. \tag{2.14}$$

Example 2.13. *To illustrate the abstract results of this section, we apply them to the following basic example: Let* $\Omega \subset \mathbb{R}^d$ *be a polygonal domain with boundary* Γ*. We consider the scalar wave equation with nonlinear forcing and damping terms and homogeneous Dirichlet boundary conditions*

$$\begin{aligned} u_{tt}(t,\mathbf{x}) + \big(u_t(t,\mathbf{x})\big)^3 - \Delta u(t,\mathbf{x}) &= \big(u(t,\mathbf{x})\big)^2, && t \ge 0, \mathbf{x} \in \Omega, && (2.15a)\\ u(t,\mathbf{x}) &= 0, && t \ge 0, \mathbf{x} \in \partial\Omega, && (2.15b)\\ u(0,\mathbf{x}) = u^0(\mathbf{x}), \qquad u_t(0,\mathbf{x}) &= v^0(\mathbf{x}), && \mathbf{x} \in \overline{\Omega}. && (2.15c) \end{aligned}$$

The weak formulation of (2.15) *is of the form* (2.13) *with* $V = H^1_0(\Omega), H = L^2(\Omega)$*, the usual* $L^2(\Omega)$ *scalar product* m*,*

$$\tilde{a}\big(v,w\big) = a\big(v,w\big) = \big(\nabla v, \nabla w\big)_{L^2(\Omega)}, \qquad \mathcal{D}(v) = v^3, \qquad F(v) = v^2, \quad \text{and} \quad f = 0.$$

This example fits into the setting of Assumption 2.12. It is straightforward to see that parts a), b), c), and e) of Assumption 2.12 are satisfied with $c_G = \beta_{\mathrm{qm}} = 0$*, while part d) is proven for a more general application in Lemma 6.4.*

To apply the results from Section 2.1, we rewrite (2.13) as a first-order evolution equation.

Formulation as an evolution equation We identify H with its dual space H^*, such that we have a Gelfand triple

$$V \overset{d}{\hookrightarrow} H \cong H^* \overset{d}{\hookrightarrow} V^*. \tag{2.16}$$

We define the operator $\mathcal{A} \in \mathcal{L}(V, V^*)$ associated to the bilinear form a via

$$\langle \mathcal{A}v, w\rangle_{V^* \times V} := a\big(v,w\big) \qquad \text{for all } v,w \in V. \tag{2.17}$$

The equation (2.13) can then be written as an evolution equation in V^*:

$$\begin{aligned} u'' + \mathcal{D}(u') + \mathcal{A}u &= F(u) + f, \qquad t \ge 0,\\ u(0) = u^0, \quad u'(0) &= v^0. \end{aligned} \tag{2.18}$$

By $A\colon D(A) \to H$ we denote the restriction of $\mathcal{A}$ to H, i.e.,

$$D(A) = \{v \in V \mid \mathcal{A}v \in H\}, \qquad \text{and} \quad Av = \mathcal{A}v \quad \text{for all } v \in D(A).$$

The restriction of (2.18) to H is then given by

$$\begin{aligned} u'' + \mathcal{D}(u') + Au &= F(u) + f, \qquad t \ge 0,\\ u(0) = u^0, \quad u'(0) &= v^0. \end{aligned} \tag{2.19}$$

By construction, a solution of (2.19) is also a solution of (2.18) and hence of (2.13).

First-order formulation We now rewrite (2.19) as a first-order evolution equation and show that it fits into the setting of Section 2.1. Therefore, let $u' = v$ and define

$$x = \begin{bmatrix} u \\ v \end{bmatrix}, \quad \mathcal{S}(x) = \begin{bmatrix} -v \\ Au + \mathcal{D}(v) \end{bmatrix}, \quad G(x) = \begin{bmatrix} 0 \\ F(u) \end{bmatrix}, \quad g = \begin{bmatrix} 0 \\ f \end{bmatrix}, \quad x^0 = \begin{bmatrix} u^0 \\ v^0 \end{bmatrix}. \tag{2.20}$$

The nonlinear operator $\mathcal{S}$ is defined on its domain $D(\mathcal{S}) = D(A) \times V$. With $X = V \times H$, (2.19) has the form (2.1). Since V is dense in H, $D(A)$ is dense in V and, hence, $D(\mathcal{S})$ is dense in X.

Lemma 2.14. *The nonlinear operator $\mathcal{S}$ is maximal and quasi-monotone with constant*

$$c_{\mathrm{qm}} = \frac{1}{2} c_G C_{H,V} + \beta_{\mathrm{qm}}.$$

Proof. Let $x_1 = [u_1, v_1]^\intercal, x_2 = [u_2, v_2]^\intercal \in X = V \times H$. Using the definition of A and the properties from Assumption 2.12 we then calculate

$$\begin{aligned} p\big(\mathcal{S}(x_1) - \mathcal{S}(x_2), x_1 - x_2\big) &= -\tilde{a}\big(v_1 - v_2, u_1 - u_2\big) + m\big(A(u_1 - u_2) + \mathcal{D}(v_1) - \mathcal{D}(v_2), v_1 - v_2\big) \\ &= -\tilde{a}\big(v_1 - v_2, u_1 - u_2\big) + a\big(u_1 - u_2, v_1 - v_2\big) + m\big(\mathcal{D}(v_1) - \mathcal{D}(v_2), v_1 - v_2\big) \\ &\geq -c_G m\big(v_1 - v_2, u_1 - u_2\big) - \beta_{\mathrm{qm}} \|v_1 - v_2\|_m^2 \\ &\geq -c_G \|v_1 - v_2\|_m \|u_1 - u_2\|_m - \beta_{\mathrm{qm}} \|v_1 - v_2\|_m^2 \\ &\geq -c_G C_{H,V} \|u_1 - u_2\|_{\tilde{a}} \|v_1 - v_2\|_m - \beta_{\mathrm{qm}} \|v_1 - v_2\|_m^2 \\ &\geq -\frac{1}{2} c_G C_{H,V} \left(\|u_1 - u_2\|_{\tilde{a}}^2 + \|v_1 - v_2\|_m^2 \right) - \beta_{\mathrm{qm}} \|v_1 - v_2\|_m^2 \\ &\geq -\left(\frac{1}{2} c_G C_{H,V} + \beta_{\mathrm{qm}} \right) \|x_1 - x_2\|_X^2. \end{aligned}$$

This proves the quasi-monotonicity.

We prove the maximality similar as in the proof of [Vitillaro, 2017, Theorem 4.1]. We have to show that there exists a $\lambda > c_{\mathrm{qm}}$ s.t. $\operatorname{range}(\lambda + \mathcal{S}) = X$. This is equivalent to proving that for every $[h_1, h_2]^\intercal \in V \times H = X$ there exists a solution $[v, w]^\intercal \in D(A) \times V = D(\mathcal{S})$ of the stationary problem

$$\lambda v - w = h_1, \tag{2.21a}$$

$$\lambda w + Av + \mathcal{D}(w) = h_2. \tag{2.21b}$$

From (2.21a) we obtain

$$v = \frac{1}{\lambda} (w + h_1). \tag{2.22}$$

We would like to plug this into (2.21b) and solve for w, but since w and h_1 are not in $D(A)$, we replace A by $\mathcal{A}$ before doing so. We obtain

$$\lambda w + \frac{1}{\lambda} \mathcal{A} w + \mathcal{D}(w) = h_2 - \frac{1}{\lambda} \mathcal{A} h_1 := h \tag{2.23}$$

with $h \in V^*$.

We will prove that $T = \lambda + \frac{1}{\lambda}\mathcal{A} + \mathcal{D} \colon V \to V^*$ is surjective for $\lambda > c_{\mathrm{qm}}$ large enough. In this case, there exists a $w \in V$ s.t. (2.23) is satisfied. By defining v via (2.22), we have that

$$\lambda w + \mathcal{A} v + \mathcal{D}(w) = h_2 \iff \mathcal{A} v = h_2 - \lambda w - \mathcal{D}(w).$$

Since $h_2 - \lambda w - \mathcal{D}(w) \in H$, it follows by definition that $v \in D(A)$ and that (2.21b) is satisfied.

Hence, it remains to prove the surjectivity of T. We use [Barbu, 2010, Corollary 2.3] stating that operators (from a reflexive Banach space to its dual space) that are continuous, monotone, and coercive are surjective, and hence will show that T has these three properties.

For this, we rewrite the operator as $T = T_1 + T_2$ with

$$T_1 = \frac{1}{\lambda}\left(\frac{\lambda^2}{2} + \mathcal{A}\right), \qquad T_2 = \frac{\lambda}{2} + \mathcal{D}.$$

We now choose a fixed

$$\lambda > \max\{c_{\mathrm{qm}}, \sqrt{2c_G}, 2\beta_{\mathrm{qm}}\}$$

and then have

- $T \in C(V, V^*)$ as the sum of continuous operators,
- T is monotone as the sum of monotone operators,
- T is coercive, i.e.,

  $$\frac{\langle T(v), v\rangle_{V^*\times V}}{\|v\|_{\tilde{a}}} \to \infty \quad \text{for } \|v\|_{\tilde{a}} \to \infty,$$

  which can be seen by the following calculation:

  $$\begin{aligned}\langle T(v), v\rangle_{V^*\times V} &= \langle T_1(v), v\rangle_{V^*\times V} + \langle T_2(v), v\rangle_{V^*\times V}\\ &\geq \frac{1}{\lambda}\|v\|_{\tilde{a}}^2 + \langle T_2(v) - T_2(0), v - 0\rangle_{V^*\times V} + \langle T_2(0), v\rangle_{V^*\times V}\\ &\geq \frac{1}{\lambda}\|v\|_{\tilde{a}}^2 - \|v\|_{\tilde{a}}\|T_2(0)\|_{V^*},\end{aligned}$$

  where we used that, due to the choice of λ, T_1 is coercive and T_2 is monotone. Hence, we have

  $$\frac{\langle T(v), v\rangle_{V^*\times V}}{\|v\|_{\tilde{a}}} \geq \frac{1}{\lambda}\|v\|_{\tilde{a}} - \|T_2(0)\|_{V^*} \to \infty \quad \text{for } \|v\|_{\tilde{a}} \to \infty.$$

□

The following corollary shows that the first-order formulation of (2.19) fits into the setting of Section 2.1.1.

Corollary 2.15. *Assumption 2.12 implies that the first-order formulation of (2.19) satisfies Assumption 2.3.*

Proof. By Lemma 2.14, we have that Assumption 2.3 a) is satisfied. Assumption 2.3 b) and c) follow directly by Assumption 2.12 d) and e). □

Corollary 2.16. *Let $\left[u^0, v^0\right]^\mathsf{T} \in D(A) \times V$. Then, the second-order evolution equation (2.18) and, hence, (2.13) are locally wellposed.*

Proof. Follows with Corollary 2.15 directly by Theorem 2.4. □

2.2.2 Space discretization

Let $(V_h)_h$ be a family of finite dimensional vector spaces related to a discretization parameter h. In each $V_h \in (V_h)_h$, we consider the following discretized version of (2.13):

$$\begin{aligned} m_h\big(u_h'', \varphi_h\big) + m_h\big(\mathcal{D}_h(u_h'), \varphi_h\big) + a_h\big(u_h, \varphi_h\big) &= m_h\big(F_h(u_h), \varphi_h\big) + m_h\big(f_h, \varphi_h\big), \quad \text{for all } \varphi_h \in V_h, t \geq 0 \\ u_h(0) = u_h^0, \qquad u_h'(0) &= v_h^0. \end{aligned} \tag{2.24}$$

Here, $m_h, a_h, \mathcal{D}_h, F_h$ and f_h are approximations of their continuous counterparts for which we assume that they satisfy similar properties:

Assumption 2.17. *All constants in the following statements are independent of h.*

a) The bilinear form m_h is a scalar product on V_h. We denote V_h equipped with this scalar product m_h by H_h and the induced norm by $\|\cdot\|_{m_h}$.

b) The bilinear form $a_h\colon V_h \times V_h \to \mathbb{R}$ is symmetric and there exists a constant $\widehat{c}_G \geq 0$ s.t.

$$\tilde{a}_h \coloneqq a_h + \widehat{c}_G m_h$$

is a scalar product on V_h with induced norm $\|\cdot\|_{\tilde{a}_h}$. In the following, we equip V_h with $\tilde{a}_h$.

c) The nonlinearity $\mathcal{D}_h\colon V_h \to H_h$ is continuous and quasi-monotone with constant $\widehat{\beta}_{\mathrm{qm}}$.

d) The nonlinearity $F_h\colon V_h \to H_h$ is locally Lipschitz-continuous with constant $\widehat{L}_\rho$.

e) The inhomogeneity satisfies $f_h \in W^{1,1}_{loc}([0,\infty); H_h)$.

f) There exists a constant $\widehat{C}_{H,V} > 0$ s.t.

$$\|v_h\|_{m_h} \leq \widehat{C}_{H,V} \|v_h\|_{\tilde{a}_h} \qquad \text{for all } v_h \in V_h. \tag{2.25}$$

Example 2.13 (continued)**.** *Let $(\mathcal{T}_h)_h$ be a quasi-uniform family of matching simplicial triangulations of Ω (cf. Definition C.1). For each $\mathcal{T}_h \in (\mathcal{T}_h)_h$ let V_h be the standard linear finite element space over $\mathcal{T}_h$, i.e., V_h is the space of piecewise linear functions defined on $\mathcal{T}_h$. This is a conformal finite element method since $V_h \subset V$ and we set*

$$m_h \coloneqq m, \qquad a_h \coloneqq a. \tag{2.26a}$$

Further, we define the discretizations of the nonlinearities via

$$m\big(\mathcal{D}_h(v_h), w_h\big) \coloneqq \int_\Omega v_h^3 w_h \,\mathrm{d}\mathbf{x}, \tag{2.26b}$$

$$m\big(F_h(v_h), w_h\big) \coloneqq \int_\Omega v_h^2 w_h \,\mathrm{d}\mathbf{x} \tag{2.26c}$$

for all $v_h, w_h \in V_h$. Note that we can evaluate these integrals exactly by a quadrature formula of order ≥ 5. This discretization fits into the framework of Assumption 2.17. As in the continuous case, it is easy to see, that the parts a), b), c), e), and f) are satisfied. Part e) is proven in Lemma 6.6 for a more general example.

We define $A_h, \in \mathcal{L}(V_h; V_h)$ via

$$m_h(A_h v_h, w_h) \coloneqq a_h(v_h, w_h) \quad \text{for all } v_h, w_h \in V_h.$$

Then, (2.24) can be written as an evolution equation in V_h:

$$\begin{aligned} u_h'' + \mathcal{D}_h(u_h') + A_h u_h &= F_h(u_h) + f_h, \qquad t \geq 0 \\ u_h(0) = u_h^0, \qquad u_h'(0) &= v_h^0. \end{aligned} \tag{2.27}$$

As in the continuous case, we can rewrite this as a first-order equation. We define the finite dimensional Hilbert space $X_h = V_h \times H_h$ and set

$$x_h = \begin{bmatrix} u_h \\ v_h \end{bmatrix}, \quad \mathcal{S}_h(x_h) = \begin{bmatrix} -v_h \\ A_h u_h + \mathcal{D}_h(v_h) \end{bmatrix}, \quad G_h(x_h) = \begin{bmatrix} 0 \\ F_h(u_h) \end{bmatrix}, \qquad g_h = \begin{bmatrix} 0 \\ f_h \end{bmatrix}, \qquad x_h^0 = \begin{bmatrix} u_h^0 \\ v_h^0 \end{bmatrix}. \tag{2.28}$$

Then, (2.27) has the form (2.4).

Corollary 2.18. *Assumption 2.17 implies that the first-order formulation of (2.27) satisfies Assumption 2.7. Furthermore, (2.3) holds true with $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}}$.*

Proof. As in Lemma 2.14, we obtain that $\mathcal{S}_h$ satisfies (2.3) with constant $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}}$. The other assumptions from Assumption 2.7 follow directly by Assumption 2.17. □

As in the first-order case, we require the existence of appropriate operators to relate continuous and discrete functions:

Assumption 2.19.

a) There exists a lift operator $\mathcal{L}_h^V \in \mathcal{L}(V_h; V)$ satisfying

$$\|\mathcal{L}_h^V v_h\|_m \leq \widehat{C}_H \|v_h\|_{m_h}, \qquad \|\mathcal{L}_h^V v_h\|_{\tilde{a}} \leq \widehat{C}_V \|v_h\|_{\tilde{a}_h}, \tag{2.29}$$

for all $v_h \in V_h$ with constants $\widehat{C}_H, \widehat{C}_V > 0$ independent of h.

b) There exists an interpolation operator $I_h \in \mathcal{L}(Z^V; V_h)$, defined on a dense subspace Z^V of V, which satisfies

$$\|I_h\|_{H_h \leftarrow Z^V} \leq \widehat{C}_{I_h} \tag{2.30}$$

with a constant $\widehat{C}_{I_h} > 0$ independent of h.

Example 2.13 (continued)**.** *Since the discretization in our example is conformal, i.e., we have $V_h \subset V$, we can set $\mathcal{L}_h^V = \mathrm{I}$. Recall that we have*

$$\|v_h\|_{m_h} = \|v_h\|_m, \qquad \|v_h\|_{\tilde{a}_h} = \|v_h\|_{\tilde{a}}, \quad \textit{for all } v_h \in V_h$$

and hence, Assumption 2.19 a) is trivially satisfied. Further, we choose $Z^V = H^2(\Omega) \subset C(\Omega)$ and define $I_h \colon H^2(\Omega) \to V_h$ as the standard Lagrange interpolation operator in the basis nodes of the triangulation. The Lagrange interpolation satisfies (2.30).

We now analyze the space discretization error by applying the theory of Section 2.1.2. Therefore, we have to specify the appearing operators.

Definition 2.20.

a) The ***adjoint lift operators*** $\mathcal{L}_h^{V*} \colon V \to V_h$ *and* $\mathcal{L}_h^{H*} \colon H \to H_h$ *w.r.t. the scalar products of* V *and* H *are defined via*

$$\begin{aligned} m_h\big(\mathcal{L}_h^{H*} v, w_h\big) &\coloneqq m\big(v, \mathcal{L}_h^V w_h\big) \quad \text{for all } v \in H, w_h \in H_h, \\ \tilde{a}_h\big(\mathcal{L}_h^{V*} v, w_h\big) &\coloneqq \tilde{a}\big(v, \mathcal{L}_h^V w_h\big) \quad \text{for all } v \in V, w_h \in V_h. \end{aligned} \tag{2.31}$$

b) We define the ***first-order lift operator*** $\mathcal{L}_h \colon X_h \to X$ *by*

$$\mathcal{L}_h \begin{bmatrix} v_h \\ w_h \end{bmatrix} \coloneqq \begin{bmatrix} \mathcal{L}_h^V v_h \\ \mathcal{L}_h^V w_h \end{bmatrix}.$$

c) We define the ***first-order reference operator*** $J_h \colon Z \to X_h$ *by*

$$J_h \begin{bmatrix} v \\ w \end{bmatrix} \coloneqq \begin{bmatrix} \mathcal{L}_h^{V*} v \\ I_h w \end{bmatrix} \tag{2.32}$$

on $Z = V \times Z^V \overset{d}{\hookrightarrow} X$.

Remark 2.21. *We use* I_h *instead of* $\mathcal{L}_h^{H*}$ *in the second component of the reference operator because the adjoint lift operator only leads to suboptimal error bounds.*

Lemma 2.22. *The first-order lift and reference operators from Definition 2.20 satisfy Assumption 2.8 with* $\widehat{C}_X = \max\{\widehat{C}_V, \widehat{C}_H\}$ *and* $\widehat{C}_{J_h} = \max\{\widehat{C}_V, \widehat{C}_{I_h}\}$.

Proof. This is a direct consequence of Assumption 2.19. □

For $z = [v, w]^\intercal \in Z$, the remainder terms are given by

$$R_h(z) = \mathcal{L}_h^* \mathcal{S}(z) - \mathcal{S}_h J_h(z) = \begin{bmatrix} -(\mathcal{L}_h^{V*} - I_h) w \\ \mathcal{L}_h^{H*}(Av + \mathcal{D}(w)) - \big(A_h \mathcal{L}_h^{V*} v + \mathcal{D}_h(I_h w)\big) \end{bmatrix}, \tag{2.33a}$$

$$r_h(z) = \mathcal{L}_h^* G(z) - G_h(J_h z) = \begin{bmatrix} 0 \\ \mathcal{L}_h^{H*} F(v) - F_h(\mathcal{L}_h^{V*} v) \end{bmatrix}. \tag{2.33b}$$

The norm of $r_h(z)$ is obviously given by

$$\|r_h(z)\|_{X_h} = \|\mathcal{L}_h^{H*} F(v) - F_h(\mathcal{L}_h^{V*} v)\|_{m_h}, \qquad z = [v, w]^\intercal \in Z. \tag{2.34}$$

To bound the remainder of the monotone operator, we make use of the following errors in the scalar products, that are for $v_h, w_h \in V_h$ defined via

$$\begin{aligned} \Delta m\big(v_h, w_h\big) &\coloneqq m\big(\mathcal{L}_h^V v_h, \mathcal{L}_h^V w_h\big) - m_h\big(v_h, w_h\big), \\ \Delta \tilde{a}\big(v_h, w_h\big) &\coloneqq \tilde{a}\big(\mathcal{L}_h^V v_h, \mathcal{L}_h^V w_h\big) - \tilde{a}_h\big(v_h, w_h\big). \end{aligned} \tag{2.35}$$

Lemma 2.23. *Let Assumptions 2.12 and 2.17 be satisfied. Then, for $z = [v, w]^\mathsf{T} \in Z$, the remainder of the monotone operator can be bounded by*

$$\begin{aligned} \|R_h(z)\|_{X_h} \le C\Big(&\max_{\|\varphi_h\|_{\tilde{a}_h}=1} \big|\Delta\tilde{a}\big(I_h w, \varphi_h\big)\big| + \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \big|\Delta\tilde{a}\big(I_h v, \varphi_h\big)\big| \\ &+ \max_{\|\psi_h\|_{m_h}=1} \big|\Delta m\big(I_h v, \psi_h\big)\big| + \|(\mathrm{I} - \mathcal{L}_h^V I_h)v\|_{\tilde{a}} \\ &+ \|\big(\mathrm{I} - \mathcal{L}_h^V I_h\big)\, w\|_{\tilde{a}} + \big\|\mathcal{L}_h^{H*}\mathcal{D}(w) - \mathcal{D}_h(I_h w)\big\|_{m_h}\Big), \end{aligned} \tag{2.36}$$

i.e., against errors in the scalar products, interpolation errors, and the discretization error of the nonlinear damping term.

Proof. The proof works similar to the proof of [Hipp et al., 2019, Lemma 4.7]. We use the identity

$$\|R_h(z)\|_{X_h} = \max_{\|y_h\|_{X_h}=1} p_h\big(R_h(z), y_h\big).$$

For $y_h = [\varphi_h, \psi_h]^\mathsf{T} \in X_h$ with $\|y_h\|_{X_h} = 1$ we obtain by (2.33a)

$$\begin{aligned} p_h\big(R_h(z), y_h\big) = &- \tilde{a}_h\big((\mathcal{L}_h^{V*} - I_h)w, \varphi_h\big) + m_h\big(\mathcal{L}_h^{H*}(Av + \mathcal{D}(w)) - \big(A_h\mathcal{L}_h^{V*}v + \mathcal{D}_h(I_h w)\big), \psi_h\big) \\ = &- \Big(\tilde{a}\big(w, \mathcal{L}_h^V\varphi_h\big) - \tilde{a}_h\big(I_h w, \varphi_h\big)\Big) + \Big(a\big(v, \mathcal{L}_h^V\psi_h\big) - a_h\big(\mathcal{L}_h^{V*}v, \psi_h\big)\Big) \\ &+ m_h\big(\mathcal{L}_h^{H*}\mathcal{D}(w) - \mathcal{D}_h(I_h w), \psi_h\big), \end{aligned} \tag{2.37}$$

and bound the three summands separately. For the first, we obtain by (2.35), (2.29) and $\|\varphi_h\|_{\tilde{a}_h} \le 1$

$$\begin{aligned} \tilde{a}\big(w, \mathcal{L}_h^V\varphi_h\big) - \tilde{a}_h\big(I_h w, \varphi_h\big) &= \tilde{a}\big(w, \mathcal{L}_h^V\varphi_h\big) - \tilde{a}\big(\mathcal{L}_h^V I_h w, \mathcal{L}_h^V\varphi_h\big) + \Delta\tilde{a}\big(I_h w, \varphi_h\big) \\ &\le \|\big(\mathrm{I} - \mathcal{L}_h^V I_h\big)\, w\|_{\tilde{a}} \|\mathcal{L}_h^V\varphi_h\|_{\tilde{a}} + \big|\Delta\tilde{a}\big(I_h w, \varphi_h\big)\big| \\ &\le \widehat{C}_V \|\big(\mathrm{I} - \mathcal{L}_h^V I_h\big)\, w\|_{\tilde{a}} + \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \big|\Delta\tilde{a}\big(I_h w, \varphi_h\big)\big|. \end{aligned} \tag{2.38}$$

The second summand in (2.37) can be bounded with the definitions of $\tilde{a}, \tilde{a}_h$, $\|\psi_h\|_{m_h} \le 1$ and (2.14), (2.25), (2.29), (2.31) and (2.35) by

$$\begin{aligned} a\big(v, \mathcal{L}_h^V\psi_h\big) - a_h\big(\mathcal{L}_h^{V*}v, \psi_h\big) &= \tilde{a}\big(v, \mathcal{L}_h^V\psi_h\big) - \tilde{a}_h\big(\mathcal{L}_h^{V*}v, \psi_h\big) - \big(c_G m\big(v, \mathcal{L}_h^V\psi_h\big) - \widehat{c}_G m_h\big(\mathcal{L}_h^{V*}v, \psi_h\big)\big) \\ &\le \max\{c_G, \widehat{c}_G\} \big|m\big(v, \mathcal{L}_h^V\psi_h\big) - m_h\big(\mathcal{L}_h^{V*}v, \psi_h\big)\big| \\ &\le \max\{c_G, \widehat{c}_G\} \Big(\big|m\big((\mathrm{I} - \mathcal{L}_h^V I_h)v, \mathcal{L}_h^V\psi_h\big)\big| + \big|\Delta m\big(I_h v, \psi_h\big)\big| \\ &\qquad + m_h\big((I_h - \mathcal{L}_h^{V*})v, \psi_h\big)\Big) \\ &\le \max\{c_G, \widehat{c}_G\} \Big(\widehat{C}_H C_{H,V} \|(\mathrm{I} - \mathcal{L}_h^V I_h)v\|_{\tilde{a}} + \max_{\|\psi_h\|_{m_h}=1} \big|\Delta m\big(I_h v, \psi_h\big)\big| \\ &\qquad + \widehat{C}_{H,V} \|(I_h - \mathcal{L}_h^{V*})v\|_{\tilde{a}_h}\Big). \end{aligned}$$

We further estimate similar to (2.38)

$$\begin{aligned} \|(I_h - \mathcal{L}_h^{V*})v\|_{\tilde{a}_h} &= \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \tilde{a}_h\big((I_h - \mathcal{L}_h^{V*})v, \varphi_h\big) \\ &= \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \tilde{a}_h\big(I_h v, \varphi_h\big) - \tilde{a}\big(v, \mathcal{L}_h^V\varphi_h\big) \\ &\le \widehat{C}_V \|\big(\mathrm{I} - \mathcal{L}_h^V I_h\big)\, v\|_{\tilde{a}} + \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \big|\Delta\tilde{a}\big(I_h v, \varphi_h\big)\big| \end{aligned}$$

Finally, the last summand in (2.37) is bounded by

$$m_h\big(\mathcal{L}_h^{H*}\mathcal{D}(w) - \mathcal{D}_h(I_h w), \psi_h\big) \le \|\mathcal{L}_h^{H*}\mathcal{D}(w) - \mathcal{D}_h(I_h w)\|_{m_h}$$

and the assertion follows by collecting all terms. □

With the results we have already obtained in this section, the following two theorems are now direct consequences of Theorem 2.10 and Theorem 2.11. The first one is a space discretization error bound under the assumption, that the numerical approximation u_h exists on $[0,T]$.

Theorem 2.24. *Let Assumptions 2.12, 2.17 and 2.19 be satisfied and u be the strong solution of (2.19) on $[0,T]$ with $u, u', u'' \in L^\infty([0,T];Z^V)$. Further, assume that the semidiscrete solution u_h of (2.27) exists on $[0,T]$. Then, for all $t \in [0,T]$, the lifted semidiscrete solution satisfies the error bound*

$$\|\mathcal{L}_h^V u_h(t) - u(t)\|_{\tilde{a}} + \|\mathcal{L}_h^V u_h'(t) - u'(t)\|_m \le C \mathrm{e}^{(\widehat{L}_{\rho_h} + \widehat{c}_{\mathrm{qm}})t}(1+t)\sum_{i=1}^{5} E_{h,i} \tag{2.39}$$

with a constant C that is independent of h and t. The other constants are given by $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}}$,

$$\rho_h = \max\left\{\widehat{C}_V \|u\|_{L^\infty([0,T];V)}, \|u_h\|_{L^\infty([0,T];V_h)}\right\},$$

and the abstract space discretization errors

$$\begin{aligned}
E_{h,1} =& \|u_h^0 - \mathcal{L}_h^{V*}u^0\|_{\tilde{a}_h} + \|v_h^0 - I_h v^0\|_{m_h} + \|\mathcal{L}_h^{H*}f - f_h\|_{L^\infty([0,T];H)},\\
E_{h,2} =& \|\mathcal{L}_h^{H*}\mathcal{D}(u') - \mathcal{D}_h(I_h u')\|_{L^\infty([0,T];H_h)},\\
E_{h,3} =& \|\mathcal{L}_h^{H*}F(u) - F_h(\mathcal{L}_h^{V*}u)\|_{L^\infty([0,T];H_h)},\\
E_{h,4} =& \|(\mathrm{I} - \mathcal{L}_h^V I_h)u\|_{L^\infty([0,T];V)} + \|(\mathrm{I} - \mathcal{L}_h^V I_h)u'\|_{L^\infty([0,T];V)} + \|(\mathrm{I} - \mathcal{L}_h^V I_h)u''\|_{L^\infty([0,T];H)},\\
E_{h,5} =& \Big\| \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \Delta\tilde{a}\big(I_h u, \varphi_h\big)\Big\|_{L^\infty(0,T)} + \Big\| \max_{\|\psi_h\|_{m_h}=1} \Delta m\big(I_h u, \psi_h\big)\Big\|_{L^\infty(0,T)}\\
&+ \Big\| \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \Delta\tilde{a}\big(I_h u', \varphi_h\big)\Big\|_{L^\infty(0,T)} + \Big\| \max_{\|\psi_h\|_{m_h}=1} \Delta m\big(I_h u'', \psi_h\big)\Big\|_{L^\infty(0,T)}.
\end{aligned} \tag{2.40}$$

Proof. We apply Theorem 2.10 to the first-order formulations of (2.19) and (2.27). By Corollarys 2.15 and 2.18 and Lemma 2.22 all assumptions are satisfied. Note that (2.12) would also require that u_h' and u' are bounded by ρ_h. But, since the nonlinearities G and G_h in the first-order formulations of (2.19) and (2.27) only depend on the first component, it is sufficient that u and u_h are bounded to exploit the local Lipschitz continuity.

The error estimate (2.6) yields

$$\begin{aligned}
\|\mathcal{L}_h^V u_h(t) - u(t)\|_{\tilde{a}} + \|\mathcal{L}_h^V u_h'(t) - u'(t)\|_m &\le 2\left(\|\mathcal{L}_h^V u_h(t) - u(t)\|_{\tilde{a}}^2 + \|\mathcal{L}_h^V u_h'(t) - u'(t)\|_m^2\right)^{\frac{1}{2}}\\
&= 2\|\mathcal{L}_h x_h(t) - x(t)\|_X\\
&\le 2\widehat{C}_X \mathrm{e}^{\left(\widehat{L}_{T,\rho_h} + \widehat{c}_{\mathrm{qm}}\right)t} E_h(t) + 2\|(\mathrm{I} - \mathcal{L}_h J_h)x(t)\|_X
\end{aligned}$$

with

$$\begin{aligned}
E_h(t) = \big\|x_h^0 - J_h x^0\big\|_{X_h} &+ t\big\|(\mathcal{L}_h^* - J_h)x'\big\|_{L^\infty([0,T];X_h)}\\
&+ t\|R_h(x)\|_{L^\infty([0,T];X_h)} + t\|r_h(x)\|_{L^\infty([0,T];X_h)} + t\|\mathcal{L}_h^* g - g_h\|_{L^\infty([0,T];X_h)}.
\end{aligned}$$

It remains to bound the different terms against $E_{h,i}, i = 1, \dots, 5$. By the remainder bounds (2.34) and (2.36) we obtain for all $t \in [0, T]$

$$\|R_h(x(t))\|_{X_h} \leq C(E_{h,2} + E_{h,4} + E_{h,5}), \qquad \|r_h(x(t))\|_{X_h} \leq CE_{h,3},$$

and by the definition of J_h and $\mathcal{L}_h^*$ we further have

$$\left\|x_h^0 - J_h x^0\right\|_{X_h} + \|\mathcal{L}_h^* g - g_h\|_{L^\infty([0,T];X_h)} \leq CE_{h,1}.$$

For the reference error we have for all $t \in [0, T]$

$$\|(\mathrm{I} - \mathcal{L}_h J_h)x(t)\|_X \leq \|(\mathrm{I} - \mathcal{L}_h^V \mathcal{L}_h^{V*})u(t)\|_{\tilde{a}} + \|(\mathrm{I} - \mathcal{L}_h^V I_h)u'(t)\|_m \tag{2.41}$$

where the second summand is bounded by $E_{h,4}$. For the first summand, we obtain as in (2.38)

$$\begin{aligned} \|(\mathrm{I} - \mathcal{L}_h^V \mathcal{L}_h^{V*})u\|_{\tilde{a}} &\leq \|(\mathrm{I} - \mathcal{L}_h^V I_h)u\|_{\tilde{a}} + \|\mathcal{L}_h^V (I_h - \mathcal{L}_h^{V*})u\|_{\tilde{a}} \\ &\leq CE_{h,4} + \widehat{C}_V \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \Big(\tilde{a}_h\big(I_h u, \varphi_h\big) - \tilde{a}\big(u, \mathcal{L}_h^V \varphi_h\big)\Big) \\ &\leq CE_{h,4} + \widehat{C}_V^2 \|\big(\mathrm{I} - \mathcal{L}_h^V I_h\big)\, u\|_{\tilde{a}} + \widehat{C}_V \max_{\|\varphi_h\|_{\tilde{a}_h}=1} \big|\Delta\tilde{a}\big(I_h u, \varphi_h\big)\big| \\ &\leq C(E_{h,4} + E_{h,5}). \end{aligned} \tag{2.42}$$

Similarly, we finally bound

$$\begin{aligned} \|(\mathcal{L}_h^* - J_h)x'\|_{X_h} &\leq \|(\mathcal{L}_h^{H*} - I_h)u''\|_{m_h} \\ &\leq \widehat{C}_H \|\big(\mathrm{I} - \mathcal{L}_h^V I_h\big)\, u''\|_m + \max_{\|\psi_h\|_{m_h}=1} \big|\Delta m\big(I_h u'', \psi_h\big)\big| \\ &\leq C(E_{h,4} + E_{h,5}). \end{aligned}$$

□

As in Section 2.1, we can conclude existence and convergence of the numerical solution under an additional consistency assumption.

Corollary 2.25. *Let Assumptions 2.12, 2.17 and 2.19 be satisfied and u be the strong solution of (2.27) on $[0, T]$ with $u, u', u'' \in L^\infty([0, T]; Z^V)$. We then define*

$$\rho := 2\widehat{C}_V \|u\|_{L^\infty([0,T];V)}.$$

Further, let $E_{h,i} \to 0$ for $h \to 0$, $i = 1, \dots, 5$. Then, there exists $h^* > 0$, s.t. u_h exists in $[0, T]$ for all $h < h^*$ with

$$\|u_h\|_{L^\infty([0,T];V_h)} \leq \rho.$$

Additionally, the error bound (2.39) holds true with $\rho_h = \rho$ and the lifted semidiscrete solution converges to the continuous solution, i.e.,

$$\lim_{h\to 0} \big(\|\mathcal{L}_h^V u_h(t) - u(t)\|_{\tilde{a}} + \|\mathcal{L}_h^V u_h'(t) - u'(t)\|_m\big) = 0, \qquad t \in [0, T].$$

Proof. This is a direct consequence of Theorems 2.11 and 2.24. □

Finally, we now can prove an error bound for our illustrative example:

Example 2.13 (continued)**.** *We now illustrate for our example equation* (2.15)*, with corresponding space discretization* (2.26)*, the application of Theorem 2.24 and Corollary 2.25 to a concrete equation. Let $u \in C^2([0,T];H^2(\Omega))$ be the solution of* (2.15)*. In the following, we estimate the error terms $E_{h,i}, i = 1,\dots,5$ from* (2.40)*. Since we are in the case of a conformal discretization with*

$$\tilde{a}_h = \tilde{a}, \qquad m_h = m, \qquad \text{and} \quad \mathcal{L}_h^V = \mathrm{I},$$

the error term $E_{h,5}$, containing the errors in the bilinear forms, vanishes. Further, note that in this case we have

$$\mathcal{L}_h^{H*} = \Pi_{L^2}, \qquad \mathcal{L}_h^{H*} = \Pi_{H^1},$$

where Π_{L^2} and Π_{H^1} denote the L^2- and H^1-orthogonal projections, respectively. We discretize the initial values via

$$u_h^0 := I_h u^0, \qquad v_h^0 := I_h v^0.$$

Then, by usual interpolation and projection error bounds and $f = 0$, we obtain

$$E_{h,1}, E_{h,4} \leq Ch.$$

It remains to bound the discretization errors of $\mathcal{D}$ and F, and we bound exemplary $E_{h,2}$. We have by the definition of $\mathcal{D}_h$ and the usual interpolation error bounds for all $t \in [0,T]$

$$\begin{aligned}
\|\mathcal{L}_h^{H*}\mathcal{D}(u') - \mathcal{D}_h(I_h u')\|_{H_h} &= \max_{\|w_h\|_{L^2(\Omega)}=1} \big(\Pi_{L^2}\mathcal{D}(u') - \mathcal{D}_h(I_h u'), w_h\big)_{L^2(\Omega)} \\
&= \max_{\|w_h\|_{L^2(\Omega)}=1} \int_\Omega \big((u')^3 - (I_h u')^3\big) w_h \,\mathrm{d}\mathbf{x} \\
&\leq \|(u')^3 - (I_h u')^3\|_{L^2(\Omega)} \\
&\leq \|(u')^2 + u' I_h u' + (I_h u')^2\|_{L^\infty(\Omega)} \|u' - I_h u'\|_{L^2(\Omega)} \\
&\leq C\big(\|u'\|_{L^\infty(\Omega)}, \|u'\|_{H^2(\Omega)}\big) Ch^2.
\end{aligned}$$

Note that $\|u'\|_{L^\infty(\Omega)}$ is bounded due to the continuous embedding $H^2(\Omega) \hookrightarrow C(\Omega)$, cf. Theorem A.4. Hence, we obtain $E_{h,2} \leq Ch^2$ and similarly $E_{h,3} \leq Ch^2$.

We already showed in the previous parts of the example, that Assumptions 2.12, 2.17 and 2.19 are satisfied. Recall that we have $V = H^1(\Omega)$ and $H = L^2(\Omega)$. Hence, we can apply Corollary 2.25 and have that, for h sufficiently small, the linear finite element approximation u_h satisfies the error bound

$$\|u_h(t) - u(t)\|_{H^1(\Omega)} + \|u_h'(t) - u'(t)\|_{L^2(\Omega)} \leq C\mathrm{e}^{\widehat{L}_\rho t}(1+t)h,$$

with a constant C independent of h and t.

CHAPTER 3

Runge–Kutta time discretization of first- and second-order evolution equations

In this chapter, we introduce and analyze Runge–Kutta time discretization schemes for abstract evolution equations in the framework of the unified error analysis from Chapter 2.

The main goal in this chapter is to approximate the solution u of the continuous second-order evolution equation (2.19). Since Runge–Kutta methods are usually formulated for first-order equations, we start by considering Runge–Kutta methods applied to the first-order formulation (2.1) of (2.19), cf. (2.20). By $\tau > 0$ we denote the time step size and set $t_n \coloneqq n\tau, n \geq 0$. The iterates $x^n = [u^n, v^n]^\mathsf{T}$ of the Runge–Kutta method applied to the first-order formulation of (2.19) then satisfy $u^n \approx u(t_n), v^n \approx u'(t_n)$.

Based on results from Hansen [2006b], we prove order q time discretization error bounds for coercive and algebraically stable Runge–Kutta methods of stage order q (cf. Definitions B.3 to B.5). A short introduction to Runge–Kutta methods and a summary of the results from Hansen [2006b] that are necessary for our analysis can be found in Appendix B.

Outline In Section 3.1, we analyze Runge–Kutta methods applied to first-order equations in the framework of the unified error analysis of Section 2.1.1. In Section 3.2, we then use these results to prove error bounds for Runge–Kutta methods applied to second-order equations in the framework of Section 2.2.1.

3.1 Runge–Kutta methods for first-order evolution equations

A Runge–Kutta method with coefficients $\mathbf{b} = (b_i)_{i=1}^s, \mathbf{c} = (c_i)_{i=1}^s, \mathcal{A} = (a_{ij})_{i,j=1}^s$ applied to the evolution equation (2.1) has the form

$$\begin{aligned} X^{ni} &= x^n + \tau \sum_{j=1}^s a_{ij}\big(-\mathcal{S}(X^{nj}) + G(X^{nj}) + g(t_n + c_j\tau)\big), \qquad i = 1, \dots, s, \\ x^{n+1} &= x^n + \tau \sum_{i=1}^s b_i \left(-\mathcal{S}(X^{ni}) + G(X^{ni}) + g(t_n + c_i\tau)\right). \end{aligned}$$

We first prove an error bound under the assumption that bounded Runge–Kutta iterations exist. Afterwards, we prove the existence of the iterations for sufficiently small τ. The error bound is a direct application of the error bound from Hansen [2006b] stated in Theorem B.12, since the framework in this paper is very similar to our framework.

Theorem 3.1. *Let Assumption 2.3 be satisfied, $q \in \mathbb{N}$, $x \in C^{q+1}([0,T];X)$ be the solution of (2.1), and x^n, for $t_n \in [0,T]$, be the approximations obtained by an algebraically stable and coercive Runge–Kutta method of stage order q. Further, we define*

$$\rho \coloneqq \max\Big\{ \|x\|_{L^\infty([0,T];X)}, \max_{t_n \le T} \|x^n\|_X, \max_{\substack{t_n \le T \\ i=1,\dots,s}} \|X^{ni}\|_X \Big\} \tag{3.1}$$

and assume that τ satisfies the step size restriction

$$\tau(c_{\mathrm{qm}} + L_\rho) < \alpha_{\mathrm{RK}}. \tag{3.2}$$

Here, α_{RK} is the coercivity constant of the Runge–Kutta method, cf. Definition B.5, and c_{qm} and L_ρ are given in Assumption 2.3.

Then, the error bound

$$\|x^n - x(t_n)\|_X \le C \frac{\mathrm{e}^{C_{\mathrm{RK}} C_{\tau,\rho}^2 (c_{\mathrm{qm}} + L_\rho) t_n} - 1}{C_{\mathrm{RK}}(c_{\mathrm{qm}} + L_\rho)} \tau^q$$

holds true with a constant C_{RK} only depending on the coefficients of the Runge–Kutta method, a constant C which depends on x, T and the Runge–Kutta method, but is independent of n and τ, and the constant

$$C_{\tau,\rho} = \big(\alpha_{\mathrm{RK}} - \tau(c_{\mathrm{qm}} + L_\rho)\big)^{-1}.$$

Proof. We use a standard trick and replace G in (2.1) by G_ρ, where G_ρ is globally Lipschitz continuous with constant L_ρ and satisfies $G_\rho(y) = G(y)$ for all $\|y\|_X \le \rho$. Note that due to (3.1), we have that x is also a solution of the modified equation with corresponding Runge–Kutta approximation x^n, i.e., we now consider (B.1) with $\mathcal{F}(t,x) = -S(x) + G_\rho(x)$. We then have that Assumption B.6 is satisfied with $c_{\mathrm{qm},\mathcal{F}} = c_{\mathrm{qm}} + L_\rho$ and can apply Theorem B.12.

This immediately gives the assertion. □

Theorem 3.2. *Let Assumption 2.3 be satisfied, $q \in \mathbb{N}$, and let $x \in C^{q+1}([0,T];X)$ be the solution of (2.1). We define*

$$\rho \coloneqq 2\|x\|_{L^\infty([0,T];X)}$$

and consider an algebraically stable and coercive Runge–Kutta method of stage order q.

Then, there exists a $\tau^* > 0$, s.t. for all $\tau < \tau^*$ the Runge–Kutta method applied to (2.1) yields unique approximations $x^n \in X$ with $\|x^n\|_X \leq \rho$, $t_n \in [0,T]$, which satisfy the error bound (4.7).

Proof. As in the proof of Theorem 3.1, we replace G in (2.1) by a function G_ρ that coincides with G on $\{y \in X \mid \|y\|_X \leq \rho\}$ and is globally Lipschitz continuous with constant L_ρ. Then, for the modified equation

$$x' + \mathcal{S}(x) = G_\rho(x) + g \tag{3.3}$$

Assumption B.6 is satisfied with $c_{\mathrm{qm},\mathcal{F}} = c_{\mathrm{qm}} + L_\rho$. Due to $\|x(t)\|_X \leq \rho$ for all $t \in [0,T]$, we further have that x is also a solution of (3.3). We obtain by Lemma B.8 that under the step size restriction (3.2) there exist unique Runge–Kutta iterations $x^n, t_n \in [0,T]$, to (3.3) and by Theorem B.12, that the error bound (4.7) holds true.

It remains to show that, for τ sufficiently small, we have $\|x^n\|_X, \|X^{ni}\|_X \leq \rho$ for all $t_n \leq T$ and $i = 1,\dots,s$, since then, the x^n are also the Runge–Kutta approximations to the original equation (2.1).

Using the bound (4.7), we can conclude that for $\tau^* > 0$ sufficiently small and all $\tau < \tau^*$ we have

$$\|x^n\|_X \leq \|x^n - x(t_n)\|_X + \|x(t_n)\|_X \leq 2\frac{\rho}{2} = \rho.$$

For the inner stages we calculate

$$\|X^{ni}\|_X \leq \|X^{ni} - \overline{X^{ni}}\| + \|\overline{X^{ni}} - x(t_n + c_i\tau)\| + \|x(t_n + c_i\tau)\|_X.$$

Here, $\overline{X^{ni}}, i = 1,\dots,s$, denote the inner stages of the Runge–Kutta method applied to the modified equation (3.3) starting from $x(t_n)$ at time t_n. By the local error and the stability bounds (B.5) and (B.6), respectively, we obtain

$$\begin{aligned} \|\overline{X^{ni}} - x(t_n + c_i\tau)\|_X &\leq C C_{\tau,\rho}\tau^{q+1}, \\ \|X^{ni} - \overline{X^{ni}}\|_X &\leq C_{\mathrm{RK}} C_{\tau,\rho}\|x^n - x(t_n)\|_X \leq C\big(\mathrm{e}^{C_{\mathrm{RK}} C_{\tau,\rho} t_n} - 1\big)\tau^q. \end{aligned}$$

Hence, by possibly further reducing τ^*, we have for all $\tau < \tau^*$ and $i = 1,\dots,s$

$$\|X^{ni}\|_X \leq \rho.$$

□

3.2 Runge–Kutta methods for second-order evolution equations

We now use Theorems 3.1 and 3.2 to prove error bounds for Runge–Kutta methods applied to the first-order formulation of the second-order equation (2.19).

Corollary 3.3. *Let Assumption 2.12 be satisfied, $q \in \mathbb{N}$, and let $u \in C^{q+2}([0,T];H) \cap C^{q+1}([0,T];V)$ be the solution of (2.19). Further, let u^n, v^n, for $t_n \in [0,T]$, be the approximations obtained by an algebraically stable and coercive Runge–Kutta method of stage order q applied to the first-order formulation* *of (2.19). By α_{RK} we denote the coercivity constant of the Runge–Kutta method, cf. Definition B.5, and we define*

$$\rho := \max\big\{\|u\|_{L^\infty([0,T];V)}, \max_{t_n\le T}\|u^n\|_{\tilde a}, \max_{\substack{t_n\le T\\ i=1,\dots,s}}\|U^{ni}\|_{\tilde a}\big\},$$

where U^{ni} denotes the first component of the inner Runge–Kutta stages. If τ satisfies the step size restriction

$$\tau(c_{\mathrm{qm}}+L_\rho) < \alpha_{\mathrm{RK}}$$

with $c_{\mathrm{qm}} = \frac{1}{2}c_G C_{H,V} + \beta_{\mathrm{qm}}$, the error bound

$$\|u^n - u(t_n)\|_{\tilde a} + \|v^n - u'(t_n)\|_m \le C\frac{\mathrm{e}^{C_{\mathrm{RK}}C_{\tau,\rho}^2(c_{\mathrm{qm}}+L_\rho)t_n} - 1}{C_{\mathrm{RK}}(c_{\mathrm{qm}}+L_\rho)}\tau^q \tag{3.4}$$

holds true with a constant C_{RK} only depending on the coefficients of the Runge–Kutta method, a constant C which depends on x, T and the Runge–Kutta method, but is independent of n and τ, and the constant

$$C_{\tau,\rho} = \big(\alpha_{\mathrm{RK}} - \tau(c_{\mathrm{qm}}+L_\rho)\big)^{-1}.$$

Proof. This is a direct application of Theorem 3.1, since by Corollary 2.15, we have that the first-order formulation of (2.19) satisfies Assumption 2.3. Note that we only need bounds on the first components of the exact and the numerical solution, since G in the first-order formulation of (2.19) only depends on the first component. Hence, it is sufficient that u, u^n, and U^{ni} are bounded to exploit the local Lipschitz continuity. □

Corollary 3.4. *Let Assumption 2.12 be satisfied, $q\in\mathbb{N}$, and let $u \in C^{q+2}([0,T];H)\cap C^{q+1}([0,T];V)$ be the solution of (4.1). We define*

$$\rho := 2\|u\|_{L^\infty([0,T];V)}$$

and consider an algebraically stable and coercive Runge–Kutta method of stage order q.

Then, there exists a $\tau^* > 0$, s.t. for all $\tau < \tau^*$ the Runge–Kutta method yields unique approximations $u^n \in V, v^n \in H$ with $\|u^n\|_{\tilde a} \le \rho, n \ge 0, t_n \in [0,T]$ which satisfy the error bound (3.4).

Proof. This is a direct application of Theorem 3.2 with the same arguments as in the proof of Corollary 3.3. □

Remark 3.5 (BDF methods)**.** *The paper Hansen [2006a] contains an error analysis for BDF methods in the same framework as used in Hansen [2006b] for the analysis of Runge–Kutta methods. This can be adapted with the same techniques as presented in this chapter to our framework to obtain error bounds for BDF methods applied to (2.1) and (2.18).*

CHAPTER 4

An implicit-explicit (IMEX) scheme for semilinear second-order evolution equations

In this section, we present and analyze an efficient IMEX time integration scheme for semilinear second-order evolution equations in the setting of Section 2.2.1. By semilinear we mean that the nonlinear part of the evolution equation is Lipschitz continuous. The scheme is a combination of the implicit Crank–Nicolson method and the explicit leapfrog scheme. We show wellposedness of the scheme, comment on the efficiency, and, as the main result of this section, prove a second-order error bound.

This chapter mainly presents the content of [Hochbruck and Leibold, 2021, Section 2 (2.2-2.5)]. We always refer to the corresponding results in this paper.

Outline In Section 4.1, we introduce the analytical framework in which we consider and analyze the IMEX scheme. Since the construction and analysis of the IMEX scheme is based on the Crank–Nicolson method, we present the numerical analysis of the Crank–Nicolson method in Section 4.2. Section 4.3 is devoted to the construction of the IMEX scheme of which we prove wellposedness in Section 4.4. For the error analysis, it is advantageous to consider a first-order formulation of the IMEX scheme, which we derive in Section 4.5. Finally, we prove a second-order error bound for the IMEX scheme in Section 4.6.

4.1 Analytical setting

We consider the evolution equation (2.18) in the setting from section Section 2.2 in the semilinear case, i.e., where $Bv \coloneqq \mathcal{D}(v)$ is a linear operator. The equation then takes the form

$$u'' + Bu' + Au = F(u) + f, \quad t \geq 0, \qquad u(0) = u^0, \quad u'(0) = v^0. \tag{4.1}$$

In the first-order formulation (2.20), we thus have

$$\mathcal{S}(x) = Sx, \qquad \text{where} \qquad S = \begin{bmatrix} 0 & -\mathrm{I} \\ A & B \end{bmatrix}$$

is a linear operator. The corresponding first-order equation (2.1) in this case is of the form

$$x' + Sx = G(x) + g, \quad t \geq 0, \qquad x(0) = x_0. \tag{4.2}$$

Remark 4.1. *Note that part c) of Assumption 2.12 translates for (4.1) to: The operator $B \in \mathcal{L}(V; H)$ is quasi-monotone, i.e., there is a constant $\beta_{\mathrm{qm}} \geq 0$ s.t.*

$$m\big(Bv, v\big) \geq -\beta_{\mathrm{qm}} \|v\|_m^2. \tag{4.3}$$

As in Chapter 3 we set $t_n \coloneqq n\tau, n \geq 0$, and denote by $x^n = [u^n, v^n]^\mathsf{T}$ the iterates of a time discretization scheme applied to (4.2) or (4.1), respectively. Furthermore, to simplify the following presentation, we use the short notations

$$G^n = \begin{bmatrix} 0 \\ F^n \end{bmatrix} = \begin{bmatrix} 0 \\ F(u^n) + f(t_n) \end{bmatrix} = G(x^n) + g(t_n).$$

4.2 Motivation: The Crank–Nicolson scheme

Since we derive the IMEX scheme as an adaption of the Crank–Nicolson scheme, and the error analysis is based on the Crank–Nicolson error analysis, we start by recalling the Crank–Nicolson scheme and some of its properties. The Crank–Nicolson scheme is a time integration method for first-order equations, see, e.g., [Hairer and Wanner, 2010, Section IV.3]. Applied to the first-order formulation (4.2) of (4.1) it is of the form

$$x^{n+1} = x^n + \frac{\tau}{2}\big(- S(x^n + x^{n+1}) + G^n + G^{n+1}\big) \tag{4.4}$$

and can be written as

$$R_+ x^{n+1} = R_- x^n + \frac{\tau}{2}\left(G^n + G^{n+1}\right) \quad \text{with} \quad R_\pm = I \pm \frac{\tau}{2} S. \tag{4.5}$$

The operators $R_\pm$ have the following properties:

Lemma 4.2 ([Hochbruck and Leibold, 2021, Lemma 2.4])**.** *Let Assumption 2.12 be satisfied and $c_{\mathrm{qm}} = \frac{1}{2} c_G C_{H,V} + \beta_{\mathrm{qm}}$ with $C_{H,V}$ defined in (2.14). Then, for $\tau c_{\mathrm{qm}} < 2$, the following assertions hold true:*

a) R_+ is invertible with $\|R_+^{-1}\|_{X \leftarrow X} \leq 1$ and $R_+^{-1} x \in D(S)$ for all $x \in X$.

b) $R \coloneqq R_+^{-1} R_-$ has a continuous extension on X satisfying $\|R\|_{X\leftarrow X} \leq \mathrm{e}^{\tau c_{\mathrm{qm}}}$.

Proof. By Lemma 2.14 we have that S is maximal and quasi-monotone with constant c_{qm}. This implies the stated properties of $R_\pm$, as shown in the proof of [Hipp, 2017, Lemma 2.14]. □

In the following, we assume $\tau c_{\mathrm{qm}} < 2$, s.t. Lemma 4.2 is valid. This allows us to apply R_+^{-1} to (4.5) and we obtain

$$x^{n+1} = Rx^n + \frac{\tau}{2} R_+^{-1} \left(G^n + G^{n+1} \right). \tag{4.6}$$

We will now prove an error bound for the Crank–Nicolson scheme. This was done in Hipp [2017] for the linear and in Leibold [2017] for the semilinear case. The idea of the proof is based on Sturm [2017], where the Crank–Nicolson scheme applied to Maxwell equations was analyzed.

Theorem 4.3. *Let Assumption 2.12 be satisfied, $u \in C^4([0,T];H) \cap C^3([0,T];V)$ be the solution of (4.1), and $x^n = [u^n, v^n]^\intercal$, $t_n \in [0,T]$, be the approximations obtained by the Crank–Nicolson scheme (4.4). Further we set*

$$\rho \coloneqq \max \left\{ \|u\|_{L^\infty([0,T];V)}, \max_{t_n \leq T} \|u^n\|_{\tilde{a}} \right\}$$

and assume that τ satisfies the step size restriction

$$\tau < \min \left\{ \frac{2}{c_{\mathrm{qm}}}, \frac{1}{L_\rho} \right\}$$

with $c_{\mathrm{qm}} = \frac{1}{2} c_G C_{H,V} + \beta_{\mathrm{qm}}$.

Then, the error bound

$$\|u^n - u(t_n)\|_{\tilde{a}} + \|v^n - u'(t_n)\|_m \leq C \mathrm{e}^{\left(c_{\mathrm{qm}} + \frac{L_\rho}{1 - L_\rho \tau}\right) t_n} \tau^2 \tag{4.7}$$

holds true with a constant C which depends on x and T but is independent of n and τ.

Proof. We use the notation

$$\widetilde{x}^n = x(t_n), \qquad \widetilde{G}^n = G(\widetilde{x}^n) + g(t_n),$$

where $x = [u, u']^\intercal$ is the exact solution of (4.2), and denote the error by

$$e^n = x^n - \widetilde{x}^n.$$

The proof consists of three main steps.

(a) Error recursion. We insert the exact solution into the Crank–Nicolson scheme (4.4) and obtain

$$\widetilde{x}^{n+1} = \widetilde{x}^n + \frac{\tau}{2} \left(-S(\widetilde{x}^n + \widetilde{x}^{n+1}) + \widetilde{G}^n + \widetilde{G}^{n+1} \right) - \delta_{\mathrm{CN}}^{n+1} \tag{4.8}$$

where $\delta_{\mathrm{CN}}^{n+1}$ is a defect. By the differential equation (4.2) we have

$$\frac{\tau}{2} \left(x'(t_{n+1}) + x'(t_n) \right) = \frac{\tau}{2} \left(-S(\widetilde{x}^n + \widetilde{x}^{n+1}) + \widetilde{G}^n + \widetilde{G}^{n+1} \right),$$

and by comparing to (4.8), we derive the following representation of the defect:

$$\delta_{\mathrm{CN}}^{n+1} = \frac{\tau}{2}\left(x'(t_{n+1}) + x'(t_n)\right) - \left(\widetilde{x}^{n+1} - \widetilde{x}^{n}\right) = \frac{\tau}{2}\left(x'(t_{n+1}) + x'(t_n)\right) - \int_{t_n}^{t_{n+1}} x'(s)\,\mathrm{d}s. \tag{4.9}$$

Similar to (4.6), we can reformulate (4.8) to

$$\widetilde{x}^{n+1} = R\widetilde{x}^{n} + \frac{\tau}{2}R_+^{-1}\left(\widetilde{G}^{n} + \widetilde{G}^{n+1}\right) + R_+^{-1}\delta_{\mathrm{CN}}^{n+1}. \tag{4.10}$$

By subtracting this equation from (4.6), we obtain the error recursion

$$e^{n+1} = Re^{n} + \frac{\tau}{2}R_+^{-1}\left(G^{n} - \widetilde{G}^{n} + G^{n+1} - \widetilde{G}^{n+1}\right) + R_+^{-1}\delta_{\mathrm{CN}}^{n+1}. \tag{4.11}$$

(b) Stability. By solving the error recursion (4.11) with $e^0 = 0$, we obtain

$$e^{n} = \sum_{m=1}^{n} R^{n-m}\left(\frac{\tau}{2}R_+^{-1}\left(G^{m} - \widetilde{G}^{m} + G^{m-1} - \widetilde{G}^{m-1}\right) + R_+^{-1}\delta_{\mathrm{CN}}^{m}\right).$$

Taking the norm, using the triangle inequality, and $\|R\|_{X\leftarrow X} \le e^{\tau c_{\mathrm{qm}}}$, $\|R_+^{-1}\|_{X\leftarrow X} \le 1$ from Lemma 4.2 yields

$$\|e^{n}\|_X \le \tau\sum_{m=1}^{n} \mathrm{e}^{(n-m)\tau c_{\mathrm{qm}}}\left(\frac{1}{2}\left\|G^{m} - \widetilde{G}^{m}\right\|_X + \frac{1}{2}\|G^{m-1} - \widetilde{G}^{m-1}\|_X\right) + \sum_{m=1}^{n} \mathrm{e}^{(n-m)\tau c_{\mathrm{qm}}}\left\|\delta_{\mathrm{CN}}^{m}\right\|_X.$$

Since we have $\|u(t)\|_{\tilde{a}}, \|u^n\|_{\tilde{a}} \le \rho$ for all $t_n, t \in [0,T]$, we can employ the local Lipschitz continuity of G and end up with

$$\mathrm{e}^{-n\tau c_{\mathrm{qm}}}\|e^{n}\|_X \le L_\rho\tau\sum_{m=1}^{n}\mathrm{e}^{-m\tau c_{\mathrm{qm}}}\|e^{m}\|_X + \sum_{m=1}^{n}\left\|\delta_{\mathrm{CN}}^{m}\right\|_X.$$

By applying Grönwall's lemma stated in Lemma A.1 b) and multiplying by $\mathrm{e}^{n\tau c_{\mathrm{qm}}}$, we obtain for $\tau < 1/L_\rho$

$$\|e^{n}\|_X \le \mathrm{e}^{\left(c_{\mathrm{qm}} + \frac{L_\rho}{1-L_\rho\tau}\right)t_n}\sum_{m=1}^{n}\|\delta_{\mathrm{CN}}^{m}\|_X. \tag{4.12}$$

(c) Defect. The Crank–Nicolson defect (4.9) consists of the quadrature error of the trapezoidal rule, which is due to our regularity assumptions bounded by

$$\|\delta_{\mathrm{CN}}^{m}\|_X \le C\tau^3\|x^{(3)}\|_{L^\infty([t_m,t_{m-1}];X)} \le C\tau^3\left(\|u^{(3)}\|_{L^\infty([t_m,t_{m-1}];V)} + \|u^{(4)}\|_{L^\infty([t_m,t_{m-1}];H)}\right) \le C\tau^3. \tag{4.13}$$

Inserting this into (4.12) finally yields

$$\begin{aligned}\|u^{n} - u(t_n)\|_{\tilde{a}} + \|v^{n} - u'(t_n)\|_m \le \sqrt{2}\|x^{n} - x(t_n)\|_X = \sqrt{2}\|e^{n}\|_X &\le C\mathrm{e}^{\left(c_{\mathrm{qm}} + \frac{L_\rho}{1-L_\rho\tau}\right)t_n}\sum_{m=1}^{n}\tau^3 \\ &\le C\mathrm{e}^{\left(c_{\mathrm{qm}} + \frac{L_\rho}{1-L_\rho\tau}\right)t_n}t_n\tau^2.\end{aligned}$$

□

4.3 Construction of the IMEX scheme

In the following, we explain how the IMEX scheme can be derived by combining the Crank–Nicolson with the leapfrog scheme. To do so, we state now a formulation of the Crank–Nicolson scheme that exploits the second-order structure of (4.1).

Lemma 4.4 ([Hochbruck and Leibold, 2021, Lemma 2.5])**.** *The Crank–Nicolson scheme* (4.4) *can equivalently be rewritten in a half-full-half step formulation*

$$\begin{aligned} v^{n+\frac{1}{2}} &= v^n - \frac{\tau}{2}Au^n - \frac{\tau^2}{4}Av^{n+\frac{1}{2}} - \frac{\tau}{2}Bv^{n+\frac{1}{2}} + \frac{\tau}{4}\big(F^n + F^{n+1}\big), && \text{(4.14a)}\\ u^{n+1} &= u^n + \tau v^{n+\frac{1}{2}}, && \text{(4.14b)}\\ v^{n+1} &= v^{n+\frac{1}{2}} - \frac{\tau}{2}Au^n - \frac{\tau^2}{4}Av^{n+\frac{1}{2}} - \frac{\tau}{2}Bv^{n+\frac{1}{2}} + \frac{\tau}{4}\big(F^n + F^{n+1}\big). && \text{(4.14c)} \end{aligned}$$

Proof. With

$$v^{n+\frac{1}{2}} := \frac{1}{2}\left(v^n + v^{n+1}\right), \tag{4.15}$$

the two components of (4.4) have the form

$$\begin{aligned} u^{n+1} &= u^n + \tau v^{n+\frac{1}{2}},\\ v^{n+1} &= v^n - \frac{\tau}{2}A(u^n + u^{n+1}) - \tau Bv^{n+\frac{1}{2}} + \frac{\tau}{2}(F^n + F^{n+1}). \end{aligned}$$

The first equation is identical to (4.14b). In the second equation, we eliminate u^{n+1} using the first one and obtain

$$v^{n+1} = v^n - \tau Au^n - \frac{\tau^2}{2}Av^{n+\frac{1}{2}} - \tau Bv^{n+\frac{1}{2}} + \frac{\tau}{2}\big(F^n + F^{n+1}\big),$$

which can be expressed equivalently by the two half steps (4.14a) and (4.14c). □

The leapfrog or Störmer–Verlet scheme is an explicit time integration scheme for second-order differential equations of the form $y'' = \phi(t, y)$, cf. Hairer et al. [2006]. Applied to (4.1) with $A = B = 0$, the scheme can be expressed in a half-full-half step formulation similar to (4.14) via

$$\begin{aligned} v^{n+\frac{1}{2}} &= v^n + \frac{\tau}{2}F^n,\\ u^{n+1} &= u^n + \tau v^{n+\frac{1}{2}},\\ v^{n+1} &= v^{n+\frac{1}{2}} + \frac{\tau}{2}F^{n+1}. \end{aligned}$$

By combining the Crank–Nicolson scheme for the linear part of (4.1) with the leapfrog scheme for the nonlinear part we obtain the following IMEX scheme:

$$\begin{aligned} v^{n+\frac{1}{2}} &= v^n - \frac{\tau}{2}Au^n - \frac{\tau^2}{4}Av^{n+\frac{1}{2}} - \frac{\tau}{2}Bv^{n+\frac{1}{2}} + \frac{\tau}{2}F^n, && \text{(4.16a)}\\ u^{n+1} &= u^n + \tau v^{n+\frac{1}{2}}, && \text{(4.16b)}\\ v^{n+1} &= v^{n+\frac{1}{2}} - \frac{\tau}{2}Au^n - \frac{\tau^2}{4}Av^{n+\frac{1}{2}} - \frac{\tau}{2}Bv^{n+\frac{1}{2}} + \frac{\tau}{2}F^{n+1}. && \text{(4.16c)} \end{aligned}$$

Remark 4.5 ([Hochbruck and Leibold, 2021, Remark 2.6]). *By subtracting (4.16a) from (4.16c), we obtain an equivalent representation of v^{n+1}*

$$v^{n+1} = -v^n + 2v^{n+\frac{1}{2}} + \frac{\tau}{2}\left(F^{n+1} - F^n\right) \tag{4.16d}$$

which is computationally more efficient because of the elimination of the operators A and B.

The implementation is comprised by solving the linear system in (4.16a), and then computing (4.16b) and (4.16d). Altogether, each time step requires the solution of one linear system, one application of A and one evaluation of the nonlinearity. Note that F^{n+1} can be reused in the next time step.

4.4 Wellposedness of the IMEX scheme

The linear system that has to be solved in (4.16a) is of the form

$$Q_+ v^{n+\frac{1}{2}} = v^n - \frac{\tau}{2}Au^n + \frac{\tau}{2}F^n, \tag{4.17}$$

with $Q_\pm \colon D(A) \to H$ given by

$$Q_\pm = \mathrm{I} \pm \frac{\tau}{2}B \pm \frac{\tau^2}{4}A.$$

These operators play an important role in the analysis of the IMEX method and satisfy the following properties:

Lemma 4.6 ([Hochbruck and Leibold, 2021, Lemma 2.7]). *Let Assumption 2.12 be satisfied and*

$$\frac{\tau^2}{2}c_G + \tau\beta_{\mathrm{qm}} \leq 1. \tag{4.18}$$

Then, the operator Q_+ is invertible and its inverse $Q_+^{-1} \colon H \to D(A)$ satisfies the bounds

$$\left\|\left(\frac{\tau}{2}B + \frac{\tau^2}{4}A\right)Q_+^{-1}\right\|_{H\leftarrow H} \leq 1, \tag{4.19a}$$

$$\left\|Q_+^{-1}\right\|_{V\leftarrow H} \leq \frac{\sqrt{2}}{\tau}, \tag{4.19b}$$

$$\left\|Q_- Q_+^{-1}\right\|_{H\leftarrow H} \leq \mathrm{e}^{\frac{\tau^2}{2}c_G + \tau\beta_{\mathrm{qm}}}. \tag{4.19c}$$

Proof. By $b\colon V \times V$ we denote the bilinear form associated to the operator B, i.e.,

$$b(v, w) = m(Bv, w) \qquad \text{for all } v, w \in V.$$

The quasi-monotonicity of B (4.3) transfers directly to b. Together with Assumption 2.12 b) and the step size restriction (4.18) we see that the bilinear form

$$m + \frac{\tau}{2}b + \frac{\tau^2}{4}a = \underbrace{\left(1 - \frac{\tau}{2}\beta_{\mathrm{qm}} - \frac{\tau^2}{4}c_G\right)}_{\geq 0} m + \frac{\tau}{2}(b + \beta_{\mathrm{qm}}m) + \frac{\tau^2}{4}\tilde{a} \colon V \times V \to \mathbb{R}$$

is coercive, cf. Definition A.2, as the sum of the coercive bilinear form $\frac{\tau^2}{4}\tilde{a}$ and two monotone bilinear forms. Hence, by the Lax–Milgram lemma (Theorem A.3), we have that for a given $v \in H \subset V^*$ there exists a unique $z \in V$ such that

$$m(z,w) + \frac{\tau}{2} b(z,w) + \frac{\tau^2}{4} a(z,w) = \langle v, w\rangle_{V^*\times V} = m(v,w) \qquad \text{for all } w \in V,$$

where we used the Gelfand triple structure (2.16). Using (2.17), we can rewrite this equivalently as

$$\frac{\tau^2}{4}\langle \mathcal{A}z, w\rangle_{V^*\times V} = \frac{\tau^2}{4} a(z,w) = m\big(v - z - \frac{\tau}{2}Bz, w\big) = \langle v - z - \frac{\tau}{2}Bz, w\rangle_{V^*\times V} \qquad \text{for all } w \in V.$$

Thus, we have

$$\frac{\tau^2}{4}\mathcal{A}z = v - z - \frac{\tau}{2}Bz \in H$$

which implies $z \in D(A)$ and $Q_+ z = (\mathrm{I} + \frac{\tau}{2}B + \frac{\tau^2}{4}A)z = v$. This proves that Q_+ is invertible.

We now show the bounds (4.19): Let $v \in H$ and set $z = Q_+^{-1}v \in D(A)$. Using Assumption 2.12 b), c), and the step size restriction (4.18), we obtain

$$\begin{aligned}
\|v\|_m^2 &= \Big\|\Big(\mathrm{I} + \frac{\tau}{2}B + \frac{\tau^2}{4}A\Big)z\Big\|_m^2 \\
&= \|z\|_m^2 + \Big\|\Big(\frac{\tau}{2}B + \frac{\tau^2}{4}A\Big)z\Big\|_m^2 + 2m\Big(\Big(\frac{\tau}{2}B + \frac{\tau^2}{4}A\Big)z, z\Big) \\
&= \Big(1 - \frac{\tau^2}{2}c_G - \tau\beta_{\mathrm{qm}}\Big)\|z\|_m^2 + \Big\|\Big(\frac{\tau}{2}B + \frac{\tau^2}{4}A\Big)z\Big\|_m^2 \\
&\qquad + 2\frac{\tau}{2}m\big((B + \beta_{\mathrm{qm}}\,\mathrm{I})z, z\big) + 2\frac{\tau^2}{4}m\big((A + c_G\,\mathrm{I})z, z\big) \\
&\geq \Big\|\Big(\frac{\tau}{2}B + \frac{\tau^2}{4}A\Big)Q_+^{-1}v\Big\|_m^2 + \frac{\tau^2}{2}\|Q_+^{-1}v\|_{\tilde{a}}^2.
\end{aligned}$$

This directly implies the bounds (4.19a) and (4.19b).

We further note that $B + \frac{\tau}{2}A$ is maximal and quasi-monotone with constant $\beta_{\mathrm{qm}} + \frac{\tau}{2}c_G$. The bound (4.19c) then follows by Lemma 4.2 with S replaced by $B + \frac{\tau}{2}A$. □

By Lemma 4.6 we obtain directly the wellposedness of the IMEX scheme:

Corollary 4.7 ([Hochbruck and Leibold, 2021, Corollary 2.8]). *The IMEX scheme (4.16) is wellposed in $D(A) \times H$, i.e., for $u^0 \in D(A)$ and $v^0 \in H$, the numerical approximations satisfy*

$$u^n \in D(A), \qquad v^n \in H, \qquad v^{n+\frac{1}{2}} \in D(A), \qquad n \geq 0.$$

Proof. We prove this by induction over n. The statement holds for $n = 0$ by assumption. We now assume that $u^n \in D(A)$ and $v^n \in H$ for some $n \geq 0$. By Lemma 4.6, we have that Q_+ is invertible and hence, (4.17) implies $v^{n+\frac{1}{2}} \in D(A)$. By (4.16b) and (4.16c) we then immediately obtain

$$u^{n+1} = u^n + \tau v^{n+\frac{1}{2}} \in D(A) \quad \text{and} \quad v^{n+1} = v^{n+\frac{1}{2}} - \frac{\tau}{2}Au^n - \frac{\tau^2}{4}Av^{n+\frac{1}{2}} - \frac{\tau}{2}Bv^{n+\frac{1}{2}} + \frac{\tau}{2}F^{n+1} \in H.$$

□

4.5 IMEX scheme in first-order formulation

To derive an error bound, we rewrite the IMEX scheme (4.16) as a perturbation of the first-order formulation of the Crank–Nicolson scheme (4.6). This formulation of the Crank–Nicolson scheme was used in Theorem 4.3 to prove an error bound for the Crank–Nicolson scheme; we will adapt this for the IMEX scheme. A similar idea was used in Hochbruck and Sturm [2016] to analyze the leapfrog scheme and locally implicit schemes for Maxwell equations.

Lemma 4.8 ([Hochbruck and Leibold, 2021, Lemma 2.10])**.**

a) The operators R_+^{-1} and R can be expressed via

$$R_+^{-1} = \begin{bmatrix} Q_+^{-1}\left(\mathrm{I} + \frac{\tau}{2}B\right) & \frac{\tau}{2}Q_+^{-1} \\ -\frac{2}{\tau} + \frac{2}{\tau}Q_+^{-1}(\mathrm{I} + \frac{\tau}{2}B) & Q_+^{-1} \end{bmatrix}, \tag{4.20a}$$

$$R = \begin{bmatrix} -\mathrm{I} + Q_+^{-1}\left(2\,\mathrm{I} + \tau B\right) & \tau Q_+^{-1} \\ -\frac{4}{\tau}\,\mathrm{I} + \frac{1}{\tau}Q_+^{-1}\left(4\,\mathrm{I} + 2\tau B\right) & Q_- Q_+^{-1} \end{bmatrix}. \tag{4.20b}$$

b) For all $w \in V$, we have

$$R_+^{-1}\begin{bmatrix} w \\ -Bw \end{bmatrix} = \begin{bmatrix} Q_+^{-1}w \\ -\left(B + \frac{\tau}{2}A\right)Q_+^{-1}w \end{bmatrix}. \tag{4.21}$$

c) The IMEX scheme (4.16) is equivalent to the first-order formulation

$$x^{n+1} = Rx^n + \frac{\tau}{2}R_+^{-1}\left(G^n + G^{n+1}\right) + \frac{\tau^2}{4}\begin{bmatrix} Q_+^{-1}\left(F^n - F^{n+1}\right) \\ -\left(B + \frac{\tau}{2}A\right)Q_+^{-1}\left(F^n - F^{n+1}\right) \end{bmatrix}. \tag{4.22}$$

Proof. a) First note that by Lemma 4.6 the right-hand side of (4.20a) is a well-defined mapping from $X = V \times H$ to $D(S) = D(A) \times V$. The identities (4.20) can be verified by straightforward calculations.

b) Using (4.20a), we calculate

$$\begin{aligned} R_+^{-1}\begin{bmatrix} w \\ -Bw \end{bmatrix} &= \begin{bmatrix} Q_+^{-1}\left(\mathrm{I} + \frac{\tau}{2}B\right) & \frac{\tau}{2}Q_+^{-1} \\ -\frac{2}{\tau} + \frac{2}{\tau}Q_+^{-1}(\mathrm{I} + \frac{\tau}{2}B) & Q_+^{-1} \end{bmatrix}\begin{bmatrix} w \\ -Bw \end{bmatrix} \\ &= \begin{bmatrix} Q_+^{-1}w \\ -\frac{2}{\tau}w + \frac{2}{\tau}Q_+^{-1}(\mathrm{I} + \frac{\tau}{2}B)w - Q_+^{-1}Bw \end{bmatrix}. \end{aligned}$$

For the second component, we obtain

$$\begin{aligned} -\frac{2}{\tau}w + \frac{2}{\tau}Q_+^{-1}(\mathrm{I} + \frac{\tau}{2}B)w - Q_+^{-1}Bw &= -\frac{2}{\tau}w + \frac{2}{\tau}Q_+^{-1}w \\ &= -\frac{2}{\tau}\left(\mathrm{I} + \frac{\tau}{2}B + \frac{\tau^2}{4}A\right)Q_+^{-1}w + \frac{2}{\tau}Q_+^{-1}w \\ &= -(B + \frac{\tau}{2}A)Q_+^{-1}w. \end{aligned}$$

c) We start by showing the equivalence of the IMEX scheme and (4.22) under the additional assumption $v^n, v^{n+1} \in V$. By subtracting (4.16c) from (4.16a), we obtain

$$v^{n+\frac{1}{2}} = \frac{1}{2}\left(v^n + v^{n+1}\right) + \frac{\tau}{4}\left(F^n - F^{n+1}\right), \tag{4.23}$$

which differs from the representation of $v^{n+\frac{1}{2}}$ in the Crank–Nicolson scheme by the contributions of the nonlinearity F, cf. (4.15). Inserting (4.23) into (4.16b) yields

$$u^{n+1} = u^n + \frac{\tau}{2}(v^n + v^{n+1}) + \frac{\tau^2}{4}\left(F^n - F^{n+1}\right). \tag{4.24}$$

On the other hand, by adding (4.16a) and (4.16c) and inserting (4.16b), we obtain

$$v^{n+1} = v^n - \frac{\tau}{2}A(u^n + u^{n+1}) - \tau B v^{n+\frac{1}{2}} + \frac{\tau}{2}\left(F^n + F^{n+1}\right).$$

Hence, with (4.23) we have

$$v^{n+1} = v^n - \frac{\tau}{2}A(u^n + u^{n+1}) - \frac{\tau}{2}B\left(v^n + v^{n+1}\right) + \frac{\tau}{2}\left(F^n + F^{n+1}\right) - \frac{\tau^2}{4}B\left(F^n - F^{n+1}\right). \tag{4.25}$$

Using the definition (4.5) of $R_\pm$, we can express (4.24) and (4.25) simultaneously as

$$R_+ x^{n+1} = R_- x^n + \frac{\tau}{2}\left(G^n + G^{n+1}\right) + \frac{\tau^2}{4}\begin{bmatrix} F^n - F^{n+1} \\ -B\left(F^n - F^{n+1}\right)\end{bmatrix}. \tag{4.26}$$

Note that by (4.23) we have $F^n - F^{n+1} \in V$, since we assumed $v^n, v^{n+1} \in V$ and have by Corollary 4.7 $v^{n+\frac{1}{2}} \in D(A) \subset V$. By multiplying (4.26) with R_+^{-1} and using (4.21), we obtain the representation (4.22) of the IMEX scheme under the additional assumption $v^n, v^{n+1} \in V$. Since the IMEX scheme (4.16) as well as the equation (4.22) are well defined for $v^n, v^{n+1} \in H$, and since V is dense in H, we also obtain the equivalence of both formulations for $v^n, v^{n+1} \in H$.

□

4.6 Error bound for the IMEX scheme

As the main result of this chapter, we now present a second-order error bound for the IMEX scheme.

Theorem 4.9 ([Hochbruck and Leibold, 2021, Theorem 2.9])**.** *Let Assumption 2.12 be satisfied and let $u \in C^4([0,T];H) \cap C^3([0,T];V) \cap C^2([0,T];D(A))$ be the solution of (4.1). Then, there exists $\tau^* > 0$ s.t. for all $\tau < \tau^*$ and all $t_n \in [0,T]$ the approximations u^n from the IMEX scheme (4.16) are bounded by*

$$\|u^n\|_{\tilde{a}} \le \rho := 2\|u\|_{L^\infty([0,T];V)}. \tag{4.27}$$

Moreover, the approximations u^n, v^n satisfy for all $t_n \in [0,T]$ the error bound

$$\|u^n - u(t_n)\|_{\tilde{a}} + \|v^n - u'(t_n)\|_m \le C\mathrm{e}^{Mt_n}\tau^2 \tag{4.28}$$

with $M = c_{\mathrm{qm}} + \frac{\left(1+(3/2)^{1/2}\right)L_\rho}{1-\left(1+(3/2)^{1/2}\right)L_\rho\tau}$, $c_{\mathrm{qm}} = \frac{1}{2}c_G C_{H,V} + \beta_{\mathrm{qm}}$, and a constant C that only depends on u and T but is independent of τ and L.

Proof. For the proof of the error bound (4.28) we use the first-order formulation (4.22) of the IMEX scheme and the notation

$$\widetilde{x}^n = \begin{bmatrix}\widetilde{u}^n \\ \widetilde{v}^n\end{bmatrix} = \begin{bmatrix} u(t_n) \\ u'(t_n)\end{bmatrix}, \qquad \widetilde{G}^n = G(\widetilde{x}^n) + g(t_n) = \begin{bmatrix} 0 \\ F(\widetilde{u}^n) + f(t_n)\end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{F}^n\end{bmatrix}$$

for the exact solution u of (4.2). Further, we denote the first-order error by

$$e^n = x^n - \widetilde{x}^n.$$

Let τ be sufficiently small, such that the assumptions of Lemmas 4.2 and 4.6 are satisfied. Further, we assume for the moment that for all approximations $u^n, t_n \in [0,T]$, (4.27) is satisfied. At the end of the proof, we will show that this is valid for sufficiently small τ.

The proof consists of four main steps, where the first three steps are similar to the error proof of the Crank–Nicolson scheme from Theorem 4.3. In the last step we show the boundedness of the approximations.

(a) Error recursion. To derive a recursion for the error, we insert the exact solution into the IMEX scheme in first-order formulation (4.22) and obtain

$$\widetilde{x}^{n+1} = R\widetilde{x}^n + \frac{\tau}{2}R_+^{-1}\left(\widetilde{G}^{n+1} + \widetilde{G}^n\right) + \frac{\tau^2}{4}\begin{bmatrix} Q_+^{-1}\bigl(\widetilde{F}^n - \widetilde{F}^{n+1}\bigr) \\ -\left(B + \frac{\tau}{2}A\right)Q_+^{-1}\bigl(\widetilde{F}^n - \widetilde{F}^{n+1}\bigr)\end{bmatrix} - \Delta^{n+1} \tag{4.29}$$

with a defect Δ^{n+1} which is yet to be determined. Comparing (4.29) with (4.10), we can interpret the defect as a perturbation of the Crank–Nicolson defect $\delta_{\mathrm{CN}}^{n+1}$ via

$$\Delta^{n+1} = R_+^{-1}\delta_{\mathrm{CN}}^{n+1} + \widetilde{\delta}^{n+1}, \qquad \text{where} \quad \widetilde{\delta}^{n+1} = \frac{\tau^2}{4}\begin{bmatrix} Q_+^{-1}\bigl(\widetilde{F}^n - \widetilde{F}^{n+1}\bigr) \\ -\left(B + \frac{\tau}{2}A\right)Q_+^{-1}\bigl(\widetilde{F}^n - \widetilde{F}^{n+1}\bigr)\end{bmatrix}. \tag{4.30}$$

Subtracting (4.29) from (4.22) yields the error recursion

$$\begin{aligned} e^{n+1} = Re^n &+ \frac{\tau}{2}R_+^{-1}\left(G^{n+1} - \widetilde{G}^{n+1} + G^n - \widetilde{G}^n\right) \\ &+ \frac{\tau^2}{4}\begin{bmatrix} Q_+^{-1}\bigl(F^n - \widetilde{F}^n - F^{n+1} + \widetilde{F}^{n+1}\bigr) \\ -\left(B + \frac{\tau}{2}A\right)Q_+^{-1}\bigl(F^n - \widetilde{F}^n - F^{n+1} + \widetilde{F}^{n+1}\bigr)\end{bmatrix} + \Delta^{n+1}. \end{aligned} \tag{4.31}$$

(b) Stability. Solving the error recursion (4.31) with $e^0 = 0$ gives

$$\begin{aligned} e^n = \sum_{m=1}^{n} R^{n-m}\Biggl(&\frac{\tau}{2}R_+^{-1}\left(G^m - \widetilde{G}^m + G^{m-1} - \widetilde{G}^{m-1}\right) \\ &+ \frac{\tau^2}{4}\begin{bmatrix} Q_+^{-1}\bigl(F^{m-1} - \widetilde{F}^{m-1} - F^m + \widetilde{F}_h^m\bigr) \\ -\left(B + \frac{\tau}{2}A\right)Q_+^{-1}\bigl(F^{m-1} - \widetilde{F}^{m-1} - F^m + \widetilde{F}^m\bigr)\end{bmatrix} + \Delta^m\Biggr). \end{aligned}$$

Taking the norm, using the triangle inequality, and the bounds from Lemma 4.2 yields

$$\begin{aligned} \|e^n\|_X \leq \tau\sum_{m=1}^{n} \mathrm{e}^{(n-m)\tau c_{\mathrm{qm}}}\Biggl(&\frac{1}{2}\bigl\|G^m - \widetilde{G}^m\bigr\|_X + \frac{1}{2}\|G^{m-1} - \widetilde{G}^{m-1}\|_X \\ &+ \frac{\tau}{4}\left\|\begin{bmatrix} Q_+^{-1}\bigl(F^{m-1} - \widetilde{F}^{m-1}\bigr) \\ -\left(B + \frac{\tau}{2}A\right)Q_+^{-1}\bigl(F^{m-1} - \widetilde{F}^{m-1}\bigr)\end{bmatrix}\right\|_X \\ &+ \frac{\tau}{4}\left\|\begin{bmatrix} Q_+^{-1}\bigl(F^m - \widetilde{F}^m\bigr) \\ -\left(B + \frac{\tau}{2}A\right)Q_+^{-1}\bigl(F^m - \widetilde{F}^m\bigr)\end{bmatrix}\right\|_X\Biggr) \\ &+ \left\|\sum_{m=1}^{n} R^{n-m}\Delta^m\right\|_X. \end{aligned} \tag{4.32}$$

Since we have $\|u(t)\|_{\tilde{a}}, \|u^n\|_{\tilde{a}} \leq \rho$ for all $t_n, t \in [0,T]$, we can further investigate the terms in the first sum by employing the local Lipschitz continuity of G and F, respectively. We have

$$\big\|G^m - \widetilde{G}^m\big\|_X \leq L_\rho \|e^m\|_X, \tag{4.33}$$

and, by using $\|Q_+^{-1}\|_{V \leftarrow H} \leq \frac{\sqrt{2}}{\tau}$, $\Big\|\big(B + \frac{\tau}{2}A\big)\, Q_+^{-1}\Big\|_{H \leftarrow H} \leq \frac{2}{\tau}$ from Lemma 4.6, we obtain

$$\left\| \begin{bmatrix} Q_+^{-1}\big(F^m - \widetilde{F}^m\big) \\ -\big(B + \frac{\tau}{2}A\big)\, Q_+^{-1}\big(F^m - \widetilde{F}^m\big) \end{bmatrix} \right\|_X \leq \frac{L_\rho}{\tau}\sqrt{6}\|e^m\|_X. \tag{4.34}$$

Inserting the bounds (4.33) and (4.34) into (4.32) yields with $C_{3/2} = 1 + (3/2)^{1/2}$

$$\mathrm{e}^{-n\tau c_{\mathrm{qm}}}\|e^n\|_X \leq C_{3/2} L_\rho \tau \sum_{m=1}^{n} \mathrm{e}^{-m\tau c_{\mathrm{qm}}}\|e^m\|_X + \mathrm{e}^{-n\tau c_{\mathrm{qm}}} \left\| \sum_{m=1}^{n} R^{n-m}\Delta^m \right\|_X.$$

For $\tau < 1/(C_{3/2}L_\rho)$ we obtain, by applying Grönwall's lemma Lemma A.1 b), multiplying by $\mathrm{e}^{n\tau c_{\mathrm{qm}}}$, and inserting (4.30),

$$\begin{aligned} \|e^n\|_X &\leq \mathrm{e}^{\frac{C_{3/2}L_\rho n\tau}{1 - C_{3/2}L_\rho\tau}} \left(\left\| \sum_{m=1}^{n} R^{n-m}\left(R_+^{-1}\delta_{\mathrm{CN}}^m + \widetilde{\delta}^m\right) \right\|_X \right) \\ &\leq \mathrm{e}^{\frac{C_{3/2}L_\rho t_n}{1 - C_{3/2}L_\rho\tau}} \left(\mathrm{e}^{n\tau c_{\mathrm{qm}}} \sum_{m=1}^{n} \|\delta_{\mathrm{CN}}^m\|_X + \left\| \sum_{m=1}^{n} R^{n-m}\widetilde{\delta}^m \right\|_X \right). \end{aligned} \tag{4.35}$$

(c) Defects. We bounded the Crank–Nicolson defect already in (4.13) by

$$\|\delta_{\mathrm{CN}}^m\|_X \leq C\tau^3 \left(\|u^{(3)}\|_{L^\infty([t_m,t_{m-1}];V)} + \|u^{(4)}\|_{L^\infty([t_m,t_{m-1}];H)} \right) \leq C\tau^3.$$

To bound the additional defect arising in the IMEX scheme, we split it into

$$\widetilde{\delta}^m = \frac{\tau^2}{4} \begin{bmatrix} Q_+^{-1}\big(\widetilde{F}^{m-1} - \widetilde{F}^m\big) \\ -\big(B + \frac{\tau}{2}A\big)\, Q_+^{-1}\big(\widetilde{F}^{m-1} - \widetilde{F}^m\big) \end{bmatrix} = \widetilde{\delta}_1^m + \widetilde{\delta}_2^m$$

with

$$\widetilde{\delta}_1^m = \frac{\tau}{4} \begin{bmatrix} \tau Q_+^{-1}\big(\widetilde{F}^{m-1} - \widetilde{F}^m\big) \\ Q_- Q_+^{-1}\big(\widetilde{F}^{m-1} - \widetilde{F}^m\big) \end{bmatrix}, \qquad \widetilde{\delta}_2^m = \frac{\tau}{4} \begin{bmatrix} 0 \\ -\big(\widetilde{F}^{m-1} - \widetilde{F}^m\big) \end{bmatrix}.$$

The terms $\widetilde{\delta}_1^m$ and $\widetilde{\delta}_2^m$ are of order τ^2, which is not sufficient to obtain a global error of order two, since we loose one order of τ when summing the defects up. To gain an additional factor of τ we use a combination of both terms from two successive time steps. With the explicit representation of R from (4.20b) we obtain

$$\widetilde{\delta}_1^m + R\widetilde{\delta}_2^{m-1} = \frac{\tau}{2} \begin{bmatrix} \frac{\tau}{2} Q_+^{-1}\big(-\widetilde{F}^{m-2} + 2\widetilde{F}^{m-1} - \widetilde{F}^m\big) \\ \frac{1}{2} Q_- Q_+^{-1}\big(-\widetilde{F}^{m-2} + 2\widetilde{F}^{m-1} - \widetilde{F}^m\big) \end{bmatrix}.$$

Using this together with the bounds (4.19), the differential equation (4.1), and $B \in \mathcal{L}(V;H)$ leads to the bound

$$\begin{aligned}\|\widetilde{\delta}_1^m + \widehat{R}\widetilde{\delta}_2^{m-1}\|_X &\leq C\tau\big\|\big(-\widetilde{F}^{m-2} + 2\widetilde{F}^{m-1} - \widetilde{F}^m\big)\big\|_m \\ &\leq C\tau^3\Big\|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\big(F(u)+f)\big)\Big\|_{L^\infty([t_{m-2},t_m];H)} \\ &\leq C\tau^3\Big\|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\big(u'' + Bu' + Au\big)\Big\|_{L^\infty([t_{m-2},t_m];H)} \\ &\leq C\tau^3\left(\|u^{(4)}\|_{L^\infty([t_{m-2},t_m];H)} + \|u^{(3)}\|_{L^\infty([t_{m-2},t_m];V)} + \|Au''\|_{L^\infty([t_{m-2},t_m];H)}\right),\end{aligned} \tag{4.36}$$

and, hence,

$$\begin{aligned}\Big\|\sum_{m=1}^n R^{n-m}\widetilde{\delta}^m\Big\|_X &\leq \Big\|R^{n-1}\widetilde{\delta}_1^1 + \widetilde{\delta}_2^n + \sum_{m=2}^n R^{n-m}(\widetilde{\delta}_1^m + R\widetilde{\delta}_2^{m-1})\Big\|_X \\ &\leq \mathrm{e}^{n\tau c_{\mathrm{qm}}}\left(\|\widetilde{\delta}_1^1\|_X + \|\widetilde{\delta}_2^n\|_X + \sum_{m=2}^n \|\widetilde{\delta}_1^m + R\widetilde{\delta}_2^{m-1}\|_X\right) \\ &\leq C\mathrm{e}^{n\tau c_{\mathrm{qm}}}\tau^2.\end{aligned}$$

Inserting the bounds of the defects into (4.35) yields

$$\|e^n\|_X \leq C\mathrm{e}^{Mt_n}\tau^2.$$

This finally gives the error bound (4.28), since

$$\|u^n - u(t_n)\|_{\tilde{a}} + \|v^n - u'(t_n)\|_m \leq \sqrt{2}\|x^n - x(t_n)\|_X = \sqrt{2}\|e^n\|_X \leq C\mathrm{e}^{Mt_n}\tau^2.$$

(d) Boundedness of numerical solution. It remains to prove that for sufficiently small τ the bound (4.27) holds true, since only then the error analysis we presented so far is valid. To do so, we proceed similarly as in the proof of Theorem 3.2.

Let $F_\rho \colon V \to H$ be a function that is globally Lipschitz continuous on V with Lipschitz constant L_ρ and satisfies $F_\rho(v) = F(v)$ for all $v \in V$ with $\|v\|_{\tilde{a}} \leq \rho$. Further, let u_ρ^n be the iterates of the IMEX scheme (4.16) with F replaced by F_ρ. Note that due to the definition of ρ in (4.27) we have

$$F(u(t)) = F_\rho(u(t)) \text{ for all } t \in [0,T]$$

and u is also a solution of (4.1) when F is replaced by F_ρ. Since F_ρ is globally Lipschitz continuous, part (a) to (c) of the proof hold true for the modified equation independent of $\|u_\rho^n\|_{\tilde{a}}$ and we obtain similar to the error bound (4.28)

$$\|u_\rho^n - u(t_n)\|_{\tilde{a}} \leq C\mathrm{e}^{Mt_n}\tau^2 \tag{4.37}$$

for all $t_n \leq T$. Furthermore, C is independent of τ and we can choose $\tau^* > 0$ s.t. for all $\tau < \tau^*$ we have

$$\|u_\rho^n - u(t_n)\|_{\tilde{a}} \leq \frac{\rho}{2}$$

and, hence, by the choice of ρ, we can conclude

$$\|u_\rho^n\|_{\tilde{a}} \leq \|u_\rho^n - u(t_n)\|_{\tilde{a}} + \|u(t_n)\|_{\tilde{a}} \leq \frac{\rho}{2} + \|u(t_n)\|_{\tilde{a}} \leq \rho.$$

This implies that for all $t_n \leq T$ the iterates u_ρ^n coincide with the original iterates u^n and thereby $\|u^n\|_{\tilde{a}} = \|u_\rho^n\|_{\tilde{a}} \leq \rho$. □

CHAPTER 5

Abstract full discretization error analysis

In this chapter, we show how the unified space discretization error analysis from Chapter 2 can be combined with the error analysis of the time discretization schemes from Chapters 3 and 4 to derive fully discrete error bounds.

As in Chapter 3, we aim at approximating the solution u of the continuous second-order evolution equation (2.19). We denote the step size of the time discretization by $\tau > 0$ and set $t_n \coloneqq n\tau, n \geq 0$. The fully discrete approximations of u and u' are denoted by $u_h^n \approx u(t_n), v_h^n \approx u'(t_n)$, respectively, and are obtained by applying the time discretization schemes from Chapters 3 and 4 to the spatially discretized equation (2.27) or its first-order formulation (2.4), respectively. We further write

$$x_h^n = \begin{bmatrix} u_h^n \\ v_h^n \end{bmatrix} \approx \begin{bmatrix} u(t_n) \\ v(t_n) \end{bmatrix} = x(t_n)$$

where x solves the first-order formulation (2.1) of (2.19), cf. (2.20).

We bound the errors of the fully discrete schemes in terms of the order of the corresponding time discretization scheme and the abstract space discretization errors from Chapter 2. These error bounds can be used to derive full discretization error estimates for concrete wave equations as we explain in Chapter 6 for the wave equation with kinetic boundary conditions.

Outline In the first two sections of this chapter, we prove error bounds for fully discrete schemes where Runge–Kutta methods are used for the time discretization. We analyze discretizations of first-order evolution equations in Section 5.1 and use these results to prove error bounds for second-order wave-type equations in Section 5.2. Section 5.3 is devoted to analyze full discretizations of second-order semilinear wave-type equations where the time discretization is performed with the IMEX scheme from Chapter 4.

5.1 Runge–Kutta methods for first-order evolution equations

A Runge–Kutta method with coefficients $\mathbf{b} = (b_i)_{i=1}^s, \mathbf{c} = (c_i)_{i=1}^s, \mathcal{Q} = (a_{ij})_{i,j=1}^s$ applied to the spatially discretized equation (2.4) reads

$$\begin{aligned} X_h^{ni} &= x_h^n + \tau \sum_{j=1}^s a_{ij}\big(-\mathcal{S}_h(X_h^{nj}) + G_h(X_h^{nj}) + g_h(t_n + c_j\tau)\big), \qquad i = 1,\dots,s, \\ x_h^{n+1} &= x_h^n + \tau \sum_{i=1}^s b_i \left(-\mathcal{S}_h(X_h^{ni}) + G_h(X_h^{ni}) + g_h(t_n + c_i\tau)\right). \end{aligned} \tag{5.1}$$

To prove full discretization error bounds, we adapt the results from Hansen [2006b], that are summarized in Appendix B, to the semidiscrete setting presented in Section 2.1.2. By this, we obtain fully discrete versions of Theorems 3.1 and 3.2. Theorem 5.1 is a fully discrete error bound under the assumption that the numerical approximations are bounded, while in Theorem 5.2, under additional consistency assumptions, existence and boundedness of the numerical approximations is shown.

Theorem 5.1. *Let Assumptions 2.3, 2.7 and 2.8 be satisfied, $q \in \mathbb{N}$, and $x \in C^{q+1}([0,T];X)$ be the solution of (2.1) with $x, x' \in L^\infty([0,T];Z)$. Furthermore, let x_h^n, for $t_n \in [0,T]$, be the approximations obtained by an algebraically stable and coercive Runge–Kutta method of stage order q given by (5.1). By α_{RK} we denote the coercivity constant of the Runge–Kutta method, cf. Definition B.5, and we define*

$$\rho_h := \max\big\{\widehat{C}_{J_h}\|x\|_{L^\infty([0,T];Z)}, \max_{t_n\le T}\|x_h^n\|_{X_h}, \max_{\substack{t_n\le T\\ i=1,\dots,s}} \|X_h^{ni}\|_{X_h}\big\}. \tag{5.2}$$

If τ satisfies the step size restriction

$$\tau(\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h}) < \alpha_{\mathrm{RK}}, \tag{5.3}$$

then the error bound

$$\|\mathcal{L}_h x_h^n - x(t_n)\|_X \le C \frac{\mathrm{e}^{C_{\mathrm{RK}}\widehat{C}^2_{\tau,\rho_h}(\widehat{c}_{\mathrm{qm}}+\widehat{L}_{\rho_h})t_n} - 1}{C_{\mathrm{RK}}(\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h})} (E_h(t_n) + \tau^q) + \|(\mathrm{I} - \mathcal{L}_h J_h)x(t)\|_X \tag{5.4}$$

holds true with constants C_{RK}, only depending on the Runge–Kutta method, C, which depends on x, T and the Runge–Kutta method, but is independent of n and τ, and

$$\widehat{C}_{\tau,\rho_h} = \big(\alpha_{\mathrm{RK}} - \tau(\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h})\big)^{-1}.$$

The term $E_h(t_n)$ contains abstract space discretization errors and is defined in (2.7).

Proof. As in the proof of Theorem 2.10, we split the error between the lifted fully discrete approximation and the exact solution via $\mathcal{L}_h x_h^n - x(t_n) = \mathcal{L}_h e_h^n + (\mathcal{L}_h J_h - \mathrm{I})x(t_n)$, where

$$e_h^n = x_h^n - J_h x(t_n) \in V_h$$

is the discrete error. We then have by (2.5)

$$\|\mathcal{L}_h x_h^n - x(t_n)\|_X \le \widehat{C}_X \|e_h^n\|_{X_h} + \|(\mathcal{L}_h J_h - \mathrm{I})x(t_n)\|_X. \tag{5.5}$$

We now proceed as in Appendix B and use the same trick as in the proof of Theorem 3.1. We replace G_h in (2.4) by G_h^ρ where $G_h^\rho(y_h) = G_h(y_h)$ for all $y_h \in X_h$ with $\|y_h\|_{X_h} \leq \rho_h$, and G_h^ρ is globally Lipschitz continuous with constant $\widehat{L}_{\rho_h}$. The modified equation then reads

$$x_h' + \mathcal{S}_h(x_h) = G_h^\rho(x_h) + g_h, \qquad t \geq 0. \tag{5.6}$$

By (5.2), we have that x_h^n are also the Runge–Kutta iterations when the method is applied to (5.6). Furthermore, (5.6) fits into the setting of Appendix B with X replaced by X_h and

$$\mathcal{F} = -\mathcal{S}_h + G_h^\rho \colon X_h \to X_h.$$

Assumption B.6 is satisfied with $c_{\mathrm{qm},\mathcal{F}} = \widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h}$.

We cannot apply Theorem B.12 directly, since additionally the space discretization errors enter. Therefore, we now derive the defects in this fully discrete case. Then, we apply the local error and stability results from Appendix B to obtain the global error bound of the fully discrete scheme.

For the exact solution x of (2.1) we use the short notation

$$\widetilde{x}^n = x(t_n), \quad \widetilde{X}^{ni} = x(t_n + c_i\tau).$$

As in (B.4), we obtain for the exact solution plugged into the Runge–Kutta method

$$\widetilde{X}^{ni} = \widetilde{x}^n + \tau \sum_{j=1}^s a_{ij} \Big(-\mathcal{S}(\widetilde{X}^{nj}) + G(\widetilde{X}^{nj}) + g(t_n + c_j\tau)\Big) + \Delta_{\mathrm{RK}}^{ni}, \qquad i = 1, \dots, s, \tag{5.7a}$$

$$\widetilde{x}^{n+1} = \widetilde{x}^n + \tau \sum_{s=1}^s b_i \Big(-\mathcal{S}(\widetilde{X}^{ni}) + G(\widetilde{X}^{ni}) + g(t_n + c_i\tau)\Big) + \delta_{\mathrm{RK}}^{n+1}, \tag{5.7b}$$

where, by Lemma B.9, the defects satisfy

$$\|\Delta_{\mathrm{RK}}^{ni}\|_X, \|\delta_{\mathrm{RK}}^{n+1}\|_X \leq C\tau^{q+1}$$

with a constant $C = C(x^{(q+1)})$. By applying the adjoint lift operator $\mathcal{L}_h^*$ to (5.7b), we obtain

$$\mathcal{L}_h^* \widetilde{x}^{n+1} = \mathcal{L}_h^* \widetilde{x}^n + \tau \sum_{s=1}^s b_i \Big(-\mathcal{L}_h^* \mathcal{S}(\widetilde{X}^{ni}) + \mathcal{L}_h^* G(\widetilde{X}^{ni}) + \mathcal{L}_h^* g(t_n + c_i\tau)\Big) + \mathcal{L}_h^* \delta_{\mathrm{RK}}^{n+1},$$

which can be rewritten as

$$J_h \widetilde{x}^{n+1} = J_h \widetilde{x}^n + \tau \sum_{s=1}^s b_i \Big(-\mathcal{S}_h(J_h \widetilde{X}^{ni}) + G_h^\rho(J_h \widetilde{X}^{ni}) + g_h(t_n + c_i\tau)\Big) + \mathcal{L}_h^* \delta_{\mathrm{RK}}^{n+1} + \delta_h^{n+1},$$

where the additional defect is given by

$$\begin{aligned} \delta_h^{n+1} &= (J_h - \mathcal{L}_h^*)\,(\widetilde{x}^{n+1} - \widetilde{x}^n) \\ &\quad + \tau \sum_{s=1}^s b_i \Big(\mathcal{S}_h(J_h \widetilde{X}^{ni}) - \mathcal{L}_h^* \mathcal{S}(\widetilde{X}^{ni}) + \mathcal{L}_h^* G(\widetilde{X}^{ni}) - G_h(J_h \widetilde{X}^{ni}) + \mathcal{L}_h^* g(t_n + c_i\tau) - g_h(t_n + c_i\tau)\Big). \end{aligned}$$

Note that, due to (5.3), we have $G_h(J_h \widetilde{X}^{ni}) = G_h^\rho(J_h \widetilde{X}^{ni})$. With the identity

$$\widetilde{x}^{n+1} - \widetilde{x}^n = \int_{t_n}^{t_{n+1}} x'(s) \,\mathrm{d}s$$

and the definition of the remainders from Definition 2.9 we can bound the defect δ_h^{n+1} via

$$\begin{aligned}\|\delta_h^{n+1}\|_{X_h} \leq \tau C\Big(&\|(\mathcal{L}_h^* - J_h)x'\|_{L^\infty([t_n,t_{n+1}];X_h)} + \|R_h(x)\|_{L^\infty([t_n,t_{n+1}];X_h)} \\ &+ \|r_h(x)\|_{L^\infty([t_n,t_{n+1}];X_h)} + \|\mathcal{L}_h^* g - g_h\|_{L^\infty([t_n,t_{n+1}];X_h)}\Big),\end{aligned}$$

where C is independent of τ and h. Similarly, we obtain for $i = 1, \dots, s$ for the inner stages

$$J_h\widetilde{X}^{ni} = J_h\widetilde{x}^n + \tau\sum_{j=1}^{s} a_{ij}\left(-\mathcal{S}_h(J_h\widetilde{X}^{nj}) + G_h^\rho(J_h\widetilde{X}^{nj}) + g_h(t_n + c_j\tau)\right) + \mathcal{L}_h^*\Delta_{\mathrm{RK}}^{ni} + \Delta_h^{ni}$$

with

$$\begin{aligned}\|\Delta_h^{ni}\|_{X_h} \leq \tau C\Big(&\|(\mathcal{L}_h^* - J_h)x'\|_{L^\infty([t_n,t_{n+1}];X_h)} + \|R_h(x)\|_{L^\infty([t_n,t_{n+1}];X_h)} \\ &+ \|r_h(x)\|_{L^\infty([t_n,t_{n+1}];X_h)} + \|\mathcal{L}_h^* g - g_h\|_{L^\infty([t_n,t_{n+1}];X_h)}\Big).\end{aligned}$$

The local error bound from Lemma B.10 translates then to

$$\begin{aligned}\|\overline{x_h^{n+1}} - J_h x(t_{n+1})\|_{X_h} &\leq \|\mathcal{L}_h^*\delta_{\mathrm{RK}}^{n+1}\|_{X_h} + \|\delta_h^{n+1}\|_{X_h} + C\,(1 + C_\tau)\max_{i=1,\dots,s}\left(\|\mathcal{L}_h^*\Delta_{\mathrm{RK}}^{ni}\|_{X_h} + \|\Delta_h^{ni}\|_{X_h}\right) \\ &\leq C\,(1 + C_\tau)\Big(\tau^{q+1} + \tau\Big(\|(\mathcal{L}_h^* - J_h)x'\|_{L^\infty([t_n,t_{n+1}];X_h)} + \|R_h(x)\|_{L^\infty([t_n,t_{n+1}];X_h)} \\ &\qquad + \|r_h(x)\|_{L^\infty([t_n,t_{n+1}];X_h)} + \|\mathcal{L}_h^* g - g_h\|_{L^\infty([t_n,t_{n+1}];X_h)}\Big)\Big).\end{aligned}$$

In this case, $\overline{x_h^{n+1}}$ is defined as one step of the Runge–Kutta method applied to (5.6) and starting from $J_h x(t_n)$ at time t_n. In the second inequality we used the continuity of $\mathcal{L}_h^*$ and the bounds of the defects.

Following exactly the lines of the proof of Theorem B.12, and by the definition of E_h, we can bound the discrete error by.

$$\begin{aligned}\|e_h^n\|_{X_h} &= \|x_h^n - J_h x(t_n)\|_{X_h} \\ &\leq C\frac{\mathrm{e}^{C_{\mathrm{RK}}\widehat{C}_{\tau,\rho_h}^2(\widehat{c}_{\mathrm{qm}} + \widehat{L}_\rho)t_n} - 1}{C_{\mathrm{RK}}(\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h})}\Big(\tau^q + \Big(\|(\mathcal{L}_h^* - J_h)x'\|_{L^\infty([0,t_{n+1}];X_h)} + \|R_h(x)\|_{L^\infty([0,t_{n+1}];X_h)} \\ &\qquad + \|r_h(x)\|_{L^\infty([0,t_{n+1}];X_h)} + \|\mathcal{L}_h^* g - g_h\|_{L^\infty([0,t_{n+1}];X_h)}\Big)\Big) \\ &\leq C\frac{\mathrm{e}^{C_{\mathrm{RK}}\widehat{C}_{\tau,\rho_h}^2(\widehat{c}_{\mathrm{qm}} + \widehat{L}_\rho)t_n} - 1}{C_{\mathrm{RK}}(\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h})}\left(\tau^q + E_h(t_n)\right).\end{aligned}$$

Together with (5.5), this gives the assertion. □

Theorem 5.2. *Let Assumptions 2.3, 2.7 and 2.8 be satisfied, $q \in \mathbb{N}$, and $x \in C^{q+1}([0,T];X)$ be the solution of (2.1) with $x, x' \in L^\infty([0,T];Z)$. Further, we assume that for E_h defined in (2.7) we have $E_h(t) \to 0$ for $h \to 0$ and for all $t \in [0,T]$. We define*

$$\rho \coloneqq 2\|x\|_{L^\infty([0,T];X)}$$

and consider an algebraically stable and coercive Runge–Kutta method of stage order q.

Then, there exist $\tau^*, h^* > 0$ s.t. for all $\tau < \tau^*$ and $h < h^*$ the Runge–Kutta scheme (5.1) yields for all $t_n \in [0,T]$ unique iterations $x_h^n \in X_h$ with $\|x_h^n\|_{X_h} \leq \rho$ which satisfy the error bound (5.4) with $\rho_h = \rho$.

Proof. This can be concluded from Theorem 5.1 by following exactly the lines of the proof of Theorem 3.2. For proving the boundedness of $x_h^n, X_h^{ni}, n \geq 0, i = 1, \dots, s$, one additionally has to use that the space discretization errors collected in E_h satisfy $E_h(t) \to 0$ for $h \to 0$ and all $t \in [0, T]$. □

5.2 Runge–Kutta methods for second-order semilinear equations

Since the first-order formulations of both the continuous second-order equation (2.19) and the corresponding spatially discretized equation (2.27) fit by Corollarys 2.15 and 2.18 and Lemma 2.22 in the setting of Section 2.1, Theorems 5.1 and 5.2 transfer directly to the second-order case and we obtain the following results:

Corollary 5.3. *Let Assumptions 2.12, 2.17 and 2.19 be satisfied, $q \in \mathbb{N}$, and let $u \in C^{q+2}([0,T]; H) \cap C^{q+1}([0,T]; V)$ be the solution of (4.1) with $u, u', u'' \in L^\infty([0,T]; Z^V)$. Further, let u_h^n, v_h^n, $t_n \in [0,T]$, be the approximations obtained by an algebraically stable and coercive Runge–Kutta method of stage order q applied to the first-order formulation of the semidiscrete equation (2.27). By α_{RK} we denote the coercivity constant of the Runge–Kutta method, cf. Definition B.5, and we define*

$$\rho_h := \max \Big\{ \widehat{C}_V \|u\|_{L^\infty([0,T];V)}, \max_{t_n \leq T} \|u_h^n\|_{\tilde{a}_h}, \max_{\substack{t_n \leq T \\ i=1,\dots,s}} \|U_h^{ni}\|_{\tilde{a}_h} \Big\},$$

where U_h^{ni} denotes the first component of the inner Runge–Kutta stages. If τ satisfies the step size restriction

$$\tau(\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h}) < \alpha_{\mathrm{RK}}$$

with $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}}$, the error bound

$$\|\mathcal{L}_h^V u_h^n - u(t_n)\|_{\tilde{a}} + \|\mathcal{L}_h^V v^n - u'(t_n)\|_m \leq C \left(\frac{\mathrm{e}^{C_{\mathrm{RK}} \widehat{C}^2_{\tau,\rho_h} (\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h}) t_n} - 1}{C_{\mathrm{RK}}(\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h})} + 1 \right) \left(\tau^q + \sum_{i=1}^{5} E_{h,i} \right) \tag{5.9}$$

holds true with constants C_{RK}, that only depends on the coefficients of the Runge–Kutta method, C, which depends on u, T and the Runge–Kutta method, but is independent of n and τ, and

$$\widehat{C}_{\tau,\rho_h} = \big(\alpha_{\mathrm{RK}} - \tau(\widehat{c}_{\mathrm{qm}} + \widehat{L}_{\rho_h}) \big)^{-1}.$$

The constants $E_{h,i}$ contain the abstract space discretization errors and are given in (2.40).

Proof. This follows directly from Theorem 5.1. As we mentioned above the corollary, all assumptions are satisfied and we have $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}}$. The first-order space discretization errors contained in E_h and $\|(\mathrm{I} - \mathcal{L}_h J_h) x(t)\|_X$ can be bounded against $E_{h,i}, i = 1, \dots, 5$, as shown in the proof of Theorem 2.24. □

Corollary 5.4. *Let Assumptions 2.12, 2.17 and 2.19 be satisfied, $q \in \mathbb{N}$, and let $u \in C^{q+2}([0,T]; H) \cap C^{q+1}([0,T]; V)$ be the solution of (4.1) with $u, u', u'' \in L^\infty([0,T]; Z^V)$. We then define*

$$\rho := 2\widehat{C}_V \|u\|_{L^\infty([0,T];V)}$$

and consider an algebraically stable and coercive Runge–Kutta method of stage order q. Further let the space discretization error terms $E_{h,i}$ defined in (2.40) satisfy $E_{h,i} \to 0$ for $h \to 0$, $i = 1, \dots, 5$.

Then, there exist $\tau^*, h^* > 0$ s.t. for all $\tau < \tau^*$ and $h < h^*$ the Runge–Kutta method yields for all $t_n \in [0, T]$ unique iterations $u_h^n, v_h^n \in V_h$ with $\|u_h^n\|_{\tilde{a}_h} \le \rho$ which satisfy the error bound (5.9) with $\rho_h = \rho$.

Proof. Follows directly by Theorem 5.2 with the same arguments as in the proof of Corollary 5.3. □

Remark 5.5 (BDF methods)**.** *Based on the time discretization error analysis of BDF methods from Hansen [2006a], and by using the same techniques as presented in this section, it is also possible to prove full discretization error bounds for BDF methods in our setting (cf. also Remark 3.5).*

5.3 IMEX scheme for semilinear second-order evolution equations

In this section, we present the results from [Hochbruck and Leibold, 2021, Section 3]. We consider the IMEX scheme from Chapter 4 applied to the spatially discretized second-order evolution equation (2.27) in the semilinear case, i.e., where $B = \mathcal{D} \in \mathcal{L}(V; H)$ and, hence, $B_h = \mathcal{D}_h \in \mathcal{L}(V_h; H_h)$ are linear operators. The evolution equation (2.27) is then of the form

$$u_h'' + B_h u_h' + A_h u_h = F_h(u_h) + f_h, \quad t \ge 0, \qquad u_h(0) = u_h^0, \quad u_h'(0) = v_h^0, \tag{5.10}$$

and with the first-order operator

$$S_h = \mathcal{S}_h = \begin{bmatrix} 0 & -\mathrm{I} \\ A_h & B_h \end{bmatrix} \in \mathcal{L}(X_h; X_h)$$

in (2.28), the corresponding first-order reformulation for $x_h = [u_h, v_h]^\intercal$ reads

$$x_h' + S_h x = G_h(x) + g_h, \quad t \ge 0, \qquad x_h(0) = x_h^0.$$

To simplify the presentation, we use the short notations

$$G_h^n = \begin{bmatrix} 0 \\ F_h^n \end{bmatrix} := \begin{bmatrix} 0 \\ F_h(u_h^n) + f_h(t_n) \end{bmatrix} = G_h(x_h^n) + g_h(t_n).$$

The IMEX scheme (4.16) applied to (5.10) then reads

$$v_h^{n+\frac{1}{2}} = v_h^n - \frac{\tau}{2} A_h u_h^n - \frac{\tau^2}{4} A_h v_h^{n+\frac{1}{2}} - \frac{\tau}{2} B_h v_h^{n+\frac{1}{2}} + \frac{\tau}{2} F_h^n, \tag{5.11a}$$

$$u_h^{n+1} = u_h^n + \tau v_h^{n+\frac{1}{2}}, \tag{5.11b}$$

$$v_h^{n+1} = v_h^{n+\frac{1}{2}} - \frac{\tau}{2} A_h u_h^n - \frac{\tau^2}{4} A_h v_h^{n+\frac{1}{2}} - \frac{\tau}{2} B_h v_h^{n+\frac{1}{2}} + \frac{\tau}{2} F_h^{n+1}, \tag{5.11c}$$

As in the continuous case, we can replace (5.11c) by the more efficient update

$$v_h^{n+1} = -v_h^n + 2 v_h^{n+\frac{1}{2}} + \frac{\tau}{2} \left(F_h^{n+1} - F_h^n \right). \tag{5.11d}$$

Analogously toChapter 4, we define the operators

$$\begin{aligned}\widehat{Q}_\pm &:= \mathrm{I} \pm \frac{\tau}{2} B_h \pm \frac{\tau^2}{4} A_h \colon V_h \to V_h,\\ \widehat{R}_\pm &:= \mathrm{I} \pm \frac{\tau}{2} S_h \qquad\qquad : X_h \to X_h,\\ \widehat{R} &:= \widehat{R}_+^{-1} \widehat{R}_-.\end{aligned}$$

Since Assumption 2.17 is similar to Assumption 2.12 in the continuous case with constants independent of h, Lemmas 4.2 and 4.6 transfer directly to the discrete case with the continuous constants replaced by the discrete ones.

As for the Runge–Kutta methods, we now first prove an abstract error result that depends on the bound of the numerical solution. In Theorem 5.8, we then show that, for sufficiently small τ and h and under additional consistency assumptions for the space discretization, the fully discrete approximations are bounded in terms of the exact solution only.

Theorem 5.6 ([Hochbruck and Leibold, 2021, Theorem 3.3]). *Let Assumptions 2.12, 2.17 and 2.19 be satisfied and let $u \in C^4([0,T];H) \cap C^3([0,T];V) \cap C^2([0,T];D(A))$ be the solution of (4.1) with $u, u', u'' \in L^\infty([0,T];Z^V)$. Further, let u_h^n, v_h^n, $t_n \in [0,T]$, be the approximations obtained by the fully discrete IMEX scheme (5.11) and set*

$$\rho_h := \max \Big\{ \widehat{C}_V \|u\|_{L^\infty([0,T];V)}, \max_{t_n \le T} \|u_h^n\|_{\tilde{a}_h} \Big\}.$$

If τ satisfies the step size restriction

$$\max\{\tau\big(1 + (3/2)^{1/2}\big)\widehat{L}_{\rho_h}, \tau\frac{\widehat{c}_{\mathrm{qm}}}{2}, \frac{\tau^2}{2}\widehat{c}_G + \tau\widehat{\beta}_{\mathrm{qm}}\} < 1$$

with $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}}$ then, for all $n > 0$ with $t_n \le T$, the error bound

$$\|\mathcal{L}_h^V u_h^n - u(t_n)\|_{\tilde{a}} + \|\mathcal{L}_h^V v_h^n - u'(t_n)\|_m \le \; C\mathrm{e}^{\widehat{M}t_n} \left(\sum_{i=1}^{5} E_{h,i} + \tau^2 \right) \tag{5.13}$$

holds true with

$$\widehat{M} = \widehat{c}_{\mathrm{qm}} + \frac{\big(1 + (3/2)^{1/2}\big)\widehat{L}_{\rho_h}}{1 - \big(1 + (3/2)^{1/2}\big)\widehat{L}_{\rho_h}\tau}$$

and a constant C that only depends on T and u but which is independent of τ, h, and $\widehat{L}$. The constants $E_{h,i}$ contain the abstract space discretization errors and are defined in (2.40).

Proof. This proof follows the lines of the proof of Theorem 4.9 but we additionally have to consider the errors arising from the space discretization. These errors were already bounded against $E_{h,i}, i = 1, \dots, 5,$ in the proof of Theorem 2.24. As in Theorem 4.9, the proof relies on the first-order formulation of the IMEX scheme and we use the notation

$$\widetilde{x}^n = \begin{bmatrix} \widetilde{u}^n \\ \widetilde{v}^n \end{bmatrix} = \begin{bmatrix} u(t_n) \\ u'(t_n) \end{bmatrix}, \quad \widetilde{G}_h^n = G_h(J_h \widetilde{x}^n) + g_h(t_n) = \begin{bmatrix} 0 \\ F_h(\mathcal{L}_h^{V*}\widetilde{u}^n) + f_h(t_n) \end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{F}_h^n \end{bmatrix},$$

where J_h is the first order reference operator defined in (2.32). The proof consists of four main steps.

(a) Splitting of the error. As in the proof of Theorem 2.10, we split the error via

$$\mathcal{L}_h x_h^n - \widetilde{x}^n = \mathcal{L}_h e_h^n + (\mathcal{L}_h J_h - \mathrm{I})\widetilde{x}^n, \quad \text{where} \quad e_h^n = x_h^n - J_h \widetilde{x}^n \in X_h$$

is the fully discrete error. Due to the continuity of the lift operator, and by (2.41) and (2.42) we have

$$\|\mathcal{L}_h x_h^n - \widetilde{x}^n\|_X \leq C\|e_h^n\|_{X_h} + \|(\mathcal{L}_h J_h - \mathrm{I})\widetilde{x}^n\|_X \leq C\left(\|e_h^n\|_{X_h} + E_{h,4} + E_{h,5}\right). \tag{5.14}$$

In the next three steps, we proceed as in the proof of Theorem 4.9 to bound the discrete error $\|e_h^n\|_{X_h}$.

(b) Error recursion for e_h^n. Analogously to Lemma 4.8 we can rewrite the fully discrete scheme 5.11 as

$$x_h^{n+1} = \widehat{R} x_h^n + \frac{\tau}{2}\widehat{R}_+^{-1}(G_h^n + G_h^{n+1}) + \frac{\tau^2}{4}\begin{bmatrix} \widehat{Q}_+^{-1}(F_h^n - F_h^{n+1}) \\ -\left(B_h + \frac{\tau}{2}A_h\right)\widehat{Q}_+^{-1}(F_h^n - F_h^{n+1}) \end{bmatrix}. \tag{5.15}$$

To derive an error recursion, we insert $J_h x$ into the fully discrete IMEX scheme (5.15) and obtain

$$J_h\widetilde{x}^{n+1} = \widehat{R} J_h \widetilde{x}^n + \frac{\tau}{2}\widehat{R}_+^{-1}\left(\widetilde{G}_h^{n+1} + \widetilde{G}_h^n\right) + \frac{\tau^2}{4}\begin{bmatrix} \widehat{Q}_+^{-1}(\widetilde{F}_h^n - \widetilde{F}_h^{n+1}) \\ -\left(B_h + \frac{\tau}{2}A_h\right)\widehat{Q}_+^{-1}(\widetilde{F}_h^n - \widetilde{F}_h^{n+1}) \end{bmatrix} - \Delta_h^{n+1} \tag{5.16}$$

with a defect Δ_h^{n+1}. Similar to (4.30), we can interpret Δ_h^{n+1} as a perturbation of the defect $\Delta_{\mathrm{CN},h}^{n+1}$ of the fully discrete Crank–Nicolson scheme, i.e.,

$$\Delta_h^{n+1} = \Delta_{\mathrm{CN},h}^{n+1} + \widetilde{\delta}_h^{n+1}, \qquad \widetilde{\delta}_h^{n+1} = \frac{\tau^2}{4}\begin{bmatrix} \widehat{Q}_+^{-1}(\widetilde{F}_h^n - \widetilde{F}_h^{n+1}) \\ -\left(B_h + \frac{\tau}{2}A_h\right)\widehat{Q}_+^{-1}(\widetilde{F}_h^n - \widetilde{F}_h^{n+1}) \end{bmatrix},$$

where $\Delta_{\mathrm{CN},h}^{n+1}$ satisfies

$$J_h\widetilde{x}^{n+1} = \widehat{R} J_h \widetilde{x}^n + \frac{\tau}{2}\widehat{R}_+^{-1}\left(\widetilde{G}_h^{n+1} + \widetilde{G}_h^n\right) - \Delta_{\mathrm{CN},h}^{n+1}. \tag{5.17}$$

To determine $\Delta_{\mathrm{CN},h}^{n+1}$, we note that by (4.8) we have

$$\mathcal{L}_h^*\widetilde{x}^{n+1} = \mathcal{L}_h^*\widetilde{x}^n + \frac{\tau}{2}\left(-\mathcal{L}_h^* S(\widetilde{x}^n + \widetilde{x}^{n+1}) + \mathcal{L}_h^*\widetilde{G}^n + \mathcal{L}_h^*\widetilde{G}^{n+1}\right) - \mathcal{L}_h^*\delta_{\mathrm{CN}}^{n+1}. \tag{5.18}$$

Using the remainder terms from Definition 2.9, we set

$$\delta_h^{n+1} = -\left(J_h - \mathcal{L}_h^*\right)(\widetilde{x}^{n+1} - \widetilde{x}^n) + \frac{\tau}{2}R_h(\widetilde{x}^{n+1} + \widetilde{x}^n) - \frac{\tau}{2}\big(r_h(t_{n+1}, \widetilde{x}^{n+1}) + r_h(t_n, \widetilde{x}^n)\big). \tag{5.19}$$

Then, (5.18) can be expressed equivalently as

$$\begin{aligned} J_h\widetilde{x}^{n+1} &= J_h\widetilde{x}^n + \frac{\tau}{2}\left(-S_h J_h(\widetilde{x}^n + \widetilde{x}^{n+1}) + \widetilde{G}_h^n + \widetilde{G}_h^{n+1}\right) - \delta_h^{n+1} - \mathcal{L}_h^*\delta_{\mathrm{CN}}^{n+1} \\ \iff \widehat{R}_+ J_h\widetilde{x}^{n+1} &= \widehat{R}_- J_h\widetilde{x}^n + \frac{\tau}{2}\left(\widetilde{G}_h^{n+1} + \widetilde{G}_h^n\right) - \delta_h^{n+1} - \mathcal{L}_h^*\delta_{\mathrm{CN}}^{n+1}. \end{aligned} \tag{5.20}$$

By applying $\widehat{R}_+^{-1}$ to (5.20), we see that in (5.17) we have

$$\Delta_{\mathrm{CN},h}^{n+1} = \widehat{R}_+^{-1}\delta_h^{n+1} + \widehat{R}_+^{-1}\mathcal{L}_h^*\delta_{\mathrm{CN}}^{n+1}.$$

Subtracting (5.16) from (5.15) yields the error recursion

$$\begin{aligned} e_h^{n+1} &= \widehat{R} e_h^n + \frac{\tau}{2}\widehat{R}_+^{-1}\left(G_h^{n+1} - \widetilde{G}_h^{n+1} + G_h^n - \widetilde{G}_h^n\right) \\ &\quad + \frac{\tau^2}{4}\begin{bmatrix} \widehat{Q}_+^{-1}(F_h^n - \widetilde{F}_h^n - F_h^{n+1} + \widetilde{F}_h^{n+1}) \\ -\left(B_h + \frac{\tau}{2}A_h\right)\widehat{Q}_+^{-1}(F_h^n - \widetilde{F}_h^n - F_h^{n+1} + \widetilde{F}_h^{n+1}) \end{bmatrix} + \Delta_h^{n+1}. \end{aligned} \tag{5.21}$$

(c) Stability. Analogously to part (b) of the proof of Theorem 4.9, we can employ the stability of the scheme and obtain from (5.21) the bound

$$\begin{aligned}\|e_h^n\|_{X_h} &\leq \mathrm{e}^{\frac{C_{3/2}\widehat{L_{\rho_h}}n\tau}{1-C_{3/2}\widehat{L_{\rho_h}}\tau}} \left(\left\| \widehat{R}^n e_h^0 + \sum_{m=1}^{n} \widehat{R}^{n-m} \Delta_h^m \right\|_{X_h} \right) \\ &\leq \mathrm{e}^{\frac{C_{3/2}\widehat{L_{\rho_h}}t_n}{1-C_{3/2}\widehat{L_{\rho_h}}\tau}} \left(\mathrm{e}^{n\tau\widehat{c}_{\mathrm{qm}}} \left(\|e_h^0\|_{X_h} + \sum_{m=1}^{n} \left(\|\delta_h^m\|_{X_h} + \|\mathcal{L}_h^* \delta_{\mathrm{CN}}^m\|_{X_h} \right) \right) + \left\| \sum_{m=1}^{n} \widehat{R}^{n-m} \widetilde{\delta}_h^m \right\|_{X_h} \right).\end{aligned} \tag{5.22}$$

Note that in contrast to the semidiscrete case we cannot employ $e_h^0 = 0$ here.

(d) Defects. We now bound the different defects from (5.22). The initial error e_h^0 is bounded by

$$\|e_h^0\|_{X_h} \leq CE_{h,1}.$$

For the defect containing the space discretization errors we obtain from (5.19)

$$\left\|\delta_h^m\right\|_{X_h} = \tau \left\| \frac{1}{\tau} \int_{t_{m-1}}^{t_m} \left(J_h - \mathcal{L}_h^*\right) x'(s) \,\mathrm{d}s + \tfrac{1}{2} R_h(\widetilde{x}^{n+1} + \widetilde{x}^n) - \frac{1}{2} \big(r_h(t_{n+1}, \widetilde{x}^{n+1}) + r_h(t_n, \widetilde{x}^n) \big) \right\|_{X_h}.$$

All of these terms were already bounded in the proof of Theorem 2.24 and we have

$$\left\|\delta_h^m\right\|_{X_h} \leq C\tau \sum_{i=1}^{5} E_{h,i}.$$

The Crank–Nicolson defect was bounded in (4.13) and with the continuity of the adjoint lift we obtain

$$\|\mathcal{L}_h^* \delta_{\mathrm{CN}}^m\|_{X_h} \leq C \|\delta_{\mathrm{CN}}^m\|_X \leq C\tau^3 \left(\|u^{(3)}\|_{L^\infty([t_m, t_{m-1}]; V)} + \|u^{(4)}\|_{L^\infty([t_m, t_{m-1}]; H)} \right) \leq C\tau^3.$$

To bound the additional IMEX defect we split it into

$$\widetilde{\delta}_h^m = \frac{\tau^2}{4} \begin{bmatrix} \widehat{Q}_+^{-1} \big(\widetilde{F}_h^{m-1} - \widetilde{F}_h^m\big) \\ -\left(B_h + \frac{\tau}{2} A_h\right) \widehat{Q}_+^{-1} \big(\widetilde{F}_h^{m-1} - \widetilde{F}_h^m\big) \end{bmatrix} = \widetilde{\delta}_{h,0}^m + \widetilde{\delta}_{h,1}^m + \widetilde{\delta}_{h,2}^m$$

with

$$\begin{aligned}\widetilde{\delta}_{h,0}^m &= \frac{\tau^2}{4} \begin{bmatrix} \widehat{Q}_+^{-1} \big(\widetilde{F}_h^{m-1} - \mathcal{L}_h^{H*} \widetilde{F}^{m-1} - \widetilde{F}_h^m + \mathcal{L}_h^{H*} \widetilde{F}^m\big) \\ -\left(B_h + \frac{\tau}{2} A_h\right) \widehat{Q}_+^{-1} \big(\widetilde{F}_h^{m-1} - \mathcal{L}_h^{H*} \widetilde{F}^{m-1} - \widetilde{F}_h^m + \mathcal{L}_h^{H*} \widetilde{F}^m \end{bmatrix}, \\ \widetilde{\delta}_{h,1}^m &= \frac{\tau}{4} \begin{bmatrix} \tau \widehat{Q}_+^{-1} \mathcal{L}_h^{H*} \big(\widetilde{F}^{m-1} - \widetilde{F}^m\big) \\ \widehat{Q}_- \widehat{Q}_+^{-1} \mathcal{L}_h^{H*} \big(\widetilde{F}^{m-1} - \widetilde{F}^m\big) \end{bmatrix}, \\ \widetilde{\delta}_{h,2}^m &= \frac{\tau}{4} \begin{bmatrix} 0 \\ -\mathcal{L}_h^{H*} \big(\widetilde{F}^{m-1} - \widetilde{F}^m\big) \end{bmatrix},\end{aligned}$$

where we used the additional notation $\widetilde{F}^m = F(\widetilde{u}^m) + f(t_n)$. Note that $\widetilde{\delta}_{h,1}^m$ and $\widetilde{\delta}_{h,2}^m$ are similar to $\widetilde{\delta}_1^m$ and $\widetilde{\delta}_2^m$ in the proof of Theorem 4.9 while $\widetilde{\delta}_{h,0}^m$ is an additional defect in the fully discrete case. Using the bounds from Lemma 4.6 for $\widehat{Q}_+^{-1}$, we have

$$\|\widetilde{\delta}_{h,0}^m\|_{X_h} \leq C\tau(E_{h,1} + E_{h,3}). \tag{5.23}$$

As in the semidiscrete case, the terms $\widetilde{\delta}^m_{h,1}$ and $\widetilde{\delta}^m_{h,2}$ are only of order τ^2, and we use a combination of both terms from two successive time steps to gain an additional factor of τ. With the explicit representation of $\widehat{R}$ analogous to that of R in (4.20b), we obtain

$$\widetilde{\delta}^m_{h,1} + \widehat{R}\widetilde{\delta}^{m-1}_{h,2} = \frac{\tau}{2}\begin{bmatrix} \frac{\tau}{2}\widehat{Q}_+^{-1}\big(-\widetilde{F}^{m-2} + 2\widetilde{F}^{m-1} - \widetilde{F}^m\big) \\ \frac{1}{2}\widehat{Q}_-\widehat{Q}_+^{-1}\big(-\widetilde{F}^{m-2} + 2\widetilde{F}^{m-1} - \widetilde{F}^m\big) \end{bmatrix}.$$

Using this together with the bounds from Lemma 4.6 for $\widehat{Q}_+^{-1}$ and $\widehat{Q}_-\widehat{Q}_+^{-1}$, the continuity of the adjoint lift operator, and (4.36), leads to the bound

$$\begin{aligned}
\|\widetilde{\delta}^m_{h,1} + \widehat{R}\widetilde{\delta}^{m-1}_{h,2}\|_{X_h} &\leq C\tau\big\|\mathcal{L}_h^{H*}\big(-\widetilde{F}^{m-2} + 2\widetilde{F}^{m-1} - \widetilde{F}^m\big)\big\|_{m_h} \\
&\leq C\tau^3\left\|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\big(F(u) + f\big)\right\|_{L^\infty([t_{m-2},t_m];H)} \\
&\leq C\tau^3\left(\|u^{(4)}\|_{L^\infty([t_{m-2},t_m];H)} + \|u^{(3)}\|_{L^\infty([t_{m-2},t_m];V)} + \|Au''\|_{L^\infty([t_{m-2},t_m];H)}\right),
\end{aligned}$$

and, hence, together with (5.23)

$$\begin{aligned}
\Big\|\sum_{m=1}^n \widehat{R}^{n-m}\widetilde{\delta}^m_h\Big\|_{X_h} &\leq \Big\|\sum_{m=1}^n \widehat{R}^{n-m}\left(\widetilde{\delta}^m_{h,0} + \widetilde{\delta}^m_{h,1} + \widetilde{\delta}^m_{h,2}\right)\Big\|_{X_h} \\
&\leq C\mathrm{e}^{n\tau\widehat{c}_{\mathrm{qm}}}(E_{h,1} + E_{h,3}) + \Big\|\widehat{R}^{n-1}\widetilde{\delta}^1_{h,1} + \widetilde{\delta}^n_{h,2} + \sum_{m=2}^n \widehat{R}^{n-m}\big(\widetilde{\delta}^m_{h,1} + \widehat{R}\widetilde{\delta}^{m-1}_{h,2}\big)\Big\|_{X_h} \\
&\leq \mathrm{e}^{n\tau\widehat{c}_{\mathrm{qm}}}\left(C(E_{h,1} + E_{h,3}) + \|\widetilde{\delta}^1_{h,1}\|_{X_h} + \|\widetilde{\delta}^n_{h,2}\|_{X_h} + \sum_{m=2}^n \|\widetilde{\delta}^m_{h,1} + \widehat{R}\widetilde{\delta}^{m-1}_{h,2}\|_{X_h}\right) \\
&\leq C\mathrm{e}^{n\tau\widehat{c}_{\mathrm{qm}}}\left(E_{h,1} + E_{h,3} + \tau^2\right).
\end{aligned}$$

Inserting the bounds of all defects into (5.22) yields

$$\|e_h^n\|_X \leq C\mathrm{e}^{\widehat{M}t_n}\left(\sum_{i=1}^5 E_{h,i} + \tau^2\right). \tag{5.24}$$

Finally, the error bound (5.13) follows from

$$\|\mathcal{L}_h^V u_h^n - u(t_n)\|_{\tilde{a}} + \|\mathcal{L}_h^V v_h^n - u'(t_n)\|_m \leq \sqrt{2}\|\mathcal{L}_h x_h^n - x(t_n)\|_X,$$

(5.14), and (5.24). □

Remark 5.7.

a) Similar to Theorem 5.6, it is also possible to show a full discretization error bound for the Crank–Nicolson scheme. In this case, the error recursion (4.31) simplifies to

$$e_h^{n+1} = \widehat{R}e_h^n + \frac{\tau}{2}\widehat{R}_+^{-1}\left(G_h^{n+1} - \widetilde{G}_h^{n+1} + G_h^n - \widetilde{G}_h^n\right) + \delta_h^{n+1} + \widehat{R}_+^{-1}\mathcal{L}_h^*\delta_{\mathrm{CN}}^{n+1}$$

and the assertion of Theorem 5.6 holds with $1 + (3/2)^{1/2}$ replaced by 1 in the step size restriction and the error bound.

b) The step size restriction in Theorem 4.9 is not a CFL condition, since it only depends on constants that are independent of the mesh width h. Note that in the monotone case, where $\widehat{c}_G = \widehat{\beta}_{\mathrm{qm}} = 0$, the step size is only restricted by the Lipschitz constant $\widehat{L}_\rho$, which is usual for the time integration of semilinear problems.

Theorem 5.8 ([Hochbruck and Leibold, 2021, Corollary 3.5]). *Let Assumptions 2.12, 2.17 and 2.19 be satisfied and let* $u \in C^4([0,T];H) \cap C^3([0,T];V) \cap C^2([0,T];D(A))$ *be the solution of* (4.1) *with* $u, u', u'' \in L^\infty([0,T];Z^V)$. *Further, let the space discretization error terms* $E_{h,i}$ *defined in* (2.40) *satisfy* $E_{h,i} \overset{h\to 0}{\longrightarrow} 0$ *for* $i = 1, \dots, 5$.

Then, there exist $\tau^*, h^* > 0$ *s.t. for all* $h < h^*, \tau < \tau^*$ *the iterations* u_h^n, v_h^n *of the fully discrete IMEX scheme* (5.11) *satisfy*

$$\max_{t_n \le T} \|u_h^n\|_{\tilde{a}_h} \le \rho := 2\widehat{C}_V \|u\|_{L^\infty([0,T];V)}, \tag{5.25}$$

and the error bound (5.13) *holds true with* $\rho_h = \rho$.

Proof. We only have to prove the bound (5.25) for τ and h sufficiently small, then the other assertions follow immediately from Theorem 5.6. This can be proven similar to part (d) in the proof of Theorem 4.9:

Let $u_h^{\rho,n}$ be the iterates of the IMEX scheme (5.11) with F_h replaced by F_h^ρ, where $F_h^\rho \colon V_h \to H_h$ is a function that is globally Lipschitz continuous on V_h with Lipschitz constant L_ρ and satisfies

$$F_h^\rho(v_h) = F_h(v_h) \text{ for all } v_h \in V_h \text{ with } \|v_h\|_{\tilde{a}_h} \le \rho.$$

Due to (5.25), we have

$$F_h(\mathcal{L}_h^{V*} u(t)) = F_h^\rho(\mathcal{L}_h^{V*} u(t)) \text{ for all } t \in [0,T].$$

Hence, as in the proof of Theorem 5.6, we obtain similar to the bound of the first component in (5.24)

$$\|u_h^{\rho,n} - \mathcal{L}_h^{V*} \widetilde{u}_h^n\|_{\tilde{a}_h} \le C e^{\widehat{M} t_n} \left(\sum_{i=0}^{4} E_{h,i} + \tau^2 \right)$$

for all $t_n \le T$. Since C is independent of h and τ, and $E_{h,i} \overset{h\to 0}{\longrightarrow} 0$ for $i = 1, \dots 5$, we can choose $h^*, \tau^* > 0$ s.t., for all $h < h^*, \tau < \tau^*$, we have

$$\|u_h^{\rho,n} - \mathcal{L}_h^{V*} \widetilde{u}_h^n\|_{\tilde{a}_h} \le \frac{\rho}{2}.$$

Hence, we obtain together with (5.25)

$$\|u_h^{\rho,n}\|_{\tilde{a}_h} \le \|u_h^{\rho,n} - \mathcal{L}_h^{V*} \widetilde{u}_h^n\|_{\tilde{a}_h} + \|\mathcal{L}_h^{V*} \widetilde{u}_h^n\|_{\tilde{a}_h} \le \frac{\rho}{2} + \widehat{C}_V \|\widetilde{u}_h^n\|_{\tilde{a}} \le \rho.$$

This implies that for all $t_n \le T$ we have $u_h^{\rho,n} = u_h^n$ and therefore also $\|u_h^n\|_{\tilde{a}_h} = \|u_h^{\rho,n}\|_{\tilde{a}_h} \le \rho$. □

CHAPTER 6

Wave equation with kinetic boundary conditions and nonlinear forcing and damping

In this chapter, we use the abstract results from the previous chapters for the numerical analysis of a wave equation with kinetic boundary conditions. We present a non-conforming finite element space discretization and show that both the equation and the space discretization fit in the setting of the unified error analysis presented in Chapter 2. By using the abstract error bounds presented there, we prove a space discretization error bound of order p for a discretization with order p finite elements. Moreover, we use the abstract time discretization analysis from Chapters 3 to 5 to analyze time and full discretization errors for the wave equation with kinetic boundary conditions.

On the one hand, these are new results for the numerical analysis of the wave equation with kinetic boundary conditions. But on the other hand, this chapter also aims to show exemplarily the application of the abstract theory of this thesis to a concrete example.

The wave equation with kinetic boundary conditions was also considered in Hipp [2017] in the linear case and in Hochbruck and Leibold [2020, 2021] in the semilinear case. In this thesis, we additionally add nonlinear damping terms and extend the numerical analysis to this nonlinear case.

Outline We introduce the analytical setting and the wave equation with kinetic boundary conditions in Section 6.1. In Section 6.2, we present a suitable finite element space discretization for which we prove an error bound in Section 6.3. Then, in Section 6.4, we study time and full discretization errors for algebraically stable Runge–Kutta methods and our IMEX scheme. Finally, in Section 6.5, we comment on the implementation of the different schemes and present numerical experiments.

6.1 Analytical equation

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with C^{p+1} boundary $\Gamma = \partial\Omega$ for $d \in \{2, 3\}$ and some $p \in \mathbb{N}$.

We consider the wave equation with kinetic boundary conditions given by

$$u_{tt} + \big(\alpha_\Omega(\mathbf{x}) + \beta_\Omega(\mathbf{x}) \cdot \nabla\big) u_t + \mathcal{D}_\Omega(\mathbf{x}, u_t) - \Delta u = F_\Omega(\mathbf{x}, u) + f_\Omega(t, \mathbf{x}), \qquad \text{for } t \geq 0, \mathbf{x} \in \Omega, \tag{6.1a}$$

$$u_{tt} + \partial_{\mathbf{n}} u + \mathcal{D}_\Gamma(\mathbf{x}, u_t) - \Delta_\Gamma u = F_\Gamma(\mathbf{x}, u) + f_\Gamma(t, \mathbf{x}), \qquad \text{for } t \geq 0, \mathbf{x} \in \Gamma, \tag{6.1b}$$

$$u(0, \mathbf{x}) = u^0(\mathbf{x}), \qquad u_t(0, \mathbf{x}) = v^0(\mathbf{x}), \qquad \text{in } \overline{\Omega}. \tag{6.1c}$$

Here, we have suppressed the arguments $(t, \mathbf{x})$ of the unknown u.

Physical motivation A physical example for kinetic boundary conditions in the case $d = 2$ can be found in [Goldstein, 2006, Section 5]. In this paper, they were derived by considering a vibrating membrane where its boundary carries a mass density and is subject to linear tension. An example for this situation is the membrane of a bass drum with a whole in the interior having a thick border. In addition to this inner border, Γ then also consist of the outer boundary of the membrane on which, e.g., Dirichlet boundary conditions can be posed (cf. also Vitillaro [2017]).

As shown in [Nicaise, 2017, Section 3.2], kinetic boundary conditions can also serve as an effective model for the interaction of an acoustic wave with a thin boundary layer with distinctive elastic or damping properties, and where the wave length is large compared to the width of the boundary layer.

In (6.1), $\mathcal{D}_\Omega, \mathcal{D}_\Gamma$ are nonlinear damping terms and F_Ω, F_Γ are nonlinear forcing terms in the interior of the domain and on its boundary, respectively. Further, $(\alpha_\Omega + \beta_\Omega \cdot \nabla)$ is a linear damping term in the interior.

Remark 6.1.

a) Similar to the linear damping term in the interior, it is possible to add a linear damping term on the boundary (cf. Hipp [2017]). We do not include it here for the sake of readability.

b) In contrast to Hochbruck and Leibold [2020, 2021], in our case the nonlinear forcing terms are not allowed to depend on time. This case is not covered by the abstract framework in Section 2.2 (cf. Remark 2.5).

c) Wellposedness and stability of (6.1), without the linear damping term but in more general spaces than we consider in this thesis, was analyzed in Vitillaro [2017].

In the following, we rewrite (6.1) in a variational formulation and pose suitable assumptions such that it fits into the setting of Section 2.2.

Assumption 6.2.

a) The nonlinearities satisfy $F_\Omega \in C^1(\overline{\Omega} \times \mathbb{R}; \mathbb{R})$, $\mathcal{D}_\Omega \in C(\overline{\Omega} \times \mathbb{R}; \mathbb{R})$, $F_\Gamma \in C^1(\Gamma \times \mathbb{R}; \mathbb{R})$, and $\mathcal{D}_\Gamma \in C(\Gamma \times \mathbb{R}; \mathbb{R})$.

b) There exist

$$1 \le \zeta_\Omega \begin{cases} < \infty, & d = 2, \\ \le 3, & d = 3, \end{cases} \qquad \text{and} \qquad 1 \le \zeta_\Gamma < \infty,$$

and a constant $C > 0$ such that for all $\mathbf{x} \in \Omega$ and all $\xi \in \mathbb{R}$

$$\begin{aligned} |F_\Omega(\mathbf{x}, \xi)| &\le C(1 + |\xi|^{\zeta_\Omega}), \\ |\partial_2 F_\Omega(\mathbf{x}, \xi)| &\le C(1 + |\xi|^{\zeta_\Omega - 1}), \\ |\mathcal{D}_\Omega(\mathbf{x}, \xi)| &\le C(1 + |\xi|^{\zeta_\Omega}), \end{aligned} \tag{6.2}$$

and for all $\mathbf{x} \in \Gamma$ and all $\xi \in \mathbb{R}$

$$\begin{aligned} |F_\Gamma(\mathbf{x}, \xi)| &\le C(1 + |\xi|^{\zeta_\Gamma}), \\ |\partial_2 F_\Gamma(\mathbf{x}, \xi)| &\le C(1 + |\xi|^{\zeta_\Gamma - 1}), \\ |\mathcal{D}_\Gamma(\mathbf{x}, \xi)| &\le C(1 + |\xi|^{\zeta_\Gamma}), \end{aligned}$$

hold true. Here, ∂_2 denotes the derivative w.r.t. the second variable, i.e., w.r.t. ξ.

c) There exists a constant $c' \ge 0$ s.t. for all $\mathbf{x} \in \Omega$ and all $\xi \in \mathbb{R}$ we have

$$\partial_2 \mathcal{D}_\Omega(\mathbf{x}, \xi) \ge -c',$$

and for all $\mathbf{x} \in \Gamma$ and all $\xi \in \mathbb{R}$

$$\partial_2 \mathcal{D}_\Gamma(\mathbf{x}, \xi) \ge -c'$$

holds true.

d) The inhomogeneities satisfy $f_\Omega \in W^{1,1}_{loc}([0, \infty); C(\overline{\Omega}))$, $f_\Gamma \in W^{1,1}_{loc}([0, \infty); C(\Gamma))$.

e) The coefficients satisfy $\alpha_\Omega \in C(\overline{\Omega})$ is non-negative, $\beta_\Omega \in C^1(\overline{\Omega})^d$, and

$$\alpha_\Omega - \frac{1}{2} \operatorname{div} \beta_\Omega \ge 0 \quad \text{in } \Omega, \qquad \beta_\Omega \cdot \mathbf{n} \ge 0 \quad \text{on } \Gamma.$$

To derive a weak formulation, we multiply (6.1a) by a test function $\varphi \in C^\infty(\overline{\Omega})$, integrate over Ω, and use Gauss' Theorem. This yields for all $t \ge 0$

$$\begin{aligned} \big(u_{tt}, \varphi\big)_{L^2(\Omega)} + \big((\alpha_\Omega + \beta_\Omega \cdot \nabla)\, u_t, \varphi\big)_{L^2(\Omega)} + \big(\mathcal{D}_\Omega(\cdot, u_t), \varphi\big)_{L^2(\Omega)} - \big(\partial_{\mathbf{n}} u, \varphi\big)_{L^2(\Gamma)} + \big(\nabla u, \nabla \varphi\big)_{L^2(\Omega)} \\ = \big(F_\Omega(\cdot, u), \varphi\big)_{L^2(\Omega)} + \big(f_\Omega(t, \cdot), \varphi\big)_{L^2(\Omega)}. \end{aligned} \tag{6.3}$$

On Γ, the following version of Gauss' Theorem holds true for all $v \in H^2(\Gamma), w \in H^1(\Gamma)$ (cf. [Kashiwabara et al., 2015, (3.1)]):

$$-\int_\Gamma (\Delta_\Gamma v) w \, \mathrm{d}s = \int_\Gamma \nabla_\Gamma v \cdot \nabla_\Gamma w \, \mathrm{d}s.$$

Thus, from the boundary equation (6.1b) we obtain similarly to (6.3)

$$\begin{aligned} \big(u_{tt}, \varphi\big)_{L^2(\Gamma)} + \big(\partial_{\mathbf{n}} u, \varphi\big)_{L^2(\Gamma)} + \big(\mathcal{D}_\Gamma(\cdot, u_t), \varphi\big)_{L^2(\Gamma)} + \big(\nabla_\Gamma u, \nabla_\Gamma \varphi\big)_{L^2(\Gamma)} \\ = \big(F_\Gamma(\cdot, u), \varphi\big)_{L^2(\Gamma)} + \big(f_\Gamma(t, \cdot), \varphi\big)_{L^2(\Gamma)}. \end{aligned} \tag{6.4}$$

By adding (6.3) and (6.4), we end up with

$$m\big(u_{tt},\varphi\big) + m\big(\mathcal{D}(u_t),\varphi\big) + a\big(u,\varphi\big) = m\big(F(u),\varphi\big) + m\big(f(t),\varphi\big), \qquad \text{for } t \geq 0, \tag{6.5}$$

where

$$\begin{aligned} m\big(v,\varphi\big) &= \int_\Omega v\varphi \,\mathrm{d}\mathbf{x} + \int_\Gamma v\varphi \,\mathrm{d}s, && \text{(6.6a)}\\ a\big(v,\varphi\big) &= \int_\Omega \nabla v\cdot\nabla\varphi \,\mathrm{d}\mathbf{x} + \int_\Gamma \nabla_\Gamma v\cdot\nabla_\Gamma\varphi \,\mathrm{d}s, && \text{(6.6b)}\\ m\big(\mathcal{D}(v),\varphi\big) &= \int_\Omega \big((\alpha_\Omega + \beta_\Omega\cdot\nabla)\, v + \mathcal{D}_\Omega(\mathbf{x},v)\big)\varphi \,\mathrm{d}\mathbf{x} + \int_\Gamma \mathcal{D}_\Gamma(\mathbf{x},v)\varphi \,\mathrm{d}s, && \text{(6.6c)}\\ m\big(F(v),\varphi\big) &= \int_\Omega F_\Omega(\mathbf{x},v)\varphi \,\mathrm{d}\mathbf{x} + \int_\Gamma F_\Gamma(\mathbf{x},v)\varphi \,\mathrm{d}s, && \text{(6.6d)}\\ m\big(f,\varphi\big) &= \int_\Omega f_\Omega\varphi \,\mathrm{d}\mathbf{x} + \int_\Gamma f_\Gamma\varphi \,\mathrm{d}s. && \text{(6.6e)} \end{aligned}$$

To obtain from (6.5) a well-defined weak formulation, we have to specify suitable Hilbert spaces on which the objects from (6.6) are defined. Therefore, we set

$$H \coloneqq L^2(\Omega)\times L^2(\Gamma) \quad \text{and} \quad V \coloneqq H^1(\Omega;\Gamma), \tag{6.6f}$$

where

$$H^k(\Omega;\Gamma) \coloneqq \{v \in H^k(\Omega) \mid \gamma(v) \in H^k(\Gamma)\}, \quad k \geq 1,$$

and γ denotes the Dirichlet trace operator. As shown in [Kashiwabara et al., 2015, Lemma 2.5], the spaces $H^k(\Omega;\Gamma)$ are Hilbert spaces w.r.t. the scalar product

$$\big(v,w\big)_{H^k(\Omega;\Gamma)} = \big(v,w\big)_{H^k(\Omega)} + \big(\gamma(v),\gamma(w)\big)_{H^k(\Gamma)}.$$

Further, in the proof of [Hipp, 2017, Corollary 6.7] it was shown that V is densely embedded into H via the embedding

$$v \mapsto [v,\gamma(v)]^\intercal\,.$$

By definition, m is the standard scalar product on H and $\tilde{a} = a + m$ is the scalar product on V.

Remark 6.3. *We use the following conventions:*

a) From now on and as in Section 2.2.1, we understand the weak solution $u\colon [0,T]\times\Omega \to \mathbb{R}$ of (6.1) on a time interval $[0,T]$ as a function $u\colon [0,T] \to V = H^1(\Omega;\Gamma)$. This is common, when reformulating a time-dependent partial differential equation into a weak formulation or an evolution equation.

b) For a function $v = [v_\Omega, v_\Gamma]^\intercal \in L^2(\Omega)\times L^2(\Gamma)$, we use the notation

$$\int_\Omega v\,\mathrm{d}\mathbf{x} \coloneqq \int_\Omega v_\Omega\,\mathrm{d}\mathbf{x}, \quad \textit{and} \quad \int_\Gamma v\,\mathrm{d}s \coloneqq \int_\Omega v_\Gamma\,\mathrm{d}s,$$

since by the domain of the integral, it is clear which component of v is used. Similarly, we use for $v \in H^k(\Omega;\Gamma)$, $k \geq 1$, the notation

$$\int_\Gamma v\,\mathrm{d}s \coloneqq \int_\Gamma \gamma(v)\,\mathrm{d}s.$$

In the next lemma, we verify that the weak formulation of (6.1) fits into the setting of Section 2.2.1.

Lemma 6.4 (Weak formulation of (6.1))**.** *Let Assumption 6.2 hold true. Then, with the objects defined in (6.6), the weak formulation of (6.1) is of the form (2.13). Further, Assumption 2.12 is satisfied with* $c_G = 1, \beta_{\mathrm{qm}} = c'$ *and*

$$L_\rho = C\left(1 + \rho^{\zeta_\Omega - 1} + \rho^{\zeta_\Gamma - 1}\right).$$

Here, ζ_Ω *and* ζ_Γ *are given in Assumption 6.2 c), and the constant* C *is independent of* ρ.

Proof. We derived in (6.5), that the weak formulation of (6.1) is of the form (2.13). In the following, we show that Assumption 2.12 is satisfied.

We already noticed that Assumption 2.12 a) and b) are satisfied. Assumption 2.12 e) follows directly by Assumption 6.2 d), since $C(\overline{\Omega}) \subset L^2(\Omega)$ and $C(\Gamma) \subset L^2(\Gamma)$. Part d) was proven in [Leibold, 2017, Lemma 4.2], where the semilinear wave equation with kinetic boundary conditions was considered, but since it is written in German, we recall the proof here: By Corollary A.6 and the growth conditions from Assumption 6.2 b), we have that

$$v \mapsto F_\Omega(\cdot, v) \in C(H^1(\Omega); L^2(\Omega)), \qquad v \mapsto F_\Gamma(\cdot, v) \in C(H^1(\Gamma), L^2(\Gamma))$$

and, hence, $F \in C(V; H)$. Now let $v, w \in V$ with $\|v\|_{\tilde{a}}, \|w\|_{\tilde{a}} \leq \rho$. By the definition (6.6d) of F we have

$$\|F(v) - F(w)\|_m^2 = \|F_\Omega(\cdot, v) - F_\Omega(\cdot, w)\|_{L^2(\Omega)}^2 + \|F_\Gamma(\cdot, v) - F_\Gamma(\cdot, w)\|_{L^2(\Gamma)}^2.$$

With Hölder's inequality, the Sobolev embedding theorem (cf. Theorem A.4), and the growth condition (6.2) we obtain

$$\begin{aligned}
\|F_\Omega(\cdot, v) - F_\Omega(\cdot, w)\|_{L^2(\Omega)} &= \|\int_0^1 \Big(\partial_2 F_\Omega(\cdot, w + \xi(v - w))\Big)(v - w)\,\mathrm{d}\xi\|_{L^2(\Omega)} \\
&\leq \int_0^1 \|\big(\partial_2 F_\Omega(\cdot, w + \xi(v - w))\big)(v - w)\|_{L^2(\Omega)}\,\mathrm{d}\xi \\
&\leq \int_0^1 \|\partial_2 F_\Omega(\cdot, w + \xi(v - w))\|_{L^{\frac{2\zeta_\Omega}{\zeta_\Omega - 1}}(\Omega)} \|(v - w)\|_{L^{2\zeta_\Omega}(\Omega)}\,\mathrm{d}\xi \\
&\leq \sup_{\|\varphi\|_{\tilde{a}} \leq \rho} \|\partial_2 F_\Omega(\cdot, \varphi)\|_{L^{\frac{2\zeta_\Omega}{\zeta_\Omega - 1}}(\Omega)} \|(v - w)\|_{H^1(\Omega)} \\
&\leq \sup_{\|\varphi\|_{\tilde{a}} \leq \rho} \|C(1 + |\varphi|^{\zeta_\Omega - 1})\|_{L^{\frac{2\zeta_\Omega}{\zeta_\Omega - 1}}(\Omega)} \|(v - w)\|_{\tilde{a}} \\
&\leq C\Big(1 + \sup_{\|\varphi\|_{\tilde{a}} \leq \rho} \|\varphi\|_{L^{2\zeta_\Omega}(\Omega)}^{\zeta_\Omega - 1}\Big)\|(v - w)\|_{\tilde{a}} \\
&\leq C\Big(1 + \sup_{\|\varphi\|_{\tilde{a}} \leq \rho} \|\varphi\|_{H^1(\Omega)}^{\zeta_\Omega - 1}\Big)\|(v - w)\|_{\tilde{a}} \\
&\leq C\Big(1 + \rho^{\zeta_\Omega - 1}\Big)\|(v - w)\|_{\tilde{a}},
\end{aligned}$$

and similarly

$$\|F_\Gamma(\cdot, v) - F_\Gamma(\cdot, w)\|_{L^2(\Gamma)} \leq C\Big(1 + \rho^{\zeta_\Gamma - 1}\Big)\|(v - w)\|_{\tilde{a}}.$$

This proves the Lipschitz continuity.

It remains to prove Assumption 2.12 c). As for the nonlinear forcing terms, we obtain by Corollary A.6 and the growth conditions from Assumption 6.2 b)

$$v \mapsto \mathcal{D}_\Omega(\cdot, v) \in C(H^1(\Omega); L^2(\Omega)), \qquad v \mapsto \mathcal{D}_\Gamma(\cdot, v) \in C(H^1(\Gamma), L^2(\Gamma)).$$

Furthermore, we have

$$v \mapsto (\alpha_\Omega + \beta_\Omega \cdot \nabla)\, v \in \mathcal{L}(H^1(\Omega); L^2(\Omega)), \tag{6.7}$$

and, hence, $\mathcal{D} \in C(V; H) \subset C(V; V^*)$ as a sum of continuous functions.

In [Hipp, 2017, Lemma 6.3] the monotonicity, i.e., the quasi-monotonicity with constant 0, of the linear damping term (6.7) was shown. We now prove the quasi-monotonicity of the nonlinear damping term

$$v \mapsto [\mathcal{D}_\Omega(\cdot, v), \mathcal{D}_\Gamma(\cdot, v)]^\intercal : V \to H.$$

Then, $\mathcal{D}$ is quasi-monotone as the sum of two quasi-monotone functions.

For $v, w \in V$ we have

$$\begin{aligned}
m\Big(\begin{bmatrix} \mathcal{D}_\Omega(\cdot, v) - F_\Omega(\cdot, w) \\ \mathcal{D}_\Gamma(\cdot, v) - \mathcal{D}_\Gamma(\cdot, w) \end{bmatrix}, v - w \Big) &= \int_\Omega \big(\mathcal{D}_\Omega(\mathbf{x}, v(\mathbf{x})) - \mathcal{D}_\Omega(\mathbf{x}, w(\mathbf{x}))\big)(v(\mathbf{x}) - w(\mathbf{x}))\,\mathrm{d}\mathbf{x} \\
&\quad + \int_\Gamma \big(\mathcal{D}_\Gamma(\mathbf{x}, v(\mathbf{x})) - \mathcal{D}_\Gamma(\mathbf{x}, w(\mathbf{x}))\big)(v(\mathbf{x}) - w(\mathbf{x}))\,\mathrm{d}s \\
&= \int_\Omega \Big(\int_0^1 \partial_2 \mathcal{D}_\Omega\big(\mathbf{x}, v(\mathbf{x}) + \xi(w(\mathbf{x}) - v(\mathbf{x}))\big)\,\mathrm{d}\xi \Big)(v(\mathbf{x}) - w(\mathbf{x}))^2\,\mathrm{d}\mathbf{x} \\
&\quad + \int_\Gamma \Big(\int_0^1 \partial_2 \mathcal{D}_\Gamma\big(\mathbf{x}, v(\mathbf{x}) + \xi(w(\mathbf{x}) - v(\mathbf{x}))\big)\,\mathrm{d}\xi \Big)(v(\mathbf{x}) - w(\mathbf{x}))^2\,\mathrm{d}s \\
&\geq -c' \Big(\int_\Omega (v(\mathbf{x}) - w(\mathbf{x}))^2\,\mathrm{d}\mathbf{x} + \int_\Gamma (v(\mathbf{x}) - w(\mathbf{x}))^2\,\mathrm{d}s \Big) \\
&= -c' \|v - w\|_m^2.
\end{aligned}$$

Since the linear damping term is monotone, the quasi-monotonicity constant β_{qm} of $\mathcal{D}$ is equal to the quasi-monotonicity constant of the nonlinear damping term given by c'. □

By Lemma 6.4 and Corollary 2.16, we have that that the variational formulation of (6.1) is locally wellposed.

6.2 Finite element space discretization

To discretize (6.1) in space, we use the bulk-surface finite element method with isoparametric elements of order p. The method was introduced in Elliott and Ranner [2013] and we give a short summary on the construction and important properties in Appendix C. The bulk-surface finite element method was also used to discretize the wave equations with kinetic boundary conditions in the linear and semilinear case (cf. Hipp [2017], Hochbruck and Leibold [2020]).

Let $(\mathcal{T}_h)_h$ be a quasi-uniform family of triangulations of Ω consisting of isoparametric elements of order p with corresponding finite element spaces $V_h = V_{h,p}^{\Omega}$ as defined in (C.1). We recall that the computational domain is denoted by $\Omega_h = \bigcup_{K\in\mathcal{T}_h} K \approx \Omega$ with boundary $\Gamma_h \approx \Gamma$. To discretize the nonlinearities, we make use of the nodal interpolation operators $I_{h,\Omega}\colon C(\overline{\Omega}) \to V_h$ and $I_{h,\Gamma}\colon C(\Gamma) \to V_{h,p}^{\Gamma}$, where $V_{h,p}^{\Gamma}$ is the corresponding surface finite element space (cf. (C.2)). Additionally, we make use of two elementwise defined quadrature formulas

$$\sum_{\Omega_h} \cdot \,\mathrm{d}\mathbf{x}\colon C(\overline{\Omega_h}) \to \mathbb{R}, \qquad\qquad \sum_{\Gamma_h} \cdot \,\mathrm{d}s\colon C(\Gamma_h) \to \mathbb{R}$$

that approximate the integrals $\int_{\Omega_h} \cdot \,\mathrm{d}\mathbf{x}$ and $\int_{\Gamma_h} \cdot \,\mathrm{d}s$, respectively. We require that the quadrature formulas have positive weights and are of order greater than $2p$, s.t. we have for all $v_h, w_h \in V_h$

$$m_h\big(v_h, w_h\big) = \sum_{\Omega_h} v_h w_h \,\mathrm{d}\mathbf{x} + \sum_{\Gamma_h} v_h w_h \,\mathrm{d}s. \tag{6.8}$$

For the discretization of the nonlinear damping terms, we have to pose the additional assumption

$$\mathcal{D}_\Omega \in C(\widehat{\Omega} \times \mathbb{R}; \mathbb{R}) \quad \text{and} \quad \mathcal{D}_\Gamma \in C(\widehat{\Gamma} \times \mathbb{R}; \mathbb{R}),$$

where $\widehat{\Omega}, \widehat{\Gamma} \subset \mathbb{R}^d$ are open neighborhoods of $\overline{\Omega}$ and Γ, respectively. This ensures tha for h sufficiently small we have $\Omega_h \subset \widehat{\Omega}$, $\Gamma_h \subset \widehat{\Gamma}$ and, hence,

$$\mathcal{D}_\Omega \in C(\Omega_h \times \mathbb{R}; \mathbb{R}) \quad \text{and} \quad \mathcal{D}_\Gamma \in C(\Gamma_h \times \mathbb{R}; \mathbb{R}). \tag{6.9}$$

The spatial discretization of (6.1) with isoparametric finite elements is then given by (2.24), where we define the discretized quantities via

$$m_h\big(v_h, w_h\big) := \int_{\Omega_h} v_h w_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_h} v_h w_h \,\mathrm{d}s, \tag{6.10a}$$

$$a_h\big(v_h, w_h\big) := \int_{\Omega_h} \nabla v_h \cdot \nabla w_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_h} \nabla_{\Gamma_h} u_h \cdot \nabla_{\Gamma_h} w_h \,\mathrm{d}s, \tag{6.10b}$$

$$m_h\big(\mathcal{D}_h(v_h), w_h\big) := \int_{\Omega_h} \big((I_{h,\Omega}\alpha_\Omega) v_h + (I_{h,\Omega}\beta_\Omega) \cdot \nabla v_h\big) w_h \,\mathrm{d}\mathbf{x} + \sum_{\Omega_h} \mathcal{D}_\Omega(\mathbf{x}, v_h) w_h \,\mathrm{d}\mathbf{x} + \sum_{\Gamma_h} \mathcal{D}_\Gamma(\mathbf{x}, v_h) w_h \,\mathrm{d}s, \tag{6.10c}$$

$$m_h\big(F_h(v_h), w_h\big) := \int_{\Omega_h} \big(I_{h,\Omega} F_\Omega\big(\cdot, v_h^\ell\big)\big)\, w_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_h} \big(I_{h,\Gamma} F_\Gamma\big(\cdot, v_h^\ell\big)\big)\, w_h \,\mathrm{d}s, \tag{6.10d}$$

$$m_h\big(f_h, w_h\big) := \int_{\Omega_h} (I_{h,\Omega} f_\Omega)\, w_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_h} (I_{h,\Gamma} f_\Gamma)\, w_h \,\mathrm{d}s, \tag{6.10e}$$

for all $v_h, w_h \in V_h$. Here, $v_h^\ell \in C(\Omega)$ denotes the lifted version of v_h defined in (C.3).

Remark 6.5.

a) *The nodal interpolation only requires function evaluations in the basis nodes $a_1, \dots, a_N$ of the finite element space. Since these are invariant under the lift operator, the computation of v_h^ℓ is not necessary. The lift is only needed for the definition of F_h since the interpolation operator acts on functions over Ω.*

b) The definition of $\mathcal{D}_h$ relies on (6.9)*, since the quadrature points are in general not contained in Ω or Γ, respectively.*

c) The use of the quadrature formulas in the definition of $\mathcal{D}_h$ is required to prove that $\mathcal{D}_h$ is quasi-monotone (Lemma 6.6). It is possible to discretize F in the same way, but the definition via the interpolation is more efficient with respect to the implementation. We discuss this in Section 6.5.1.

We now prove, that this discretization fits into the abstract setting of Section 2.2.2.

Lemma 6.6. *Let Assumption 6.2 hold true and additionally let*

$$I_{h,\Omega}\alpha_\Omega \geq 0, \quad I_{h,\Omega}\alpha_\Omega - \frac{1}{2}\operatorname{div} I_{h,\Omega}\beta_\Omega \geq 0 \quad \text{in } \Omega, \qquad I_{h,\Omega}\beta_\Omega \cdot \mathbf{n} \geq 0 \quad \text{on } \Gamma. \tag{6.11}$$

Then, the bulk-surface finite element space discretization of (6.1) *satisfies Assumption 2.17 with $\widehat{c}_G = 1$, $\widehat{\beta}_{\mathrm{qm}} = c'$, $\widehat{C}_{H,V} = 1$, and*

$$\widehat{L}_\rho = C\left(1 + \rho^{\zeta_\Omega - 1} + \rho^{\zeta_\Gamma - 1}\right).$$

Here, ζ_Ω and ζ_Γ are from Assumption 6.2 c), and the constant C is independent of ρ.

Proof. Assumption 2.17 a) and b) are trivially satisfied, since we have

$$V_h \subset H^1(\Omega_h;\Gamma_h) \hookrightarrow L^2(\Omega_h) \times L^2(\Gamma_h),$$

and $\tilde{a}_h = a_h + m_h$ and m_h are the corresponding scalar products on these spaces and, hence, also scalar products on the subspace V_h. Part f) follows directly by the definition of $\tilde{a}_h = a_h + m_h$.

Assumption 2.17 d) was proven in Hochbruck and Leibold [2020] where the semilinear case was considered. Further, Assumption 2.17 e) follows from Assumption 6.2 d) and the continuity of the interpolation operators, cf. Lemma C.4.

It remains to prove c). As in the continuous case, we can split $\mathcal{D}_h$, in the linear part, that was also considered in Hipp [2017], and the nonlinear part. For the linear part, given for for $v_h, w_h \in V_h$ by

$$\int_{\Omega_h} \left((I_{h,\Omega}\alpha_\Omega)v_h + (I_{h,\Omega}\beta_\Omega)\cdot\nabla v_h\right) w_h \,\mathrm{d}\mathbf{x},$$

it was shown in the proof of [Hipp, 2017, Theorem 7.4], that under the assumption (6.11) it is continuous and monotone, i.e., quasi-monotone with constant 0.

We now prove the continuity of the nonlinear part of $\mathcal{D}_h$. Since V_h is a finite dimensional vector space, we have for $v_h, w_h \in V_h$

$$\begin{aligned}
&\sup_{\|\varphi_h\|_{m_h}=1}\left(\sum_{\Omega_h}\left(\mathcal{D}_\Omega(\cdot,v_h) - \mathcal{D}_\Omega(\cdot,w_h)\right)\varphi_h \,\mathrm{d}\mathbf{x} + \sum_{\Gamma_h}\left(\mathcal{D}_\Gamma(\cdot,v_h) - \mathcal{D}_\Gamma(\cdot,w_h)\right)\varphi_h \,\mathrm{d}s\right)\\
&\quad = \max_{i=1}^{N}\frac{1}{\|\phi_i\|_{m_h}}\left(\sum_{\Omega_h}\left(\mathcal{D}_\Omega(\cdot,v_h) - \mathcal{D}_\Omega(\cdot,w_h)\right)\phi_i \,\mathrm{d}\mathbf{x} + \sum_{\Gamma_h}\left(\mathcal{D}_\Gamma(\cdot,v_h) - \mathcal{D}_\Gamma(\cdot,w_h)\right)\phi_i \,\mathrm{d}s\right),
\end{aligned}$$

where $\phi_i, i = 1,\dots,N$, are the nodal basis functions of V_h. For $v_h \to w_h$ in V_h, we have for all $i = 1,\dots,N$

$$\sum_{\Omega_h}\left(\mathcal{D}_\Omega(\cdot,v_h) - \mathcal{D}_\Omega(\cdot,w_h)\right)\phi_i \,\mathrm{d}\mathbf{x} + \sum_{\Gamma_h}\left(\mathcal{D}_\Gamma(\cdot,v_h) - \mathcal{D}_\Gamma(\cdot,w_h)\right)\phi_i \,\mathrm{d}s \to 0,$$

where we used (6.9) and the fact that convergence in V_h implies pointwise convergence. This proves the continuity of $\mathcal{D}_h$.

The proof of the quasi-monotonicity works similar to the continuous case, cf. the proof of Lemma 6.4. Let $v_h, w_h \in V_h$. The nonlinear part of $m_h\big(\mathcal{D}_h(v_h) - \mathcal{D}_h(w_h), v_h - w_h\big)$ then satisfies

$$\begin{aligned}
\sum_{\Omega_h} (\mathcal{D}_\Omega(\mathbf{x}, v_h) - \mathcal{D}_\Omega(\mathbf{x}, w_h))\,(v_h - w_h)\,\mathrm{d}\mathbf{x} + \sum_{\Gamma_h} (\mathcal{D}_\Gamma(\mathbf{x}, v_h) - \mathcal{D}_\Gamma(\mathbf{x}, w_h))\,(v_h - w_h)\,\mathrm{d}s& \\
= \sum_{\Omega_h} \bigg(\int_0^1 \partial_2 \mathcal{D}_\Omega\big(\mathbf{x}, v_h + \xi(w_h - v_h)\big)\,\mathrm{d}\xi \bigg) (v_h - w_h)^2\,\mathrm{d}\mathbf{x}& \\
+ \sum_{\Gamma_h} \bigg(\int_0^1 \partial_2 \mathcal{D}_\Gamma\big(\mathbf{x}, v_h + \xi(w_h - v_h)\big)\,\mathrm{d}\xi \bigg) (v_h - w_h)^2\,\mathrm{d}s& \\
\geq - c' \bigg(\sum_{\Omega_h} (v_h - w_h)^2\,\mathrm{d}\mathbf{x} + \sum_{\Gamma_h} (v_h - w_h)^2\,\mathrm{d}s \bigg)& \\
= - c' \| v_h - w_h \|_{m_h}^2,&
\end{aligned}$$

where we used Assumption 6.2 c) and (6.8). This finishes the proof. □

Remark 6.7. *Note that by Assumption 6.2 e) and the interpolation error bound (C.7), we have that (6.11) is at least asymptotically satisfied for $h \to 0$.*

6.3 Space discretization error bound

To apply the abstract error results from Section 2.2, we have to specify the operators appearing in this context.

Definition 6.8.

a) The lift operator $\mathcal{L}_h^V \in \mathcal{L}(V_h; V)$ is defined via

$$\mathcal{L}_h^V v_h := v_h^\ell \qquad \textit{for all } v_h \in V_h$$

with v_h^ℓ defined in (C.3).

b) We set $Z^V := H^2(\Omega; \Gamma)$.

c) We define the interpolation operator via $I_h := I_{h,\Omega}$.

Lemma 6.9. *The operators defined in Definition 6.8 satisfy Assumption 2.19 with*

$$\widehat{C}_V = \max\{C_{\Omega,\Omega_h}, C_{\Gamma,\Gamma_h}\},$$

where C_{Ω,Ω_h} and C_{Γ,Γ_h} are given in Lemma C.2.

Proof. This follows directly by Lemma C.2, and Lemma C.5 a) with $k = 1$. □

In the following, we bound the different error terms arising in the abstract error results of Sections 2.2.2, 5.2 and 5.3. To do so, we first state the required regularity of the exact solution.

Assumption 6.10.

a) Let $T > 0$, $\alpha_\Omega \in H^p(\overline{\Omega})$ and $\beta_\Omega \in H^p(\overline{\Omega})^d$. For the inhomogeneities and the nonlinear damping terms we assume the additional regularity

$$f_\Omega \in L^\infty\big([0,T]; H^{\max\{2,p\}}(\Omega)\big), \quad f_\Gamma \in L^\infty\big([0,T]; H^{\max\{2,p\}}(\Gamma)\big), \tag{6.12a}$$

$$\mathcal{D}_\Omega \in C^{\max\{2,p\}}(\widehat{\Omega} \times \mathbb{R}; \mathbb{R}), \quad \mathcal{D}_\Gamma \in C^{\max\{2,p\}}(\widehat{\Gamma} \times \mathbb{R}; \mathbb{R}), \tag{6.12b}$$

where $\widehat{\Omega}$ and $\widehat{\Gamma}$ are defined as prior to (6.9). Furthermore, we assume that the strong solution u of (6.1) exists on $[0,T]$ and satisfies

$$u, u' \in L^\infty\big([0,T]; H^{p+1}(\Omega;\Gamma)\big), \qquad u'' \in L^\infty\big([0,T]; H^{\max\{2,p\}}(\Omega;\Gamma)\big), \tag{6.12c}$$

$$F_\Omega(\cdot, u(t)) \in L^\infty\big([0,T]; H^{\max\{2,p\}}(\Omega)\big), \qquad F_\Gamma(\cdot, u(t)) \in L^\infty\big([0,T]; H^{\max\{2,p\}}(\Gamma)\big), \tag{6.12d}$$

$$\mathcal{D}_\Omega(\cdot, u'(t)) \in L^\infty\big([0,T]; H^{\max\{2,p\}}(\Omega)\big), \qquad \mathcal{D}_\Gamma(\cdot, u'(t)) \in L^\infty\big([0,T]; H^{\max\{2,p\}}(\Gamma)\big). \tag{6.12e}$$

We then set

$$\rho := 2\max\{C_{\Omega,\Omega_h}, C_{\Gamma,\Gamma_h}\} \|u\|_{L^\infty([0,T];H^1(\Omega;\Gamma))} \tag{6.13}$$

with C_{Ω,Ω_h} and C_{Γ,Γ_h} given in Lemma C.2.

b) Let the discrete initial values satisfy

$$\|u_h^0 - I_h u^0\|_{H^1(\Omega;\Gamma)} + \|v_h^0 - I_h v^0\|_{L^2(\Omega)\times L^2(\Gamma)} \le C_{\mathrm{iv}} h^p$$

with a constant C_{iv} independent of h.

Lemma 6.11. *Let Assumption 6.10 be satisfied and u be the strong solution of (6.1) on $[0,T]$. Then, for the space discretization with isoparametric finite elements of order p, the error terms defined in (2.40) satisfy $E_{h,i} \le Ch^p, i = 1,\dots,5$, where the constant $C = C(u)$ is independent of h.*

Proof. The terms $E_{h,i}$ for $i = 1,4,5$ arise already in the linear case and were bounded in [Hipp, 2017, Theorem 7.4] under Assumption 6.10 by

$$E_{h,i} \le Ch^p.$$

The discretization error of the Lipschitz continuous nonlinearity $E_{h,3}$ was bounded in [Hochbruck and Leibold, 2020, proof of Theorem 2.7]. But since this was based on the additional regularity assumption $u \in L^\infty\big([0,T]; H^4(\Omega;\Gamma)\big)$, we recall the proof here and show that this assumption is not necessary, if instead (6.12d) is satisfied.

Since most of the bounds in Assumption 6.10 only hold true for almost all $t \in [0,T]$, we keep in mind that the following calculations are only valid for almost all $t \in [0,T]$, but this is sufficient for our purpose. By the local Lipschitz continuity of F_h we have

$$\begin{aligned} E_{h,3} &= \|\mathcal{L}_h^{H*} F(\cdot,u) - F_h(\cdot, \mathcal{L}_h^{V*} u)\|_{L^\infty([0,T];H_h)} \\ &\le \|\mathcal{L}_h^{H*} F(\cdot,u) - F_h(\cdot, I_h u)\|_{L^\infty([0,T];H_h)} + \|F_h(\cdot, I_h u) - F_h(\cdot, \mathcal{L}_h^{V*} u)\|_{L^\infty([0,T];H_h)} \\ &\le \|\mathcal{L}_h^{H*} F(\cdot,u) - F_h(\cdot, I_h u)\|_{L^\infty([0,T];H_h)} + \widehat{L}_\rho \|(I_h - \mathcal{L}_h^{V*})u\|_{L^\infty([0,T];V_h)}. \end{aligned} \tag{6.14}$$

In the following, let $t \in [0,T]$ and we use the short notation $u = u(t)$. We start by bounding the second summand in (6.14). We have with (2.38)

$$\begin{aligned}\|(I_h - \mathcal{L}_h^{V*})u\|_{\tilde{a}_h} &= \max_{\|w_h\|_{\tilde{a}_h}=1} \big(\tilde{a}_h\big(I_h u, w_h\big) - \tilde{a}\big(u, \mathcal{L}_h^V w_h\big)\big) \\ &\le \widehat{C}_V \|\big(\mathrm{I} - \mathcal{L}_h^V I_h\big)\, u\|_{\tilde{a}} + \max_{\|w_h\|_{\tilde{a}_h}=1} \big|\Delta\tilde{a}\big(I_h u, w_h\big)\big| \\ &\le C(E_{h,4} + E_{h,5}) \\ &\le Ch^p.\end{aligned}$$

To bound the first summand in (6.14), we derive using the definition of F and F_h

$$\begin{aligned}\|\mathcal{L}_h^{H*}F(u) - F_h(I_h u)\|_{m_h} &= \sup_{\|w_h\|_{m_h}=1} m_h\big(\mathcal{L}_h^{H*}F(u) - F_h(I_h u), w_h\big) \\ &= \sup_{\|w_h\|_{m_h}=1} \Big(m\big(F(u), \mathcal{L}_h^V w_h\big) - m_h\big(F_h(I_h u), w_h\big)\Big) \\ &= \sup_{\|w_h\|_{m_h}=1} \bigg(\int_\Omega F_\Omega(\mathbf{x}, u(\mathbf{x}))w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\Omega_h} I_{h,\Omega}F_\Omega(\cdot, (I_{h,\Omega}u)^\ell)(\mathbf{x})w_h(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\qquad + \int_\Gamma F_\Gamma(\mathbf{x}, u(\mathbf{x}))w_h^\ell(\mathbf{x})\,\mathrm{d}s - \int_{\Gamma_h} I_{h,\Gamma}F_\Gamma(\cdot, (I_{h,\Gamma}u)^\ell)(\mathbf{x})w_h(\mathbf{x})\,\mathrm{d}s\bigg).\end{aligned}$$

Let $w_h \in V_h$ with $\|w_h\|_{m_h} = 1$. For the error in Ω we obtain

$$\begin{aligned}&\int_\Omega F_\Omega(\mathbf{x}, u(\mathbf{x}))w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\Omega_h} I_{h,\Omega}F_\Omega(\cdot, (I_{h,\Omega}u)^\ell)(\mathbf{x})w_h(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\quad= \int_\Omega F_\Omega(\mathbf{x}, u(\mathbf{x}))w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\Omega_h} I_{h,\Omega}F_\Omega(\cdot, u)(\mathbf{x})w_h(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\quad= \int_\Omega F_\Omega(\mathbf{x}, u(\mathbf{x}))w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_\Omega \left(I_{h,\Omega}F_\Omega(\cdot, u)\right)^\ell(\mathbf{x})w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\qquad+ \int_\Omega \left(I_{h,\Omega}F_\Omega(\cdot, u)\right)^\ell(\mathbf{x})w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\Omega_h} I_{h,\Omega}F_\Omega(\cdot, u)(\mathbf{x})w_h(\mathbf{x})\,\mathrm{d}\mathbf{x},\end{aligned} \tag{6.15}$$

where we used the definition of the nodal interpolation in the first step: The inner interpolation and the lift can be omitted, since the outer interpolation only depends on the function values at the interpolation nodes which are invariant under the inner interpolation and the lift.

The first term on the right hand side of (6.15) can be bounded by

$$\begin{aligned}&\int_\Omega F_\Omega(\mathbf{x}, u(\mathbf{x}))w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_\Omega (I_{h,\Omega}F_\Omega(\cdot, u))^\ell(\mathbf{x})w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\qquad\qquad \le \Big\|F_\Omega(\cdot, u) - (I_{h,\Omega}F_\Omega(\cdot, u))^\ell\Big\|_{L^2(\Omega)}\|w_h^\ell\|_{L^2(\Omega)} \\ &\qquad\qquad \le C_{\Omega,\Omega_h}Ch^p\big\|F_\Omega(\cdot, u)\big\|_{H^{\max\{2,p\}}(\Omega)},\end{aligned}$$

where we used (C.5), Lemma C.2, and $\|w_h\|_{L^2(\Omega)} \le \|w_h\|_{m_h} = 1$. Since $I_{h,\Omega}F_\Omega\big(\cdot, u\big) \in V_{h,p}^\Omega$ and $I_{h,\Omega} \in \mathcal{L}(H^2(\Omega); L^2(\Omega_h))$, we can bound the second term in (6.15) using (C.4a) by

$$\begin{aligned}&\int_\Omega \left(I_{h,\Omega}F_\Omega(\cdot, u)\right)^\ell(\mathbf{x})w_h^\ell(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\Omega_h} I_{h,\Omega}F_\Omega(\cdot, u)(\mathbf{x})w_h(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\qquad\qquad \le Ch^p\big\|I_{h,\Omega}F_\Omega(\cdot, u)\big\|_{L^2(\Omega_h)}\|w_h\|_{L^2(\Omega_h)} \\ &\qquad\qquad \le Ch^p\big\|F_\Omega(\cdot, u)\big\|_{H^2(\Omega)}.\end{aligned}$$

The error term on Γ can be bounded analogously and we obtain

$$E_{h,3} \leq Ch^p.$$

It remains to bound the discretization error of the nonlinear damping term

$$E_{h,2} = \|\mathcal{L}_h^{H*}\mathcal{D}(u') - \mathcal{D}_h(I_h u')\|_{L^\infty([0,T];H_h)}.$$

Therefore, let $t \in [0,T]$ and we denote $v = u'(t)$. We then have, due to (6.6c) and (6.10c),

$$\begin{aligned}
&\|\mathcal{L}_h^{H*}\mathcal{D}(v) - \mathcal{D}_h(I_h v)\|_{m_h} \\
&\quad= \sup_{\|w_h\|_{m_h}=1} \Big(m\big(\mathcal{D}(v), \mathcal{L}_h^V w_h\big) - m_h\big(\mathcal{D}_h(I_h v), w_h\big)\Big) \\
&\quad= \sup_{\|w_h\|_{m_h}=1} \bigg(\int_\Omega (\alpha_\Omega + \beta_\Omega \cdot \nabla)\, v w_h^\ell \,\mathrm{d}\mathbf{x} - \int_{\Omega_h} ((I_{h,\Omega}\alpha_\Omega) I_{h,\Omega} v + (I_{h,\Omega}\beta_\Omega) \cdot \nabla I_{h,\Omega} v)\, w_h \,\mathrm{d}\mathbf{x} \\
&\qquad + \int_\Omega \mathcal{D}_\Omega(\cdot, v) w_h^\ell \,\mathrm{d}\mathbf{x} - \sum_{\Omega_h} \mathcal{D}_\Omega(\cdot, I_{h,\Omega} v) w_h \,\mathrm{d}\mathbf{x} + \int_\Gamma \mathcal{D}_\Gamma(\cdot, v) w_h^\ell \,\mathrm{d}s - \sum_{\Gamma_h} \mathcal{D}_\Gamma(\cdot, I_{h,\Gamma} v) w_h \,\mathrm{d}s \bigg)
\end{aligned}$$

We again bound only the error terms in Ω; the surface error terms can be bounded analogously. We start with the linear term. In [Hipp, 2017, Theorem 7.4] it was proven that for all $w_h \in V_h$ with $\|w_h\|_{m_h} = 1$ we have under Assumption 6.10

$$\begin{aligned}
&\int_\Omega (\alpha_\Omega + \beta_\Omega \cdot \nabla)\, v w_h^\ell \,\mathrm{d}\mathbf{x} - \int_{\Omega_h} ((I_{h,\Omega}\alpha_\Omega) I_{h,\Omega} v + (I_{h,\Omega}\beta_\Omega) \cdot \nabla I_{h,\Omega} v)\, w_h \,\mathrm{d}\mathbf{x} \\
&\qquad\qquad \leq C(h^p + \|\alpha_\Omega - (I_{h,\Omega}\alpha_\Omega)^\ell\|_{L^\infty(\Omega)} + \|\beta_\Omega - (I_{h,\Omega}\beta_\Omega)^\ell\|_{L^\infty(\Omega)^d}).
\end{aligned}$$

We can bound the whole term by $\mathcal{O}(h^p)$ by using the L^∞ interpolation result (C.7a) with $k + 1 = \max\{2, p\}$.

To bound the nonlinear damping term, we have to consider

$$\begin{aligned}
\int_\Omega \mathcal{D}_\Omega(\cdot, v) w_h^\ell \,\mathrm{d}\mathbf{x} - \sum_{\Omega_h} \mathcal{D}_\Omega(\cdot, I_{h,\Omega} v) w_h \,\mathrm{d}\mathbf{x} &= \int_\Omega \mathcal{D}_\Omega(\cdot, v) w_h^\ell \,\mathrm{d}\mathbf{x} - \int_\Omega (I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v))^\ell\, w_h^\ell \,\mathrm{d}\mathbf{x} \\
&\quad + \int_\Omega (I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v))^\ell\, w_h^\ell \,\mathrm{d}\mathbf{x} - \int_{\Omega_h} (I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v))\, w_h \,\mathrm{d}\mathbf{x} \\
&\quad + \int_{\Omega_h} (I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v))\, w_h \,\mathrm{d}\mathbf{x} - \sum_{\Omega_h} \mathcal{D}_\Omega(\cdot, I_{h,\Omega} v) w_h \,\mathrm{d}\mathbf{x}
\end{aligned} \tag{6.16}$$

for all $w_h \in V_h$ with $\|w_h\|_{m_h} = 1$. We bound the three terms on the right hand side of (6.16) separately. The first term is an interpolation error and can be bounded by

$$\begin{aligned}
\int_\Omega \mathcal{D}_\Omega(\cdot, v) w_h^\ell \,\mathrm{d}\mathbf{x} - \int_\Omega (I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v))^\ell\, w_h^\ell \,\mathrm{d}\mathbf{x} &\leq \Big\|\mathcal{D}_\Omega(\cdot, v) - (I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v))^\ell\Big\|_{L^2(\Omega)} \|w_h^\ell\|_{L^2(\Omega)} \\
&\leq Ch^p,
\end{aligned}$$

where we used $\mathcal{D}_\Omega(\cdot, v) \in H^{\max\{2,p\}}(\Omega)$, the interpolation bound (C.5a), and the continuity of the lift. The second term in (6.16) is a geometric error that, by (C.4a) and the continuity of the lift, can be bounded via

$$\int_\Omega (I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v))^\ell\, w_h^\ell \,\mathrm{d}\mathbf{x} - \int_{\Omega_h} I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v) w_h \,\mathrm{d}\mathbf{x} \leq Ch^p \|I_{h,\Omega}\mathcal{D}_\Omega(\cdot, v)\|_{L^2(\Omega_h)} \leq Ch^p \|\mathcal{D}_\Omega(\cdot, v)\|_{H^2(\Omega)}.$$

For the last term, we use that $I_{h,\Omega}\mathcal{D}_\Omega(\cdot,v) = I_{h,\Omega}\left(\mathcal{D}_\Omega(\cdot,I_{h,\Omega}v)^\ell\right) \in V_{h,p}^\Omega$, since the inner interpolation and the lift do not change the values in the interpolation points. Furthermore, we obtain by the discrete Hölder inequality

$$\begin{aligned}
\int_{\Omega_h} \big(I_{h,\Omega}\mathcal{D}_\Omega(\cdot,v)\big) w_h \,\mathrm{d}\mathbf{x} &- \sum_{\Omega_h} \mathcal{D}_\Omega(\cdot,I_{h,\Omega}v) w_h \,\mathrm{d}\mathbf{x} \\
&= \sum_{\Omega_h} \left(I_{h,\Omega}\left(\mathcal{D}_\Omega(\cdot,I_{h,\Omega}v)^\ell\right)\right) w_h \,\mathrm{d}\mathbf{x} - \sum_{\Omega_h} \mathcal{D}_\Omega(\cdot,I_{h,\Omega}v) w_h \,\mathrm{d}\mathbf{x} \\
&\le \left(\sum_{\Omega_h} \left(I_{h,\Omega}\left(\mathcal{D}_\Omega(\cdot,I_{h,\Omega}v)^\ell\right) - \mathcal{D}_\Omega(\cdot,I_{h,\Omega}v)\right)^2 \mathrm{d}\mathbf{x}\right)^{\frac{1}{2}} \left(\sum_{\Omega_h} w_h^2 \,\mathrm{d}\mathbf{x}\right)^{\frac{1}{2}} \\
&\le \sigma(\Omega_h) \Big\| I_{h,\Omega}\big(\mathcal{D}_\Omega(\cdot,I_{h,\Omega}v)^\ell\big) - \mathcal{D}_\Omega(\cdot,I_{h,\Omega}v) \Big\|_{L^\infty(\Omega_h)} \|w_h\|_{L^2(\Omega_h)} \\
&\le Ch^p \sum_{K\in\mathcal{T}_h} \|\mathcal{D}_\Omega(\cdot,I_{h,\Omega}v)\|_{H^{\max\{2,p\}}(K)},
\end{aligned}$$

where $\sigma(\Omega_h)$ denotes the measure of Ω_h, and we additionally used that the order of the quadrature formula is greater than $2p$ and the L^∞ interpolation error bound (C.7a).

It remains to bound $\|\mathcal{D}_\Omega(\cdot,I_{h,\Omega}v)\|_{H^{\max\{2,p\}}(K)}$ for all $K \in \mathcal{T}_h$. Since $\mathcal{D}_\Omega$ is sufficiently smooth (cf. (6.9)), we have that $\mathcal{D}_\Omega$ and all of its partial derivatives up to order $\max\{2,p\}$ are bounded on bounded sets. Hence, we obtain by the chain and the product rule

$$\|\mathcal{D}_\Omega(\cdot,I_{h,\Omega}v)\|_{H^{\max\{2,p\}}(K)} \le C\big(\|I_{h,\Omega}v\|_{L^\infty(K)}\big)\|I_{h,\Omega}v\|_{H^{\max\{2,p\}}(K)},$$

where the constant $C\big(\|I_{h,\Omega}v\|_{L^\infty(K)}\big)$ depends on $\mathcal{D}_\Omega$ and its derivatives, and $\|I_{h,\Omega}v\|_{L^\infty(K)}$. The constant is bounded, since we have by the continuity of the interpolation w.r.t. the L^∞ norm (cf. Lemma C.4) and the Sobolev embedding theorem (Theorem A.4)

$$\|I_{h,\Omega}v\|_{L^\infty(K)} \le C\|v\|_{L^\infty(K)} \le C\|v\|_{H^2(K)}.$$

Finally we have by Lemma C.2 and (C.6a)

$$\begin{aligned}
\|I_{h,\Omega}v\|_{H^{\max\{2,p\}}(K)} &\le C\|(I_{h,\Omega}v)^\ell\|_{H^{\max\{2,p\}}(K^\ell)} \\
&\le C\left(\|(I_{h,\Omega}v)^\ell - v\|_{H^{\max\{2,p\}}(K^\ell)} + \|v\|_{H^{\max\{2,p\}}(K^\ell)}\right) \\
&\le C\|v\|_{H^{\max\{2,p\}}(K^\ell)}.
\end{aligned}$$

This concludes the proof. □

The space discretization error bound of the finite element space discretization follows now directly from Corollary 2.25.

Theorem 6.12. *Let Assumption 6.2 be satisfied and u be the solution of (6.1) on $[0,T]$. Further, let Assumption 6.10 be satisfied and let the condition (6.11) hold true for all h sufficiently small. Then, there exists $h^* > 0$ s.t. for all $h < h^*$ the spatial approximation u_h of u, obtained with the bulk-surface finite element method of order p, exists on $[0,T]$ and satisfies the error bound*

$$\|u_h^\ell(t) - u(t)\|_{H^1(\Omega;\Gamma)} + \|(u_h')^\ell(t) - u'(t)\|_{L^2(\Omega)\times L^2(\Gamma)} \le C\mathrm{e}^{(\widehat{L}_\rho + \frac{1}{2} + c')t}(1+t)h^p$$

with $\widehat{L}_\rho$ from Lemma 6.6, ρ from (6.13), and a constant C independent of h.

Proof. We have by Lemmas 6.4, 6.6 and 6.9 that all assumptions of Corollary 2.25 are satisfied with $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}} = \frac{1}{2} + c'$. The assertion follows then directly from Corollary 2.25, since we have shown in Lemma 6.11 that all error terms appearing in the abstract error bound (2.39) are bounded by $\mathcal{O}(h^p)$. □

6.4 Time and full discretization error bounds

With all the results obtained in Chapter 6 so far, it is now straightforward to prove time- and full-discretization error bounds for the wave equation with kinetic boundary conditions, using the abstract results from Chapters 3 to 5.

We first consider the time discretization with Runge–Kutta schemes and afterwards the IMEX time discretization in the semilinear case.

Theorem 6.13. *Let Assumption 6.2 be satisfied and for some $q \in \mathbb{N}$*

$$u \in C^{q+2}([0,T]; L^2(\Omega) \times L^2(\Gamma)) \cap C^{q+1}([0,T]; H^1(\Omega;\Gamma))$$

be the solution of (6.1). We consider an algebraically stable and coercive Runge–Kutta method of stage order q. By α_{RK} we denote the coercivity constant of the Runge–Kutta method(cf. Definition B.5).

a) There exists $\tau^* > 0$ s.t. for all $\tau < \tau^*$ the Runge–Kutta method applied to the wave equation with kinetic boundary conditions (6.1) yields for all $t_n = n\tau \le T$ unique approximations u^n, v^n. Further, the error bound

$$\|u^n - u(t_n)\|_{H^1(\Omega;\Gamma)} + \|v^n - u'(t_n)\|_{L^2(\Omega)\times L^2(\Gamma)} \le C \frac{\mathrm{e}^{C_{\mathrm{RK}} C_{\tau,\rho}^2 (L_\rho + \frac{1}{2} + c') t_n} - 1}{C_{\mathrm{RK}}(L_\rho + \frac{1}{2} + c')} \tau^q$$

is satisfied with the Lipschitz constant L_ρ from Lemma 6.4,

$$\rho = \|u\|_{L^\infty([0,T];H^1(\Omega;\Gamma))},$$

a constant C_{RK} only depending on the coefficients of the Runge–Kutta method, a constant C which depends on u, T and the Runge–Kutta method, but is independent of τ, and the constant

$$C_{\tau,\rho} = \big(\alpha_{\mathrm{RK}} - \tau(L_\rho + \frac{1}{2} + c')\big)^{-1}.$$

b) Let additionally Assumption 6.10 be satisfied and let (6.11) hold true for all h sufficiently small. Then, there exist $\tau^*, h^* > 0$ s.t. for all $\tau < \tau^*$ and $h < h^*$ the fully discrete approximations u_h^n, v_h^n obtained by the bulk-surface finite element method of order p and the Runge–Kutta method exist uniquely for all $t_n = n\tau \le T$ and satisfy the error bound

$$\|(u_h^n)^\ell - u(t_n)\|_{H^1(\Omega;\Gamma)} + \|(v_h^n)^\ell - u'(t_n)\|_{L^2(\Omega)\times L^2(\Gamma)} \le C \frac{\mathrm{e}^{C_{\mathrm{RK}} C_{\tau,\rho}^2 (\widehat{L}_\rho + \frac{1}{2} + c') t_n} - 1}{C_{\mathrm{RK}}(\widehat{L}_\rho + \frac{1}{2} + c')} (\tau^q + h^p)$$

with $\widehat{L}_\rho$ from Lemma 6.6 and ρ from (6.13). The constant C_{RK} depends only on the coefficients of the Runge–Kutta method, the constant C depends on u, T and the Runge–Kutta method, but is independent of τ, and the constant $C_{\tau,\rho}$ is given by

$$C_{\tau,\rho} = \big(\alpha_{\mathrm{RK}} - \tau(\widehat{L}_\rho + \frac{1}{2} + c')\big)^{-1}.$$

Proof.

a) This is a direct application of Corollary 3.4. By Lemma 6.4, we have that all assumptions are satisfied and that $c_{\mathrm{qm}} = \frac{1}{2} c_G C_{H,V} + \beta_{\mathrm{qm}} = \frac{1}{2} + c'$.

b) By Lemmas 6.4, 6.6 and 6.9, we have that all assumptions of Corollary 5.4 are satisfied and that $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}} = \frac{1}{2} + c'$. Corollary 5.4 directly implies the assertion, since we bounded the space discretization error terms appearing in (5.9) in Lemma 6.11 by $\mathcal{O}(h^p)$.

□

Theorem 6.14. *Let Assumption 6.2 be satisfied and*

$$u \in C^4([0,T]; L^2(\Omega) \times L^2(\Gamma)) \cap C^3([0,T]; H^1(\Omega;\Gamma)) \cap C^2([0,T]; H^2(\Omega;\Gamma)) \tag{6.17}$$

be the solution of (6.1) in the semilinear case, i.e., with $\mathcal{D}_\Omega = \mathcal{D}_\Gamma = 0$.

a) There exists $\tau^* > 0$ s.t. for all $\tau < \tau^*$ the iterations u^n, v^n of the IMEX scheme (4.16) applied to the wave equation with kinetic boundary conditions (6.1) satisfy for all $t_n \le T$ the error bound

$$\|u^n - u(t_n)\|_{H^1(\Omega;\Gamma)} + \|v^n - u'(t_n)\|_{L^2(\Omega)\times L^2(\Gamma)} \le C\mathrm{e}^{Mt_n}\tau^2$$

with $M = \frac{1}{2} + \frac{\left(1+(3/2)^{1/2}\right)L_\rho}{1-\left(1+(3/2)^{1/2}\right)L_\rho\tau}$, L_ρ from Lemma 6.4,

$$\rho = \|u\|_{L^\infty([0,T];H^1(\Omega;\Gamma))},$$

and a constant C that only depends on T and u but is independent of τ.

b) Let additionally Assumption 6.10 be satisfied and let (6.11) hold true for all h sufficiently small. Then, there exist $\tau^*, h^* > 0$, s.t. for all $\tau < \tau^*$ and $h < h^*$ the fully discrete approximations u_h^n, v_h^n obtained by the bulk-surface finite element method of order p and the IMEX scheme satisfy for all $t_n = n\tau \le T$ the error bound

$$\|(u_h^n)^\ell - u(t_n)\|_{H^1(\Omega;\Gamma)} + \|(v^n)^\ell - u'(t_n)\|_{L^2(\Omega)\times L^2(\Gamma)} \le \; C\mathrm{e}^{\widehat{M}t_n}\left(h^p + \tau^2\right)$$

with $\widehat{M} = \frac{1}{2} + \frac{\left(1+(3/2)^{1/2}\right)\widehat{L}_{T,\rho}}{1-\left(1+(3/2)^{1/2}\right)\widehat{L}_\rho\tau}$, $\widehat{L}_\rho$ from Lemma 6.6, ρ from (6.13), and a constant C that only depends on T and u but which is independent of τ and h.

Proof. Note that we have $D(A) = H^2(\Omega;\Gamma)$, and, hence, by (6.17) in terms of the abstract framework

$$u \in C^4([0,T]; H) \cap C^3([0,T]; V) \cap C^2([0,T]; D(A)). \tag{6.18}$$

a) This is a direct application of Theorem 4.9. By Lemma 6.4, we have that all assumptions are satisfied and, since we are in the semilinear case, $\mathcal{D}_\Omega = \mathcal{D}_\Gamma = 0$ implies $c_{\mathrm{qm}} = \frac{1}{2} c_G C_{H,V} + \beta_{\mathrm{qm}} = \frac{1}{2}$.

b) We apply Theorem 5.8. All assumptions are satisfied by Lemmas 6.4, 6.6 and 6.9 and we have $\widehat{c}_{\mathrm{qm}} = \frac{1}{2}\widehat{c}_G \widehat{C}_{H,V} + \widehat{\beta}_{\mathrm{qm}} = \frac{1}{2}$. This directly implies the assertion, since, in Lemma 6.11 we bounded the space discretization error terms appearing in (5.9) by $\mathcal{O}(h^p)$.

□

6.5 Numerical experiments

In this section, we illustrate the theoretical results of this thesis with numerical experiments for the wave equation with kinetic boundary conditions (6.1) on the unit disc $\Omega = B_1(0) \subset \mathbb{R}^2$. Before presenting the results, we give some details about the implementation.

6.5.1 Implementation details

We used version 9.2 of the `C++` finite element library `deal.II` Arndt et al. [2020], Bangerth et al. [2007]) to implement our numerical experiments. The source code to reproduce the experiments is available on `https://doi.org/10.5445/IR/1000130223`. We ran the experiments on a computer with an i5 processor (3.5 GHz) and 16 GB RAM.

Remark 6.15. *The finite element library* `deal.II` *supports only quadrilateral and hexahedral elements. Simplicial elements, for which we prove our results, are not implemented. However, we emphasize that our theoretical results rely on the construction and approximation properties from Elliott and Ranner [2013, 2020] where only the case of simplicial elements is considered. Thus, if these results can be transfered to the quadrilateral case, then our results transfer as well, but this it out of the scope of this thesis. Note that for standard Lagrange finite elements without domain approximation, it is well-known that finite element spaces consisting of quadrilateral and hexahedral elements have the same approximation orders as the corresponding spaces consisting of simplicial elements.*

Finite element space discretization

Let $\{\phi_1, \dots, \phi_N\}$ be the nodal basis of the finite element space V_h with $\dim(V_h) = N$. For a finite element function $v_h \in V_h$ we denote the corresponding coefficient by $\mathbf{v} \in \mathbb{R}^N$. Further, we denote by $\mathbf{M} \in \mathbb{R}^{N\times N}$ the mass matrix and by $\mathbf{A} \in \mathbb{R}^{N\times N}$ the stiffness matrix corresponding to the bilinear forms m_h, a_h, respectively. This means

$$\mathbf{M}_{i,j} = m_h\big(\phi_i, \phi_j\big), \qquad \mathbf{A}_{i,j} = a_h\big(\phi_i, \phi_j\big), \qquad i, j \in \{1, \dots, N\}.$$

The representations of the nonlinearities $\mathcal{D}_h, F_h$, and the inhomogeneity f_h with respect to the nodal basis are denoted by $\mathbf{D}, \mathbf{F} \colon \mathbb{R}^N \to \mathbb{R}^N$, and $\mathbf{f} \colon [0, \infty) \to \mathbb{R}^N$, respectively. The spatially discretized wave equation with kinetic boundary conditions (2.24) is then equivalent to the ordinary differential equation

$$\mathbf{M}\mathbf{u}'' + \mathbf{D}(\mathbf{u}') + \mathbf{A}\mathbf{u} = \mathbf{F}(\mathbf{u}) + \mathbf{f}, \qquad t \geq 0, \tag{6.19a}$$

$$\mathbf{u}(0) = \mathbf{u^0}, \qquad \mathbf{u}'(0) = \mathbf{v^0}. \tag{6.19b}$$

In our implementation, we discretized the initial values via $u_h^0 = I_h u^0, v_h^0 = I_h v_h^0$ and by $\mathbf{u^0}, \mathbf{v^0}$ we denote the corresponding coefficient vectors. In this case, Assumption 6.10 b) is satisfied.

We shortly comment on how we compute the discretization of the nonlinearities (6.10c) and (6.10d). Let $\phi_1^\Gamma, \dots, \phi_{N_\Gamma}^\Gamma$ be the nodal basis functions of the boundary finite element space $V_{h,p}^\Gamma$. Further, let $\mathbf{M}_\Omega \in \mathbb{R}^{N\times N}$ be the mass matrix corresponding to the $L^2(\Omega)$ scalar product on V_h and $\mathbf{M}_\Gamma \in \mathbb{R}^{N_\Gamma \times N_\Gamma}$

be the mass matrix corresponding to the $L^2(\Gamma)$ scalar product on $V_{h,p}^{\Gamma}$. Then, for $\mathbf{v} \in \mathbb{R}^N$, we compute $\mathbf{F}(\mathbf{v})$ by

$$\mathbf{F}(\mathbf{v}) = \mathbf{M}_\Omega \mathbf{F}_\Omega(\mathbf{v}) + \mathbf{M}_\Gamma \mathbf{F}_\Gamma(\mathbf{v})$$

with

$$\mathbf{F}_\Omega(\mathbf{v})_i = F_\Omega(\mathbf{v}_i), \quad i = 1, \dots, N, \qquad \mathbf{F}_\Gamma(\mathbf{v})_j = F_\Gamma((\gamma \mathbf{v})_j), \quad j = 1, \dots, N_\Gamma.$$

Here, $\gamma \mathbf{v} \in \mathbb{R}^{N_\Gamma}$ is the coefficient vector of $v_h\big|_\Gamma$ with respect to the basis $\phi_1^\Gamma, \dots, \phi_{N_\Gamma}^\Gamma$.

The computation of $\mathcal{D}_h$ as in (6.10c) is computationally quite expensive since it requires the application of quadrature rules to $\mathcal{D}_\Omega(v_h)$ and $\mathcal{D}_\Gamma(v_h)$ and, hence, the evaluation of these functions outside the nodal basis. Therefore, we implemented the discretization of $\mathcal{D}$ similar to the discretization of F via

$$\mathbf{D}(\mathbf{v}) = \mathbf{B}\mathbf{v} + \mathbf{M}_\Omega \mathbf{D}_\Omega(\mathbf{v}) + \mathbf{M}_\Gamma \mathbf{D}_\Gamma(\mathbf{v}), \qquad \mathbf{v} \in \mathbb{R}^N, \tag{6.20}$$

where $\mathbf{D}_\Omega$ and $\mathbf{D}_\Gamma$ are defined similarly to $\mathbf{F}_\Omega$ and $\mathbf{F}_\Gamma$, respectively, and $\mathbf{B} \in \mathbb{R}^{N \times N}$ is the matrix corresponding to the bilinear form

$$b_h\big(v_h, w_h\big) = \int_{\Omega_h} \big((I_{h,\Omega}\alpha_\Omega) v_h + (I_{h,\Omega}\beta_\Omega) \cdot \nabla v_h\big)\, w_h \,\mathrm{d}\mathbf{x}.$$

Since the nonlinear part has to be evaluated in every time step, this reduces the computational effort significantly. Despite not being covered by our analysis, our numerical experiments show that using (6.20) does not decrease the order of convergence.

Note that the numerical computation of the lift of a finite element function is very laborious. Therefore, in our numerical experiments we refrain from computing the error from the error bound in Theorem 6.12, but instead consider the discrete error

$$\mathbf{E}(t) \coloneqq \|u_h(t) - u(t)\big|_{\Omega_h}\|_{H^1(\Omega_h;\Gamma_h)} + \|u_h'(t) - u'(t)\big|_{\Omega_h}\|_{L^2(\Omega_h)\times L^2(\Gamma_h)}.$$

For computing $\mathbf{E}$, we evaluate the integrals with a quadrature rule of sufficiently high order such that the quadrature error is negligible. The restriction of u to Ω_h is possible since we are running our experiments on the unit disc and, hence, have $\Omega_h \subset \Omega$ for all $h > 0$.

Runge–Kutta methods

A Runge–Kutta method with coefficients $\mathbf{b} = (b_i)_{i=1}^s, \mathbf{c} = (c_i)_{i=1}^s, \mathcal{A} = (a_{ij})_{i,j=1}^s$ applied to (6.19) is of the form

$$\begin{aligned}
&\left.\begin{aligned}
\mathbf{U}^{ni} &= \mathbf{u}^n + \tau \sum_{j=1}^s a_{ij} \mathbf{V}^{nj}, \\
\mathbf{M}\mathbf{V}^{ni} &= \mathbf{M}\mathbf{v}^n + \tau \sum_{j=1}^s a_{ij} \big(- \mathbf{D}(\mathbf{V}^{nj}) - \mathbf{A}\mathbf{U}^{nj} + \mathbf{F}(\mathbf{U}^{nj}) + \mathbf{f}(t_n + c_j\tau)\big),
\end{aligned}\right\} \quad i = 1, \dots, s, \\
&\mathbf{u}^{n+1} = \mathbf{u}^n + \tau \sum_{i=1}^s b_i \mathbf{V}^{ni}, \\
&\mathbf{M}\mathbf{v}^{n+1} = \mathbf{M}\mathbf{u}^n + \tau \sum_{i=1}^s b_i \big(- \mathbf{D}(\mathbf{V}^{ni}) - \mathbf{A}\mathbf{U}^{ni} + \mathbf{F}(\mathbf{U}^{ni}) + \mathbf{f}(t_n + c_i\tau)\big).
\end{aligned} \tag{6.21}$$

Note that for a general implicit Runge–Kutta method the solution of a nonlinear system of equations of dimension $2sN$ is required in every time step to compute the inner stages in (6.21). We implemented the implicit midpoint method, i.e., the 1-stage Gauß method, which can be written in the form

$$\begin{aligned}\mathbf{u}^{n+1} &= \mathbf{u}^n + \frac{\tau}{2}(\mathbf{v}^n + \mathbf{v}^{n+1}),\\ \mathbf{M}\mathbf{v}^{n+1} &= \mathbf{M}\mathbf{v}^n + \frac{\tau}{2}\mathbf{A}(\mathbf{u}^n + \mathbf{u}^{n+1}) - \tau\mathbf{D}\Big(\frac{1}{2}(\mathbf{v}^n + \mathbf{v}^{n+1})\Big) + \tau\mathbf{F}\Big(\frac{1}{2}(\mathbf{u}^n + \mathbf{u}^{n+1})\Big) + \mathbf{f}\Big(t_n + \frac{\tau}{2}\Big).\end{aligned}$$

This method is algebraical stable, coercive, and of stage order $q = 1$. Note that by plugging the first equation into the second one, we only have to solve a nonlinear equation of dimension N for $\mathbf{v}^{n+1}$ given by

$$\begin{aligned}\Big(\mathbf{M} + \frac{\tau^2}{4}\mathbf{A}\Big)\mathbf{v}^{n+1} + \tau\mathbf{D}\Big(\frac{1}{2}(\mathbf{v}^n + \mathbf{v}^{n+1})\Big) - \tau\mathbf{F}\Big(\mathbf{u}^n + \frac{\tau}{4}(\mathbf{v}^{n+1} + \mathbf{v}^n)\Big)&\\ = \Big(\mathbf{M} - \frac{\tau^2}{4}\mathbf{A}\Big)\mathbf{v}^n - \tau\mathbf{A}\mathbf{u}^n + \tau\mathbf{f}\Big(t_n + \frac{\tau}{2}\Big).&\end{aligned}$$

This equation is solved with the simplified Newton method where we choose the tolerances such that the total error is not affected. Then, $\mathbf{u}^{n+1}$ can be computed explicitly.

For comparison, we also implemented the classical Runge–Kutta scheme. This is an explicit scheme of order four that is suited for hyperbolic problems because its stability region contains an interval on the imaginary axis. The scheme is given by the coefficients

$$\mathbf{b} = \begin{bmatrix}\frac{1}{6}\\ \frac{1}{3}\\ \frac{1}{3}\\ \frac{1}{6}\end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix}0\\ \frac{1}{2}\\ \frac{1}{2}\\ 0\end{bmatrix}, \qquad \mathcal{Q} = \begin{bmatrix}0 & & & \\ \frac{1}{2} & 0 & & \\ 0 & \frac{1}{2} & 0 & \\ 0 & 0 & 1 & 0\end{bmatrix}.$$

We implemented it using mass lumping (cf. [Zienkiewicz et al., 2013, Section 12.2.4]) to obtain a fully explicit scheme.

IMEX and Crank–Nicolson scheme in the semilinear case

In the following, we discuss the implementation of the IMEX and the Crank–Nicolson scheme in the semilinear case, i.e., for $\mathcal{D}_\Omega = \mathcal{D}_\Gamma = 0$. Due to (6.20), we see that in this case the ordinary differential equation (6.19a) reduces to

$$\mathbf{M}\mathbf{u}'' + \mathbf{B}\mathbf{u}' + \mathbf{A}\mathbf{u} = \mathbf{F}(\mathbf{u}) + \mathbf{f}, \qquad t \geq 0. \tag{6.22}$$

This was already presented in [Hochbruck and Leibold, 2021, Section 4.4].

IMEX scheme The fully discrete IMEX scheme (5.11) applied to the semilinear equation (6.22) reads

$$\mathbf{M}\mathbf{v}^{n+\frac{1}{2}} = \mathbf{M}\mathbf{v}^n - \frac{\tau}{2}\mathbf{A}\mathbf{u}^n - \frac{\tau^2}{4}\mathbf{A}\mathbf{v}^{n+\frac{1}{2}} - \frac{\tau}{2}\mathbf{B}\mathbf{v}^{n+\frac{1}{2}} + \frac{\tau}{2}\mathbf{F}^n, \tag{6.23a}$$

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \tau\mathbf{v}^{n+\frac{1}{2}}, \tag{6.23b}$$

$$\mathbf{M}\mathbf{v}^{n+1} = -\mathbf{M}\mathbf{v}^n + 2\mathbf{M}\mathbf{v}^{n+\frac{1}{2}} + \frac{\tau}{2}\left(\mathbf{F}^{n+1} - \mathbf{F}^n\right) \tag{6.23c}$$

with

$$\mathbf{F}^n = \mathbf{F}(\mathbf{u}^n) + \mathbf{f}(t_n).$$

The linear system in (6.23a) is of the form

$$\mathbf{Q}_+ \mathbf{v}^{n+\frac{1}{2}} = \mathbf{M}\mathbf{v}^n - \frac{\tau}{2}\mathbf{A}\mathbf{u}^n + \frac{\tau}{2}\mathbf{F}^n, \qquad \mathbf{Q}_+ = \mathbf{M} + \frac{\tau}{2}\mathbf{B} + \frac{\tau^2}{4}\mathbf{A}. \tag{6.24}$$

Since we perform runtime comparisons between the IMEX and the Crank–Nicolson scheme to compare the efficiency, we go into more detail about the implementation. We solve the linear system (6.24) with the GMRES solver provided by `deal.II` and either a sparse incomplete LU or a geometric multigrid preconditioner. For the measurement of the error in the GMRES iterations, the residual r with corresponding coefficient vector $\mathbf{r}$ is used. A suitable stopping criteria would be

$$\|r\|_{\tilde{a}_h} \leq \tau^2 \,\mathrm{tol},$$

where tol is a given tolerance. Then, in (6.23b) the error in $\mathbf{u}^{n+1}$ caused by the solution of the linear system measured in $\|\cdot\|_{\tilde{a}_h}$ is of order τ^3 which corresponds to the local error of the IMEX scheme. However, the computation of $\|r\|_{\tilde{a}_h}$ is rather expensive. Thus, we use the stopping criterion

$$\|r\|_{h,2} = \|\mathbf{r}\|_{h,2} \leq \tau^2 \,\mathrm{tol}$$

in the grid dependent scaled Euclidean norm $\|\cdot\|_{h,2} = h^{d/2}\|\cdot\|_2$. This is significantly more efficient since $\|\mathbf{r}\|_{h,2}$ is available within the GMRES algorithm at no additional cost. The criterion worked well in our numerical experiments as we show in Section 6.5.3. We always use tol $= 0.01$ in our numerical examples, which was chosen by experiment s.t. the errors caused by solving the linear systems do not affect the overall order of convergence.

Note that in the IMEX scheme (6.23) only $\mathbf{M}\mathbf{v}^{n+1}, n \geq 0$, is required so that we only compute $\mathbf{M}\mathbf{v}^{n+1}$ but not $\mathbf{v}^{n+1}$ itself.

Crank–Nicolson scheme In our experiments, we compare the IMEX scheme with the Crank–Nicolson scheme (4.4). Applied to the first-order formulation of the semilinear equation (6.22), it is of the form

$$\mathbf{M}\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n + \frac{\tau}{2}(\mathbf{M}\mathbf{v}^n + \mathbf{M}\mathbf{v}^{n+1}), \tag{6.25a}$$

$$\mathbf{M}\mathbf{v}^{n+1} = \mathbf{M}\mathbf{v}^n - \frac{\tau}{2}\mathbf{A}(\mathbf{u}^n + \mathbf{u}^{n+1}) - \frac{\tau}{2}\mathbf{B}(\mathbf{v}^n + \mathbf{v}^{n+1}) + \frac{\tau}{2}(\mathbf{F}^n + \mathbf{F}^{n+1}) \tag{6.25b}$$

with $\mathbf{F}^n = \mathbf{F}(\mathbf{u}^n) + \mathbf{f}(t_n)$. These are two coupled nonlinear equations. However, by plugging (6.25b) into (6.25a) to eliminate $\mathbf{M}\mathbf{v}^{n+1}$, and using (6.25a) to replace $\mathbf{v}^n + \mathbf{v}^{n+1}$ in (6.25b), the scheme can be written in the form

$$\mathbf{Q}_+\mathbf{u}^{n+1} - \frac{\tau^2}{4}\mathbf{F}^{n+1} = (\mathbf{M} + \frac{\tau}{2}\mathbf{B} - \frac{\tau^2}{4}\mathbf{A})\mathbf{u}^n + \tau\mathbf{M}\mathbf{v}^n + \frac{\tau^2}{4}\mathbf{F}^n, \tag{6.26a}$$

$$\mathbf{M}\mathbf{v}^{n+1} = \mathbf{M}\mathbf{v}^n - \frac{\tau}{2}\mathbf{A}(\mathbf{u}^n + \mathbf{u}^{n+1}) + \mathbf{B}(\mathbf{u}^n - \mathbf{u}^{n+1}) + \frac{\tau}{2}\big(\mathbf{F}^n + \mathbf{F}^{n+1}\big) \tag{6.26b}$$

with $\mathbf{Q}_+$ given in (6.24). The formulation (6.26) has the advantage that we only have to solve the nonlinear equation (6.26a) since (6.26b) can then be computed explicitly. We solve the nonlinear equation (6.26a) with a simplified Newton method where we approximate the Jacobian by $\mathbf{Q}_+$ and stop the Newton scheme when the update $\Delta\mathbf{u}$ satisfies $\|\Delta\mathbf{u}\|_{h,2} \leq \tau^3\widetilde{\mathrm{tol}}$ with a given tolerance $\widetilde{\mathrm{tol}}$. In the numerical examples, we use $\widetilde{\mathrm{tol}} = 0.1$ which is chosen based on experiments such that the Newton errors do not affect the overall convergence of the Crank–Nicolson scheme. The matrix vector products appearing in (6.26a) and (6.26b) are computed only once and are stored in temporary vectors. This is also done for all terms that can be reused in the next time step. As in the IMEX scheme, we only compute and store $\mathbf{M}\mathbf{v}^{n+1}$ and not $\mathbf{v}^{n+1}$.

6.5.2 Experiments for the nonlinear damped case

In this section, we consider the wave equation with kinetic boundary conditions (6.1) on the unit circle $\Omega = B_1(0) \subset \mathbb{R}^2$ with

$$\begin{aligned} \mathcal{D}_\Omega(v) &= v^3, & \mathcal{D}_\Gamma &\equiv 0, \\ F_\Omega(v) &= |v|v, & F_\Gamma(v) &= v^3, \\ f_\Omega(t,\mathbf{x}) &= -\left(4\pi^2 + |\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2|\right)\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2 + \left(2\pi\cos(2\pi t)\mathbf{x}_1\mathbf{x}_2\right)^3, \\ f_\Gamma(t,\mathbf{x}) &= -4\pi^2\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2 + 6\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2 - \left(\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2\right)^3 \end{aligned}$$

for $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]^\mathsf{T}$. Further, we set $\alpha_\Omega = \beta_\Omega = 0$ and choose the initial values

$$u(0,\mathbf{x}) = 0, \quad u_t(0,\mathbf{x}) = 2\pi\mathbf{x}_1\mathbf{x}_2.$$

In this case, the exact solution is given by

$$u(t,\mathbf{x}) = \sin(2\pi t)\mathbf{x}_1\mathbf{x}_2$$

which can be verified by a straightforward calculation using $\Delta_\Gamma(\mathbf{x}_1\mathbf{x}_2) = -4\mathbf{x}_1\mathbf{x}_2$ and $\partial_\mathbf{n}(\mathbf{x}_1\mathbf{x}_2) = 2\mathbf{x}_1\mathbf{x}_2$ on the unit circle $\partial\Omega$.

In Figure 6.1, the space discretization error is plotted against the maximal mesh width h when discretizing with isoparametric elements of orders $p = 1$ and $p = 2$. For the time discretization, we use the implicit midpoint rule with sufficiently small time step size such that the time discretization error is negligible. One can observe that the space discretization error converges with order p which is consistent with the error bound from Theorem 6.12.

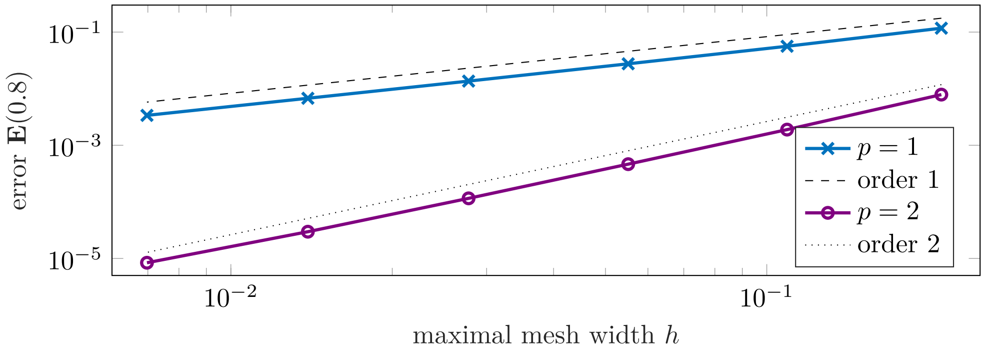

Figure 6.1: Error $\mathbf{E}(0.8)$ of the isoparametric finite element discretization with order $p = 1$ and $p = 2$ plotted against the maximal mesh width h

In Figure 6.2, the error of the implicit midpoint rule is plotted against the time step size τ. We observe that the scheme converges with order two until the plateau of the space discretization error is reached. Although this is one order larger than predicted by Theorem 6.13, as the stage order of the implicit midpoint rule is $q = 1$, this is not a contradiction, since Theorem 6.13 states a worst case error bound. However, it does raise the question whether the estimate is too pessimistic. For example, in Hochbruck et al. [2018] an error bound for Runge–Kutta methods applied to quasilinear hyperbolic evolution equations of order $q+1$ is proven. However, the framework considered in Hochbruck et al. [2018] is not suitable for nonlinear damping terms since they are not quasilinear.

As a further test, we perform the same experiment with $f_\Omega = f_\Gamma \equiv 0$. In this case, we do not know the exact solution and, therefore, test against a reference solution that we calculate with a smaller time step size and on a finer grid. The results are shown in Figure 6.3. We see that in this case the implicit midpoint rule converges with order one. This might suggest that the error bound from Theorem 6.13 is sharp. However, note that in this case we do not know whether the exact solution u satisfies the regularity assumptions of Theorem 6.13. A more detailed analysis under which conditions Runge–Kutta methods converge with order $q + 1$ or q is beyond the scope of this thesis.

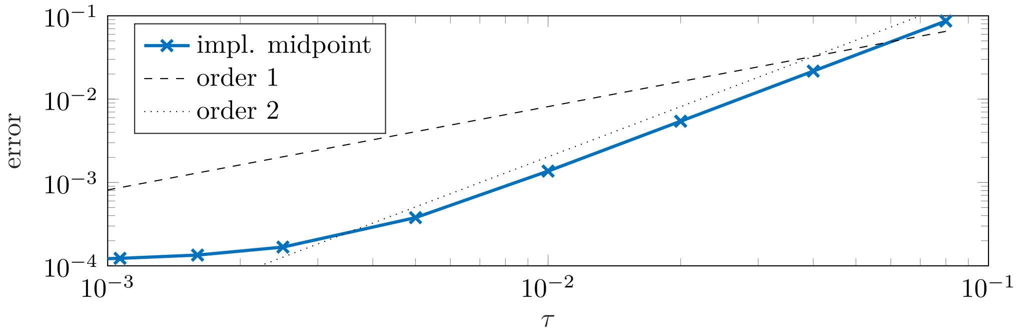

Figure 6.2: Error $\mathbf{E}(0.8)$ of the implicit midpoint rule plotted against the step size τ

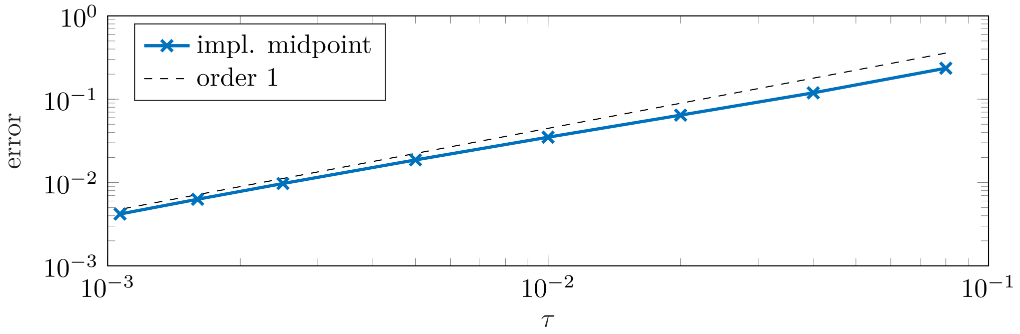

Figure 6.3: Error $\mathbf{E}(0.8)$ of the implicit midpoint rule in the case $f_\Omega = f_\Gamma \equiv 0$ plotted against the step size τ

6.5.3 Experiments for the IMEX scheme in the semilinear case

The experiments presented in this section are taken from [Hochbruck and Leibold, 2021, Section 4.5]. We consider the wave equation with kinetic boundary conditions (6.1) on the unit circle $\Omega = B_1(0) \subset \mathbb{R}^2$ in the semilinear case, i.e., with $\mathcal{D}_\Omega = \mathcal{D}_\Gamma \equiv 0$. Further, we set

$$\begin{aligned} \alpha_\Omega &\equiv 1, & \beta_\Omega(\mathbf{x}) &= \mathbf{x}, \\ F_\Omega(u) &= |v|v, & F_\Gamma(v) &= v^3, \\ f_\Omega(t,\mathbf{x}) &= -\left(4\pi^2 + |\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2|\right)\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2 + 6\pi\cos(2\pi t)\mathbf{x}_1\mathbf{x}_2, \\ f_\Gamma(t,\mathbf{x}) &= -4\pi^2\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2 + 6\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2 - \left(\sin(2\pi t)\mathbf{x}_1\mathbf{x}_2\right)^3, \end{aligned}$$

and choose the initial values

$$u(0,\mathbf{x}) = 0, \quad u_t(0,\mathbf{x}) = 2\pi\mathbf{x}_1\mathbf{x}_2.$$

As in Section 6.5.2, f_Ω and f_Γ are chosen such that the exact solution is given by

$$u(t,\mathbf{x}) = \sin(2\pi t)\mathbf{x}_1\mathbf{x}_2.$$

For the following experiments we always use isoparametric elements of order $p = 2$ for the space discretization. In Figure 6.4, the errors of the IMEX, the Crank–Nicolson, and the classical Runge–Kutta scheme are plotted against the time-step size τ for a coarse ($h \approx 0.014$) and a fine ($h \approx 0.007$) space discretization, respectively. As predicted by Theorem 6.14, the IMEX and also the Crank–Nicolson scheme converge with order two until the space discretization error is reached. The classical Runge–Kutta scheme is only stable under a strong CFL condition and, in this case, the error reaches immediately the space discretization error plateau. The Crank–Nicolson scheme is only considered for the coarse grid, since on the fine grid it is computational very expensive.

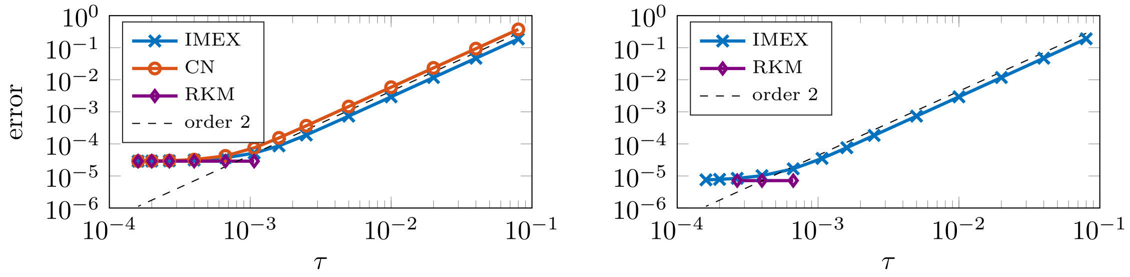

Figure 6.4: Error $\mathbf{E}(0.8)$ of the IMEX scheme, the Crank–Nicolson scheme, and the classical Runge–Kutta method plotted against step size τ for coarse space discretization (328 193 degrees of freedom, left) and fine space discretization (1 311 745 degrees of freedom, right)

We now compare the efficiency of the different schemes for our test case. In Figure 6.5, the errors of the schemes are plotted against the runtime for the same coarse and fine space discretization as in Figure 6.4. We observe that the IMEX scheme is significantly faster than the Crank–Nicolson scheme. For obtaining errors of the magnitude of the space discretization error plateau, the classical Runge–Kutta scheme is more efficient than the IMEX scheme, but the IMEX scheme is faster than the the Runge–Kutta scheme if less accuracy is sufficient. Additionally, the Runge–Kutta method has the disadvantage that the stability limit in applications is not exactly known and, hence, there is a risk that it will not be stable if the time step size is chosen too large, or the effort is unnecessarily high if the time step size is too small. Further, we see that for the fine space discretization and large time step sizes the usage of the multigrid preconditioner is quite efficient.

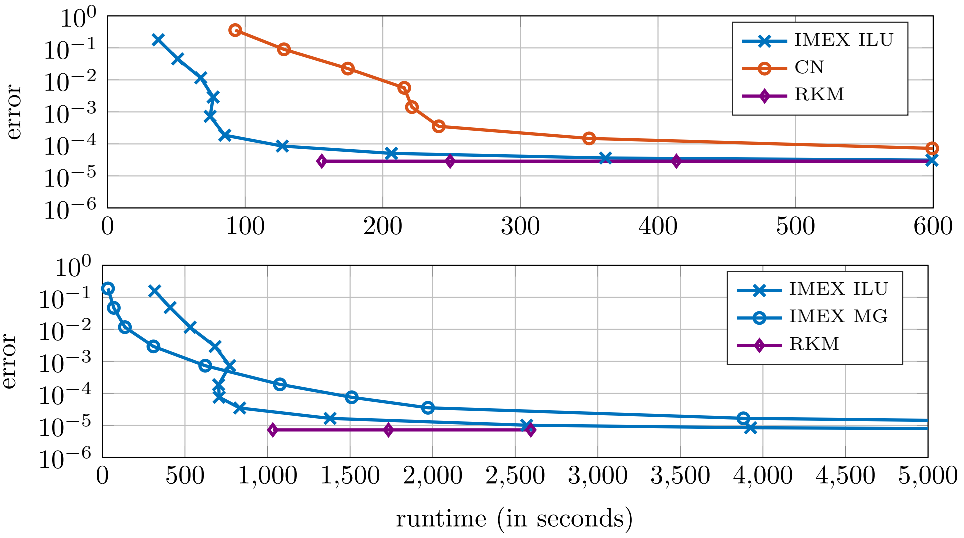

Figure 6.5: Error $\mathbf{E}(0.8)$ of the IMEX scheme, solved with GMRES and ILU/multigrid(MG, F-cycle with 8 levels) preconditioner, the Crank–Nicolson scheme, and the classical Runge–Kutta method plotted against runtime for coarse space discretization (328 193 degrees of freedom, top) and fine space discretization (1 311 745 degrees of freedom, bottom)

Finally, Figure 6.6 shows a comparison of the runtimes of the IMEX scheme when using the different stopping criteria for the GMRES solver discussed in Section 6.5.1, namely using $\|r\|_{\tilde{a}_h}$ or $\|r\|_{h,2}$ to estimate the error, respectively. It can be seen that the effort of computing the (better suited) $\|r\|_{\tilde{a}_h}$ is too high and does not pay off. It is noticeable that the runtimes, when using $\|r\|_{\tilde{a}_h}$, are not monotonically decreasing. So far, we are not aware yet where this phenomenon stems from.

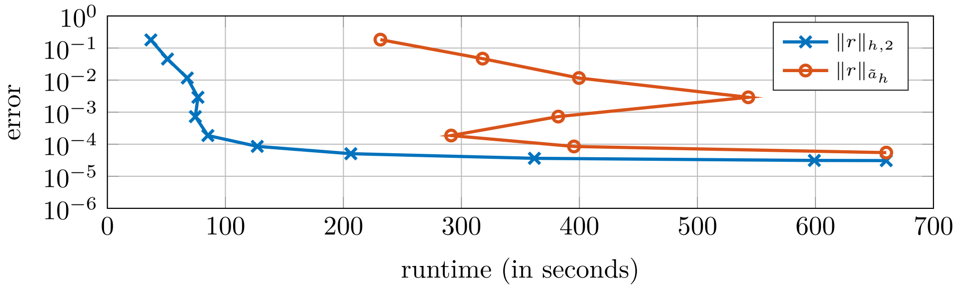

Figure 6.6: Error $\mathbf{E}(0.8)$ of the IMEX scheme plotted against the runtime when using the two different error estimates as stopping criteria for the GMRES scheme as discussed in Section 6.5.1 for a coarse space discretization (328 193 degrees of freedom)

CHAPTER 7

Outlook

This thesis provides some starting points for further research which we will briefly discuss these here.

We developed a unified error analysis for nonlinear wave-type equations in a general setting. However, so far, we only applied these results to a specific wave equation with kinetic boundary conditions. We emphasize that these abstract results can be used in future research to analyze discretizations of other equations that have not been considered so far. Especially in our CRC project, we are interested in the numerical analysis of other types of nonlinear and non-trivial boundary conditions. We plan to accomplish this using the theory developed in this dissertation.

Further, we analyzed a novel and efficient implicit-explicit scheme for semilinear second-order wave-type equations. As mentioned in the introduction, we are not aware of other implicit-explicit schemes in the literature that also exploit the structure of such equations. Thus, it would be interesting to investigate whether it is possible to systematically derive and analyze higher-order implicit-explicit methods which are tailor-made for this type of equation.

Finally, we encountered in our numerical analysis that under some conditions Runge–Kutta methods might converge one order faster than predicted by our theory (cf. Section 6.5.2). Thus, the question arises whether it is possible to refine the error analysis in order to recognize under which conditions this is the case.

Bibliography

R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003. ISBN 0-12-044143-8.

G. Akrivis, M. Crouzeix, and C. Makridakis. Implicit-explicit multistep methods for quasilinear parabolic equations. *Numer. Math.*, 82(4):521–541, 1999. ISSN 0029-599X. doi: 10.1007/s002110050429. URL https://doi.org/10.1007/s002110050429.

D. Arndt, W. Bangerth, B. Blais, T. C. Clevenger, M. Fehling, A. V. Grayver, T. Heister, L. Heltai, M. Kronbichler, M. Maier, P. Munch, J.-P. Pelteret, R. Rastak, I. Thomas, B. Turcksin, Z. Wang, and D. Wells. The `deal.II` library, version 9.2. *Journal of Numerical Mathematics*, 28(3):131–146, 2020. doi: 10.1515/jnma-2020-0043. URL https://dealii.org/deal92-preprint.pdf.

U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton. Implicit-explicit methods for time-dependent partial differential equations. *SIAM J. Numer. Anal.*, 32(3):797–823, 1995. ISSN 0036-1429. doi: 10.1137/0732037. URL https://doi.org/10.1137/0732037.

U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, 25(2-3):151–167, 1997. ISSN 0168-9274. doi: 10.1016/S0168-9274(97)00056-1. URL https://doi.org/10.1016/S0168-9274(97)00056-1. Special issue on time integration (Amsterdam, 1996).

T. Aubin. *Nonlinear analysis on manifolds. Monge-Ampère equations*, volume 252 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, 1982. ISBN 0-387-90704-1. doi: 10.1007/978-1-4612-5734-9. URL https://doi.org/10.1007/978-1-4612-5734-9.

W. Bangerth, R. Hartmann, and G. Kanschat. deal.II—a general-purpose object-oriented finite element library. *ACM Trans. Math. Software*, 33(4):Art. 24, 27, 2007. ISSN 0098-3500. doi: 10.1145/1268776.1268779. URL https://doi.org/10.1145/1268776.1268779.

V. Barbu. *Nonlinear differential equations of monotone types in Banach spaces*. Springer Monographs in Mathematics. Springer, New York, 2010. ISBN 978-1-4419-5541-8. doi: 10.1007/978-1-4419-5542-5. URL https://doi.org/10.1007/978-1-4419-5542-5.

S. Boscarino. Error analysis of IMEX Runge-Kutta methods derived from differential-algebraic systems. *SIAM J. Numer. Anal.*, 45(4):1600–1621, 2007. ISSN 0036-1429. doi: 10.1137/060656929. URL https://doi.org/10.1137/060656929.

S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008. ISBN 978-0-387-75933-3. doi: 10.1007/978-0-387-75934-0. URL https://doi.org/10.1007/978-0-387-75934-0.

I. Chueshov, M. Eller, and I. Lasiecka. On the attractor for a semilinear wave equation with critical exponent and nonlinear boundary dissipation. *Comm. Partial Differential Equations*, 27(9-10):1901–1951, 2002. ISSN 0360-5302. doi: 10.1081/PDE-120016132. URL https://doi.org/10.1081/PDE-120016132.

K. Dekker and J. G. Verwer. *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, volume 2 of *CWI Monographs*. North-Holland Publishing Co., Amsterdam, 1984. ISBN 0-444-87634-0.

K. Disser, M. Meyries, and J. Rehberg. A unified framework for parabolic equations with mixed boundary conditions and diffusion on interfaces. *J. Math. Anal. Appl.*, 430(2):1102–1123, 2015. ISSN 0022-247X. doi: 10.1016/j.jmaa.2015.05.041. URL https://doi.org/10.1016/j.jmaa.2015.05.041.

C. M. Elliott and T. Ranner. Finite element analysis for a coupled bulk-surface partial differential equation. *IMA J. Numer. Anal.*, 33(2):377–402, 2013. ISSN 0272-4979. doi: 10.1093/imanum/drs022. URL https://doi.org/10.1093/imanum/drs022.

C. M. Elliott and T. Ranner. A unified theory for continuous-in-time evolving finite element space approximations to partial differential equations in evolving domains. *IMA Journal of Numerical Analysis*, 11 2020. ISSN 0272-4979. doi: 10.1093/imanum/draa062. URL https://doi.org/10.1093/imanum/draa062. draa062.

E. Emmrich, D. Šiška, and M. Thalhammer. On a full discretisation for nonlinear second-order evolution equations with monotone damping: construction, convergence, and error estimates. *Found. Comput. Math.*, 15(6):1653–1701, 2015. ISSN 1615-3375. doi: 10.1007/s10208-014-9238-4. URL https://doi.org/10.1007/s10208-014-9238-4.

J. Frank, W. Hundsdorfer, and J. G. Verwer. On the stability of implicit-explicit linear multistep methods. *Appl. Numer. Math.*, 25(2-3):193–205, 1997. ISSN 0168-9274. doi: 10.1016/S0168-9274(97)00059-7. URL https://doi.org/10.1016/S0168-9274(97)00059-7. Special issue on time integration (Amsterdam, 1996).

D. J. Gardner, J. E. Guerra, F. P. Hamon, D. R. Reynolds, P. A. Ullrich, and C. S. Woodward. Implicit–explicit (IMEX) Runge–Kutta methods for non-hydrostatic atmospheric models. *Geosci. Model Dev.*, 11(4):1497, 2018.

H. Goldberg, W. Kampowsky, and F. Tröltzsch. On Nemytskij operators in L_p-spaces of abstract functions. *Math. Nachr.*, 155:127–140, 1992. ISSN 0025-584X. doi: 10.1002/mana.19921550110. URL https://doi.org/10.1002/mana.19921550110.

G. R. Goldstein. Derivation and physical interpretation of general boundary conditions. *Adv. Differential Equations*, 11 (4):457–480, 2006. ISSN 1079-9389. URL https://projecteuclid.org/euclid.ade/1355867704.

P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 69 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. ISBN 978-1-611972-02-3. doi: 10.1137/1.9781611972030.ch1. URL https://doi.org/10.1137/1.9781611972030.ch1. Reprint of the 1985 original [MR0775683], With a foreword by Susanne C. Brenner.

E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010. ISBN 978-3-642-05220-0. doi: 10.1007/978-3-642-05221-7. URL https://doi.org/10.1007/978-3-642-05221-7. Stiff and differential-algebraic problems, Second revised edition, paperback.

E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006. ISBN 3-540-30663-3; 978-3-540-30663-4. Structure-preserving algorithms for ordinary differential equations.

E. Hansen. Convergence of multistep time discretizations of nonlinear dissipative evolution equations. *SIAM J. Numer. Anal.*, 44(1):55–65, 2006a. ISSN 0036-1429. doi: 10.1137/040610362. URL https://doi.org/10.1137/040610362.

E. Hansen. Runge-Kutta time discretizations of nonlinear dissipative evolution equations. *Math. Comp.*, 75(254):631–640, 2006b. ISSN 0025-5718. doi: 10.1090/S0025-5718-05-01866-1. URL https://doi.org/10.1090/S0025-5718-05-01866-1.

D. Hipp. *A unified error analysis for spatial discretizations of wave-type equations with applications to dynamic boundary conditions*. PhD thesis, Karlsruher Institut für Technologie (KIT), 2017. URL https://publikationen.bibliothek.kit.edu/1000070952.

D. Hipp, M. Hochbruck, and C. Stohrer. Unified error analysis for nonconforming space discretizations of wave-type equations. *IMA J. Numer. Anal.*, 39(3):1206–1245, 2019. ISSN 0272-4979. doi: 10.1093/imanum/dry036. URL https://doi.org/10.1093/imanum/dry036.

M. Hochbruck and J. Leibold. Finite element discretization of semilinear acoustic wave equations with kinetic boundary conditions. *Electron. Trans. Numer. Anal.*, 53:522–540, 2020. doi: 10.1553/etna_vol53s522. URL https://doi.org/10.1553/etna_vol53s522.

M. Hochbruck and J. Leibold. An implicit-explicit time discretization scheme for second-order semilinear wave equations with application to dynamic boundary conditions. *Numer. Math.*, 2021. ISSN 0945-3245. doi: 10.1007/s00211-021-01184-w. URL https://doi.org/10.1007/s00211-021-01184-w.

M. Hochbruck and B. Maier. Error analysis for space discretizations of quasilinear wave-type equations. CRC 1173 Preprint 2021/2, Karlsruhe Institute of Technology, jan 2021. URL https://www.waves.kit.edu/downloads/CRC1173_Preprint_2021-2.pdf.

M. Hochbruck and A. Sturm. Error analysis of a second-order locally implicit method for linear Maxwell's equations. *SIAM J. Numer. Anal.*, 54(5):3167–3191, 2016. ISSN 0036-1429. doi: 10.1137/15M1038037. URL https://doi.org/10.1137/15M1038037.

M. Hochbruck, T. Pažur, and R. Schnaubelt. Error analysis of implicit Runge-Kutta methods for quasilinear hyperbolic evolution equations. *Numer. Math.*, 138(3):557–579, 2018. ISSN 0029-599X. doi: 10.1007/s00211-017-0914-6. URL https://doi.org/10.1007/s00211-017-0914-6.

M. Hochbruck, B. Maier, and C. Stohrer. Heterogeneous multiscale method for Maxwell's equations. *Multiscale Model. Simul.*, 17(4):1147–1171, 2019. ISSN 1540-3459. doi: 10.1137/18M1234072. URL https://doi.org/10.1137/18M1234072.

W. Hundsdorfer and S. J. Ruuth. IMEX extensions of linear multistep methods with general monotonicity and boundedness properties. *J. Comput. Phys.*, 225(2):2016–2042, 2007. ISSN 0021-9991. doi: 10.1016/j.jcp.2007.03.003. URL https://doi.org/10.1016/j.jcp.2007.03.003.

S. Y. Kadioglu, D. A. Knoll, R. B. Lowrie, and R. M. Rauenzahn. A second order self-consistent IMEX method for radiation hydrodynamics. *J. Comput. Phys.*, 229(22):8313 – 8332, 2010. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2010.07.019. URL http://www.sciencedirect.com/science/article/pii/S0021999110004122.

T. Kashiwabara, C. M. Colciago, L. Dedè, and A. Quarteroni. Well-posedness, regularity, and convergence analysis of the finite element approximation of a generalized Robin boundary value problem. *SIAM J. Numer. Anal.*, 53(1):105–126, 2015. ISSN 0036-1429. doi: 10.1137/140954477. URL https://doi.org/10.1137/140954477.

W. Layton and C. Trenchea. Stability of two IMEX methods, CNLF and BDF2-AB2, for uncoupling systems of evolution equations. *Appl. Numer. Math.*, 62(2):112–120, 2012. ISSN 0168-9274. doi: 10.1016/j.apnum.2011.10.006. URL https://doi.org/10.1016/j.apnum.2011.10.006.

W. Layton, Y. Li, and C. Trenchea. Recent developments in IMEX methods with time filters for systems of evolution equations. *J. Comput. Appl. Math.*, 299:50–67, 2016. ISSN 0377-0427. doi: 10.1016/j.cam.2015.09.038. URL https://doi.org/10.1016/j.cam.2015.09.038.

J. Leibold. Semilineare Wellengleichungen mit dynamischen Randbedingungen. Master's thesis, Karlsruhe Institue of Technology, 2017. URL http://na.math.kit.edu/download/thesis/2017-Leibold.pdf.

J.-F. Lemieux, D. A. Knoll, M. Losch, and C. Girard. A second-order accurate in time IMplicit–EXplicit (IMEX) integration scheme for sea ice dynamics. *J. Comput. Phys.*, 263:375–392, 2014. ISSN 0021-9991. doi: 10.1016/j.jcp.2014.01.010. URL https://doi.org/10.1016/j.jcp.2014.01.010.

B. Maier. *Error analysis for space and time discretizations of quasilinear wave-type equations*. PhD thesis, Karlsruher Institut für Technologie (KIT), 2020. URL https://publikationen.bibliothek.kit.edu/1000120935.

S. Nicaise. Convergence and stability analyses of hierarchic models of dissipative second order evolution equations. *Collect. Math.*, 68(3):433–462, 2017. ISSN 0010-0757. doi: 10.1007/s13348-017-0192-8. URL https://doi.org/10.1007/s13348-017-0192-8.

Y. Qin. *Integral and discrete inequalities and their applications. Vol. I*. Birkhäuser/Springer, [Cham], 2016. ISBN 978-3-319-33300-7; 978-3-319-33301-4. doi: 10.1007/978-3-319-33304-5_8. URL https://doi.org/10.1007/978-3-319-33304-5_8. Linear inequalities.

R. E. Showalter. *Monotone operators in Banach space and nonlinear partial differential equations*, volume 49 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1997. ISBN 0-8218-0500-2. doi: 10.1090/surv/049. URL https://doi.org/10.1090/surv/049.

A. Stern and E. Grinspun. Implicit-explicit variational integration of highly oscillatory problems. *Multiscale Model. Simul.*, 7(4):1779–1794, 2009. ISSN 1540-3459. doi: 10.1137/080732936. URL https://doi.org/10.1137/080732936.

A. Sturm. *Locally Implicit Time Integration for Linear Maxwell's Equations*. PhD thesis, Karlsruher Institut für Technologie (KIT), 2017. URL https://publikationen.bibliothek.kit.edu/1000069341.

A. H. van Zuijlen and H. Bijl. Implicit and explicit higher order time integration schemes for structural dynamics and fluid-structure interaction computations. *Comput. Struct.*, 83(2-3):93–105, 2005.

E. Vitillaro. Strong solutions for the wave equation with a kinetic boundary condition. In *Recent trends in nonlinear partial differential equations. I. Evolution problems*, volume 594 of *Contemp. Math.*, pages 295–307. Amer. Math. Soc., Providence, RI, 2013. doi: 10.1090/conm/594/11793.

E. Vitillaro. On the wave equation with hyperbolic dynamical boundary conditions, interior and boundary damping and source. *Arch. Ration. Mech. Anal.*, 223(3):1183–1237, 2017. ISSN 0003-9527. doi: 10.1007/s00205-016-1055-2. URL https://doi.org/10.1007/s00205-016-1055-2.

M. Zhang and R. D. Skeel. Cheap implicit symplectic integrators. *Appl. Numer. Math.*, 25(2-3):297–302, 1997. ISSN 0168-9274. doi: 10.1016/S0168-9274(97)00066-4. URL https://doi.org/10.1016/S0168-9274(97)00066-4. Special issue on time integration (Amsterdam, 1996).

O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu. *The finite element method: its basis and fundamentals*. Elsevier/Butterworth Heinemann, Amsterdam, seventh edition, 2013. ISBN 978-1-85617-633-0. doi: 10.1016/C2009-0-24909-9. URL https://doi.org/10.1016/C2009-0-24909-9.

APPENDIX A

Collection of auxiliary results

In this chapter we collect some results that are necessary for this thesis.

Lemma A.1 (Grönwall's lemma)**.**

a) ***Continuous case:*** *Let* $\Phi\colon [0,T] \to \mathbb{R}$ *and* $M, \alpha \geq 0$ *s.t. for all* $t \in [0,T]$

$$0 \leq \Phi(t) \leq M + \alpha \int_0^t \Phi(s)\,\mathrm{d}s.$$

Then,

$$\Phi(t) \leq M\mathrm{e}^{\alpha t}$$

holds true for all $t \in [0,T]$*.*

b) ***Discrete case:*** *Let* $\tau > 0$ *and* $M, \alpha \geq 0$ *with* $\alpha\tau < 1$*. If* $\{\varepsilon_n\}_n$ *is a non-negative sequence with*

$$\varepsilon_n \leq M + \alpha\tau \sum_{j=1}^{n} \varepsilon_j \qquad \text{for } n = 0, \ldots, N.$$

Then,

$$\varepsilon_n \leq M(1-\alpha\tau)^{-n} \leq M\mathrm{e}^{\frac{\alpha n \tau}{1-\alpha\tau}}$$

holds true for all $n = 0, \ldots, N$*.*

Proof.

a) Cf. [Qin, 2016, Theorem 1.1.2].

b) The first inequality follows by induction and the second with $(1-s)^{-1} \leq \mathrm{e}^{\frac{s}{1-s}}$ for all $s \in \mathbb{R} \setminus \{1\}$.

□

Definition A.2 (Coercivity). *Let V be a Hilbert space. A bilinear form $\Lambda\colon V \times V \to \mathbb{R}$ is called* ***coercive****, if there exits a constant $\alpha > 0$, s.t. for all $v \in V$*

$$\Lambda(v, v) \geq \alpha \|v\|_V^2.$$

Theorem A.3 (Lax-Milgram theorem). *Let V be a Hilbert space and let $\Lambda\colon V \times V \to \mathbb{R}$ be a bounded and coercive bilinear form. Then, for every $\ell \in V^*$ the equation*

$$\Lambda(v, w) = \langle l, w \rangle_{V^* \times V} \quad \textit{for all } w \in V$$

possess a unique solution $v \in V$.

Proof. Cf. [Brenner and Scott, 2008, Theorem 2.7.7]. □

Theorem A.4 (Sobolev embedding theorem). *Let $\mathcal{M} \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary* ***or*** *let $\mathcal{M} \subset \mathbb{R}^{d+1}$ be a d-dimensional compact C^1 manifold. Then, there exist continuous embeddings*

$$H^1(\mathcal{M}) \hookrightarrow L^q(\mathcal{M}) \qquad \textit{for all} \quad q \in \begin{cases} \leq \frac{2d}{d-2}, & d \geq 3, \\ < \infty, & d = 1, 2. \end{cases}$$

Furthermore, for $d \in \{1, 2, 3\}$, there exists a continuous embedding

$$H^2(\mathcal{M}) \hookrightarrow C(\overline{\mathcal{M}}). \tag{A.1}$$

Proof. Cf. [Adams and Fournier, 2003, Theorem 4.12] and [Aubin, 1982, Chapter 2]. □

Theorem A.5 (Continuity of functions in L^q-spaces). *Let $\mathcal{M} \subset \mathbb{R}^d$ be a measurable set and let $\phi \in C(\mathcal{M} \times \mathbb{R}; \mathbb{R})$ satisfy for some $q \in (1, \infty)$ the growth condition*

$$|\phi(\mathbf{x}, \xi)| \leq C(1 + |\xi|^{\frac{q}{2}}) \qquad \textit{for almost all } \mathbf{x} \in \mathcal{M} \textit{ and all } \xi \in \mathbb{R}.$$

Then,

$$\Phi(v)(\mathbf{x}) := \phi(\mathbf{x}, v(\mathbf{x})), \qquad \mathbf{x} \in \mathcal{M}.$$

defines a function $\Phi \in C(L^q(\mathcal{M}); L^2(\mathcal{M}))$.

Proof. Cf. [Goldberg et al., 1992, Theorem 4]. □

As a direct combination of Theorems A.4 and A.5 we obtain the following corollary:

Corollary A.6. *Let $d \in \{1, 2, 3\}$ and let $\mathcal{M} \subset \mathbb{R}^d$ be a bounded domain with Lipschitz-boundary* ***or*** *let $\mathcal{M} \subset \mathbb{R}^{d+1}$ be a d-dimensional compact C^1-manifold. Further let $\phi \in C(\mathcal{M} \times \mathbb{R}; \mathbb{R})$ satisfy for some*

$$\zeta \begin{cases} < \infty, & d \leq 2, \\ \leq 3, & d = 3. \end{cases}$$

the growth condition

$$|\phi(\mathbf{x}, \xi)| \leq C(1 + |\xi|^{\zeta}) \qquad \textit{for almost all } \mathbf{x} \in \mathcal{M} \textit{ and all } \xi \in \mathbb{R}.$$

Then

$$\Phi(v)(\mathbf{x}) := \phi(v(\mathbf{x})), \qquad \mathbf{x} \in \mathcal{M},$$

defines a function $\Phi \in C(H^1(\mathcal{M}); L^2(\mathcal{M}))$.

APPENDIX B

Runge–Kutta time discretization of nonlinear dissipative evolution equations

In this chapter, we recall results of Hansen [2006b] that we use for the error analysis of Runge–Kutta methods in Sections 3.1 and 5.1. Hansen [2006b] generalizes the classical B-convergence theory for Runge–Kutta approximations of stiff ordinary differential equations (cf. Dekker and Verwer [1984]) to evolution equations in infinite-dimensional Hilbert spaces.

We start by giving a short introduction of Runge–Kutta methods applied to a first-order initial value problem

$$x' = \mathcal{F}(x) + g, \quad t \geq 0, \qquad x^0 = x(0), \tag{B.1}$$

in a Hilbert space X with $\mathcal{F}\colon D(\mathcal{F}) \to X$ and $g\colon [0,\infty) \to X$.

Let τ denote the time step size and $t_n \coloneqq n\tau$ for $n \in \mathbb{N}_0$. A Runge–Kutta method applied to (B.1) has for $n \geq 0$ the form

$$X^{ni} = x^n + \tau \sum_{j=1}^{s} a_{ij}\mathcal{F}(X^{nj}) + g(t_n + c_j\tau), \qquad i = 1, \dots, s, \tag{B.2a}$$

$$x^{n+1} = x^n + \tau \sum_{i=1}^{s} b_i\mathcal{F}(X^{ni}) + g(t_n + c_i\tau), \tag{B.2b}$$

with coefficients $\mathcal{A} = (a_{ij})_{i,j=1}^s, \mathbf{b} = (b_i)_{i=1}^s, \mathbf{c} = (c_i)_{i=1}^s$.

We recall some basic properties of Runge–Kutta methods.

Definition B.1 (Local error)**.** *The* ***local error*** *of a Runge–Kutta method (or more generally of a one step method) is defined as*

$$\overline{x^{n+1}} - x(t_{n+1}),$$

where $\overline{x^{n+1}}$ *is obtained by applying one step of the method to* (B.1) *starting from* $x(t_n)$ *at time* t_n*.*

Definition B.2 (Classical order)**.** *A Runge–Kutta method (or in general a one step scheme) has* ***classical order*** $p \geq 0$*, if for sufficiently smooth* $\mathcal{F}$ *and solutions* x *of* (B.1) *the local error satisfies*

$$\|\overline{x^{n+1}} - x(t_{n+1})\|_X \in \mathcal{O}(\tau^{p+1}).$$

Even for ordinary differential equations, Runge–Kutta methods suffer from order reduction in the stiff case, i.e., for nonlinearities $\mathcal{F}$ with large Lipschitz constants. Since in (B.1) $\mathcal{F}$ can even be an unbounded operator, it is not possible to prove convergence with the full order of the scheme. Instead, in Hansen [2006b], convergence with the so called stage order is proven.

Definition B.3 (Stage order)**.** *A Runge–Kutta method has stage order* $q \geq 0$*, if the coefficients satisfy the conditions*

$$\sum_{i=1}^{s} b_i^{k-1} = \frac{1}{k}, \qquad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \quad k = 1, \dots, q.$$

Important properties of Runge–Kutta methods, which are required for the analysis in Hansen [2006b], are algebraic stability and coercivity.

Definition B.4 (Algebraic stability)**.** *A Runge–Kutta method is called* ***algebraically stable*** *if*

a) $b_i \geq 0$ *for* $i = 1 \dots, s$*, and*

b) the matrix $M = (b_i a_{ij} + b_j a_{ji} - b_i b_j)_{i,j=1}^s$ *is positive semidefinite.*

Definition B.5 (Coercivity)**.** *We call a Runge–Kutta method* ***coercive*** *if the Runge–Kutta matrix* $\mathcal{A}$ *is invertible and there exists a diagonal matrix* $D > 0$ *s.t.*

$$\frac{1}{2}\big(D\mathcal{A}^{-1} + (D\mathcal{A}^{-1})^T\big)$$

is positive definite with smallest Eigenvalue $\alpha_{\mathrm{RK}} > 0$*.*

In Hansen [2006b], Runge–Kutta methods applied to an equation of the form (B.1) (without inhomogeneity g) are analyzed. We now summarize the results and incorporate the presence of the inhomogeneity g. The analysis in Hansen [2006b] relies on the following assumptions on the continuous equation and the Runge–Kutta method:

Assumption B.6.

a) The Runge–Kutta method (B.2) *given by* $\mathcal{A}, \mathbf{b}, \mathbf{c}$ *is algebraically stable, coercive with constant* $\alpha_{\mathrm{RK}} > 0$*, and has stage order* q *for some* $q \in \mathbb{N}$*.*

b) The function $-\mathcal{F}$ *is quasi-monotone with constant* $c_{\mathrm{qm},\mathcal{F}} \geq 0$ *and maximal.*

c) There exists a $T > 0$ *s.t. the solution of* (B.1) *satisfies* $x \in C^{q+1}([0,T]; X)$*.*

d) The step size τ *satisfies the step size restriction*

$$\tau c_{\mathrm{qm},\mathcal{F}} < \alpha_{\mathrm{RK}}$$

such that we can define

$$C_\tau := \frac{1}{\alpha_{\mathrm{RK}} - \tau c_{\mathrm{qm},\mathcal{F}}} > 0. \tag{B.3}$$

Remark B.7.

a) Assumption B.6 a) is satisfied for, e.g., s-stage Gauss and Radau IIA methods with stage order $q = s$ (cf. Hairer and Wanner [2010]).

b) By part b) of Assumption B.6, we have that (B.1) is globally well posed (cf. [Showalter, 1997, Theorem IV.4.1]).

We now recall the results presented in Hansen [2006b]. In a first step, the wellposedness of the Runge–Kutta method (B.2) is shown.

Lemma B.8. *Let Assumption B.6 hold true. Then the Runge–Kutta method (B.2) is globally wellposed, i.e., the approximations x^n uniquely exist for all $n \in \mathbb{N}$.*

Proof. This is shown in [Hansen, 2006b, Lemma 7.1, Theorem 5.2]. Note that the argument, that (B.2a) can be solved for the inner stages $X^{ni}, i = 1, \dots, s$, is not affected by the inhomogeneity g, since it only appears as an additional right-hand side. □

In the next step, the defects of the scheme are bounded. The exact solution of (B.1) inserted into (B.2) reads

$$\begin{aligned}
\widetilde{X}^{ni} &= \widetilde{x}^n + \tau \sum_{j=1}^{s} a_{ij} \Big(\mathcal{F}(\widetilde{X}^{nj}) + g(t_n + c_j\tau) \Big) + \Delta_{\mathrm{RK}}^{ni}, \qquad i = 1, \dots, s, \\
\widetilde{x}^{n+1} &= \widetilde{x}^n + \tau \sum_{s=1}^{s} b_i \Big(\mathcal{F}(\widetilde{X}^{ni}) + g(t_n + c_i\tau) \Big) + \delta_{\mathrm{RK}}^{n+1},
\end{aligned} \tag{B.4}$$

where we used the notation

$$\widetilde{x}^n = x(t_n), \quad \widetilde{X}^{ni} = x(t_n + c_i\tau),$$

and $\Delta_{\mathrm{RK}}^{ni}, \delta_{\mathrm{RK}}^{n+1}$ are the defects.

Lemma B.9 (Defects)**.** *Let Assumption B.6 hold true. Then, for all $i = 1, \dots, s$ and $n \in \mathbb{N}_0$ s.t. $t_n \le T$, with T given in Assumption B.6 c), the defects are bounded by*

$$\|\Delta_{\mathrm{RK}}^{ni}\|_X, \|\delta_{\mathrm{RK}}^{n+1}\|_X \le C\tau^{q+1},$$

with a constant C only depending on $x^{(q+1)}$ and on the coefficients of the Runge–Kutta method.

Proof. This follows from Assumption B.6 c) by Taylor expansion of the exact solution. □

The local error defined in Definition B.1 can be bounded in terms of the defect.

Lemma B.10 (Local error)**.** *Let Assumption B.6 be satisfied. Then, for all $n \in \mathbb{N}$ s.t. $t_n \le T$, with T given in Assumption B.6 c), the local error is bounded by*

$$\|\overline{x^{n+1}} - x(t_{n+1})\|_X \le \|\delta_{\mathrm{RK}}^{n+1}\|_X + C\,(1 + C_\tau) \max_{i=1,\dots,s} \|\Delta_{\mathrm{RK}}^{ni}\|_X \le C\,(1 + C_\tau)\,\tau^{q+1}.$$

Further, the inner stages $\overline{X^{ni}}$ of $\overline{x^{n+1}}$ satisfy for all $t_n \le T$ and $i = 1, \dots, s$ the bound

$$\|\overline{X^{ni}} - x(t_n + c_i\tau)\|_X \le CC_\tau \max_{i=1,\dots,s} \|\Delta_{\mathrm{RK}}^{ni}\|_X \le CC_\tau\tau^{q+1}. \tag{B.5}$$

The generic constant C depends on $x^{(q+1)}$ and the coefficients of the Runge–Kutta method but are independent of $n, \tau, \alpha_{\mathrm{RK}}$, and $c_{\mathrm{qm},\mathcal{F}}$. Further, C_τ is given in (B.3).

Proof. This follows from [Hansen, 2006b, Theorem 6.1]. In the notation of this paper we have

$$m_D[A^{-1}] = \alpha_{\mathrm{RK}}, \qquad hM_X[f] = \tau c_{\mathrm{qm},\mathcal{F}},$$

and $L_{D,X}[\mathcal{A}^{-1}]$ only depends on the coefficients of the Runge–Kutta method. Note that the proof in Hansen [2006b] also works in the presence of the inhomogeneity g, since g vanishes when computing the difference $\overline{x^n} - x(t_n)$ by subtracting (B.4) from the Runge–Kutta scheme (B.2). □

As a last step for the proof of a global error estimate, stability of the scheme is shown.

Lemma B.11 (Stability)**.** *Let Assumption B.6 hold true and let x^n, y^n, for some $n \in \mathbb{N}$, be the approximations obtained by a coercive and algebraically stable Runge–Kutta method applied to (B.1) with starting values x^0 and y^0, respectively. Then, the stability bound*

$$\|x^{n+1} - y^{n+1}\|_X \leq (1 + \tau C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}}) \|x^n - y^n\|_X$$

holds true. Further, the corresponding inner stages satisfy for all $t_n \leq T$ and $i = 1, \dots, s$

$$\|X^{ni} - Y^{ni}\|_X \leq C_{\mathrm{RK}} C_\tau \|x^n - y^n\|_X. \tag{B.6}$$

The constant C_τ is given in (B.3) and C_{RK} depends only on the coefficients of the Runge–Kutta method.

Proof. The result follows by the proof of [Hansen, 2006b, Theorem 7.2]. As in Lemma B.10 we have in the notation of Hansen [2006b]

$$m_D[A^{-1}] = \alpha_{\mathrm{RK}} \quad \text{and} \quad hM_X[f] = \tau c_{\mathrm{qm},\mathcal{F}}.$$

Further, C_0 in [Hansen, 2006b, Theorem 7.2] is given by

$$C_0 = C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}}.$$

Additionally, the inhomogeneity does not affect the proof since g vanishes in the difference of the approximations. □

Finally, by combining the local error bound of Lemma B.10 with the stability result from Lemma B.11 one obtains a global error bound.

Theorem B.12 (Global error)**.** *Let Assumption B.6 hold true and x^n be the approximations obtained by a coercive and algebraically stable Runge–Kutta method applied to (B.1). Then, for all $t_n \leq T$ with T from Assumption B.6, the error bound*

$$\|x^n - x(t_n)\|_X \leq C \frac{\mathrm{e}^{C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}} t_n} - 1}{C_{\mathrm{RK}} c_{\mathrm{qm},\mathcal{F}}} \tau^q$$

holds true, where the constants C_{RK} and C_τ are given in Lemma B.11 and C is independent of τ and n.

Proof. The proof can be found in [Hansen, 2006b, Corollary 7.3]. But, since we have a slightly different representation of the constants in the error bound, we recall it.

We split the error via

$$\|x^{n+1} - x(t_{n+1})\|_X \leq \|x^{n+1} - \overline{x^{n+1}}\|_X + \|\overline{x^{n+1}} - x(t_{n+1})\|_X.$$

By using the local error and the stability bound from Lemmas B.10 and B.11 we obtain

$$\|x^{n+1} - x(t_{n+1})\|_X \leq (1 + \tau C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}})\|x^n - x(t_n)\|_X + C\,(1 + C_\tau)\,\tau^{q+1}.$$

Solving this error recursion using $x^0 = x(t_0)$ and

$$\sum_{i=0}^{n} (1 + C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}} \tau)^i \leq \frac{\mathrm{e}^{C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}} t_{n+1}} - 1}{\tau C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}}}$$

yields

$$\begin{aligned}
\|x^{n+1} - x(t_{n+1})\|_X &\leq C \frac{\mathrm{e}^{C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}} t_{n+1}} - 1}{C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}}}\,(1 + C_\tau)\,\tau^q \\
&\leq C \frac{\left(\mathrm{e}^{C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}} t_{n+1}} - 1\right)(\alpha_{\mathrm{RK}} - \tau c_{\mathrm{qm},\mathcal{F}} + 1)\,(\alpha_{\mathrm{RK}} - \tau c_{\mathrm{qm},\mathcal{F}})^2}{C_{\mathrm{RK}} c_{\mathrm{qm},\mathcal{F}}\,(\alpha_{\mathrm{RK}} - \tau c_{\mathrm{qm},\mathcal{F}})} \tau^q \\
&\leq C \frac{\left(\mathrm{e}^{C_{\mathrm{RK}} C_\tau^2 c_{\mathrm{qm},\mathcal{F}} t_{n+1}} - 1\right)(\alpha_{\mathrm{RK}} + 1)\,\alpha_{\mathrm{RK}}}{C_{\mathrm{RK}} c_{\mathrm{qm},\mathcal{F}}} \tau^q,
\end{aligned}$$

where we additionally used the definition of C_τ (B.3). □

APPENDIX C

Bulk-surface isoparametric finite element method

As in Section 6.2, we assume that for $d \in \{2, 3\}$ and $p \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ is a bounded domain with C^{p+1} boundary $\Gamma = \partial\Omega$.

Since Γ is smooth, the domain Ω cannot be represented exactly by a polygonal mesh. Instead, it has to be approximated by a computational domain Ω_h with boundary Γ_h. Isoparametric finite element spaces of order p are based on piecewise polynomials of order p and the domain Ω is approximated with order p as well. This is due to the fact that using only piecewise linear, i.e., polygonal approximations of the boundary would not lead to the desired convergence order p.

For the kinetic boundary conditions considered in Chapter 6, we also need a suitable boundary finite element space. Such a discretization using a combination of a finite element space for the domain (bulk) Ω and a corresponding finite element space for the boundary (surface) Γ is called bulk-surface finite element method.

In this chapter, we give a short introduction to the bulk-surface finite element method with isoparametric elements that was introduced in Elliott and Ranner [2013]. We recall the construction of the meshes and the finite element spaces and also some properties that are necessary for our error analysis in Section 6.2. Some of the results are taken from the more general paper Elliott and Ranner [2020].

Simplicial and exact Triangulation

The construction of the finite element spaces starts with a family $(\Omega_h^{\#})_h$ of polygonal approximations $\Omega_h^{\#}$ of Ω. Furthermore, let $(\mathcal{T}_h^{\#})_h$ be a corresponding family of simplicial triangulations of $(\Omega_h^{\#})_h$, i.e.,

$$\overline{\Omega_h^{\#}} = \bigcup_{K^{\#} \in \mathcal{T}_h^{\#}} K^{\#} \quad \text{with closed simplices } K^{\#} \in \mathcal{T}_h^{\#}.$$

We denote by

$$h = \max\{\operatorname{diam}(K^{\#}) \mid K^{\#} \in \mathcal{T}_h^{\#}\}$$

the maximal mesh width. The analysis relies on the following properties of the triangulations:

Definition C.1 (Quasi-uniform family of matching simplicial triangulations)**.** *A triangulation $\mathcal{T}_h^{\#} \in (\mathcal{T}_h^{\#})_h$ is called **matching triangulation**, if for any $K^{\#} \in \mathcal{T}_h^{\#}$ with vertices $\{a_1, \dots a_k\}$, the set $\partial K^{\#} \cap \partial \widehat{K}^{\#}$ for any $\widehat{K}^{\#} \in \mathcal{T}_h^{\#}, \widehat{K}^{\#} \neq K^{\#}$, is the convex hull of a (possibly empty) subset of $\{a_0, \dots, a_k\}$.*

*The family $(\mathcal{T}_h^{\#})_h$ is called **quasi-uniform**, if there exists a constant $\rho_{\mathrm{qu}} > 0$ independent of h such that for all $\mathcal{T}_h^{\#} \in (\mathcal{T}_h^{\#})_h$*

$$\min\{\rho_{K^{\#}} \mid K^{\#} \in \mathcal{T}_h^{\#}\} > \rho_{\mathrm{qu}} h$$

holds true. Here, $\rho_{K^{\#}}$ denotes the radius of the largest d-dimensional ball that is contained in $K^{\#}$.

In the following, we require that the family $(\mathcal{T}_h^{\#})_h$ is quasi-uniform and consists of matching triangulations $\mathcal{T}_h^{\#} \in (\mathcal{T}_h^{\#})_h$ which further satisfy the following conditions:

- The vertices of $\mathcal{T}_h^{\#}$ that lie on $\Gamma^{\#} = \partial \Omega_h^{\#}$ also lie on Γ.
- Each $K^{\#} \in \mathcal{T}_h^{\#}$ has at most one face on the boundary $\Gamma^{\#}$.
- The mesh width h is sufficiently small such that for all $\mathbf{x} \in \Gamma^{\#}$ a unique normal projection $\pi_n(\mathbf{x}) \in \Gamma$ exists, i.e., $x - \pi_n(x)$ is orthogonal to the tangent plane of Γ in $\pi_n(x)$ (cf. [Elliott and Ranner, 2013, Section 2.1]).

Construction of isoparametric finite element spaces

We now construct the isoparametric finite element spaces starting from a triangulation $\mathcal{T}_h^{\#}$. Keep in mind that this is done for all triangulations of the family $(\mathcal{T}_h^{\#})_h$, we hence obtain a family of finite element spaces.

Using the normal projection π_n, it is possible to construct an exact triangulation $\mathcal{T}_h^e$ of Ω based on $\mathcal{T}_h^{\#}$, i.e., $\bigcup_{K^e \in \mathcal{T}_h^e} K^e = \overline{\Omega}$ (cf. [Elliott and Ranner, 2013, Section 4.1.2]): All internal mesh elements $K^e \in \mathcal{T}_h^e$, i.e., all elements K^e with at most one point on Γ, coincide with the corresponding elements in $\mathcal{T}_h^{\#}$. The other elements are modified in such a way that the domain Ω is triangulated exactly. Furthermore, each element $K^e \in \mathcal{T}_h^e$ can be described by a C^{p+1}-transformation $F_{K^e}^e : \widehat{K} \to K^e$ where $\widehat{K}$ is the d-dimensional unit simplex.

The computational domain Ω_h is constructed by interpolating the exact triangulation. Let $\widehat{\phi}_1, \dots \widehat{\phi}_{n_p}$ be a Lagrangian basis of the polynomial space $\mathbb{P}_p(\widehat{K})$ corresponding to basis nodes $\widehat{a}_1, \dots \widehat{a}_{n_p}$. Then, we define for $K^e \in \mathcal{T}_h^e$ the interpolation F_{K^e} of $F_{K^e}^e$ via

$$F_{K^e}(\widehat{x}) \coloneqq \sum_{i=1}^{n_p} F_{K^e}^e(\widehat{a}_i)\widehat{\phi}_i(\widehat{x})$$

and the element $K \coloneqq F_{K^e}(\widehat{K}) \approx K^e$. By this, we obtain a triangulation $\mathcal{T}_h \coloneqq \{K = F_{K^e}(\widehat{K}) \mid K^e \in \mathcal{T}_h^e\}$ with the corresponding computational domain $\overline{\Omega_h} \coloneqq \bigcup_{K \in \mathcal{T}_h} K$.

The isoparametric bulk finite element space of order p is then defined via

$$V_{h,p}^\Omega \coloneqq \big\{v_h \in C(\Omega_h) \mid v_h\big|_K = \widehat{v}_h \circ (F_{K^e})^{-1} \text{ with } \widehat{v}_h \in \mathbb{P}_p(\widehat{K}) \text{ for all } K \in \mathcal{T}_h\big\}, \tag{C.1}$$

and the corresponding surface finite element space, for discretizing functions living on the boundary Γ, is given by

$$V_{h,p}^\Gamma \coloneqq \big\{\vartheta_h \in C(\Gamma_h) \mid \vartheta_h = v_h\big|_{\Gamma_h} \text{ with } v_h \in V_{h,p}^\Omega\big\}. \tag{C.2}$$

We have that $V_{h,p}^\Gamma$ is a finite element space over the triangulation

$$\mathcal{T}_h^\Gamma \coloneqq \mathcal{T}_h\big|_{\Gamma_h} \coloneqq \{F = K \cap \Gamma_h \mid K \in \mathcal{T}_h\}$$

of Γ_h and it is defined in such a way that it satisfies the relation

$$\gamma(V_{h,p}^\Omega) = V_{h,p}^\Gamma,$$

where γ is the usual Dirichlet trace operator. Using the finite element space $V_{h,p}^\Omega$ together with the surface finite element space $V_{h,p}^\Gamma$ to discretize differential equations in a domain coupled to a boundary differential equation, as in the situation of Chapter 6, is called the bulk-surface finite element method.

In the following, we denote by $\phi_1, \dots, \phi_N \in V_{h,p}^\Omega$, $N = \dim(V_{h,p}^\Omega)$, the nodal basis functions of $V_{h,p}^\Omega$ and the set of the corresponding basis nodes by $\mathfrak{A} \coloneqq \{a_1, \dots, a_N\} \subset \overline{\Omega} \cap \overline{\Omega_h}$. Similarly, $\phi_1^\Gamma, \dots, \phi_{N_\Gamma}^\Gamma \in V_{h,p}^\Gamma$, $N_\Gamma = \dim(V_{h,p}^\Gamma)$, are the nodal basis functions of $V_{h,p}^\Gamma$ in the basis nodes $\mathfrak{B} \coloneqq \{b_1, \dots, b_{N_\Gamma}\} \subset \mathfrak{A}$.

Lift operator

As in [Elliott and Ranner, 2013, Section 4.2], we define the element-wise smooth homeomorphism $G_h \colon \Omega_h \to \Omega$ via

$$G_h\big|_K \coloneqq F_{K^e}^e \circ (F_{K^e})^{-1} \qquad \text{for all } K^e \in \mathcal{T}_h^e \text{ and } K = F_{K^e}(\widehat{K}).$$

This allows to define for $v_h \in V_{h,p}^\Omega, \vartheta_h \in V_{h,p}^\Gamma$ lifted versions $v_h^\ell \in C(\overline{\Omega}), \vartheta_h^\ell \in C(\Gamma)$ via

$$v_h^\ell \coloneqq v_h \circ G_h^{-1} \qquad \text{and} \qquad \vartheta_h^\ell \coloneqq \vartheta_h \circ G_h^{-1}. \tag{C.3}$$

In [Elliott and Ranner, 2020, Lemmas 5.3 and 7.3] the following element-wise norm equivalences related to the lift are shown.

Lemma C.2. *There exists $C_{\Omega,\Omega_h} > c_{\Omega,\Omega_h} > 0$, $C_{\Gamma,\Gamma_h} > c_{\Gamma,\Gamma_h} > 0$ independent of h s.t. for all $v_h \in V^{\Omega}_{h,p}$, $\vartheta_h \in V^{\Gamma}_{h,p}$, $k = 0, 1, \dots, p+1$, and $K_\Omega \in \mathcal{T}_h$, $K_\Gamma \in \mathcal{T}^{\Gamma}_h$ we have*

$$c_{\Omega,\Omega_h}\|v_h\|_{H^k(K_\Omega)} \le \|v^\ell_h\|_{H^k(K^\ell_\Omega)} \le C_{\Omega,\Omega_h}\|v_h\|_{H^k(K_\Omega)},$$
$$c_{\Gamma,\Gamma_h}\|\vartheta_h\|_{H^k(K_\Gamma)} \le \|\vartheta^\ell_h\|_{H^k(K^\ell_\Gamma)} \le C_{\Gamma,\Gamma_h}\|\vartheta_h\|_{H^k(K_\Gamma)},$$

where $K^\ell_\Omega = G_h(K_\Omega), K^\ell_\Gamma = G_h(K_\Gamma)$. By construction, the lift additionally preserves the L^∞ norm, i.e.,

$$\|v^\ell_h\|_{L^\infty(K^\ell_\Omega)} = \|v_h\|_{L^\infty(K_\Omega)},$$
$$\|\vartheta^\ell_h\|_{L^\infty(K^\ell_\Gamma)} = \|\vartheta_h\|_{L^\infty(K_\Gamma)}.$$

We further have the following bounds of the geometric errors stemming from approximating the domain (cf. [Elliott and Ranner, 2013, proof of Lemma 6.2]).

Lemma C.3. *For $u_h, \varphi_h \in V^{\Omega}_{h,p}$ and $\vartheta_h, \psi_h \in V^{\Gamma}_{h,p}$, the following bounds hold true:*

$$|\int_\Omega u^\ell_h \varphi^\ell_h \,\mathrm{d}\mathbf{x} - \int_{\Omega_h} u_h \varphi_h \,\mathrm{d}\mathbf{x}| \le Ch^p \|u_h\|_{L^2(\Omega_h)} \|\varphi_h\|_{L^2(\Omega_h)}, \tag{C.4a}$$
$$|\int_\Omega \nabla u^\ell_h \nabla \varphi^\ell_h \,\mathrm{d}\mathbf{x} - \int_{\Omega_h} \nabla u_h \nabla \varphi_h \,\mathrm{d}\mathbf{x}| \le Ch^p \|\nabla u_h\|_{L^2(\Omega_h)} \|\nabla \varphi_h\|_{L^2(\Omega_h)}, \tag{C.4b}$$
$$|\int_\Gamma \vartheta^\ell_h \psi^\ell_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_h} \vartheta_h \psi_h \,\mathrm{d}\mathbf{x}| \le Ch^{p+1} \|\vartheta_h\|_{L^2(\Gamma_h)} \|\psi_h\|_{L^2(\Gamma_h)}, \tag{C.4c}$$
$$|\int_\Gamma \nabla_\Gamma \vartheta^\ell_h \nabla_\Gamma \psi^\ell_h \,\mathrm{d}\mathbf{x} - \int_{\Gamma_h} \nabla_{\Gamma_h} \vartheta_h \nabla_{\Gamma_h} \psi_h \,\mathrm{d}\mathbf{x}| \le Ch^{p+1} \|\nabla_{\Gamma_h} \vartheta_h\|_{L^2(\Gamma_h)} \|\nabla_{\Gamma_h} \psi_h\|_{L^2(\Gamma_h)}. \tag{C.4d}$$

Interpolation

By $I_{h,\Omega} \colon C(\overline{\Omega}) \to V^{\Omega}_{h,p}$ and $I_{h,\Gamma} \colon C(\Gamma) \to V^{\Gamma}_{h,p}$ we denote the usual Lagrange interpolation operators in the domain and on the boundary, respectively. Note that the interpolations are well defined since, by construction of the finite element spaces, the Lagrange nodal basis points lie in Ω or on Γ, respectively.

Lemma C.4. *The interpolation operators $I_{h,\Omega}$ and $I_{h,\Gamma}$ are continuous with respect to $\|\cdot\|_\infty$, i.e., for all $u \in C(\overline{\Omega}), \vartheta \in C(\Gamma)$ we have*

$$\|I_{h,\Omega} u\|_\infty \le C_\infty \|u\|_\infty, \qquad \|I_{h,\Gamma} \vartheta\|_\infty \le C_\infty \|\vartheta\|_\infty$$

with $C_\infty = \|\sum_{i=1}^N |\phi_i|\|_\infty$. This constant is independent of h since the number of non-vanishing basis functions on each mesh element only depends on the polynomial degree p.

We further have the following interpolation error bounds, which follow from [Elliott and Ranner, 2020, Theorem 4.28, Theorem 5.9] for the bulk and [Elliott and Ranner, 2020, Theorem 6.24, Theorem 7.10] for the surface, respectively:

Lemma C.5. *Let $1 \leq k \leq p$.*

a) Globally, the interpolation operators satisfy for all $v \in H^{k+1}(\Omega)$, and $\vartheta \in H^{k+1}(\Gamma)$ the error bounds

$$\|v - (I_{h,\Omega}v)^\ell\|_{L^2(\Omega)} + h\|v - (I_{h,\Omega}v)^\ell\|_{H^1(\Omega)} \leq Ch^{k+1}\|v\|_{H^{k+1}(\Omega)}, \tag{C.5a}$$

$$\|\vartheta - (I_{h,\Gamma}\vartheta)^\ell\|_{L^2(\Gamma)} + h\|\vartheta - (I_{h,\Gamma}\vartheta)^\ell\|_{H^1(\Gamma)} \leq Ch^{k+1}\|\vartheta\|_{H^{k+1}(\Gamma)}, \tag{C.5b}$$

with a constant C independent of h.

b) Locally, on each element $K_\Omega \in \mathcal{T}_h^\Omega$, $K_\Gamma \in \mathcal{T}_h^\Gamma$, the interpolation operators satisfy for all $0 \leq r \leq k$ and all $v \in H^{k+1}(K_\Omega^\ell), \vartheta \in H^{k+1}(K_\Gamma^\ell)$, the error bounds

$$\|v - (I_{h,\Omega}v)^\ell\|_{H^r(K_\Omega^\ell)} \leq Ch^{k+1-r}\|v\|_{H^{k+1}(K_\Omega^\ell)}, \tag{C.6a}$$

$$\|\vartheta - (I_{h,\Gamma}\vartheta)^\ell\|_{H^r(K_\Gamma^\ell)} \leq Ch^{k+1-r}\|\vartheta\|_{H^{k+1}(K_\Gamma^\ell)}, \tag{C.6b}$$

with a constant C independent of h.

c) Locally, on each element $K_\Omega \in \mathcal{T}_h^\Omega$, $K_\Gamma \in \mathcal{T}_h^\Gamma$, and for every $v_h \in H^{k+1}(K_\Omega), \vartheta_h \in H^{k+1}(K_\Gamma)$, the L^∞ error bounds

$$\|v_h - I_{h,\Omega}v_h^\ell\|_{L^\infty(K_\Omega)} \leq Ch^{k+1}\|v_h\|_{H^{k+1}(K_\Omega)}, \tag{C.7a}$$

$$\|\vartheta_h - I_{h,\Gamma}\vartheta_h^\ell\|_{L^\infty(K_\Gamma)} \leq Ch^{k+1}\|\vartheta_h\|_{H^{k+1}(K_\Gamma)} \tag{C.7b}$$

hold true with a constant C independent of h.