Error Analysis of Exponential Integrators for Nonlinear Wave-Type Equations

Zur Erlangung des akademischen Grades

DOKTOR DER NATURWISSENSCHAFTEN

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

DISSERTATION

von

Benjamin Dörich

Tag der mündlichen Prüfung: 24. Februar 2021

1. Referentin:	Prof. Dr. Marlis Hochbruck
2. Referent:	Prof. Dr. Roland Schnaubelt
3. Referent:	Univ.-Prof. Dr. Alexander Ostermann

Acknowledgement

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

I continue in German.

Mein größter Dank gebührt meiner Betreuerin Prof. Dr. Marlis Hochbruck. Sie war es, die mich in meinen ersten Numerik-Vorlesungen für das Fach begeistert hat. Nach meiner Masterarbeit hatte sie mir das Vertrauen entgegengebracht und mich darin bestärkt, mir eine Promotion zuzutrauen. Ich bin mir nicht sicher, ob ich heute sonst hier wäre.

Weiter bedanke ich mich bei Prof. Dr. Roland Schnaubelt, bei dem ich nicht nur meine Bachelorarbeit geschrieben habe, sondern dem ich zu jeder Zeit meine Fragen zur Funktionalanalysis stellen konnte. Nachdem er schon meine Masterarbeit referiert hatte, freut es mich, ihn auch als Gutachter für diese Dissertation gewinnen zu können. Ebenso danke ich Prof. Dr. Wolfgang Reichel für die Zeit, die er sich für meine Fragen zu elliptischen Problemen genommen hat.

Dank gilt auch Univ.-Prof. Dr. Alexander Ostermann, der sich bereit erklärt hat, die Begutachtung dieser Arbeit zu übernehmen.

Ich möchte die Gelegenheit weiter nutzen, um mich sowohl bei Prof. Dr. Tobias Jahnke zu bedanken, der mich in seiner Vorlesung zu Finiten Elementen für die numerische Analysis begeistert hat, als auch bei meinen Kollegen aus der Analysis, Simon Kohler und Konstantin Zerulla, die sich immer die Zeit genommen haben, mir mit ihrem Wissen weiterzuhelfen.

Nicht genug danken kann ich meiner ganzen Arbeitsgruppe, die das Lehren und Forschen in den letzten drei Jahren unendlich bereichert hat. Nicht nur die Kompaktseminare und die fachlichen Diskussionen, sondern auch die unzähligen Kaffeerunden und Mittagessen haben diese letzten drei Jahre zu etwas sehr besonderem gemacht. Bei Christian und Mathias bedanke ich mich für die Geduld und die Unterstützung bei meinen IT-Fragen. Gerade für die letzten Monate, möchte ich mich bei Bernhard, Constantin und Jan bedanken für die Zeit, diese Arbeit Korrektur zu lesen und sie dadurch so aufzuwerten.

Dankbar bin ich auch meinen Eltern für ihre Zuversicht und ihr Vertrauen, dass ich meinen Weg gehen werde, und meinen Schwestern dafür, dass sie schon so lange meine größten Vorbilder sind.

Ich weiß gar nicht, wie ich dem Beitrag gerecht werden kann, den du, Samira, zu dieser Arbeit hattest. Du hast mich aufgebaut, wenn es einmal nicht so lief, und dich mit mir gefreut, wenn ich ein Problem gelöst hatte. Du hast mich abgelenkt, wenn ich zu verbissen war, und mir die Ruhe gegeben, wenn es voran gehen musste. Für dein Gespür bin ich dir so unendlich dankbar. Diese Arbeit widme ich dir!

Abstract

This thesis is concerned with the time integration of certain classes of nonlinear evolution equations in Hilbert spaces by exponential integrators. We aim to prove error bounds which can be established by including only quantities given by a wellposedness result. In the first part, we consider semilinear wave equations and introduce a class of first- and second-order exponential schemes. A standard error analysis is not possible due to the lack of regularity. We have to employ appropriate filter functions as well as the integration by parts and summation by parts formulas in order to obtain optimal error bounds. In the second part, we propose two exponential integrators of first and second order applied to a class of quasilinear wave-type equations. By a detailed investigation of the differentiability of the right-hand side we derive error bounds in different norms. In the framework we can treat quasilinear Maxwell's equations in full space and on a smooth domain as well as a class of quasilinear wave equations. In both parts, we include numerical examples to confirm our theoretical findings.

Contents

CHAPTER 1

Motivation and Introduction

In the natural sciences many phenomena are modeled by ordinary (ODEs) and partial differential equations (PDEs). They arise from physical models and describe how certain processes take place. This means that, once we know the solution of the differential equation, we have a precise knowledge of what will happen in the considered system. In this thesis we focus on a specific class of PDEs, namely wave-type problems, in particular on the wave equation and Maxwell's equations. Wave equations model for example the propagation of sound or the vibration of a membrane and the foundations of classical electromagnetism are laid by Maxwell's equations.

Because of these important applications, scientists and engineers have been encouraged for centuries to predict the physical processes by finding the solutions of the given equations. For the specific models this can sometimes be done even analytically, i.e., one can derive an explicit formula for the solution. Typical strategies are separation of variables as well as the application of Fourier and Laplace transforms, respectively.

However, in most cases such a direct approach is not possible and only an approximation of the exact solution can be found. To do this on a computer, we need to turn the problems, which are continuous in space and time, into finite dimensional problems. Often the method of lines technique is used where first the spatial domain is discretized and the time variable remains continuous. This leads to a generally large system of ODEs which in a second step has to be discretized in time. In this thesis we restrict ourselves to the discretization in time and work in an abstract function space which is an important first step towards the analysis of fully discrete methods. We note that for practical implementations a discretization in space is necessary.

The thesis comprises two parts where we treat different classes of nonlinear evolution equations. However, both parts share two common features:

We aim at error bounds that only rely on the regularity of the solution which can be directly derived from the problem. This means given certain data of the problem, we first conclude uniqueness and existence of a solution and derive its regularity and a priori bounds. This information then enters the

error analysis where all appearing constants can be traced back to known data. This approach is usually referred to as an error analysis in terms of the data.

Moreover, we only study exponential integrators which have become quite popular in the last decades. They are constructed from the variation-of-constants formula and treat the linear part of the equation exactly. Hence, more information of the exact solution is incorporated in the numerical method which improves the numerical approximation even if the exact solution is not smooth.

1.1 Semilinear problems

In the first part we are interested in solving abstract wave equations such as for example the cubic wave equation or the sine-Gordon equation. In particular, we study the time integration in an abstract Hilbert space framework and focus on error bounds that can be established under physically realistic assumptions, in particular finite-energy conditions.

If the equation is posed on a finite dimensional space, for example after the discretization by finite differences, there is some literature available, see e.g., [35, Chap. XIII.] for an overview. García-Archilla, Sanz-Serna and Skeel [21], Grimm and Hochbruck [30], Hochbruck and Lubich [38], and Sanz-Serna [66] studied exponential (or trigonometric) integrators for such equations. These methods were shown to be second-order convergent and unconditionally stable, i.e., the constants do not depend on the Lipschitz constant of the discretized differential operator. Remarkably, the error analysis is performed under a finite-energy condition only and does not make use of bounds on the second time derivative. The key ingredient are certain matrix functions that act as filters which remove resonances in the local error. If these filters are chosen appropriately, they ensure cancellations in the global error such that the local error is of the same order as the global error.

In the recent paper [8] by Buchholz et al. and the PhD thesis [10] by Buchholz, a completely new technique was introduced to prove and extend the above mentioned results in the ODE case. The trigonometric integrator was reformulated as a Strang splitting applied to a modified problem and the proof was divided into two steps. First, the error introduced from the modified problem is bounded. Afterwards, using ideas from [46] by Jahnke and Lubich and [54] by Lubich, a specific representation of the defect was derived that separates terms of order three from the leading error terms which are of order two. The order three terms are then summed up in the standard way and the filters are employed for the leading error terms. In this way, a modified Lady Windermere's fan argument together with the above-mentioned cancellations lead to the global error of order two.

We close with a brief overview of further work on exponential integration schemes for the time integration of semilinear wave equations. Baumstark, Faou and Schratz [7] studied the time integration of a sine-Gordon equation that depends on a parameter $c \to \infty$ which induces high oscillations in time. They construct methods that allow for error bounds independent of c. A spatially discretized wave equation with periodic boundary conditions is considered by Gauckler in [22]. The right-hand side is chosen as a polynomial such that the smooth coefficients together with the algebra structure allow for optimal error bounds. Gauckler et al. extended the approach in [23] to the quasilinear case. For linear and semilinear evolution equations in [36, 37] exponential splitting methods are analyzed by Hansen and Ostermann where the error bounds rely on commutator bounds of the splitting operators. Crouseilles, Einkemmer and Massot [17] compare the stability of different types of exponential integrators applied to Vlasov-

Poisson and drift-kinetic equations. In a recent preprint, Caliari et al. [13] apply rational exponential integrators to hyperbolic and oscillatory PDEs.

Aim and main results

We present rigorous error bounds for the time discretization of abstract wave equations with exponential integrators under low regularity assumptions on the data in the first part of this thesis. Due to this lack of regularity we do not consider higher order methods. We note that most of the material has already been published in [9].

The methods are constructed as follows. We introduce filter functions and replace the right-hand side of the equation by a filtered variant which leads to an averaged equation. We then take an exponential method, which we call the underlying scheme, and apply it to the averaged equation. This new method is the averaged scheme which is analyzed in the first part of this thesis. For the underlying schemes we refer to the papers by Celledoni, Cohen and Owren [15] and Hochbruck and Ostermann [39, 40] on exponential Runge–Kutta methods, by Hochbruck, Leibold and Ostermann [45] on Lawson methods and by Wang, Wu and Xia [71, 72] on extended Runge–Kutta–Nyström methods.

The analysis is performed on a whole scale of Sobolev spaces and covers different boundary conditions. Thus, we can treat a large class of examples, and we included all computations to verify that they fit into the abstract framework. In particular, our framework covers non-constant coefficients for the differential operator and power bounded nonlinearities for which the admissible polynomial degree is determined by the spatial dimension and the corresponding Sobolev embeddings. Most importantly, the framework admits for a classical wellposedness result which is the only regularity that enters the constants in the error bounds. Up to our knowledge this has only been done before by Gauckler [22] for periodic boundary conditions where a far richer structure is available.

The error analysis applies to a large class of first- and second-order exponential integrators. For the presented error bounds of the second-order methods we first provide a detailed characterization of the filter functions, the averaged solution, as well as for the defects. This allows us to derive the estimates in a unified way. Finally, we obtain the error bounds of the first-order methods under even lower regularity assumptions for the approximation of mild solutions.

Outline

The first part of the thesis is organized as follows. In Chapter 2, we introduce the analytical framework, discuss the wave equation in the second-order formulation and briefly illustrate it with an example. We reformulate the equation as a first-order system and recall some basic semigroup theory which leads us to a standard wellposedness result. We conclude this chapter by introducing a functional calculus for skew-adjoint operators in Hilbert spaces.

We proceed with a review on exponential integrators in Chapter 3. We first explain the general construction of such schemes and derive the two simplest methods of first order. Further, we present the second-order methods which serve as the underlying schemes for our averaged methods.

The core of the first part is Chapter 4. We begin with an informal overview of the main concepts appearing in the error analysis and sketch the main results. For the purpose of illustration we added a numerical experiment and a class of examples which are covered by the presented theory. In the

subsequent sections we refine the analytical framework, introduce the filter functions and estimate the error stemming from the averaging procedure. We finally establish the desired error bounds for the different methods in different norms.

All technical calculations which are necessary to fit the concrete examples into the abstract framework are collected in Appendix A.

1.2 Quasilinear problems

The second part of the thesis is concerned with the time integration of the quasilinear evolution equations posed in a Hilbert space by exponential integrators.

In a very general framework, Kato proved in [50, 51] the wellposedness of such equations. This generality does not only cover symmetric hyperbolic system of first order, but also the wave equation, the Kortweg-de Vries equation and many more interesting examples. In order to relax the assumptions on the initial data and make the results of Kato easier to apply, the framework was refined in the PhD thesis [61] by Müller with a focus on certain quasilinear wave and Maxwell's equations. Due to the wide range of applications of quasilinear equations in the modeling of different phenomena, a large literature on their numerical treatment emerged over the last few years which we will present in the following.

We first mention the papers by Kanda [47], Kobayashi [52], and Takahashi [69], where the wellposedness of general nonlinear evolution equations is studied. The aim was to construct solutions via difference approximations which are in principle the implicit and semi-implicit Euler method. The nonlinear semigroups are generated by multi-valued, dissipative operators on some Banach space. For the special case of quasilinear problems, Crandall and Souganidis [16] give a different proof of Kato's results also via difference approximations. However, the convergence rates in these papers are only of order 1/2. The first results for the time integration of quasilinear hyperbolic problems with optimal order were derived by Hochbruck and Pažur in [41]. They employed completely new techniques and proved error bounds of order one for the mentioned variants of the Euler method. Building upon this, Hochbruck, Pažur and Schnaubelt [44] and Kovács and Lubich [53] extended the techniques to coercive and algebraically stable Runge–Kutta methods. A similar framework was used by Maier in the thesis [56] where finite element methods were combined with the leapfrog method and Runge–Kutta schemes in order to prove full discretization error bounds.

In the case of the one-dimensional quasilinear wave equation equipped with periodic boundary condition, Gauckler et al. [23] used trigonometric integrators for the time integration. By a sophisticated stability analysis, the authors proved second-order error bounds in time and further treated the full discretization with pseudo-spectral methods. The Westervelt equation in two and three dimensions was studied by Antonietti et al. [6] and Nikolić and Wohlmuth [62] where the discretization in space was performed with continuous and discontinuous Galerkin (dG) methods. Absorbing boundary conditions for this equation were treated in [60] by Muhr, Nikolić and Wohlmuth. For parabolic problems, Casas and Chrysafinos [14] combined linear finite elements in space with a dG method of order zero in time and showed error bounds under low regularity assumptions.

Before we turn to our contributions, we emphasize the influence of the following two series of papers. The original Kato framework was used by Kovács and Lubich in [53] and the refined Kato framework, which is also the basis of our error analysis, was used by Hochbruck, Pažur and Schnaubelt in [41, 44].

In these related frameworks, the time integration by algebraically stable and coercive Runge–Kutta schemes was considered and error bounds were derived. In particular, their stability estimates were the starting point for our stability analysis. The idea for the method and the idea how to represent the local error comes mainly from the papers [26–28] by González and Thalhammer. They considered quasilinear parabolic equations and constructed and analyzed several exponential integration schemes. However, the analyticity of the semigroup is a key ingredient in their analysis such that we cannot apply it to wave-type equations where the semigroup usually is generated by a skew-adjoint operator.

Aim and main results

In the second part of this thesis we present rigorous error bounds for the time discretization of quasilinear wave-type evolution equations with exponential integrators. In this framework we treat quasilinear wave equations and Maxwell's equations simultaneously. We extend the framework of possible examples and thus provide an extension of the wellposedness result by Müller [61] to more general right-hand sides.

We propose two exponential integrators which are of first and second order and derive error bounds in different norms. Similar to the first part of the thesis our main error bounds only rely on the regularity obtained from the wellposedness result. Compared to the methods from Hochbruck, Pažur and Schnaubelt [41, 44] and Kovács and Lubich [53], we hence relax the assumptions on the regularity of the exact solution. A key ingredient is the precise knowledge of the differentiability of the data evaluated at smooth functions, which we formulate as assumptions. For the specific examples, we postpone the proofs to the appendix.

Up to our knowledge there are only two papers that treat exponential integrators for quasilinear wave-type equations. Pototschnig et al. [64] consider an application from physical optics and perform numerical experiments. Gauckler et al. [23] only treat the one-dimensional wave equation with periodic boundary conditions. Hence, this is the first result concerning error bounds on exponential integrators for this general class of quasilinear wave-type evolution equations.

Outline

The second part of the thesis is structured as follows. In Chapter 5, we introduce the analytical framework and explain the two main examples fitting into it. All necessary assumptions for the error analysis are presented in an abstract way in order to treat the examples at once. We conclude the chapter with the extension of a known wellposedness result which is the foundation of the following error analysis.

Next, we review some numerical methods for quasilinear problems in Chapter 6. In the first two sections, we present the methods for wave-type problems by Hochbruck, Pažur and Schnaubelt [41, 44] and Kovács and Lubich [53] and for parabolic problems by González and Thalhammer [26–28]. Further, we explain the error bounds derived by Gauckler et al. [23] for the special case of the quasilinear wave equation in one spatial dimension. We also comment on a numerical comparison by Pototschnig et al. [64] of exponential integrators and classical time integration schemes for quasilinear Maxwell's equations which clearly indicates that exponential integrators can be highly competitive.

Finally, in Chapter 7 we propose our new methods, state the main results and prove them in the subsequent sections. Further, we provide a numerical example where we combine our methods with a finite element method in space. As a possible further application of the technique used in the proofs we establish error bounds also in stronger norms. However, they cannot be derived from the wellposedness

result proven earlier but need additional regularity.

As in part I, we postpone the calculations to show that the wave equation and Maxwell's equations fit into the abstract framework to Appendix B.

Notation

In this section, we introduce the notation used throughout the thesis.

Differential operators Let $\Omega \subseteq \mathbb{R}^d$ be some domain with $d \in \{1,2,3\}$ and consider sufficiently smooth functions $f\colon \Omega \to \mathbb{R}$ and $g\colon \Omega \subseteq \mathbb{R}^3 \to \mathbb{R}^3$. We define for $x = \begin{pmatrix} x_1 \dots x_d \end{pmatrix}$ the gradient

$$\nabla f(x) = \sum_{i=1}^{d} \partial_{x_i} f(x)\,.$$

Further, for $g = \begin{pmatrix} g_1, g_2, g_3 \end{pmatrix}^T$ we define the divergence by

$$\operatorname{div} g = \partial_{x_1} g_1 + \partial_{x_2} g_2 + \partial_{x_3} g_3\,,$$

and the curl-operator by

$$\operatorname{curl} g = \begin{pmatrix} \partial_{x_2} g_3 - \partial_{x_3} g_2 \\ \partial_{x_3} g_1 - \partial_{x_1} g_3 \\ \partial_{x_1} g_2 - \partial_{x_2} g_1 \end{pmatrix}.$$

The Laplacian Δ is given by

$$\Delta f = \operatorname{div}(\nabla f) = \sum_{i=1}^{d} \partial_{x_i}^2 f\,.$$

Spaces For Banach spaces X, Y, $\|\cdot\|_X$ denotes the norm on X and $\mathcal{L}(X,Y)$ the set of all bounded operators $T\colon X \to Y$ equipped with the standard operator norm $\|T\|_{Y \leftarrow X}$. We use the abbreviation $\mathcal{L}(X) \coloneqq \mathcal{L}(X,X)$. If X is a Hilbert space $\langle \cdot, \cdot \rangle_X$ denotes the scalar product on X.

We denote for a radius $r > 0$ the ball around zero in X by

$$\mathcal{B}_X(r) \coloneqq \{x \in X \mid \|x\|_X \leq r\}\,,$$

and if X can be written as a product $X = X_1 \times X_2$ we denote by denote by $\pi_i\colon X \to X$ the projection onto the i-th component of the product space X, $i = 1,2$, i.e., for $x = \begin{pmatrix} x_1, x_2 \end{pmatrix}^T \in X$

$$\pi_1 x = \begin{pmatrix} x_1 \\ 0 \end{pmatrix}, \qquad \pi_2 x = \begin{pmatrix} 0 \\ x_2 \end{pmatrix}.$$

Further, $C^k(X,Y)$ is the space of all k-times Fréchet-differentiable functions from X to Y. We write $W^{k,p}(\Omega)$, $k \in \mathbb{N}_0$, $1 \leq p \leq \infty$, for the Sobolev space of order k with all (weak) derivatives in $L^p(\Omega)$ and abbreviate $H^k(\Omega) \coloneqq W^{k,2}(\Omega)$. For multi-indices $\alpha, \beta \in \mathbb{N}^\ell$ we write $\alpha \leq \beta$ if $\alpha_i \leq \beta_i$ for all $i = 1, \dots, \ell$.

Calculus on Banach spaces We use several times the theory of differentiation and integration in Banach spaces. We refer the reader to [4, Section IV.3] for results concerning the differentiation. Since all integrals can be understood as Riemann integrals, the necessary theory can be found in [5, Chapter VI].

Part I

On averaged exponential integrators for semilinear wave equations with solutions of low regularity

CHAPTER 2

Analytical framework - semilinear problems

In this chapter we lay the foundations of the first part of this thesis. After recalling some basic facts from functional analysis, we introduce the semilinear wave equation in first- and second-order formulation. The last section contains the most important properties of the functional calculus for self-adjoint operators.

Since we deal with differential operators, we need to extend the classical operator theory concerning bounded linear operators. The proper generalization is given by closed linear operators. Such operators A are in general not defined on the full space but rather only on a subset $\mathcal{D}(A)$ of a Hilbert space X which we call the domain of A. We further say that an operator is densely defined if $\overline{\mathcal{D}(A)} = X$ holds.

Definition 2.1. *Let $A\colon \mathcal{D}(A) \subseteq X \to X$ be a linear operator. We call A a closed operator if the following implication holds:*

For any sequences $(x_n)_n$ in $\mathcal{D}(A)$ and $(y_n)_n$ in X given by $y_n \coloneqq Ax_n$ with

$$x_n \to x \quad \text{and} \quad y_n \to y$$

for some $x, y \in X$ it holds

$$x \in \mathcal{D}(A) \quad \text{and} \quad Ax = y\,.$$

For such operators one can also define the adjoint operator. This has to be done slightly more carefully compared to the bounded operator case.

Definition 2.2. *Let $A\colon \mathcal{D}(A) \to X$ be a closed, linear and densely defined operator and consider the set*

$$\mathcal{D}(A^*) \coloneqq \{y \in X \mid \exists z \in X : \langle Ax, y\rangle_X = \langle x, z\rangle_X\}\,.$$

(a) *We define $A^* : \mathcal{D}(A^*) \to X$ for any $y \in \mathcal{D}(A^*)$ by $A^*y \coloneqq z$. Note that z is unique and the map is well-defined. We call A^* the adjoint of A.*

(b) *A is called self adjoint if $\mathcal{D}(A^*) = \mathcal{D}(A)$ and $A^* = A$.*

(c) *A is called skew adjoint if $\mathcal{D}(A^*) = \mathcal{D}(A)$ and $A^* = -A$.*

This next definition recalls the well-known concepts of the spectrum and the resolvent set.

Definition 2.3. *Let $A\colon \mathcal{D}(A) \to X$ be a linear, closed operator.*

(a) *The resolvent set of A is given by*

$$\rho(A) \coloneqq \{\lambda \in \mathbb{C} \mid \lambda I - A\colon \mathcal{D}(A) \to X \quad \text{is invertible}\}\,.$$

(b) *The spectrum of A is given by $\sigma(A) \coloneqq \mathbb{C} \setminus \rho(A)$.*

(c) *The operator A is called strictly positive if there is some $c_A > 0$ such that $\langle Ax, x\rangle_X \geq c_A \|x\|_X^2$.*

We finally introduce the concept of compact operators which often appears in the context of embeddings.

Definition 2.4. *Let $T\colon Y \to X$ be a linear operator. We call T a compact operator if the following holds: For any bounded sequence $(y_n)_n$ in Y, there exists a subsequence $(y_{n_j})_j$ such that the sequence $(Ty_{n_j})_j$ converges in X.*

2.1 Wave equation

The equation of interest in this first part of the thesis is the semilinear wave equation. We naturally consider it in a second-order formulation. Hence, we present the equation and all imposed assumptions on the data in this formulation. Since we prove a wellposedness result with the aid of semigroup theory, we reformulate the equation afterwards as a first-order system. In addition, all methods we propose for the time integration are applied to this formulation and we therefore also translate the assumptions into this setting.

2.1.1 Second-order formulation

Let $\mathcal{H}$ be a real, separable Hilbert space and $L\colon \mathcal{D}(L) \subseteq \mathcal{H} \to \mathcal{H}$ be a strictly positive, self-adjoint operator. By [67, Prop. 5.13] we define $L^{1/2}$ as the unique, strictly positive, self-adjoint operator that satisfies $L^{1/2}L^{1/2} = L$ and may hence introduce the intermediate space

$$V = \mathcal{D}(L^{1/2}) \quad \text{with} \quad \mathcal{D}(L) \hookrightarrow V \hookrightarrow \mathcal{H}, \qquad \|v\|_V = \left\|L^{1/2}v\right\|_{\mathcal{H}}\,, \tag{2.1}$$

with dense and continuous embeddings. In particular, we assume the existence of a constant C_{emb} such that the following bounds hold

$$\|v\|_{\mathcal{H}} \leq C_{\text{emb}} \|v\|_V\,,\ v \in V, \qquad \|q\|_V \leq C_{\text{emb}} \|q\|_{\mathcal{D}(L)}\,,\ q \in \mathcal{D}(L)\,. \tag{2.2}$$

In the first part of this thesis we consider the abstract second-order evolution equation

$$q''(t) = -Lq(t) + G(t, q(t)), \quad t \in [0, t_{\text{def}}], \qquad q(0) = q_0, \quad q'(0) = q_0', \tag{2.3}$$

in $\mathcal{H}$ and use the spaces above to reformulate it as a first-order system. In order to illustrate the abstract framework considered in the rest of the first part, we present a class of examples of semilinear wave equations.

Example 2.5 ([9, Example 3.1]). *We consider the semilinear evolution equation (2.3) in the following setting:*

(a) $\emptyset \neq \Omega \subseteq \mathbb{R}^d$ *is a convex, bounded Lipschitz domain with* $d \in \{1,2,3\}$.

(b) $L = -\operatorname{div}(\mathbf{A}\nabla)$ *with uniformly positive definite* $\mathbf{A} \in L^\infty(\Omega)^{d\times d}$.

(c) *For* $g\colon [0,t_{\mathrm{def}}] \times \Omega \times \mathbb{R} \to \mathbb{R}$ *there is some* $\alpha = (\alpha_t, \alpha_x, \alpha_y) \in \mathbb{N}^3$ *such that all partial derivatives* $\partial^\beta g$, $\beta \leq \alpha$, *exist, are continuous in* t *and* y *and bounded in* x.

(d) *There is* $\gamma > 1$ *and a constant* $C_g > 0$ *such that for all* $(t,x,y) \in [0,t_{\mathrm{def}}] \times \Omega \times \mathbb{R}$ *we have*

$$|g(t,x,y)|, |\partial_t g(t,x,y)| \leq C_g\big(1+|y|^\gamma\big)\,,$$
$$|\partial_y g(t,x,y)| \leq C_g\big(1+|y|^{\gamma-1}\big)\,.$$

For the corrected Lie Splitting (3.14) we assume in addition

$$|\partial_{yy} g(t,x,y)| \leq C_g\big(1+|y|^{\gamma-1}\big)\,.$$

For $(t,x) \in [0,t_{\mathrm{def}}] \times \Omega$ *and* $q \in V$ *we define*

$$G(t,q)(x) \coloneqq g(t,x,q(x)).$$

The most common examples fitting into this framework, are in $d = 1,2,3$ *on* $\mathcal{H} = L^2(\Omega)$ *the cubic wave equation with* $L = -\Delta$ *and* $g(q) = q^3$ *or the Sine-Gordon equation with* $g(q) = \sin(q)$. *In Table 4.1, we provide a detailed list of criteria such that the error analysis presented in this part of the thesis can be conducted.*

2.1.2 First-order system

In order to prove a wellposedness result for (2.3) we formulate it as the first-order system

$$u'(t) = Au(t) + f(t,u(t)), \qquad u = \begin{pmatrix} q \\ v \end{pmatrix}, \tag{2.4}$$

with

$$A = \begin{pmatrix} 0 & I \\ -L & 0 \end{pmatrix}, \qquad f(t,u) = \begin{pmatrix} 0 \\ G(t,q) \end{pmatrix}, \tag{2.5}$$

on the separable Hilbert space $X = V \times \mathcal{H}$ with inner product

$$\langle u_1, u_2 \rangle_X = \langle q_1, q_2 \rangle_V + \langle v_1, v_2 \rangle_{\mathcal{H}}\,.$$

The wave operator A is given on its domain $\mathcal{D}(A) = \mathcal{D}(L) \times V$ which allows for the following properties. The first one, see for example [19, Section VI.3.c], is crucial for the wellposedness of (2.4) as we see in the next section.

Lemma 2.6. *The operator* $A\colon \mathcal{D}(A) \to X$ *is skew adjoint.*

Proof. It is clear that for $u \in \mathcal{D}(A)$ it holds

$$\langle Au, u \rangle_X = -\langle u, Au \rangle_X\,,$$

i.e., A is skew-symmetric. Hence, it suffices to prove that $I \pm A : \mathcal{D}(A) \to X$ is surjective, see [73, Satz VII.2.8]. Given any $y = \big(\, y_1, y_2 \,\big)^T \in X$ we seek $u = \big(\, q, v \,\big)^T \in \mathcal{D}(A)$ such that

$$\big(I + A\big)u = y \quad \iff \quad \begin{pmatrix} q + v \\ v - Lq \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

holds. Substituting the first equation in the second gives

$$\big(I + L\big)q = y_1 - y_2 \in \mathcal{H}\,,$$

which has a solution $q \in \mathcal{D}(L)$. Setting $v = y_1 - q \in V$ gives surjectivity for $I + A$. The case $I - A$ is fully analogous. □

In many cases the embeddings in (2.1) are compact maps which makes the later appearing functional calculus more intuitive. In the next result we explain how this transfers to the first-order formulation, cf. [65, Lemma 9.20].

Lemma 2.7. *If the embeddings in (2.1) are compact, A has a compact resolvent.*

Proof. Take any sequences $(u_n)_n$ in $\mathcal{D}(A)$ and $(y_n)_n$ in X with

$$Au_n = y_n \quad \text{and} \quad \|y_n\|_X \le C \quad \text{for all } n \in \mathbb{N}\,.$$

We then have to prove that there exists a converging subsequence $(u_{n_j})_j$ in X. Looking at the single components we have

$$u_n = \begin{pmatrix} q_n \\ v_n \end{pmatrix}, \quad y_n = \begin{pmatrix} y_{1,n} \\ y_{2,n} \end{pmatrix}, \quad \begin{pmatrix} v_n \\ -Lq_n \end{pmatrix} = \begin{pmatrix} y_{1,n} \\ y_{2,n} \end{pmatrix}.$$

Since we have $\|v_n\|_V = \|y_{1,n}\|_V \le C$, there is a subsequence converging in $\mathcal{H}$ and hence there is some $v \in \mathcal{H}$ such that

$$v_{n_j} \to v \text{ in } \mathcal{H}, \quad j \to \infty\,,$$

holds. Further, by $\big\|q_{n_j}\big\|_{\mathcal{D}(L)} = \big\|y_{2,n_j}\big\|_{\mathcal{H}} \le C$ and the compact embedding into V we may extract another converging subsequence and conclude the assertion. □

In the error analysis we also prove convergence in the stronger norm

$$\|u\|_{\mathcal{D}(A)} \coloneqq \|Au\|_X\,.$$

The following lemma states that $\|\cdot\|_{\mathcal{D}(A)}$ is indeed a norm and is equivalent to the standard graph norm.

Lemma 2.8. *For $u \in \mathcal{D}(A)$ and C_{emb} from (2.2) it holds*

$$\|u\|_X \le C_{emb}\,\|Au\|_X\,.$$

Proof. For $u = \big(\, q, v \,\big)^T \in \mathcal{D}(A)$ we have by (2.2)

$$\|u\|_X^2 = \|q\|_V^2 + \|v\|_{\mathcal{H}}^2 \le C_{\mathrm{emb}}^2\big(\|q\|_{\mathcal{D}(L)}^2 + \|v\|_V^2\big) = C_{\mathrm{emb}}^2\,\|u\|_{\mathcal{D}(A)}^2\,,$$

and hence the assertion is shown. □

2.2 Wellposedness results

In this section we consider the theory of linear and semilinear evolution equations. To simplify the presentation, we restrict ourselves to wave-type equations in Hilbert spaces. All results can be found in the monographs [19, 58, 63].

2.2.1 Linear, inhomogeneous evolution equations

A preliminary step towards the semilinear equations studied in the numerical analysis is the careful treatment of the inhomogeneous evolution equation

$$u'(t) = Au(t) + f(t), \qquad u(0) = u_0\,, \tag{2.6}$$

which is done in two steps.

The homogeneous case

In the first part we investigate the homogeneous evolution equation

$$u'(t) = Au(t), \qquad u(0) = u_0\,, \tag{2.7}$$

in some Hilbert space X. If X is finite dimensional and $A \in \mathbb{C}^{n\times n}$, it is well-known that the solution of (2.7) is given by

$$u(t) = e^{tA}u_0, \qquad e^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!}A^k\,, \tag{2.8}$$

and that the series is absolutely convergent by the boundedness of A. Obviously, this construction cannot be done in the case of an unbounded operator A. In the following we will generalize the solution theory of (2.7) to this case.

Definition 2.9. *Consider a family $\big(T(t)\big)_{t\geq 0}$ of bounded linear operators on X. We call $\big(T(t)\big)_{t\geq 0}$ a strongly continuous semigroup or C_0-semigroup if the following properties are satisfied:*

(a) $T(0) = I$,

(b) $T(t)T(s) = T(t+s)$ *for all* $t, s \geq 0$,

(c) $\lim\limits_{t\to 0} T(t)x = x$ *for all* $x \in X$.

It is easily verified that the family e^{tA} from (2.8) satisfies all conditions from Definition 2.9. Given this family we can get back the matrix A by the representation

$$A = \frac{d}{dt}e^{tA}\Big|_{t=0}.$$

This can be used in order to extract an operator from any given semigroup which is called the generator.

Definition 2.10. *Let $\big(T(t)\big)_{t\geq 0}$ be a C_0-semigroup on a Hilbert space X. We define the set*

$$\mathcal{D}(A) := \{x \in X \mid \lim_{h\to 0^+} \tfrac{1}{h}\big(T(h)x - x\big) \text{ exists in } X\}$$

and define for $x \in \mathcal{D}(A)$ the operator

$$A\colon \mathcal{D}(A) \to X, \quad x \mapsto \lim_{h\to 0^+} \tfrac{1}{h}\big(T(h)x - x\big)\,,$$

called the infinitesimal generator of the semigroup $\big(T(t)\big)_{t\geq 0}$.

We present some useful results on semigroups and their generators, see, e.g., [19, Chapter II].

Proposition 2.11. *Let $(T(t))_{t\geq 0}$ be a C_0-semigroup on a Hilbert space X and $A\colon \mathcal{D}(A) \to X$ its generator. Then the following assertions hold:*

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

(b) *The map $t \mapsto T(t)x$ is continuous from $[0,\infty)$ to X for all $x \in X$.*

(c) *There exist $M \geq 1$ and $\omega \in \mathbb{R}$ such that $\|T(t)\|_{X\leftarrow X} \leq Me^{\omega t}$.*

(d) *For $x \in \mathcal{D}(A)$ it holds $T(t)x \in \mathcal{D}(A)$ and*

$$\frac{d}{dt}T(t)x = AT(t)x = T(t)Ax, \quad t \geq 0\,.$$

From the last point it is clear that for $u_0 \in \mathcal{D}(A)$ the function

$$u(t) \coloneqq T(t)u_0$$

satisfies $u \in C^1([0,\infty), X) \cap C([0,\infty), \mathcal{D}(A))$ and solves (2.7).

Proposition 2.12. *Let $(T(t))_{t\geq 0}$ be a C_0-semigroup on a Hilbert space X and $A\colon \mathcal{D}(A) \to X$ its generator. Further, let $M \geq 1$ and $\omega \in \mathbb{R}$ such that $\|T(t)\|_{X\leftarrow X} \leq Me^{\omega t}$ holds.*

If $\operatorname{Re}\lambda > \omega$, then λ is in the resolvent set $\rho(A)$ and it holds

$$\left\|(\lambda I - A)^{-n}\right\|_{X\leftarrow X} \leq \frac{M}{(\operatorname{Re}\lambda - \omega)^n}, \qquad n \geq 1\,.$$

With this proposition, we can prove one of the most important theorems going back to Hille and Yosida which gives a characterization whether an operator is a generator of a semigroup.

Theorem 2.13 (Hille-Yosida)**.** *Let $A\colon \mathcal{D}(A) \to X$ be a linear operator in a Hilbert space X and take $M \geq 1$ and $\omega \in \mathbb{R}$. Then A is the generator of a semigroup satisfying $\|T(t)\|_{X\leftarrow X} \leq Me^{\omega t}$ if and only if the following is satisfied:*

(a) *A is closed and $\mathcal{D}(A)$ is dense in X*

(b) *For any λ with $\operatorname{Re}\lambda > \omega$ it holds $\lambda \in \rho(A)$ and*

$$\left\|(\lambda I - A)^{-n}\right\|_{X\leftarrow X} \leq \frac{M}{(\operatorname{Re}\lambda - \omega)^n}, \qquad n \geq 1\,.$$

We further mention one important subclass of strongly continuous semigroups, the so-called C_0-groups. We obtain them if we replace in Definition 2.9 $t \geq 0$ by $t \in \mathbb{R}$ and $t, s \geq 0$ by $t, s \in \mathbb{R}$. A C_0-group is called unitary if one can choose $M = 1$ and $\omega = 0$. An example for such an generator is the wave operator in (2.5) as can be seen from the next result.

Theorem 2.14 (Stone)**.** *Let $A\colon \mathcal{D}(A) \to X$ be a linear, densely defined operator. Then A generates a unitary C_0-group if and only if A is skew adjoint.*

The inhomogeneous case

In the next section we turn to the inhomogeneous evolution equation (2.6) and assume throughout that A is the generator of a strongly continuous semigroup and change the notation to $T(t) = e^{tA}$. The main goal of this section is to find solutions to (2.6) and we hence clarify what a solution is.

Definition 2.15. (a) *We call u a classical solution of (2.6) on $[0, t^*)$ if u solves (2.6), $u(0) = u_0$, and*

$$u \in C^1([0, t_{\text{end}}], X) \cap C([0, t_{\text{end}}], \mathcal{D}(A)) \tag{2.9}$$

for any $t_{\text{end}} < t^*$.

(b) *Let $f \in C([0, t_{\text{def}}], X)$, then the function $u \in C([0, t_{\text{def}}], X)$ defined by*

$$u(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A} f(s)\, ds \tag{2.10}$$

is called a mild solution of (2.6). One often refers to (2.10) as the variation-of-constants formula.

We note that every classical solution of (2.6) is also a mild solution and, since the mild solution is uniquely defined by (2.10), classical solutions must be unique as well. On the other hand, if there exists a classical solution it must be given by (2.10). Hence, we need to study whether a mild solution is also a classical one. The answer can be given in terms of the regularity of f and u_0.

Proposition 2.16. *Let $u \in C([0, t_{\text{def}}], X)$ be the mild solution of (2.6). If $u_0 \in \mathcal{D}(A)$ and one of the two conditions*

$$(a)\ f \in C^1([0, t_{\text{def}}], X)\,, \qquad (b)\ f \in C([0, t_{\text{def}}], \mathcal{D}(A))\,,$$

is satisfied, then u is the classical solution of (2.6).

2.2.2 Semilinear evolution equations

We now turn to the actual equation of interest given by the semilinear evolution equation

$$u'(t) = Au(t) + f(t, u(t)), \qquad u(0) = u_0\,, \tag{2.11}$$

where A is the generator of a strongly continuous semigroup. The final task of this section is to give sufficient conditions on f in order to obtain a classical solution of (2.11) in the sense of Definition 2.15. We start with a crucial observation. Assume that we have a solution $u \in C^1([0, t_{\text{end}}], X)$ and define

$$g(t) \coloneqq f(t, u(t))\,.$$

Then u is also the solution of the inhomogeneous problem (2.6) with f replaced by g and hence we obtain by (2.10) the variation-of-constants formula in the form

$$u(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A} f(s, u(s))\, ds\,. \tag{2.12}$$

We again denote a function $u \in C([0, t_{\text{end}}], X)$ satisfying (2.12) a mild solution to (2.11). As before we see that every classical solution of (2.11) also is a mild solution. We first state a result [63, Thm. 6.1.2, 6.1.4] that guarantees the existence of mild solutions and close with the final theorem on the existence of classical solutions.

Theorem 2.17. *Let $A\colon \mathcal{D}(A) \to X$ be the generator of a strongly continuous semigroup and let $u_0 \in X$.*

(a) *If $f : [0, t_{\text{def}}] \times X \to X$ is globally Lipschitz continuous, then (2.11) has a unique mild solution on $[0, t_{\text{def}}]$.*

(b) *If $f : [0, t_{\text{def}}] \times X \to X$ is locally Lipschitz continuous, then there is a $t^* \leq t_{\text{def}}$ such that (2.11) has a unique mild solution on $[0, t^*)$.*

The proof is performed via a fixed-point argument using the representation in (2.12). This theorem directly implies the uniqueness of the classical solution. Hence, as before it remains to decide if the mild solution also is a classical one.

Theorem 2.18 ([63, Thm. 6.1.5])**.** *Let $A\colon \mathcal{D}(A) \to X$ be the generator of a strongly continuous semigroup and let $u_0 \in \mathcal{D}(A)$. Further, let $f \in C^1([0, t_{\text{def}}] \times X, X)$ be locally Lipschitz continuous. Then the mild solution of (2.12) on $[0, t^*)$ is also a classical solution. Hence, for every $0 < t_{\text{end}} < t^*$ there exists a constant $K > 0$ with*

$$\max \left\{ \|Au(t)\|_X , \|u'(t)\|_X \right\} \leq K, \qquad t \in [0, t_{\text{end}}]. \tag{2.13}$$

*In the following we refer to (2.13) as the **generalized finite-energy condition**.*

Remark 2.19. *The addition **generalized** is due to the fact that the **finite-energy condition** refers in the literature to quantities of the form*

$$\left\| \nabla q(t) \right\|_{L^2}^2 + \|q'(t)\|_{L^2}^2 \leq K^2 ,$$

which is only a special case of our framework. We comment on this in Section 4.2.2.

2.3 Functional calculus for skew-adjoint operators on Hilbert spaces

In this section we sketch the construction of a functional calculus which we need later for the construction of the filters. We briefly explain the finite dimensional case to illustrate a general functional calculus.

We follow the monograph [67] and first present the case of a continuous spectrum and afterwards explain the simplification in the case of a compact resolvent.

2.3.1 Example: The finite-dimensional case

As a first step we treat the finite dimensional case of symmetric and skew-symmetric matrices. For such a matrix B we have a decomposition with a unitary matrix U of the form

$$B = UDU^H , \qquad D = \operatorname{diag}\big(\lambda_1, \dots, \lambda_n\big) ,$$

where $\lambda_i \in \sigma(B)$ are the eigenvalues of B. Now given any function f that is defined on the spectrum $\sigma(B)$, we may define

$$f(B) \coloneqq U f(D) U^H , \quad f(D) \coloneqq \operatorname{diag}\big(f(\lambda_1), \dots, f(\lambda_n)\big) .$$

Hence, one can think of this functional calculus as a manipulation of the spectrum. However, such a decomposition is usually not available in the infinite dimensional case and we need a more involved theory in order to construct functions applied to operators.

2.3.2 The general case

In the following we present a functional calculus for skew-adjoint operators in Hilbert spaces. Such operators have their spectrum on the imaginary axis. Therefore, we may restrict ourselves to the function space

$$\mathcal{C}_b(i\mathbb{R}) \coloneqq \{h\colon i\mathbb{R} \to \mathbb{C} \mid h \text{ is continuous and } \|h\|_\infty < \infty\}.$$

One can actually treat a larger class of functions, but we avoid these technicalities since the given set is sufficient for our purposes.

The construction of the functional calculus is also based on a decomposition of the skew-adjoint operator A given by so-called spectral measures, see for example [67, Chapter 4]. We will not further comment on this, but only state the most important properties. To this end we need the two functions defined for $z \in i\mathbb{R}$

$$\mathbb{1}(z) = 1, \qquad r_\lambda(z) = \frac{1}{\lambda - z}, \ \lambda \in \mathbb{C} \setminus i\mathbb{R},$$

which both lie in $\mathcal{C}_b(i\mathbb{R})$.

Theorem 2.20. *Let $A\colon \mathcal{D}(A) \to X$ be a skew-adjoint operator on a separable Hilbert space X. Then there is a map*

$$\Psi_A\colon \mathcal{C}_b(i\mathbb{R}) \to \mathcal{L}\big(X\big), \qquad h \mapsto h(A),$$

which satisfies the following properties for $g, h \in \mathcal{C}_b(i\mathbb{R})$:

(a) *Ψ_A is linear,*
(b) *$\mathbb{1}(A) = I$ and $r_\lambda(A) = \big(\lambda - A\big)^{-1}$ for $\lambda \in \mathbb{C} \setminus i\mathbb{R}$,*
(c) *$\|h(A)\|_{X \leftarrow X} \leq \|h\|_\infty$,*
(d) *$(gh)(A) = g(A)h(A)$,*
(e) *For $x \in \mathcal{D}(A)$ it holds $h(A)x \in \mathcal{D}(A)$ and $Ah(A)x = h(A)Ax$.*
(f) *If $\widetilde{h}\colon z \mapsto zh(z) \in \mathcal{C}_b(i\mathbb{R})$, then for any $x \in X$ it holds*

$$h(A)x \in \mathcal{D}(A), \qquad \widetilde{h}(A)x = Ah(A)x.$$

Proof. All statements can be found in [67, Theorem 5.9] for the case of self-adjoint operators where one might neglect the closure of the operators as they are already closed by our restriction to bounded functions. In order to obtain the skew-adjoint case we simply consider the self-adjoint operator

$$B \coloneqq -iA\colon \mathcal{D}(A) \to X$$

and use the functional calculus $\Psi_B\colon \mathcal{C}_b(\mathbb{R}) \to \mathcal{L}\big(X\big)$ for unbounded, self-adjoint operators. We obtain the desired functional calculus by setting for $h \in \mathcal{C}_b(i\mathbb{R})$

$$\widehat{h}(z) \coloneqq h(iz), \ z \in \mathbb{R}, \qquad h(A) \coloneqq \widehat{h}(B)$$

since $\widehat{h} \in \mathcal{C}_b(\mathbb{R})$ holds. This gives the functional calculus for unbounded, skew-adjoint operators. □

In the special case of an analytic function h satisfying $h(z) = h(-z)$ and the wave operator A given in (2.5) a formal computation would lead to a block diagonal operator $h(A)$, for example using a power series expansion of h. As this property is needed in the proofs later, we will confirm this property in the following proposition.

Proposition 2.21. *Let $h \in \mathcal{C}_b(i\mathbb{R})$ be an even function, i.e., $h(z) = h(-z)$ for $z \in i\mathbb{R}$, and assume that*

$$\lim_{x \to \pm\infty} h(ix) = 0$$

holds. Further, consider the wave operator from (2.5). Then, for the projections π_i, $i \in \{1, 2\}$, onto the i-th component and $x \in X$ it holds

$$\pi_i x = 0 \quad \textit{implies} \quad \pi_i h(A) x = 0\,.$$

Proof. We prove the assertion by an approximation argument. By [70, Section 1.6] we find a sequence of even rational functions $(h_n)_n$ that convergence uniformly on $i\mathbb{R}$ to h. Hence, by the continuity of π_i it is sufficient to prove the result only for the functions h_n. Now fix $n \in \mathbb{N}$, and since h tends to zero at infinity we may decompose h_n as a finite product of functions of the type

$$\phi(z) = \frac{\alpha + \beta z^2}{\gamma - z^2}, \quad z \in i\mathbb{R}\,.$$

A direct calculation gives the assertion for ϕ and iteratively for the product h_n which closes the proof. □

2.3.3 Case of a compact resolvent

If the embeddings in (2.1) are compact, we deduced in Lemma 2.7 that A has a compact resolvent. Hence, the spectral theorem yields that A admits an orthonormal basis of eigenvectors

$$(\phi_k)_{k \in M}, \quad A\phi_k = i\lambda_k \phi_k, \quad \phi_k \in \bigcap_{j \in \mathbb{N}} \mathcal{D}(A^j)\,,$$

where $M \subseteq \mathbb{N}$ and $\lambda_k \in \mathbb{R}$. Any $x \in X$ can thus be represented as

$$x = \sum_{k \in M} \alpha_k \phi_k, \quad \alpha_k = \langle x, \phi_k \rangle_X\,,$$

with the equivalence

$$x \in \mathcal{D}(A) \iff \sum_{k \in M} |\lambda_k \alpha_k|^2 < \infty\,.$$

This enables us to define the following functional calculus on the set $\mathcal{C}_b(i\mathbb{R})$ very elegantly by

$$\Psi_A \colon \mathcal{C}_b(i\mathbb{R}) \to \mathcal{L}(X), \quad h \mapsto \begin{cases} h(A) \colon X \to X \\ x = \sum\limits_{k \in M} \alpha_k \phi_k \mapsto h(A)x = \sum\limits_{k \in M} h(i\lambda_k)\alpha_k \phi_k \end{cases},$$

which is then fully analogous to the finite dimensional case.

CHAPTER 3

Review on exponential integrators

In this chapter we recall the general idea and the construction of exponential integrators. Further, we present all the underlying methods covered in the latter error analysis. Since we are not concerned with classical integration schemes as for example Runge–Kutta methods, we do not review them, but only explain the most important differences. We will first consider methods of order one, as they illustrate the basic ideas of exponential integrators nicely. Afterwards we turn to second-order methods which will be of most interest in Chapter 4.

To begin with, we briefly recall the variation-of-constants formula already introduced in (2.12)

$$u(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A} f(s, u(s))\, ds\,.$$

which was the formula any solution of (2.4) has to satisfy. Note that the formula reduces to the fundamental theorem of calculus by setting $A = 0$ and one obtains

$$u(t) = u_0 + \int_0^t f(s, u(s))\, ds\,.$$

From this, many numerical integration schemes can be derived discretizing the integral term in a suitable way. This leads for example to Runge–Kutta or Adams–Bashforth methods. We now pursue the same idea but applied to the integral term in (2.12).

In the following we need the φ-functions which are defined for $z \in \mathbb{C}$ as

$$\varphi_{k+1}(z) \coloneqq \int_0^1 e^{(1-s)z} \frac{s^k}{k!}\, ds, \qquad k \geq 0\,. \tag{3.1}$$

With the definition $\varphi_0(z) = e^z$ they also satisfy the recursion

$$\varphi_{k+1}(z) = \frac{1}{z}\Big(\varphi_k(z) - \frac{1}{k!}\Big), \qquad z \neq 0,\ k \geq 0\,. \tag{3.2}$$

3.1 Methods of order one

We now explain the two simplest ways to construct a numerical integrator from (2.12). To this end fix some stepsize $\tau > 0$ and replace t by τ in the variation-of-constants formula. The first idea is to freeze f in the integral at $(0, u_0)$ and to obtain the approximation

$$\begin{aligned} u(\tau) &= e^{\tau A}u_0 + \int_0^\tau e^{(\tau-s)A} f(s, u(s))\, ds \\ &\approx e^{\tau A}u_0 + \int_0^\tau e^{(\tau-s)A} f(0, u(0))\, ds \\ &= e^{\tau A}u_0 + \tau\varphi_1(\tau A) f(0, u_0)\,, \end{aligned} \tag{3.3}$$

with φ_1 from (3.1). Using the notation $t_n = n\tau$ and $f_n = f\big(t_n, u_n\big)$, the idea in (3.3) leads to the exponential Euler method

$$u_{n+1} = e^{\tau A}u_n + \tau\varphi_1(\tau A) f_n\,, \tag{3.4}$$

from which we already observe some properties of exponential integrators. The coefficients of the method are analytic functions evaluated at the operator τA. This means that the unbounded part of the evolution equation is incorporated in the numerical scheme and the approximation mainly takes place in the „nice“ part of (2.12). In particular, we can see from the calculations in (3.3) that the method (3.4) is exact if f is constant.

The second possible choice is a weakened version of the idea above and tends somehow more in the direction of the classical methods that use the fundamental theorem of calculus. It starts with the variation-of-constants formula (2.12) but approximates the whole integrand. One example is to do this at the left boundary as it is done for the explicit Euler method which results in

$$\begin{aligned} u(\tau) &\approx e^{\tau A}u_0 + \int_0^\tau e^{\tau A} f(0, u_0)\, ds \\ &= e^{\tau A}u_0 + \tau e^{\tau A} f(0, u_0)\,. \end{aligned}$$

We obtain the following method which we call the Lie Splitting

$$u_{n+1} = e^{\tau A}u_n + \tau e^{\tau A} f_n = e^{\tau A}\big(u_n + \tau f_n\big)\,. \tag{3.5}$$

The name is motivated from the second representation in (3.5). We mention that this method is in general not exact for constant f.

These two methods are the simplest exponential integrators and are of stiff order 1 if applied to a sufficiently smooth solution. One can view them as prototypes of how to construct exponential integrators from (2.12).

3.2 Methods of order two

There is a rich literature on how to construct also higher order schemes from the presented first-order method. But since we want to conduct the error analysis only with respect to data, we restrict ourselves

to schemes of order less or equal 2. Whereas the methods of order 1 are applicable to general first-order systems, we now consider the first-order equation (2.4) only with the special structure given in (2.5). This is essential since we make repeated use of the maps $\pi_i \colon X \to X$, $i = 1, 2$, which are the projection onto the i-th component of the product space X. We present these methods in the following.

3.2.1 A general class of second-order exponential methods

Since we will consider a whole class of second-order methods in this thesis, we cast them in the following abstract formulation and exemplify it in the next sections. The coefficients $a, \widehat{B}_i, B_i$ are elements of $\mathcal{C}_b(i\mathbb{R})$ which will also be specified below for the different methods. For a node $c_2 \in (0, 1]$ using the additional notation $t_{n+\xi} = t_n + \xi\tau$ and $f_{n+c_2} = f\big(t_{n+c_2}, U_n\big)$ we consider the schemes in the explicit formulation

$$\begin{aligned} U_n &= e^{c_2\tau A}u_n + c_2\tau a(c_2\tau A)f_n\,, \\ u_{n+1} &= e^{\tau A}u_n + \tau\Big(\pi_1\big(\widehat{B}_1(\tau A)f_n + \widehat{B}_2(\tau A)f_{n+c_2}\big) + \pi_2\big(B_1(\tau A)f_n + B_2(\tau A)f_{n+c_2}\big)\Big)\,, \end{aligned} \tag{3.6}$$

or in the implicit formulation

$$\begin{aligned} U_n &= e^{c_2\tau A}u_n + c_2\tau a(c_2\tau A)f_{n+c_2}\,, \\ u_{n+1} &= e^{\tau A}u_n + \tau\Big(\pi_1\big(\widehat{B}_1(\tau A)f_n + \widehat{B}_2(\tau A)f_{n+c_2}\big) + \pi_2\big(B_1(\tau A)f_n + B_2(\tau A)f_{n+c_2}\big)\Big)\,. \end{aligned} \tag{3.7}$$

We assume that the coefficients satisfy the conditions

$$\begin{aligned} a(z) &= a_0 + z a_1(z)\,, \\ \widehat{B}_1(z) + \widehat{B}_2(z) &= \varphi_1(z) + z^2\,\widehat{\rho}(z)\,, \\ B_1(z) + B_2(z) &= \varphi_1(z) + z^2\,\rho(z)\,, \\ c_2\widehat{B}_2(0) = c_2 B_2(0) &= \tfrac{1}{2}\,, \end{aligned} \tag{3.8}$$

where also $a_1, \widehat{\rho}, \rho \in \mathcal{C}_b(i\mathbb{R})$. As we will see below, the conditions in (3.8) can lead to second-order error bounds. Further, the class of methods with $\widehat{\rho} = \rho = 0$ can be treated differently in the error analysis.

We now proceed and investigate the range of application of the schemes (3.6) and (3.7).

Exponential Runge–Kutta schemes We first consider general two-stage exponential Runge–Kutta methods. They are of the form

$$\begin{aligned} U_n &= e^{c_2\tau A}u_n + c_2\tau\varphi_1(c_2\tau A)f_n\,, \\ u_{n+1} &= e^{\tau A}u_n + \tau\Big(b_1(\tau A)f_n + b_2(\tau A)f_{n+c_2}\Big)\,, \end{aligned} \tag{3.9}$$

and are obtained from (3.6) letting $\widehat{B}_i = B_i = b_i$, $i = 1, 2$ and $a(z) = \varphi_1(z)$. If the coefficient functions b_1, b_2 satisfy

$$b_1(z) + b_2(z) = \varphi_1(z), \qquad c_2 b_2(0) \quad = \tfrac{1}{2}\,,$$

Hochbruck and Ostermann showed that the method is second-order convergent for parabolic problems, see [39, Theorem 4.3.]. Popular choices are $c_2 = \frac{1}{2}$, $b_1 = 0$ or $c_2 = 1$, $b_2(z) = \varphi_2(z)$.

The symmetric, but implicit exponential Runge–Kutta scheme from Celledoni, Cohen and Owren [15, Example 2.1]

$$\begin{aligned} U_n &= e^{1/2\tau A}u_n + \tfrac{\tau}{2}\varphi_1(\tfrac{\tau}{2}A)f_{n+1/2}\,, \\ u_{n+1} &= e^{\tau A}u_n + \tau\varphi_1(\tau A)f_{n+1/2}\,, \end{aligned}$$

is covered by (3.7) with $a(z) = \varphi_1(z)$, $b_1 = 0$ and $b_2(z) = \varphi_1(z)$. Obviously, both schemes satisfy (3.8). We note that those schemes are the natural generalization of the exponential Euler method (3.4).

Lawson methods A variant of the above-mentioned exponential Runge–Kutta schemes are the Lawson methods which can be obtained by a transformation of variables and applying a standard Runge–Kutta scheme. Hochbruck, Leibold and Ostermann presented a convergence analysis in [45]. We only present the methods that are of second order which take the form

$$\begin{aligned} U_n &= e^{c_2\tau A}u_n + c_2\tau e^{c_2\tau A}f_n\,, \\ u_{n+1} &= e^{\tau A}u_n + \tau\Big((1-\tfrac{1}{2c_2})e^{\tau A}f_n + \tfrac{1}{2c_2}e^{(1-c_2)\tau A}f_{n+c_2}\Big)\,. \end{aligned} \tag{3.10}$$

Note that they can also be seen as a generalization of the method (3.5) where we applied a quadrature formula to the integral term in (2.12). We have $a(z) = e^z$ and use Taylor expansion on the coefficients $\widehat{B}_i = B_i$ to obtain

$$\begin{aligned} B_1(z) &= (1-\tfrac{1}{2c_2})e^z = (1-\tfrac{1}{2c_2}) + (1-\tfrac{1}{2c_2})z + \mathcal{O}(z^2)\,, \\ B_2(z) &= \tfrac{1}{2c_2}e^{(1-c_2)z} = \tfrac{1}{2c_2} + \tfrac{1-c_2}{2c_2}z + \mathcal{O}(z^2)\,, \\ \varphi_1(z) &= 1 + \tfrac{1}{2}z + \mathcal{O}(z^2)\,. \end{aligned} \tag{3.11}$$

Thus, (3.8) is valid for any $c_2 \in (0, 1]$.

Strang splitting Another famous example fitting in the general framework is the Strang splitting applied to the first-order system (2.4) coming from the second-order formulation (2.3). It is based on the following decomposition. The exact flows φ_τ^A and φ_τ^f of the two subproblems

$$\begin{pmatrix} t' \\ u' \end{pmatrix} = \begin{pmatrix} 1 \\ Au \end{pmatrix}, \qquad \begin{pmatrix} t' \\ u' \end{pmatrix} = \begin{pmatrix} 0 \\ f(t,u) \end{pmatrix},$$

are by the special form of f in (2.5) given explicitly by

$$\varphi_\tau^A \begin{pmatrix} t_0 \\ u_0 \end{pmatrix} = \begin{pmatrix} t_0 + \tau \\ e^{\tau A}u_0 \end{pmatrix}, \qquad \varphi_\tau^f \begin{pmatrix} t_0 \\ u_0 \end{pmatrix} = \begin{pmatrix} t_0 \\ u_0 + \tau f(t_0, u_0) \end{pmatrix}.$$

We consider the Strang splitting in the variants (A, f, A) and (f, A, f) given by

$$\begin{pmatrix} t_{n+1} \\ u_{n+1} \end{pmatrix} = \varphi_{\tau/2}^A \circ \varphi_\tau^f \circ \varphi_{\tau/2}^A \begin{pmatrix} t_n \\ u_n \end{pmatrix}, \tag{3.12a}$$

$$\begin{pmatrix} t_{n+1} \\ u_{n+1} \end{pmatrix} = \varphi_{\tau/2}^f \circ \varphi_\tau^A \circ \varphi_{\tau/2}^f \begin{pmatrix} t_n \\ u_n \end{pmatrix}, \tag{3.12b}$$

respectively. Note that the (f, A, f) variant in (3.12b) is equivalent to a trigonometric integrator without filter functions, see, e.g., [35, XIII.2.2]. For (3.12a) the coefficients are given by

$$c_2 = \frac{1}{2}, \quad a(z) = \widehat{B}_1(z) = B_1(z) = 0, \quad \widehat{B}_2(z) = B_2(z) = e^{z/2}$$

and for (3.12b) by

$$c_2 = 1, \quad a(z) = \widehat{B}_1(z) = B_1(z) = \frac{1}{2}e^z, \quad \widehat{B}_2(z) = B_2(z) = \frac{1}{2}.$$

Similar computations as in (3.11) verify the conditions in (3.8).

Extended Runge–Kutta–Nyström methods The motivation to allow for the additional degree of freedom in (3.6) and (3.7) induced by the projections π_1 and π_2 comes from the class of extended Runge–Kutta–Nyström methods. So far these methods have only been considered for ordinary differential equations. For the sake of readability we present these methods for L and A being matrices, but note that this can be made rigorous. The two-stage methods considered for example by Wang, Wu and Xia in [71, 72] for problem (2.3) are given by

$$\begin{aligned} q_{n+c_2} &= \cos(c_2\tau\Lambda)q_n + c_2\tau\,\mathrm{sinc}(c_2\tau\Lambda)v_n + \tau^2 a_{21}(\tau\Lambda)G(t_n, q_n) \\ q_{n+1} &= \cos(\tau\Lambda)q_n + \tau\,\mathrm{sinc}(\tau\Lambda)v_n + \tau^2\Big(\widehat{b}_1(\tau\Lambda)G(t_n, q_n) + \widehat{b}_2(\tau\Lambda)G(t_{n+c_2}, q_{n+c_2})\Big) \\ v_{n+1} &= -\Lambda\sin(\tau\Lambda)q_n + \cos(\tau\Lambda)v_n + \tau\Big(b_1(\tau\Lambda)G(t_n, q_n) + b_2(\tau\Lambda)G(t_{n+c_2}, q_{n+c_2})\Big), \end{aligned} \tag{3.13}$$

where $\Lambda = L^{1/2}$ is the positive definite matrix square root. The main difference to the methods covered by (3.9) is the second-order formulation which allows for different choices of b_1 and $\widehat{b}_1$ whereas they are not independent in the first-order formulation. In order to see the connection to (3.6) we need some preparation.

Analogously to the definition in (3.1) we define the ψ-functions by a parameter integral.

Definition 3.1. *For $z \in \mathbb{R}$ define $\psi_0(z) := \cos(z)$ and let for $j \geq 0$*

$$\psi_{j+1}(z) := \int_0^1 \cos((1-s)z)\frac{s^j}{j!}\,ds.$$

By construction the ψ-functions are analytic and from the definition we directly obtain several properties which include a relation to the φ-functions.

Lemma 3.2. (a) *For $z \in \mathbb{R} \setminus \{0\}$ we have*

$$\psi_1(z) = \mathrm{sinc}(z), \qquad \psi_2(z) = \frac{1-\cos(z)}{z^2}, \qquad \psi_3(z) = \frac{1-\mathrm{sinc}(z)}{z^2} = \frac{z-\sin(z)}{z^3}.$$

(b) *For $j \geq 0$ it holds*

$$\psi_{j+2}(z) = \int_0^1 \frac{\sin((1-s)z)}{z}\frac{s^j}{j!}\,ds, \qquad z \in \mathbb{R}\setminus\{0\}.$$

(c) *For $j \geq 0$ it holds $\psi_j(0) = \frac{1}{j!}$.*

(d) *We have the symmetric connection*

$$\psi_j(z) = \frac{1}{2}\Big(\varphi_j(iz) + \varphi_j(-iz)\Big), \qquad j \geq 0,\ z \in \mathbb{R}.$$

(e) *We have*

$$\psi_{j+1}(z) = \frac{\psi_{j-1}(0) - \psi_{j-1}(z)}{z^2}, \qquad j \geq 1,\ z \in \mathbb{R}\setminus\{0\}.$$

We note that since the φ_j and ψ_j are analytic functions, all assertions in the lemma remain true in the limit $z = 0$ and actually also hold for $z \in \mathbb{C}$. We do not treat this case further since we only need it for matrices with spectrum on the real axis.

Proof. The properties (a),(c) and (e) are easily verified.

(b) Let $j \geq 0$ and $z \in \mathbb{R} \setminus \{0\}$. We compute

$$\begin{aligned}\int_0^1 \frac{\sin((1-s)z)}{z}\frac{s^j}{j!}\,ds &= \left[\frac{\sin((1-s)z)}{z}\frac{s^{j+1}}{(j+1)!}\right]_0^1 + \int_0^1 \cos((1-s)z)\frac{s^{j+1}}{(j+1)!}\,ds \\ &= \left[\frac{\sin((1-s)z)}{z}\frac{s^{j+1}}{(j+1)!}\right]_0^1 + \psi_{(j+1)+1}(z) \\ &= \psi_{j+2}(z)\,.\end{aligned}$$

d) The assertion is a direct consequence of Euler's formula, i.e., for $z \in \mathbb{R}$ it holds

$$\frac{1}{2}\left(e^{(1-s)iz} + e^{-(1-s)iz}\right) = \cos((1-s)z).$$

The next lemma provides another connection of the φ- and the ψ-functions in the context of matrix functions.

Lemma 3.3. *Consider the matrix* $A = \begin{pmatrix} 0 & I \\ -\Lambda^2 & 0 \end{pmatrix}$ *for some positive definite matrix* Λ. *Then the following relations hold for* $t \in \mathbb{R}$*:*

$$e^{tA} = \begin{pmatrix} \cos(t\Lambda) & t\,\mathrm{sinc}(t\Lambda) \\ -\Lambda\sin(t\Lambda) & \cos(t\Lambda) \end{pmatrix} = \begin{pmatrix} \psi_0(t\Lambda) & t\psi_1(t\Lambda) \\ -t\Lambda^2\psi_1(t\Lambda) & \psi_0(t\Lambda) \end{pmatrix}, \tag{R0}$$

$$\varphi_1(tA) = \begin{pmatrix} \psi_1(t\Lambda) & t\,\psi_2(t\Lambda) \\ -t\Lambda^2\psi_2(t\Lambda) & \psi_1(t\Lambda) \end{pmatrix}, \tag{R1}$$

$$\varphi_2(tA) = \begin{pmatrix} \psi_2(t\Lambda) & t\,\psi_3(t\Lambda) \\ -t\Lambda^2\psi_3(t\Lambda) & \psi_2(t\Lambda) \end{pmatrix}. \tag{R2}$$

Proof. The relation in (R0) is well-known and we only verify (R1). The proof of (R2) is completely analogous.

For the φ_1-function it holds $\varphi_1(z) = \frac{e^z-1}{z}$ by (3.2) and hence for $t \neq 0$ we compute

$$\begin{aligned}\varphi_1(tA) &= \frac{1}{t}\begin{pmatrix} 0 & -\Lambda^{-2} \\ I & 0 \end{pmatrix}\begin{pmatrix} \cos(t\Lambda) - I & t\,\mathrm{sinc}(t\Lambda) \\ -\Lambda\sin(t\Lambda) & \cos(t\Lambda) - I \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{sinc}(t\Lambda) & t\,(t\Lambda)^{-2}\,(I - \cos(t\Lambda)) \\ \frac{1}{t}\,(\cos(t\Lambda) - I) & \mathrm{sinc}(t\Lambda) \end{pmatrix} \\ &= \begin{pmatrix} \psi_1(t\Lambda) & t\,\psi_2(t\Lambda) \\ -t\Lambda^2\psi_2(t\Lambda) & \psi_1(t\Lambda) \end{pmatrix}.\end{aligned}$$

Note that the representation is also valid for $t = 0$. □

method	$\widehat{b}_2$	b_2
ERKN1	$\psi_2(z)$	$\psi_0(\frac{z}{2})$
ERKN2	$\psi_2(z)$	$\psi_1(z)$
ERKN3	$\frac{1}{2}\psi_1(\frac{z}{2})$	$\cos(\frac{z}{2})$
ERKN5	$\operatorname{sinc}(z)\frac{1}{2}\psi_1(\frac{z}{2})$	$\operatorname{sinc}(z)\psi_0(\frac{z}{2})$

Table 3.1: Different methods considered in [71]. The scheme ERKN4 is excluded since it is equivalent to ERKN2. We will see in next chapter that ERKN5 can be seen as Strang (3.12a) with outer filter $\operatorname{sinhc}\colon z \mapsto \frac{\sinh(z)}{z}$ and hence is covered by our error analysis.

With these results we go back and establish the connection of (3.13) and (3.6). Since we only multiply in (3.6) with a vector f that is zero in the first component and use the projections π_i, we compute

$$\pi_1 \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} 0 \\ g \end{pmatrix} = \begin{pmatrix} a_{12} g \\ 0 \end{pmatrix}, \qquad \pi_2 \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} 0 \\ g \end{pmatrix} = \begin{pmatrix} 0 \\ a_{22} g \end{pmatrix}.$$

Hence, the choice of B_i and $\widehat{B}_i$ can be traced back to find operators that satisfy

$$\widehat{B}_i(\tau A) = \begin{pmatrix} \star & \tau\widehat{b}_i \\ \star & \star \end{pmatrix}, \qquad B_i(\tau A) = \begin{pmatrix} \star & \star \\ \star & b_i \end{pmatrix}, \qquad i = 1, 2,$$

with $\widehat{b}_i, b_i$ from (3.13). With Lemma 3.3 we obtain

$$e^{\tau A} \begin{pmatrix} 0 \\ g \end{pmatrix} = \begin{pmatrix} \tau\psi_1(\tau\Lambda) g \\ \psi_0(\tau\Lambda) g \end{pmatrix}, \qquad \varphi_1(\tau A) \begin{pmatrix} 0 \\ g \end{pmatrix} = \begin{pmatrix} \tau\psi_2(\tau\Lambda) g \\ \psi_1(\tau\Lambda) g \end{pmatrix}, \qquad \varphi_2(\tau A) \begin{pmatrix} 0 \\ g \end{pmatrix} = \begin{pmatrix} \tau\psi_3(\tau\Lambda) g \\ \psi_2(\tau\Lambda) g \end{pmatrix}.$$

such that we arrive at

$$\begin{aligned} \widehat{B}_i = \varphi_j \quad &\text{implies} \quad \widehat{b}_i = \psi_{j+1}, \qquad j = 0, 1, \\ B_i = \varphi_j \quad &\text{implies} \quad b_i = \psi_j, \qquad j = 0, 1, 2. \end{aligned}$$

Those cover the common choices in (3.13), see for example [72, Table 1]. Different choices with $c_2 = \frac{1}{2}$ and $\widehat{b}_1 = b_1 = 0$ were considered in [71], see Table 3.1. Hence, if the methods are constructed with the right φ-functions, they also satisfy the conditions (3.8).

3.2.2 Further methods

In this section we present four more methods that do not fit in the general framework of (3.6) and (3.7). Nevertheless, they can be analyzed by the same techniques as the before mentioned methods such that we can also derive error bounds for these.

Corrected Lie Splitting We consider the second-order corrected Lie splitting given by

$$u_{n+1} = e^{\tau A}\Big(u_n + \tau f\big(t_{n+1/2}, u_n\big) + \tfrac{\tau^2}{2} r_f\big(t_{n+1/2}, u_n\big)\Big) \tag{3.14}$$

with the correction term

$$r_f(t,u) \coloneqq J_f\big(t,u\big)\begin{pmatrix}0\\ Au\end{pmatrix} - Af\big(t,u\big).$$

where J_f denotes the Jacobian of f.

It is inspired by a fourth-order method of this type proposed by McLachlan and Quispel in [57, 4.9.3 (c)]. Further, note that in the linear case, where $f(t,u) = Fu$, the correction term reduces to the (linear) commutator

$$r_F(t,u) = FAu - AFu = [F,A]\,u\,.$$

Hence, one can consider (3.14) as an approximation to the method

$$u_{n+1} = e^{\tau A}e^{\tau F}e^{\frac{\tau^2}{2}[F,A]}u_n\,,$$

which was considered by Suzuki in [68, (3.37)]. As far as we know there is no convergence analysis for this method.

Exponential multistep method of Adams-type The two-step exponential multistep method from Hochbruck and Ostermann [40, (2.7)]

$$\begin{aligned} u_{n+1} &= e^{\tau A}u_n + \tau\varphi_1(\tau A)f_n + \tau\varphi_2(\tau A)\,(f_n - f_{n-1})\,, \quad n \geq 1\,,\\ u_1 &= e^{\tau A}\,(u_0 + \tau f_0)\,, \end{aligned} \tag{3.15}$$

is derived from the variation-of-constants formula for the exact solution of (2.3) by approximating the nonlinearity f in the integral term by an interpolation polynomial using the last two approximations u_{n-1}, u_n.

Exponential multistep methods of Nyström-type Similarly, we consider a method that was used by Frisch, She and Thual in [20, (B 4)], namely

$$\begin{aligned} u_{n+1} &= e^{2\tau A}u_{n-1} + 2\tau e^{\tau A}f_n\,, \quad n \geq 1\,,\\ u_1 &= e^{\tau A}\big(u_0 + \tau f_0\big)\,. \end{aligned} \tag{3.16}$$

A variant of this method is given in [20, (B 5)] called the "slaved frog". It reads

$$\begin{aligned} u_{n+1} &= e^{2\tau A}u_{n-1} + 2\tau\varphi_1(2\tau A)f_n\,, \quad n \geq 1\,,\\ u_1 &= e^{\tau A}\big(u_0 + \tau f_0\big)\,. \end{aligned} \tag{3.17}$$

For $A = 0$ both methods reduce to an explicit Nyström method, cf. method (1.13') in [34].

CHAPTER 4

Error analysis for averaged exponential integrators

We now present the core of the first part of the thesis. We first explain the main ideas and the main results and afterwards turn to a rigorous description of the framework and the error analysis.

We emphasize that most of the material is taken from [9] and is extended by additional explanations, more detailed computations and also some new results not presented elsewhere. In particular, Sections 4.5.2, 4.5.3, 4.5.4, 4.7 and some bounds in Section 4.6.2 have been added.

4.1 Informal overview of methods, concepts and results

In this section we give an informal overview of the methods of interest, the main concepts, and the main results, as it is done in [9, Sect. 2], and present the analytical framework which is necessary to formulate our results rigorously in the later sections. In order to postpone all technical difficulties, we explain everything in the finite dimensional case $\dim H < \infty$. This is not the case of interest for us, but here all the approximations presented are well-defined and the statements valid. However, the appropriate function spaces to treat evolution equations and additional assumptions necessary for the error analysis are introduced in Section 4.2. We recall the second-order equation (2.3)

$$q''(t) = -Lq(t) + G(t, q(t)), \quad t \in [0, t_{\text{def}}]$$

which is the starting point of this overview.

4.1.1 Averaged differential equation

Let L be a symmetric, positive definite matrix in $\mathbb{R}^{m\times m}$ and let $\chi = \phi, \psi \colon i\mathbb{R} \to \mathbb{R}$ be even (i.e., $\chi(-z) = \chi(z)$) and analytic functions satisfying $\chi(0) = 1$. By the theory of matrix functions we can define the filter operator

$$\widetilde{\chi} = \chi(i\tau L^{1/2})$$

and with this an averaged nonlinearity

$$\widetilde{G}(t,q) \coloneqq \widetilde{\psi} G(t, \widetilde{\phi} q)\,.$$

In order to apply them to the first order system (2.4) we enlarge the filters to the block diagonal operators

$$\Phi = \begin{pmatrix} \widetilde{\phi} & 0 \\ 0 & \widetilde{\phi} \end{pmatrix}, \qquad \Psi = \begin{pmatrix} \widetilde{\psi} & 0 \\ 0 & \widetilde{\psi} \end{pmatrix},$$

and turn our attention to the *averaged* differential equation

$$\widetilde{u}'(t) = A\widetilde{u}(t) + \widetilde{f}(t, \widetilde{u}(t)), \qquad \widetilde{f}(t,\widetilde{u}) = \Psi f(t, \Phi\widetilde{u}) = \begin{pmatrix} 0 \\ \widetilde{G}(t, \widetilde{q}) \end{pmatrix}. \tag{4.1}$$

We have to make sure that the averaging has the two following properties. On the one hand the solution $\widetilde{u}$ of (4.1) should still satisfy a generalized finite-energy condition (2.13). In Lemma 4.15 we show that the modified constant $\widetilde{K}$ is independent of τ and n. On the other hand we need a relation of the original solution u and the averaged solution $\widetilde{u}$. If we denote by $\|\cdot\|_X$ the norm induced by $\langle\cdot,\cdot\rangle_X$, we prove in Theorem 4.14

$$\|u(t) - \widetilde{u}(t)\|_X \leq C\tau^2, \qquad t \in [0, t_{\text{end}}],$$

provided that ψ, ϕ satisfy a certain set of conditions. Under some less restrictive assumptions on the filters we further establish the bound

$$\|u(t) - \widetilde{u}(t)\|_X \leq C\tau, \qquad t \in [0, t_{\text{end}}]\,.$$

Having all this at hand, the averaged solution $\widetilde{u}$ is, concerning the regularity, roughly speaking as good as the original solution u and it is sufficient to prove error bounds for numerical schemes applied to (4.1) as long as only the finite-energy condition enters in the error constant.

4.1.2 Averaged methods

As we have explained above, the averaged solution $\widetilde{u}$ inherits the essential properties of u. Hence, the averaged methods are constructed by applying any of the numerical methods in Section 3.2 to the averaged equation (4.1) instead of the original one (2.4). So taking for example the Strang splitting (3.12a), the averaged variant reads

$$\begin{pmatrix} t_{n+1} \\ u_{n+1} \end{pmatrix} = \varphi^A_{\tau/2} \circ \varphi^{\widetilde{f}}_{\tau} \circ \varphi^A_{\tau/2} \begin{pmatrix} t_n \\ u_n \end{pmatrix}, \qquad u_0 = \widetilde{u}(0) = u(0)\,. \tag{4.2}$$

Actually, this is equivalent to use a modified numerical scheme where the nonlinearity f is replaced by the averaged nonlinearity $\widetilde{f}$ in (4.1). In Figure 4.1 these different views are depicted. We emphasize that the first perspective is only needed for theoretical reasons to perform the error analysis, however it allows us to analyze many different averaged methods simultaneously. When it comes to implementation, one will simply use the method in the form (4.2).

Since we can show that the difference of (2.4) and (4.1) is of order τ^2, it is natural to use methods of order 2 in order to obtain global error bounds of the same order. While this approach would work perfectly fine in the finite dimensional case, for evolution equations this is not so clear. In fact, numerical experiments show that order reduction might be a problem.

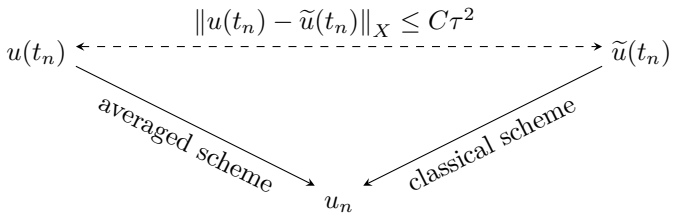

Figure 4.1: Different ways to construct an approximation u_n of the solution $u(t_n)$ of the original equation (2.4) and the solution $\widetilde{u}(t_n)$ of the averaged equations (4.1), [9, Fig. 1].

The thesis aims at rigorous error bounds of first and second order and gives a precise characterization of the framework. This includes the numerical methods, the data in terms of L and G and the choice of the filter functions.

4.1.3 Overview of results

With the ideas explained before, we prove different types of error bounds. We distinguish between error bounds in the X- and the $\mathcal{D}(A)$-norm and also between error bounds for classical and weak solutions.

Error bounds in the X-norm In Theorem 4.24 and several corollaries in Section 4.6.2 we provide the following error bounds.

(a) The Strang splitting, the exponential Runge–Kutta methods, the Lawson methods, the extended Runge–Kutta–Nyström methods and the exponential multistep methods applied to the original equation (2.4) satisfy

$$\|u(t_n) - u_n\|_X \leq C_1\tau\,.$$

(b) Using appropriate filters ϕ, ψ any method of Section 3.2 applied to the averaged equation (4.1) satisfies the bound

$$\|u(t_n) - u_n\|_X \leq C_2\tau^2\,.$$

The constants C_1, C_2 only depend on the initial value u_0, the finite energy K, properties of G, and t_{end}, but not on n and τ.

Error bounds in the $\mathcal{D}(A)$-norm Similarly, we establish error bounds in a stronger norm in Theorem 4.31 and several corollaries in Section 4.6.2.

(a) The exponential Runge–Kutta methods and the exponential multistep methods of Adams-type applied to the original equation (2.4) satisfy

$$\|u(t_n) - u_n\|_{\mathcal{D}(A)} \leq C_1\tau\,.$$

(b) Using an appropriate filter ψ, the Strang splitting, the Lawson methods, the extended Runge–Kutta–Nyström method, the corrected Lie splitting and the exponential multistep methods of Nyström-type applied to the averaged equation (4.1) satisfy the bound

$$\|u(t_n) - u_n\|_{\mathcal{D}(A)} \leq C_2\tau\,.$$

Again, the constants C_1, C_2 only depend on the initial value u_0, the finite energy K, properties of G, and t_{end}, but not on n and τ.

Error bounds for weak solutions Finally, we treat the first-order methods from Section 3.1. The key difficulty arises in only considering weak solutions of (2.4). For linear f we prove the following bound in Theorem 4.45.

Using an appropriate filter ϕ, the exponential Euler method and the Lie splitting applied to the averaged equation (4.1) satisfy the bound

$$\|u(t_n) - u_n\|_X \leq C\tau\,.$$

The constant C only depends on the initial value u_0, properties of G, and t_{end}, but not on n and τ.

Strategy All proves rely on the decomposition

$$\|u(t_n) - u_n\|_X \leq \|u(t_n) - \widetilde{u}(t_n)\|_X + \|\widetilde{u}(t_n) - u_n\|_X\,. \tag{4.3}$$

In the different scenarios we proceed in the same two steps. We first bound the term induced by the averaged equation and in the second step bound the error of the numerical method applied to (4.1).

4.1.4 Numerical example

In this section we consider one of the examples that fits in our framework, cf. Section 4.2.1, and show that one can gain something with the averaging within numerical methods. We solve a variant of the sine-Gordon equation given on the torus $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$ by

$$q''(t) = \Delta q(t) - q(t) + m_a \sin(m_i \cos(q(t)))\, q(t), \tag{4.4}$$

with $t \in [0,1]$ and $m_i, m_a \in L^\infty(\mathbb{T})$. Since one of the main difficulties in the error analysis is induced by low regularity assumptions, we construct the initial values in the following way. In order to control the regularity of the solution, we follow the approach of [45] and use a Fourier spectral method in space. We choose the Fourier coefficients for the initial values $(q_0, v_0) \in H^1(\mathbb{T}) \times L^2(\mathbb{T})$ such that

$$(q_0, v_0) \in H^1(\mathbb{T}) \times L^2(\mathbb{T}) \setminus H^{1+\epsilon}(\mathbb{T}) \times H^\epsilon(\mathbb{T})$$

holds for $\epsilon = 10^{-6}$. Although, we truncate the Fourier series for some large $N \in \mathbb{N}$ to discretize in space, the experiments in [45] show that in the limit $N \to \infty$ the $H^{1+\epsilon}(\mathbb{T}) \times H^\epsilon(\mathbb{T})$-norm is not bounded uniformly in N. By the standard semigroup theory one cannot expect to gain any regularity over time, and we are hence most likely in the situation of a solution of low regularity.

Another crucial generalization compared to the analysis in [22, 23] is that the coefficients of the right-hand side do not need to be smooth. So for example let $q \in L^2(\mathbb{T})$ and consider

$$G(q)(x) \coloneqq m_a(x) \sin(m_i(x) \cos(q))\, q\,.$$

This obviously gives $G(q)$ in $L^2(\mathbb{T})$, but we cannot improve this by additional regularity of q, i.e., that even if $q \in H^1(\mathbb{T})$ holds, there is in general no $\epsilon > 0$ with $G(q) \in H^\epsilon(\mathbb{T})$.

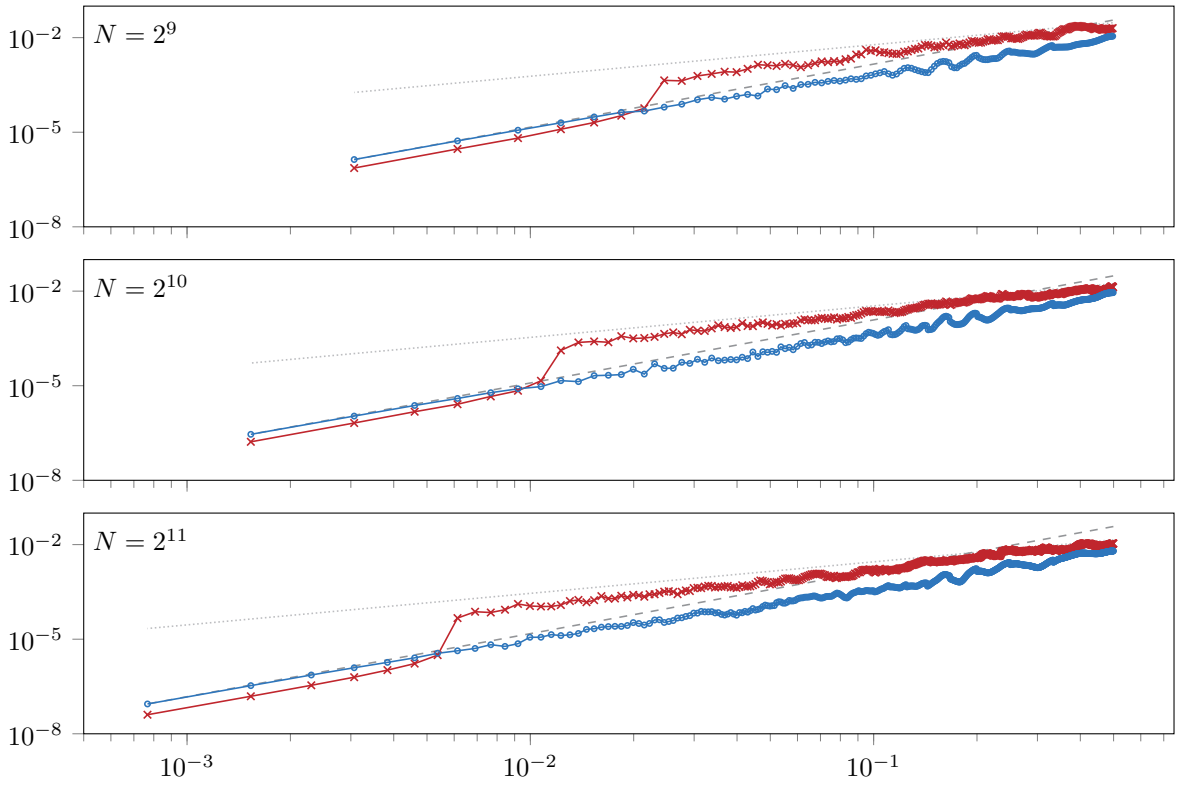

Figure 4.2: Discrete $L^\infty\Big([0,1], L^2(\mathbb{T}) \times H^{-1}(\mathbb{T})\Big)$ error (on the y-axis) of the numerical solution of (4.4) with (blue, dots) and without filters (red, crosses) plotted against the stepsize τ (on the x-axis) with N grid points. The gray lines indicate order one (dotted) and two (dashed).

As numerical method we used the Strang splitting variant (3.12a), i.e., $\big(A, \widetilde{f}, A\big)$ with $N = 2^j$, $j = 9, 10, 11$, spatial grid points. In Figure 4.2 we displayed the results using filters (blue, dots)

$$\phi(z) = \psi(z) = \operatorname{sinhc}(\tfrac{z}{2}) = \tfrac{1}{2}\Big(\varphi_1(\tfrac{z}{2}) + \varphi_1(-\tfrac{z}{2})\Big) = \frac{\sinh(z/2)}{z/2} \tag{4.5}$$

and also without filters, i.e., $\phi = \psi = 1$, (red, crosses). The code to reproduce the plots is available on https://doi.org/10.5445/IR/1000130189.

These experiments clearly indicate the above-mentioned order reduction to order one for the non-averaged scheme. However, this only happens in the stiff regime, and we briefly explain why this is the case. Later in the error analysis, a key ingredient to prevent the order reduction is the fact that the filters ϕ, ψ roughly behave like the φ_1-function, cf. (F3), in particular they are zero whenever φ_1 is zero, i.e., for $z = 2i\pi k$, $k \geq 1$. Since $\|A\|_{X \leftarrow X} \approx N/2$, we obtain for $\tau < \tau_0 \approx 4\pi/N$

$$\tau \|A\|_{X \leftarrow X} < 2\pi \tag{4.6}$$

and $\varphi_1(\tau A)$ is invertible. Hence, in this non-stiff regime it holds

$$I = \varphi_1(\tau A)\varphi_1(\tau A)^{-1}$$

and even the identity behaves like a filter. Actually, in this regime the two errors of both schemes are quite close. However, we are interested in abstract evolution equations and for unbounded operators A (4.6) cannot be achieved. Therefore, only the stiff regime is relevant, i.e., the limit $N \to \infty$.

4.2 Refined analytical framework

In this section, cf. [9, Sect. 3], we specify the assumptions necessary to prove the results mentioned in Section 4.1.3. In order to illustrate the applicability of our results, we specify the general Example 2.5 in Table 4.1. There three examples are collected where we stated, for a given Hilbert space $\mathcal{H}$, the dimension d of the domain Ω and additional assumptions on the data.

4.2.1 Second-order equation

As in Section 2.1 we begin with the second-order formulation since the equation is posed in this form. This enables us to assess and verify the assumption more easily. In the following we recall sufficient conditions on the nonlinearity G to guarantee wellposedness of the equation and to establish the error analysis presented in Sections 4.3, 4.4, 4.5, and 4.6.

Assumption 4.1 (Wellposedness)**.** *For G we have $G \in C^1([0,t_{\text{end}}] \times V, \mathcal{H})$, i.e., G is Fréchet-differentiable with Fréchet-derivative $J_G(t,q) \in \mathcal{L}\big([0,t_{\text{end}}] \times V, \mathcal{H}\big)$ for all $q \in V, t \in [0,t_{\text{end}}]$.*

In the infinite dimensional case differentiability is a subtle matter. In Example 2.5 the growth bounds guarantee that Assumption 4.1 is valid. Only additional conditions on the growth of higher order derivatives would lead to a twice Fréchet-differentiable function G. Therefore, we only assume regularity for G evaluated at a sufficiently smooth function.

Assumption 4.2 (Regularity of G evaluated at a smooth function)**.**
For $q \in C^1([0,t_{\text{end}}], V) \cap C([0,t_{\text{end}}], \mathcal{D}(L))$ we have

$$t \mapsto G(t,q(t)) \in C^1\left([0,t_{\text{end}}], V\right) \text{ with } \frac{d}{dt} G(t,q(t)) = J_G(t,q(t)) \begin{pmatrix} 1 \\ q'(t) \end{pmatrix}, \tag{A1}$$

$$t \mapsto J_G(t,q(t)) \in C^1\left([0,t_{\text{end}}], \mathcal{L}\big([0,t_{\text{end}}] \times V, \mathcal{H}\big)\right) \text{ with } C > 0 \text{ such that}$$

$$\left\| \frac{d}{dt} J_G(t,q(t)) \right\|_{\mathcal{H} \leftarrow [0,t_{\text{end}}] \times V} \leq C, \quad C = C\left(\|q(t)\|_{\mathcal{D}(L)}, \|q'(t)\|_V \right) \tag{A2}$$

Remark 4.3. *We note that (A1) is not implied by Assumption 4.1. Using the chain rule we can only conclude the weaker assertion*

$$t \mapsto G(t,q(t)) \in C^1\left([0,t_{\text{end}}], \mathcal{H}\right)$$

which is not sufficient for the error analysis.

(a) *In Example 2.5 the additional regularity $q \in C([0,t_{\text{end}}], \mathcal{D}(L))$ is sufficient to verify the Assumption (A1). This is mainly due to the fact that $\mathcal{D}(L)$ is a subset of $L^\infty(\Omega)$ in the example and the composition*

$$t \mapsto g(t,q(t))$$

is then also continuous in $L^\infty(\Omega)$.

(b) *Another approach would be to assume $G \in C^1([0,t_{\text{end}}] \times V, V)$ and the chain rule would immediately yield Assumption (A1). However, this assumption excludes many interesting nonlinearities. In Example 2.5 with $\mathcal{H} = H^{-1}(\Omega)$ and $V = L^2(\Omega)$, see Table 4.1, this would imply that G is already an affine transformation, see [25, Section 3]. Hence, not even the function $q \mapsto \sin(q)$ would be covered by the analysis.*

Finally, we need an assumption on bounds of G and J_G. They are posed on balls with radii in different norms which play an important role in the error analysis. We mainly need them when evaluating the functions at the averaged and numerical solution where not the same bounds as for the exact solution are available.

Assumption 4.4 (Regularity of G)**.** *There are constants $C = C(r)$ such that for given $r_V, r_L > 0$ and q with $\|q\|_V \le r_V$, $\|q\|_{\mathcal{D}(L)} \le r_L$, $p \in V$, and $t \in [0, t_{\text{end}}]$ the following inequalities are satisfied:*

$$\|G(t,q)\|_V \le C(r_L), \tag{A3}$$

$$\left\| J_G(t,q) \begin{pmatrix} s \\ p \end{pmatrix} \right\|_{\mathcal{H}} \le C(r_V) \left(|s| + \|p\|_V \right), \tag{A4a}$$

$$\left\| J_G(t,q) \begin{pmatrix} s \\ p \end{pmatrix} \right\|_{V} \le C(r_L) \left(|s| + \|p\|_V \right). \tag{A4b}$$

For the corrected Lie Splitting (3.14) we assume in addition for $\|p_i\|_V \le r_V$, $i = 1, 2$,

$$\left\| \left(J_G(t,p_1) - J_G(t,p_2) \right) \begin{pmatrix} 0 \\ q \end{pmatrix} \right\|_{\mathcal{H}} \le C(r_L, r_V) \|p_1 - p_2\|_V . \tag{A-CLS-1}$$

and for $\|p_i\|_{\mathcal{D}(L)} \le r_L$, $i = 1, 2$, also

$$\left\| \left(J_G(t,p_1) - J_G(t,p_2) \right) \begin{pmatrix} 0 \\ q \end{pmatrix} \right\|_{V} \le C(r_L) \|p_1 - p_2\|_{\mathcal{D}(L)} . \tag{A-CLS-2}$$

Remark 4.5. *Let G be an operator satisfying Assumptions 4.1, 4.2, and 4.4. Then for any $c \in \mathbb{R}$ the operator $G + cI$ does so, too. This allows us to treat positive semidefinite operators L, e.g., the Laplacian with Neumann or periodic boundary condition, by shifting the spectrum to the right half-plane.*

We mention that Assumptions 4.1, 4.2, and 4.4 are satisfied for the different configurations in Table 4.1, but we postpone the calculations to Appendix A.

All examples are posed with homogeneous Dirichlet boundary conditions. By possibly shifting L, we can also treat Neumann, Robin, or periodic boundary conditions, see Remark 4.5.

Higher order Sobolev spaces $\mathcal{H} = H^k(\Omega)$, $k \ge 2$, can be handled as well but the spaces and conditions for the operators and parameters become more complicated.

4.2.2 First-order equation

The exponential methods from Chapter 3 are all applied to the first-order formulation (2.4) of equation (2.3). In Section 2.1.2 we already considered the operator $A\colon \mathcal{D}(A) \to X$ in this formulation and now turn to the nonlinearity f defined in (2.5). We translate the Assumptions 4.1, 4.2, and 4.4 posed on G into this setting by means of the following three lemmas. The first one provides a classical solution of (2.4) by standard semigroup theory. All statements in the lemmas directly follow from the special structure of f and the assumptions in Section 4.2.1.

Lemma 4.6 (Wellposedness)**.** *Let G satisfy Assumption 4.1. Then $f\colon [0, t_{\text{end}}] \times X \to X$ defined in (2.4) satisfies $f \in C^1([0, t_{\text{end}}] \times X, X)$ with Fréchet derivative $J_f(t,u) \in \mathcal{L}([0, t_{\text{end}}] \times X, X)$ for all $u \in X$ and $t \in [0, t_{\text{end}}]$.*

$\mathcal{H}$	$H^{-1}(\Omega)$	$L^2(\Omega)$	$H_0^1(\Omega)$
d	$d = 1$	$d = 1, 2, 3$	$d = 1, 2, 3$
$\mathbf{A}$	–	$W^{1,\infty}(\Omega)^{d\times d}$	$C^{1,1}(\Omega)^{d\times d} \cap W^{2,\infty}(\Omega)^{d\times d}$ or $H^4(\Omega)^{d\times d}$
Ω	–	–	$\partial\Omega$ of class C^3
$\mathcal{D}(L)$	$H_0^1(\Omega)$	$H^2(\Omega) \cap H_0^1(\Omega)$	$\{q \in H^3(\Omega) \cap H_0^1(\Omega) \mid Lq \in H_0^1(\Omega)\}$
V	$L^2(\Omega)$	$H_0^1(\Omega)$	$H^2(\Omega) \cap H_0^1(\Omega)$
α	$(2, 0, 2)$	$(2, 1, 3)$	$(3, 2, 3)$
g	–	$g(t, \cdot, 0) = 0$ on $\partial\Omega$	$g(t, \cdot, 0) = 0$ on $\partial\Omega$
growth bound	$\gamma \le 2$	$\gamma \begin{cases} < \infty\,, & d = 2 \\ \le 3\,, & d = 3 \end{cases}$	–

Table 4.1: Overview on the specification of Example 2.5. An empty box corresponds to no additional assumptions on this datum.

In the error analysis it is not sufficient to only have differentiability of f in X, but we also need this in the stronger $\mathcal{D}(A)$-norm. As in Assumption 4.2 this cannot be achieved in terms of Fréchet-derivatives.

Lemma 4.7 (Regularity of f evaluated at a smooth function)**.** *Let G satisfy Assumption 4.2 and u satisfy (2.9). Then we have*

$$t \mapsto f\big(t, u(t)\big) \in C^1\left([0, t_{\text{end}}], \mathcal{D}(A)\right) \text{ with } \frac{d}{dt} f\big(t, u(t)\big) = J_f\big(t, u(t)\big) \begin{pmatrix} 1 \\ u'(t) \end{pmatrix}, \tag{A1'}$$

$$t \mapsto J_f\big(t, u(t)\big) \in C^1\left([0, t_{\text{end}}], \mathcal{L}\big([0, t_{\text{end}}] \times X, X\big)\right) \text{ with } C > 0 \text{ such that}$$

$$\left\| \frac{d}{dt} J_f\big(t, u(t)\big) \right\|_{X \leftarrow [0,t_{\text{end}}]\times X} \le C\left(\|Au(t)\|_X, \|u'(t)\|_X\right). \tag{A2'}$$

The next lemma contains two Lipschitz properties of f which easily follow from the corresponding bound on the derivative. They are crucial for the forthcoming error analysis.

Lemma 4.8 (Regularity of f)**.** *Let G satisfy Assumption 4.4. Then there are constants $C = C(r)$ such that for given $r_X, r_A > 0$ and u_i with $\|u_i\|_X \le r_X$, $\|u_i\|_{\mathcal{D}(A)} \le r_A$, $i = 1, 2$, $v \in X$, and $t \in [0, t_{\text{end}}]$ the*

following inequalities are satisfied:

$$\|f(t,u_1)\|_{\mathcal{D}(A)} \le C(r_A), \tag{A3'}$$

$$\left\| J_f(t,u_1) \begin{pmatrix} s \\ v \end{pmatrix} \right\|_X \le C(r_X)\left(|s| + \|v\|_X\right), \tag{A4a'}$$

$$\left\| J_f(t,u_1) \begin{pmatrix} s \\ v \end{pmatrix} \right\|_{\mathcal{D}(A)} \le C(r_A)\left(|s| + \|v\|_X\right), \tag{A4b'}$$

$$\left\| f(t,u_1) - f(t,u_2) \right\|_X \le C\left(r_X\right)\|u_1 - u_2\|_X, \tag{A5a'}$$

$$\|f(t,u_1) - f(t,u_2)\|_{\mathcal{D}(A)} \le C\left(r_A\right)\|u_1 - u_2\|_X. \tag{A5b'}$$

For the corrected Lie Splitting (3.14) we further have for $\|v_i\|_X \le r_X$, $i = 1,2$,

$$\left\| \left(J_f(t,v_1) - J_f(t,v_2)\right) \begin{pmatrix} 0 \\ u_1 \end{pmatrix} \right\|_X \le C(r_A, r_X)\|v_1 - v_2\|_X, \tag{A-CLS-1'}$$

and for $\|v_i\|_{\mathcal{D}(A)} \le r_A$, $i = 1,2$, *also*

$$\left\| \left(J_f(t,v_1) - J_f(t,v_2)\right) \begin{pmatrix} 0 \\ u_1 \end{pmatrix} \right\|_{\mathcal{D}(A)} \le C(r_A)\|v_1 - v_2\|_{\mathcal{D}(A)}. \tag{A-CLS-2'}$$

In Theorem 2.18 we have seen that Lemma 4.6 together with Lemma 4.8 guarantee local wellposedness of (2.4). Since our error analysis only requires assumptions on the data, we recall the following wellposedness result which is a direct consequence of Theorem 2.18.

Proposition 4.9. *Let Assumptions 4.1 and 4.4 be satisfied and take an initial value* $u_0 \in \mathcal{D}(A)$. *Then there exists a time* $t^* > 0$ *and a classical solution of (2.4) on* $[0,t^*)$ *satisfying (2.9) and the generalized finite-energy condition (2.13) for some* $K > 0$.

We note that the generalized finite-energy condition has been used before in the literature. For $u = \left(q, q'\right)$ in the situation of Example 2.5 with $\mathcal{H} = H^{-1}(\Omega)$, see Table 4.1, (2.13) implies

$$\|Au(t)\|_X^2 = \|q(t)\|_{\mathcal{D}(L)}^2 + \|q'(t)\|_V^2 = \left\|\mathbf{A}^{1/2}\nabla q(t)\right\|_{L^2}^2 + \|q'(t)\|_{L^2}^2 \le K^2,$$

which corresponds to the finite-energy condition used in [21, 30, 38, 66]. We further mention, that the bound (2.13) also implies

$$\|q''(t)\|_{\mathcal{H}} \le \|u'(t)\|_X \le K,$$

which is essential in verifying the abstract assumptions on G.

4.2.3 Filter

We finally characterize the functions which can be used as filter functions. We define them on the imaginary axis since they are applied to the skew-adjoint operator A.

Definition 4.10. *Let $\chi \in \mathcal{C}_b(i\mathbb{R})$. We call χ a filter of order m, $m = 1, 2$, if the following properties are satisfied: There exist $\vartheta, \Theta \in \mathcal{C}_b(i\mathbb{R})$ such that for all $z \in i\mathbb{R}$*

$$|\chi(z)| \leq 1\,, \tag{F1}$$
$$1 - \chi(z) = z^m \vartheta(z)\,, \tag{F2}$$
$$z\chi(z) = (e^z - 1)\Theta(z)\,. \tag{F3}$$

In addition, for $m = 2$, χ is symmetric, i.e.,

$$\chi(z) = \chi(-z)\,. \tag{F4}$$

Note that (F3) is equivalent to $\chi(z) = \varphi_1(z)\Theta(z)$.

Remark 4.11. (a) *The simplest example for a filter of order 1 is $\chi(z) = \varphi_1(z)$, where we simply have $\Theta(z) = 1$. With this in mind, one can think of a filter of order 2 as a symmetric version of the φ_1-function.*

(b) *In our example (4.5) we used the short average filter proposed in [21] which is a filter of order 2. We note that in this example $\chi(ix) = \operatorname{sinc}(\frac{x}{2})$ holds for all $x \in \mathbb{R}$, which relates our filters to the ones considered in [35, Chapter XIII.] since they are always defined on the real axis.*

In Theorem 2.20 we answered the question on how to apply such functions to unbounded operators by a functional calculus. This allows us to define a corresponding class of filter operators that we later use in the averaged schemes.

Theorem 4.12. *Let $\tau > 0$ and $\chi \in \mathcal{C}_b(i\mathbb{R})$ be a filter of order m with ϑ, Θ from Definition 4.10. Then we have*

Boundedness:
$$\|\chi(\tau A)\|_{X \leftarrow X} \leq 1 \tag{OF1}$$
$$\|\vartheta(\tau A)\|_{X \leftarrow X} \leq \|\vartheta\|_\infty, \quad \|\Theta(\tau A)\|_{X \leftarrow X} \leq \|\Theta\|_\infty$$

Smoothing: $\chi(\tau A)\colon X \to \mathcal{D}(A)$ *is continuous with*
$$\|\tau A \chi(\tau A)\|_{X \leftarrow X} \leq 2\,\|\Theta\|_\infty \tag{OF2}$$

Consistency: $\vartheta(\tau A)\colon X \to \mathcal{D}(A^m)$,
$$I - \chi(\tau A) = (\tau A)^m\, \vartheta(\tau A) \tag{OF3}$$

Cancelation:
$$(\tau A)\chi(\tau A) = (e^{\tau A} - I)\Theta(\tau A) \tag{OF4}$$

Block structure: For $m = 2$ and $i \in \{1, 2\}$
$$\pi_i x = 0 \quad \textit{implies} \quad \pi_i \chi(\tau A) x = 0\,. \tag{OF5}$$

Proof. All statements are direct consequences of Theorem 2.20 and Proposition 2.21. □

Remark 4.13. (a) *We further obtain* $\|\tau A\vartheta(\tau A)\|^2_{X\leftarrow X} \le 2\|\vartheta\|_\infty$ *for* $m = 2$ *as*

$$|z\vartheta(z)|^2 = |z^2\vartheta(z)|\,|\vartheta(z)| \le 2\|\vartheta\|_\infty \quad \text{for all } z \in i\mathbb{R}\,.$$

In particular, every second-order filter is also a filter of order 1.

(b) *The property* (OF5) *allows us to transfer the structure of* f *given in* (2.5) *to* $\widetilde{f}$ *given in* (4.1)*, in particular we have*

$$\pi_1\widetilde{f}(t,u) = \Psi\pi_1 f(t,\Phi u) = 0, \qquad \widetilde{f}(t,u) = \Psi f(t,\pi_1\Phi u) = \Psi f(t,\Phi\pi_1 u) = \widetilde{f}(t,\pi_1 u), \tag{4.7}$$

which is obviously true for $\psi = \phi = 1$.

4.3 Averaged problem

In this section we make precise what was motivated in Section 4.1.1. Since $\widetilde{f}$ also satisfies the assertion of Lemma 4.6 we conclude with Proposition 4.9 the existence of a unique classical solution $\widetilde{u}$ of (4.1) for all $\tau > 0$. However, a priori we do not know anything about the maximal existence time and the bounds on $\widetilde{u}$, $\widetilde{u}'$ and $A\widetilde{u}$ and whether they depend on the stepsize τ.

Both questions are answered in the following results. The existence time is coupled to a bound on the difference of the original solution u of (2.4) and the averaged solution $\widetilde{u}$ of (4.1). Since we need the Lipschitz continuity of f in (A5a'), we define r_X via

$$\max_{t\in[0,t_{\text{end}}]} \|u(t)\|_X \le C_{\text{emb}}K =: \tfrac{1}{2}r_X$$

with C_{emb} defined in (2.2) and K in (2.13).

Theorem 4.14 ([9, Thm. 4.1])**.** *Let Assumptions 4.1, 4.2, and 4.4 be valid and consider the averaged nonlinearity* $\widetilde{f}$ *defined in* (4.1) *with filters of order* m*. Then there is a* $\tau_0 > 0$ *and a constant* $C_{av} > 0$ *such that for all* $\tau \le \tau_0$ *and filters of order* 1 *it holds*

$$\|u(t) - \widetilde{u}(t)\|_X \le C_{av}\tau, \qquad 0 \le t \le t_{\text{end}}\,, \tag{4.8}$$

and if the filters are of order 2 *also*

$$\|u(t) - \widetilde{u}(t)\|_X \le C_{av}\tau^2, \qquad 0 \le t \le t_{\text{end}}\,. \tag{4.9}$$

The constant C_{av} *and* τ_0 *depend on* r_X*,* u_0*,* t_{end}*, the generalized finite-energy* K *from Proposition 4.9, the filter functions, and the embedding constant* C_{emb}*, but not on* τ*. In particular,* $\widetilde{u}$ *exists on* $[0, t_{\text{end}}]$ *and is bounded by*

$$\max_{t\in[0,t_{\text{end}}]} \|\widetilde{u}(t)\|_X \le \tfrac{3}{4}r_X\,.$$

Proof. We only prove the second-order bound (4.9), since (4.8) is then derived by a simplification of the presented arguments. Let $\widetilde{t^*} > 0$ be the maximal existence time of $\widetilde{u}$ and define

$$t_0 := \sup\{s \in (0,\widetilde{t^*}) \mid \max_{t\in[0,s]} \|\widetilde{u}(t)\|_X \le r_X\}\,.$$

This time is needed in order to apply the uniform bounds on $\widetilde{f}$ and f evaluated at $\widetilde{u}$. The proof is closed by proving $t_0 \geq t_{\text{end}}$.

We first observe that for $t \leq \min\{t_0, t_{\text{end}}\}$ the variation-of-constants formula yields

$$\begin{aligned} u(t) - \widetilde{u}(t) &= \int_0^t e^{(t-s)A} \left(f\big(s, u(s)\big) - \widetilde{f}\big(s, \widetilde{u}(s)\big) \right) ds \\ &= I_1(t) + I_2(t) + \int_0^t e^{(t-s)A} \left(\widetilde{f}\big(s, u(s)\big) - \widetilde{f}\big(s, \widetilde{u}(s)\big) \right) ds \end{aligned} \tag{4.10}$$

with

$$\begin{aligned} I_1(t) &= \int_0^t e^{(t-s)A} \left(I - \Psi\right) f\big(s, u(s)\big)\, ds, \\ I_2(t) &= \int_0^t e^{(t-s)A} \Psi \left(f\big(s, u(s)\big) - f\big(s, \Phi u(s)\big) \right) ds. \end{aligned}$$

By Assumption (A5a') and since $t \leq t_0$, the third term in (4.10) is bounded by

$$\left\| \int_0^t e^{(t-s)A} \left(\widetilde{f}\big(s, u(s)\big) - \widetilde{f}\big(s, \widetilde{u}(s)\big) \right) ds \right\|_X \leq C\big(r_X\big) \int_0^t \|u(s) - \widetilde{u}(s)\|_X \, ds \,,$$

where we also used the bound in (OF1). We are left to prove

$$\|I_j(t)\|_X \leq C\tau^2, \qquad j = 1, 2, \tag{4.11}$$

since these bounds are sufficient to apply a Gronwall lemma which shows the assertion for all $t \leq \min\{t_0, t_{\text{end}}\}$.

We first bound I_1 and use (OF3) and integration by parts to obtain

$$\begin{aligned} I_1(t) &= \tau^2 \int_0^t e^{(t-s)A} A^2 \vartheta(\tau A) f\big(s, u(s)\big)\, ds \\ &= \tau^2 \left[-e^{(t-s)A} A \vartheta(\tau A) f\big(s, u(s)\big) \right]_0^t \\ &\quad + \tau^2 \int_0^t e^{(t-s)A} A \vartheta(\tau A) J_f\big(s, u(s)\big) \begin{pmatrix} 1 \\ u'(s) \end{pmatrix} ds, \end{aligned} \tag{4.12}$$

where we used that $f\big(s, u(s)\big)$ is differentiable in X. By Assumptions (A3'), (A4b'), and the bound (2.13) on u' we have

$$\left\| A f\big(s, u(s)\big) \right\|_X \leq C\,(K)\,, \qquad \left\| A J_f\big(s, u(s)\big) \begin{pmatrix} 1 \\ u'(s) \end{pmatrix} \right\|_X \leq C\,(K)\,.$$

and immediately conclude (4.11) for $j = 1$.

To increase the readability we use the notation $\mathbf{u}(s, \sigma) = \sigma u(s) + (1 - \sigma)\Phi u(s)$ and the differentiability

(A1') of f to get

$$\begin{aligned}
I_2(t) &= \int_0^t e^{(t-s)A}\Psi\big(f\big(s,u(s)\big) - f\big(s,\Phi u(s)\big)\big)\,ds \\
&= \int_0^t \int_0^1 e^{(t-s)A}\Psi \tfrac{d}{d\sigma} f\big(s,\mathbf{u}(s,\sigma)\big)\,d\sigma\,ds \\
&= \int_0^t \int_0^1 e^{(t-s)A}\Psi J_f\big(s,\mathbf{u}(s,\sigma)\big)\begin{pmatrix}0\\(I-\Phi)\,u(s)\end{pmatrix} d\sigma\,ds \\
&= \int_0^t \int_0^1 e^{(t-s)A}\Psi J_f\big(s,\mathbf{u}(s,\sigma)\big)\begin{pmatrix}0\\(I-\Phi)\,e^{sA}u_0\end{pmatrix} d\sigma\,ds \\
&\quad + \int_0^t \int_0^1 e^{(t-s)A}\Psi J_f\big(s,\mathbf{u}(s,\sigma)\big)\begin{pmatrix}0\\(I-\Phi)\int_0^s e^{(s-\theta)A} f\big(\theta,u(\theta)\big)\,d\theta\end{pmatrix} d\sigma\,ds \\
&= I_{2,1}(t) + I_{2,2}(t)\,,
\end{aligned}$$

where we applied the variation-of-constants formula (2.12) again on $u(s)$ in the last step. By (OF3) and integration by parts, the first term can be rewritten as

$$\begin{aligned}
I_{2,1}(t) = \tau^2 &\left[\int_0^1 e^{(t-s)A}\Psi J_f\big(s,\mathbf{u}(s,\sigma)\big)\begin{pmatrix}0\\\vartheta(\tau A)e^{sA}Au_0\end{pmatrix} d\sigma\right]_0^t \\
&+ \tau^2 \int_0^t \int_0^1 e^{(t-s)A} A\Psi J_f\big(s,\mathbf{u}(s,\sigma)\big)\begin{pmatrix}0\\\vartheta(\tau A)e^{sA}Au_0\end{pmatrix} d\sigma\,ds \\
&- \tau^2 \int_0^t \int_0^1 e^{(t-s)A}\Psi \tfrac{d}{ds} J_f\big(s,\mathbf{u}(s,\sigma)\big)\begin{pmatrix}0\\\vartheta(\tau A)e^{sA}Au_0\end{pmatrix} d\sigma\,ds\,.
\end{aligned}$$

Hence, we have $\|I_{2,1}(t)\|_X \le C\tau^2$ by (A2'), (A4a'), and (A4b'). Concerning the term $I_{2,2}$, by assumption (A1') we also have

$$\begin{aligned}
&\int_0^s e^{(s-\theta)A} f\big(\theta,u(\theta)\big)d\theta \in \mathcal{D}(A), \\
&A\int_0^s e^{(s-\theta)A} f\big(\theta,u(\theta)\big)d\theta = \int_0^s e^{(s-\theta)A} Af\big(\theta,u(\theta)\big)d\theta.
\end{aligned}$$

Hence, again integration by parts gives

$$\begin{aligned}
&(I-\Phi)\int_0^s e^{(s-\theta)A} f\big(\theta,u(\theta)\big)\,d\theta \\
&= \tau^2\vartheta_2(\tau A)\left(\Big[-e^{(s-\theta)A} Af\big(\theta,u(\theta)\big)\,d\theta\Big]_0^s + \int_0^s e^{(s-\theta)A} AJ_f\big(\theta,u(\theta)\big)\begin{pmatrix}1\\u'(\theta)\end{pmatrix} d\theta\right)
\end{aligned}$$

and Assumptions (A3') and (A4b') yield the desired bound (4.11). Using (4.9) for $t \le \min\{t_0, t_{\mathrm{end}}\}$ we obtain for $\tau \le \tau_0 = \frac{1}{2}\big(\frac{r_X}{C_{\mathrm{av}}}\big)^{1/2}$

$$\max_{s\in[0,t]} \|\widetilde{u}(s)\|_X \le \max_{s\in[0,t]} \|u(s)\|_X + C_{\mathrm{av}}\tau^2 \le \tfrac{3}{4} r_X\,.$$

This proves $t_0 \ge t_{\mathrm{end}}$ and hence (4.9) holds on $[0, t_{\mathrm{end}}]$ for all $\tau \le \tau_0$. □

From the previous theorem we know something about the maximal existence time of $\widetilde{u}$, we have a bound on $\widetilde{u}$, and we obtained bounds on the difference of u and $\widetilde{u}$. The open question on the generalized finite-energy condition of $\widetilde{u}$ is answered in the next lemma.

Lemma 4.15 ([9, Lemma 4.2])**.** *Let Assumptions 4.1, 4.2, and 4.4 be valid and let ψ, ϕ be filters of order 1. Then there is a $\tau_0 > 0$ and a constant $\widehat{C}_{av} > 0$ such that for all $\tau \leq \tau_0$*

$$\|Au(t) - A\widetilde{u}(t)\|_X \leq \widehat{C}_{av}\tau, \quad 0 \leq t \leq t_{\text{end}}.$$

In particular, $\widetilde{u}$ satisfies the generalized finite-energy condition uniformly in $\tau \leq \tau_0$, i.e.,

$$\max\left\{\|A\widetilde{u}(t)\|_X, \|\widetilde{u}'(t)\|_X\right\} \leq \widetilde{K}, \quad 0 \leq t \leq t_{\text{end}}, \tag{4.13}$$

where τ_0 and the constants $\widehat{C}_{av}$ and $\widetilde{K}$ depend on r_X, u_0, t_{end}, the generalized finite-energy K from Proposition 4.9, the filter functions, and the embedding constant C_{emb}, but not on τ.

Proof. We proceed as in the proof of Theorem 4.14 and define t_0 by

$$t_0 \coloneqq \sup\{s \in (0, t_{\text{end}}] \mid \max_{t\in[0,s]} \|A\widetilde{u}(t)\|_X \leq 2K\}.$$

For $0 \leq t \leq t_0$, (4.10), (A5b'), and (4.8) imply

$$\begin{aligned}
\|Au(t) - A\widetilde{u}(t)\|_X &= \Big\| \int_0^t Ae^{(t-s)A}\left(f\big(s,u(s)\big) - \widetilde{f}\big(s,\widetilde{u}(s)\big)\right) ds\Big\|_X \\
&\leq \|AI_1(t)\|_X + \|AI_2(t)\|_X + C\,(2K)\int_0^t \|u(s) - \widetilde{u}(s)\|_X\, ds \\
&\leq \|AI_1(t)\|_X + \|AI_2(t)\|_X + \tau t C\,(2K)\, C_{\text{av}}.
\end{aligned}$$

We may expand the terms similarly as before, and as in (4.12) it holds

$$\begin{aligned}
AI_1(t) &= \int_0^t e^{(t-s)A}\big(I - \Psi\big)Af\big(s,u(s)\big)\, ds \\
&= \Big[-e^{(t-s)A}\big(I-\Psi\big)f\big(s,u(s)\big)\Big]_0^t + \int_0^t e^{(t-s)A}\big(I-\Psi\big)J_f\big(s,u(s)\big)\begin{pmatrix}1\\ u'(s)\end{pmatrix} \\
&= \tau\Big[-e^{(t-s)A}\vartheta(\tau A)Af\big(s,u(s)\big)\Big]_0^t + \tau\int_0^t e^{(t-s)A}\vartheta(\tau A)AJ_f\big(s,u(s)\big)\begin{pmatrix}1\\ u'(s)\end{pmatrix}.
\end{aligned}$$

where we used (OF3) for $m = 1$ and obtain a $\mathcal{O}(\tau)$ bound for $\|AI_1(t)\|_X$. Similarly it holds by (A1'),

$$\begin{aligned}
AI_2(t) &= \int_0^t \int_0^1 e^{(t-s)A}\Psi AJ_f\big(s,\mathbf{u}(s,\sigma)\big)\begin{pmatrix}0\\ (I-\Phi)\,u(s)\end{pmatrix} d\sigma\, ds \\
&= \tau\int_0^t \int_0^1 e^{(t-s)A}\Psi AJ_f\big(s,\mathbf{u}(s,\sigma)\big)\begin{pmatrix}0\\ \vartheta(\tau A)Au(s)\end{pmatrix} d\sigma\, ds,
\end{aligned}$$

which also gives a $\mathcal{O}(\tau)$-bound. By possibly reducing τ_0 we obtain the result for $0 \leq t \leq t_{\text{end}}$. This immediately implies the first bound in (4.13) and the second bound is then obtained from (4.1). □

Remark 4.16. *Note that Theorem 4.14 and Lemma 4.15 remain true for $\Psi = I$ as for this choice $I_1(t) = 0$ holds. Additionally, the proof does not require the property (F3) and the constant function $z \mapsto 1$ satisfies all the other properties in Definition 4.10. This case is of interest for methods (3.9) and (3.15). Roughly speaking, here the outer filter is replaced by the φ_1-function which behaves like a filter as we have already seen from (F3).*

By the same argument, $\Phi = I$ yields $I_2(t) = 0$ and hence the assertion. Clearly, choosing $\Phi = \Psi = I$ gives $u = \widetilde{u}$ and the bounds are trivial.

4.4 Abstract assumptions on the one-step methods

In this section we provide abstract assumptions that characterize the classes of methods which are covered by our error analysis, and we show how the methods presented before are included in the framework.

We recall that u denotes the solution of the original problem (2.4) and $\widetilde{u}$ the solution of the averaged problem (4.1). Further, we denote the numerical flow by S_τ and the defect by δ_n, i.e., a one-step method is given by

$$u_{n+1} = S_\tau(t_n, u_n), \qquad \delta_n = S_\tau\big(t_n, \widetilde{u}(t_n)\big) - \widetilde{u}(t_{n+1}). \tag{4.14}$$

We start with an assumption on the stability of the method.

Assumption 4.17 (Stability)**.** *The method applied to (4.1) is stable in the sense that for all $v \in \mathcal{D}(A)$, $w \in X$, $t \geq 0$,*

$$S_\tau(t, v) - S_\tau(t, w) = e^{\tau A}(v - w) + \tau \mathcal{J}(t, v, w), \tag{4.15}$$

where $\mathcal{J} : \mathbb{R} \times \mathcal{D}(A) \times X \to X$ is bounded by

$$\|\mathcal{J}(t, v, w)\|_X \leq C_{\mathcal{J}}\left(\|v\|_{\mathcal{D}(A)}, \|w\|_X\right)\|v - w\|_X, \qquad t \in [0, t_{\text{end}}]. \tag{4.16}$$

We note that the stronger $\mathcal{D}(A)$-norm in the above assumption does not cause any problems since we use the stability only for comparing the numerical flow starting at $\widetilde{u}(t_n)$ and at u_n and hence only $\|\widetilde{u}(t_n)\|_{\mathcal{D}(A)}$ appears when we use (4.16). The following proposition states that all the one-step methods from Chapter 3 are stable in the sense of (4.15) and (4.16).

Proposition 4.18 ([9, Prop. 5.5])**.** *Let Assumptions 4.1, 4.2, and 4.4 be satisfied.*

(a) *The general explicit exponential class (3.6) satisfies the stability Assumption 4.17.*

(b) *There is some $\tau_0 > 0$ such that the general implicit exponential class (3.7) satisfies the stability Assumption 4.17 for all $\tau \leq \tau_0$.*

(c) *The second-order variant of the Lie splitting (3.14) applied to the averaged equation (4.1) satisfies Assumption 4.17.*

We emphasize that we cannot analyze the second-order variant of the Lie splitting (3.14) without filter functions. Starting with $u_0 \in \mathcal{D}(A)$ and checking the summands of u_1, we see that all of them lie in $\mathcal{D}(A)$ except $Af(t_{1/2}, u_0) \in X$ by (A3'). Hence, we can only conclude $u_1 \in X$ and this does not allow us to define u_2.

However, if we replace f by $\widetilde{f}$, we obtain

$$r_{\widetilde{f}}(t,u) = \Psi J_f\big(t,\Phi u\big)\begin{pmatrix}0\\ A\Phi u\end{pmatrix} - A\Psi f\big(t,\Phi u\big)$$

and property (OF2) yields that $A\Phi\colon X \to X$ is bounded and thus $A\widetilde{f}(t_{1/2},u_0) \in \mathcal{D}(A)$ holds. Therefore, $u_1 \in \mathcal{D}(A)$ and the scheme is well-defined.

Proof. (a) We recall $a, B_1, B_2 \in \mathcal{C}_b(i\mathbb{R})$ and define the inner stage by

$$s_\tau^{\mathrm{ex}}(t,v) = e^{c_2\tau A}v + c_2\tau a(c_2\tau A)f(t,v) \tag{4.17}$$

and compute for $\|v\|_X, \|w\|_X \le r_X$ by (A5a')

$$\begin{aligned}\|s_\tau^{\mathrm{ex}}(t,v) - s_\tau^{\mathrm{ex}}(t,w)\|_X &= \big\|e^{c_2\tau A}(v-w) + c_2\tau a(c_2\tau A)\big(f(t,v)-f(t,w)\big)\big\|_X\\ &\le \big(1 + C(r_X)\tau\big)\|v-w\|_X\,,\end{aligned}$$

as well as

$$\|s_\tau^{\mathrm{ex}}(t,v)\|_X \le r_X + C(r_X)\tau =: r_1\,.$$

For the outer stage we hence consider with $\mathcal{J} = \pi_1\mathcal{J}_1 + \pi_2\mathcal{J}_2$ by symmetry only the case

$$\begin{aligned}\mathcal{J}_2(t_n,v,w) =& B_1(\tau A)\big(f(t_n,v) - f(t_n,w)\big)\\ &+ B_2(\tau A)\Big(f(t_{n+c_2}, s_\tau^{\mathrm{ex}}(t_n,v)) - f(t_{n+c_2}, s_\tau^{\mathrm{ex}}(t_n,w))\Big)\,.\end{aligned} \tag{4.18}$$

Taking norms and using the properties of the inner stages gives

$$\|\mathcal{J}_2(t_n,v,w)\|_X \le \Big(C(r_X) + C(r_1)\big(1+C(r_X)\tau\big)\Big)\|v-w\|_X\,.$$

For a fixed maximal stepsize $\tau_0 > 0$, r_1 is uniformly bounded by some $C(r_X)$ which closes the argument.

(b) In order to obtain stability of the implicit scheme we define for fixed $\|v\|_X \le r_X$ the fixed-point map

$$T_{v,t_n}(U) = e^{c_2\tau A}v + c_2\tau a(c_2\tau A)f(t_{n+c_2},U)\,. \tag{4.19}$$

Note that once we established stability and boundedness of the inner stage, the outer stage is handled as in part (a). We first check that for $\|U\|_X \le 2r_X$ it holds

$$\|T_{v,t_n}(U)\|_X \le r_X + C\tau\|f(t_{n+c_2},U)\|_X \le r_X + \tau C(2r_X) \le 2r_X$$

for $\tau \le \tau_0 \le \frac{r_X}{C(2r_X)}$. For the contractivity we compute for $\|U\|_X, \|V\|_X \le 2r_X$

$$\|T_{v,t_n}(U) - T_{v,t_n}(V)\|_X \le C\tau\|f(t_{n+c_2},U) - f(t_{n+c_2},V)\|_X \le \tau C(2r_X)\|U-V\|_X \le \tfrac{1}{2}\|U-V\|_X$$

for $\tau \le \tau_0 \le \frac{1}{2C(2r_X)}$. By Banach fixed-point theorem we obtain a unique solution $U^* = T_{v,t_n}(U^*)$ and define the solution map $s_\tau^{\mathrm{im}}(t_n,v) = U^*$.

In the last step we obtain with $\|v\|_X, \|w\|_X \le r_X$ and $V^* = s_\tau^{\mathrm{im}}(t_n,v)$, $W^* = s_\tau^{\mathrm{im}}(t_n,w)$

$$\begin{aligned}\|V^* - W^*\|_X &\le \|v-w\|_X + C\tau\|f(t_{n+c_2},V^*) - f(t_{n+c_2},W^*)\|_X\\ &\le \|v-w\|_X + \tau C(r_2)\|V^*-W^*\|_X\\ &\le \|v-w\|_X + \tfrac{1}{2}\|V^*-W^*\|_X\,,\end{aligned} \tag{4.20}$$

which yields $\|s_\tau^{\mathrm{im}}(t_n,v) - s_\tau^{\mathrm{im}}(t_n,w)\|_X \le 2\|v-w\|_X$.

(c) We recall the scheme (3.14) with filters

$$\begin{aligned} S_\tau(t_n,u) &= e^{\tau A}\big(u_n + \tau\widetilde{f}(t_{n+1/2},u) + \tfrac{\tau^2}{2}\Big(\Psi J_f(t,\Phi u)\begin{pmatrix}0\\ A\Phi u\end{pmatrix} - A\Psi f(t,\Phi u)\Big)\\ &= e^{\tau A}u_n + \tau e^{\tau A}\Big(\widetilde{f}(t_{n+1/2},u) + \tfrac{1}{2}\Big(\Psi J_f(t,\Phi u)\begin{pmatrix}0\\ (\tau A\Phi)u\end{pmatrix} - (\tau A\Psi)f(t,\Phi u)\Big)\Big)\,. \end{aligned}$$

Hence, the operator $\mathcal{J}$ is given by

$$\begin{aligned} \mathcal{J}(t_n,v,w) &= e^{\tau A}\Big(\widetilde{f}(t_{n+1/2},v) - \widetilde{f}(t_{n+1/2},w)\Big)\\ &\quad + \tfrac{1}{2}e^{\tau A}\Big(\Psi J_f(t,\Phi v)\begin{pmatrix}0\\ (\tau A\Phi)v\end{pmatrix} - \Psi J_f(t,\Phi w)\begin{pmatrix}0\\ (\tau A\Phi)w\end{pmatrix}\Big)\\ &\quad - \tfrac{1}{2}e^{\tau A}(\tau A\Psi)\Big(f(t,\Phi v) - f(t,\Phi w)\Big)\\ &= e^{\tau A}\Big(\mathcal{J}_1 + \mathcal{J}_2 - \mathcal{J}_3\Big)\,. \end{aligned} \tag{4.21}$$

By (OF1), (OF4), and (A5a') we directly obtain

$$\|\mathcal{J}_1\|_X + \|\mathcal{J}_3\|_X \le C(\|v\|_X, \|w\|_X)\,\|v-w\|_X\,. \tag{4.22}$$

We expand the remaining term as

$$\mathcal{J}_2 = \tfrac{1}{2}\Big(\Psi J_f(t,\Phi v) - \Psi J_f(t,\Phi w)\Big)\begin{pmatrix}0\\ (\tau A\Phi)v\end{pmatrix} + \tfrac{1}{2}\Psi J_f(t,\Phi w)\begin{pmatrix}0\\ (\tau A\Phi)(v-w)\end{pmatrix}\,.$$

Again (OF4), the bound (A-CLS-1') for the first term and (A4a') for the second term yield

$$\|\mathcal{J}_2\|_X \le C(\|v\|_{\mathcal{D}(A)}, \|w\|_X)\,\|v-w\|_X + C(\|w\|_X)\,\|v-w\|_X\,. \tag{4.23}$$

Combining (4.22) and (4.23) we have shown the condition of Assumption 4.17. □

In order to prove convergence we also need consistency of the methods. For the first-order error bounds the assumption is rather standard.

Assumption 4.19 (Consistency for order one)**.** *The method applied to the original equation (2.4) satisfies Assumption 4.17 (with $\phi = \psi = 1$) and its defect (4.14) satisfies*

$$\|\delta_n\|_X \le C\tau^2\,,$$

where $C > 0$ is independent of τ and n.

A straightforward assumption for second-order convergence would be consistency with $\|\delta_n\|_X \le C\tau^3$. Then standard arguments lead to error bounds in $\mathcal{O}(\tau^2)$. However, under the assumptions made on the data we can not expect this to hold at least in the non-averaged case as we have seen in the numerical example in Section 4.1.4.

For the averaged methods we are hence left with some terms of lower order τ^2 and some terms of the right order τ^3. To end up with a global error of order 2 we require a particular structure of the defect, which we will motivate in the following.

In Chapter 3 we have seen that most of the methods we consider are constructed from the variation-of-constants formula

$$\widetilde{u}(t_{n+1}) = e^{\tau A}\widetilde{u}(t_n) + \tau \int_0^1 e^{(1-s)\tau A}\widetilde{f}(t_n+\tau s, \widetilde{u}(t_n+\tau s))\,ds\,, \tag{4.24}$$

and the method is constructed by approximating the integral term. Hence, this defect can be expressed as some quadrature error that contains the second derivative in s of

$$f_1(s) = \tau\widetilde{f}(t_n+\tau s, \widetilde{u}(t_n+\tau s)) \qquad \text{or} \qquad f_2(s) = e^{(1-s)\tau A} f_1(s)\,,$$

depending on the precise method. The terms of order τ^3 can be treated in the standard way. However, from f_1 we obtain the second-order term

$$\tau^2 J_{\widetilde{f}}\big(t_n+\tau s, \widetilde{u}(t_n+\tau s)\big)\begin{pmatrix} 0 \\ (\tau A\Phi)\,A\widetilde{u}(t_n+\tau s)\end{pmatrix}, \tag{4.25}$$

where one τ is needed to compensate the operator $A\Phi$ which is only bounded by $C\tau^{-1}$. Additionally, f_2 gives the term

$$\tau^2\,(\tau A\Psi)\,e^{(1-s)\tau A}Af(t_n+\tau s, \Phi\widetilde{u}(t_n+\tau s)). \tag{4.26}$$

For this term property (OF4) comes into play. It allows us to carry over the local convergence order to the global error. Similar terms are obtained for the defect of the splitting scheme (3.14). We hence propose the following general structure of δ_n which also includes the integral in (4.24) and the structures in (4.25) and (4.26).

Assumption 4.20 (Structure of defects for order two)**.** *The defect δ_n defined in (4.14) of a numerical method applied to the averaged equation (4.1) is of the form*

$$\delta_n = \delta_n^{(1)} + \delta_n^{(2)} + D_n$$

with $\|D_n\|_X \le C\tau^3$, where the constant $C > 0$ is independent of τ and n. In addition, one of the following sets of conditions is satisfied:

(a) *If ϕ, ψ are filters of order 2, then there exist $w_n \in X$ and a linear map $W_n\colon X \to \mathcal{D}(A)$ which satisfy*

$$\|w_n\|_X \le C, \qquad \Big\|\frac{1}{\tau}\big(w_{n+1}-w_n\big)\Big\|_X \le C, \tag{4.27a}$$

$$\|W_n\|_{X\leftarrow X} \le C, \qquad \Big\|\frac{1}{\tau}\big(W_{n+1}-W_n\big)\Big\|_{X\leftarrow X} \le C, \tag{4.27b}$$

$$\|AW_n\|_{X\leftarrow X} \le C, \tag{4.27c}$$

with a constant C which is independent of τ and n such that $\delta_n^{(i)}$ can be written as

$$\delta_n^{(1)} = \tau^2\big(\tau A\Psi\big)w_n\,, \qquad \delta_n^{(2)} = \tau^2 W_n\big(\tau A\Phi\big)A\widetilde{u}(t_n)\,, \tag{4.28}$$

(b) *If $\psi = 1$ and ϕ is a filter of order 2, then (4.27) and (4.28) hold with $w_n = 0$ for all n.*

Remark 4.21. *If we use property (OF2) in the representation (4.28) we can conclude $\|\delta_n\|_X \le C\tau^2$. But this is precisely Assumption 4.19 and would only yield a suboptimal first-order bound in the global error.*

The following propositions shows that the methods presented in Chapter 3 satisfy the abstract assumptions on the defect.

Proposition 4.22 ([9, Prop. 5.5]). *Let Assumptions 4.1, 4.2, and 4.4 be satisfied.*

(a) *The general explicit and implicit exponential class (3.6) and (3.7) applied to the averaged equation (4.1) satisfy Assumptions 4.19 and 4.20 (a).*

(b) *If the coefficients in (3.8) are chosen such that $\rho = \widehat{\rho} = 0$, then (3.6) and (3.7) applied to the averaged equation (4.1) also satisfy Assumption 4.20 (b).*

(c) *The second-order variant of the Lie splitting (3.14) applied to the averaged equation (4.1) satisfies Assumption 4.20 (a).*

Proof. We mainly focus on part (a) and (b) of the proposition since they can be proved together. In the end we sketch the ideas of part (c).

(i) We first establish Assumption 4.20 for (a) and (b). Recall $t_{n+\xi} = t_n + \tau\xi$ and let $\widetilde{u}_{n+\xi} \coloneqq \widetilde{u}(t_{n+\xi})$ and $\widetilde{f}_{n+\xi} \coloneqq \widetilde{f}(t_{n+\xi}, \widetilde{u}_{n+\xi})$. We first consider the inner defect Δ_n of the explicit scheme

$$\begin{aligned}\Delta_n &= s_\tau^{\mathrm{ex}}\big(t_n, \widetilde{u}_n\big) - \widetilde{u}_{n+c_2} \\ &= c_2\tau\Big(a(c_2\tau A)\widetilde{f}_n - \int_0^1 e^{(1-\xi)c_2\tau A}\widetilde{f}_{n+c_2\xi}\,d\xi\Big) \\ &= c_2\tau\big(\Delta_{n,1} - \Delta_{n,2}\big)\end{aligned}$$

where we used the variation-of-constants formula. As we will only need the first component of the inner stage due to (4.7), it is sufficient to estimate $\pi_1\Delta_{n,1}$ and $\pi_1\Delta_{n,2}$. Since $\pi_1\widetilde{f}_n = 0$ by (4.7), we obtain by (A3')

$$\|\pi_1\Delta_{n,1}\|_X = \Big\|\pi_1 a(c_2\tau A)\widetilde{f}_n\Big\|_X = \tau\Big\|\pi_1 a_1(c_2\tau A)A\widetilde{f}_n\Big\|_X \le C\tau$$

with a_1 given in (3.8) and once more (4.7) gives

$$\|\pi_1\Delta_{n,2}\|_X = \Big\|\pi_1\int_0^1 e^{(1-\xi)c_2\tau A}\widetilde{f}_{n+c_2\xi}\,d\xi\Big\|_X = \Big\|\pi_1\Big(\int_0^1 e^{(1-\xi)c_2\tau A}\widetilde{f}_{n+c_2\xi}\,d\xi - \widetilde{f}_{n+c_2}\Big)\Big\|_X \le C\tau$$

by the order of the implicit Euler method, where we employ (A1') and (A3') to bound the integrand. In summary this gives

$$\|\pi_1\Delta_n\|_X \le C\tau^2\,, \tag{4.29}$$

which also holds in the case of the unfiltered problem. For the implicit scheme we obtain

$$\begin{aligned}\Delta_n &= s_\tau^{\mathrm{im}}\big(t_n, \widetilde{u}_n\big) - \widetilde{u}_{n+c_2} \\ &= c_2\tau a(c_2\tau A)\Big(\widetilde{f}(t_{n+\xi}, s_\tau^{\mathrm{im}}\big(t_n, \widetilde{u}_n\big)) - \widetilde{f}_{n+c_2}\Big) + c_2\tau\Big(a(c_2\tau A)\widetilde{f}_{n+c_2} - \int_0^1 e^{(1-\xi)c_2\tau A}\widetilde{f}_{n+c_2\xi}\,d\xi\Big) \\ &= c_2\tau a(c_2\tau A)\Big(\widetilde{f}(t_{n+\xi}, s_\tau^{\mathrm{im}}\big(t_n, \widetilde{u}_n\big)) - \widetilde{f}_{n+c_2}\Big) + c_2\tau\big(\Delta_{n,1} - \Delta_{n,2}\big)\,.\end{aligned}$$

Choosing $\tau \le \tau_0$ sufficiently small and estimating $\Delta_{n,1}$ and $\Delta_{n,2}$ as above, we obtain as in (4.20) the bound (4.29).

For $s_\tau = s_\tau^{\mathrm{ex}}$ or $s_\tau = s_\tau^{\mathrm{im}}$, this leads us to the defect

$$\begin{aligned}\delta_n = e^{\tau A}\widetilde{u}_n + \tau\Big(&\pi_1\big(\widehat{B}_1(\tau A)\widetilde{f}_n + \widehat{B}_2(\tau A)\widetilde{f}\big(t_{n+c_2}, s_\tau\big(t_n,\widetilde{u}_n\big)\big)\big)\\ &+\pi_2\big(B_1(\tau A)\widetilde{f}_n + B_2(\tau A)\widetilde{f}\big(t_{n+c_2}, s_\tau\big(t_n,\widetilde{u}_n\big)\big)\big)\Big) - \widetilde{u}_{n+1}\,.\end{aligned}$$

We then use the variation-of-constants formula (4.24) and, due to the decomposition $I = \pi_1 + \pi_2$, it is sufficient to consider the defects $\delta_{n,i} = \pi_i\delta_n$. Since both defects have an identical structure, we only consider

$$\begin{aligned}\delta_{n,2} &= \tau\pi_2\Big(B_1(\tau A)\widetilde{f}_n + B_2(\tau A)\widetilde{f}\big(t_{n+c_2}, s_\tau\big(t_n,\widetilde{u}_n\big)\big) - \int_0^1 e^{(1-\xi)\tau A}\widetilde{f}_{n+\xi}\,d\xi\Big)\\ &= \tau\pi_2 B_2(\tau A)\Big(\widetilde{f}\big(t_{n+c_2}, s_\tau\big(t_n,\widetilde{u}_n\big)\big) - \widetilde{f}_{n+c_2}\Big)\\ &\quad + \tau\pi_2\Big(B_1(\tau A)\widetilde{f}_n + B_2(\tau A)\widetilde{f}_{n+c_2} - \int_0^1 e^{(1-\xi)\tau A}\widetilde{f}_{n+\xi}\,d\xi\Big)\\ &= \tau\pi_2\widehat{I}_1 + \tau\pi_2\widehat{I}_2\,.\end{aligned} \tag{4.30}$$

Using (A5a') and the bound in (4.29), we have by (4.7) in the filtered as well as in the unfiltered case

$$\begin{aligned}\left\|\tau\pi_2\widehat{I}_1\right\|_X &\le C\tau\left\|\widetilde{f}\big(t_{n+c_2}, s_\tau\big(t_n,\widetilde{u}_n\big)\big) - \widetilde{f}_{n+c_2}\right\|_X\\ &\le C(r_X)\tau\left\|\pi_1\Big(s_\tau\big(t_n,\widetilde{u}_n\big) - \widetilde{u}_{n+c_2}\Big)\right\|_X\\ &= C(r_X)\tau\left\|\pi_1\Delta_n\right\|_X\\ &\le C\tau^3\,.\end{aligned}$$

The term $\widehat{I}_2$ is the defect of an exponential quadrature rule. Using Taylor expansion on

$$\widetilde{f}_{n+\sigma} = \widetilde{f}_n + \tau\sigma\widetilde{f}'_n + \tau^2\sigma^2\int_0^1 (1-s)\widetilde{f}''_{n+\sigma s}\,ds, \qquad \widetilde{f}^{(k)}_{n+s} \coloneqq \tfrac{d^k}{d\xi^k}\widetilde{f}(t_n+\xi,\widetilde{u}(t_n+\xi))\big|_{\xi=\tau s}\,, \tag{4.31}$$

we are able to write with the definition of the φ-functions in (3.1) and the coefficients in (3.8)

$$\begin{aligned}\widehat{I}_2 &= B_1(\tau A)\widetilde{f}_n + B_2(\tau A)\big(\widetilde{f}_n + \tau c_2\widetilde{f}'_n\big) - \int_0^1 e^{(1-\xi)\tau A}\big(\widetilde{f}_n + \tau\xi\widetilde{f}'_n\big)\,d\xi + \widehat{I}_{2,3}\\ &= \Big(B_1(\tau A) + B_2(\tau A) - \varphi_1(\tau A)\Big)\widetilde{f}_n + \tau\Big(c_2 B_2(\tau A) - \varphi_2(\tau A)\Big)\widetilde{f}'_n + \widehat{I}_{2,3}\\ &= \widehat{I}_{2,1} + \widehat{I}_{2,2} + \widehat{I}_{2,3}\,.\end{aligned}$$

where $\widehat{I}_{2,3}$ is given by

$$\widehat{I}_{2,3} = \tau^2 B_2(\tau A)c_2^2\int_0^1 (1-s)\widetilde{f}''_{n+c_2 s}\,ds - \tau^2\int_0^1 e^{(1-\xi)\tau A}\int_0^1 (1-s)\widetilde{f}''_{n+\xi s}\,ds\,d\xi\,. \tag{4.32}$$

We estimate the three terms separately. The first dominant term $\widehat{I}_{2,3}$ gives rise to the term W_n motivated in (4.25). Since the two terms of the difference have the precise same structure we further

decompose $\widehat{I}_{2,3} = \widehat{I}^A_{2,3} - \widehat{I}^B_{2,3}$ and only investigate the first part. We compute

$$\begin{aligned}\tau\pi_2\widehat{I}^A_{2,3} &= \tau^3\pi_2 B_2(\tau A)c_2^2 \int_0^1 (1-s)\widetilde{f}''_{n+c_2 s}\,ds \\ &= \tau^3\pi_2 B_2(\tau A)c_2^2 \int_0^1 (1-s)J_{\widetilde{f}}\big(t_{n+c_2 s},\widetilde{u}_{n+c_2 s}\big)\begin{pmatrix}0\\ A\Phi u'(t_{n+c_2 s})\end{pmatrix} ds + D_n^1 \\ &= \tau^2\pi_2 B_2(\tau A)c_2^2 \int_0^1 (1-s)J_{\widetilde{f}}\big(t_{n+c_2 s},\widetilde{u}_{n+c_2 s}\big)\begin{pmatrix}0\\ (\tau A\Phi)A\widetilde{u}_n\end{pmatrix} ds + D_n^1 + D_n^2\end{aligned} \tag{4.33}$$

with $\|D_n^1\|_X \le C\tau^3$ by (A2') and $\|D_n^2\|_X \le C\tau^3$ by (A3'). The term W_n^A is given by

$$W_n^A x = \pi_2 B_2(\tau A)c_2^2 \int_0^1 (1-s)J_{\widetilde{f}}\big(t_{n+c_2 s},\widetilde{u}_{n+c_2 s}\big)\begin{pmatrix}0\\ x\end{pmatrix} ds\,.$$

The properties (4.27b) and (4.27c) follow from (A2') and (A4b'). Analogously we define a linear map W_n^B with the same properties and set $W_n = W_n^A - W_n^B$.

To bound $\widehat{I}_{2,2}$ we use that by (3.8)

$$c_2 B_2(0) = \tfrac{1}{2} = \varphi_2(0)$$

holds and thus there exists $\widehat{\varphi} \in \mathcal{C}_b(i\mathbb{R})$ with

$$c_2 B_2(z) - \varphi_2(z) = z\widehat{\varphi}(z), \qquad z \in i\mathbb{R}\,.$$

From this we conclude by (A1')

$$\big\|\tau\pi_2\widehat{I}_{2,2}\big\|_X = \big\|\tau^2\pi_2\widehat{\varphi}(\tau A)\tau A\widetilde{f}'_n\big\|_X \le C\tau^3\big\|A\widetilde{f}'_n\big\|_X \le C\tau^3\,.$$

We conclude with the term $\widehat{I}_{2,1}$. If $\rho = \widehat{\rho} = 0$ in (3.8) holds, we have $\widehat{I}_{2,1} = 0$ and for part (b) Assumption 4.20 (b) is proven. In the other cases we write

$$\begin{aligned}\tau\pi_2\widehat{I}_{2,1} &= \tau\pi_2\Big(B_1(\tau A) + B_2(\tau A) - \varphi_1(\tau A)\Big)\widetilde{f}_n \\ &= \tau^3\pi_2\rho(\tau A)A^2\widetilde{f}_n \\ &= \tau^2\big(\tau A\Psi\big)w_n\end{aligned} \tag{4.34}$$

where we used (3.8) and $\pi_2 A\Psi = A\Psi\pi_1$ due to (2.5) and (OF5). Then w_n is given by

$$w_n = \rho(\tau A)\pi_2 A f(t_n, \Phi\widetilde{u}_n)\,.$$

The properties (4.27a) follow directly from (A3') and (A5b'), and Assumption 4.20 (a) is satisfied.

(ii) In order to verify Assumption 4.19, we let $\psi = \phi = 1$ and as in (4.31) we expand $\widetilde{f}_{n+s}$ only up to order 1. Then $\widehat{I}_2$ is given by

$$\begin{aligned}\widehat{I}_2 &= \tau\Big(B_1(\tau A) + B_2(\tau A) - \varphi_1(\tau A)\Big)\widetilde{f}_n \\ &\quad + \tau^2 c_2 b_2(\tau A)\int_0^1 \widetilde{f}'_{n+c_2 s}\,ds - \tau^2\int_0^1 e^{(1-\xi)\tau A}\int_0^1 \widetilde{f}'_{n+\xi s}\,ds\,d\xi\,.\end{aligned}$$

From (3.8) we have $\rho \in \mathcal{C}_b(i\mathbb{R})$ and $z \mapsto z^2\rho(z) \in \mathcal{C}_b(i\mathbb{R})$ and with the computation of Remark 4.13 we also have

$$\widetilde{\rho}\colon z \mapsto z\rho(z) \in \mathcal{C}_b(i\mathbb{R}). \tag{4.35}$$

The assertion then follows by the boundedness of $\widetilde{f}'_{n+s}$ and the fact that

$$B_1(z) + B_2(z) = \varphi_1(z) + z\,\widetilde{\rho}(z) \tag{4.36}$$

holds together with the bound (A3').

(iii) We briefly comment on the scheme (3.14). The defect can be written as

$$\begin{aligned}
\delta_n &= e^{\tau A}\Big(\widetilde{u}_n + \tau\widetilde{f}\big(t_{n+1/2},\widetilde{u}_n\big) + \tfrac{\tau^2}{2} r_{\widetilde{f}}\big(t_{n+1/2},\widetilde{u}_n\big)\Big) - \widetilde{u}_{n+1} \\
&= \int_0^\tau \tfrac{d}{d\xi}\Big(e^{\xi A}\big(\widetilde{u}_{n+1-\xi} + \xi\widetilde{f}\big(t_{n+1/2},\widetilde{u}_{n+1-\xi}\big) + \tfrac{\xi^2}{2} r_{\widetilde{f}}\big(t_{n+1/2},\widetilde{u}_{n+1-\xi}\big)\big)\Big)\,d\xi \\
&= \int_0^\tau e^{\xi A}\Big(\widetilde{f}\big(t_{n+1/2},\widetilde{u}_{n+1-\xi}\big) - \widetilde{f}_{n+1-\xi}\Big)\,d\xi \\
&\quad + \int_0^\tau \tfrac{\xi^2}{2} e^{\xi A}\Big(\tfrac{d}{d\xi} r_{\widetilde{f}}\big(t_{n+1/2},\widetilde{u}_{n+1-\xi}\big) + A r_{\widetilde{f}}\big(t_{n+1/2},\widetilde{u}_{n+1-\xi}\big)\Big)\,d\xi \\
&= \widehat{I}_3 + \widehat{I}_4\,,
\end{aligned}$$

where we used the structure of f to obtain $J_f(t,u)\begin{pmatrix}0\\ f\end{pmatrix} = 0$. In the first term $\widehat{I}_3$ we add and subtract $\tau e^{\tau/2A}\widetilde{f}_{n+1/2}$ and get the quadrature error of the midpoint rule twice. The term $\widehat{I}_4$ admits a similar structure as $\widehat{I}_{2,1}$ and hence Assumption 4.20 can be verified as before. □

4.5 Error bounds for exponential one-step methods

This section is devoted to the main results for averaged exponential one-step methods. We prove error bounds in the X-norm in Theorem 4.24, cf. [9, Thm. 6.2], and in the $\mathcal{D}(A)$-norm in Section 4.5.4.

A key ingredient is the so-called summation by parts formula

$$\sum_{j=0}^{n} a_j b_j = \sum_{j=0}^{n} a_n b_j + \sum_{j=0}^{n-1}(a_j - a_{j+1})\left(\sum_{k=0}^{j} b_k\right), \tag{4.37}$$

which also comes in the form

$$\sum_{j=0}^{n} a_{n-j} b_j = \sum_{j=0}^{n} a_j b_0 + \sum_{j=0}^{n-1}\left(\sum_{k=0}^{n-j-1} a_k\right)(b_{j+1} - b_j), \tag{4.38}$$

and can be seen as a discrete analogous of the integration by parts formula. It is verified by straightforward calculations.

4.5.1 Bounds in the X-norm

The following result corresponds to the right diagonal arrow in Figure 4.1 and is the last step towards our main theorems in this section, Theorem 4.24. It states that, given suitable filters, any one-step method of Section 3.2 applied to the averaged equation (4.1) allows for a global error of order $\mathcal{O}(\tau^2)$.

As before, u denotes the solution of the original problem (2.4) and $\widetilde{u}$ the solution of the averaged problem (4.1).

Theorem 4.23 (Global error of the averaged problem, [9, Thm. 6.1])**.** *Let Assumptions 4.1, 4.2, and 4.4 be fulfilled. Moreover, let $(u_n)_n$ be the numerical approximations of a scheme applied to the averaged equation (4.1) that satisfies Assumptions 4.17 and 4.20. Then there is a $\tau_0 > 0$ and a constant $C_e > 0$ such that for all $\tau \leq \tau_0$*

$$\|u_n - \widetilde{u}(t_n)\|_X \leq C_e\, \tau^2, \qquad 0 \leq t_n = n\tau \leq t_{\mathrm{end}}.$$

The constant C_e and τ_0 depend on u_0, t_{end}, the generalized finite-energy K from Proposition 4.9, the filter functions, and the embedding constant C_{emb}, but are independent of τ and n.

Proof. The proof makes use of the error recursion from [30] and adapts techniques from [8, Theorem 5.3].

Due to definition (4.14) of the defect δ_n, the global error $\widetilde{e}_n = \widetilde{u}(t_n) - u_n$ can be written as

$$\begin{aligned}\widetilde{e}_{n+1} &= S_\tau(t_n, \widetilde{u}(t_n)) - S_\tau(t_n, u_n) - \delta_n \\ &= e^{\tau A}\widetilde{e}_n + \tau \mathcal{J}\big(t_n, \widetilde{u}(t_n), u_n\big) - \delta_n\end{aligned}$$

by Assumption 4.17. Resolving the recursion we obtain that the global error satisfies

$$\widetilde{e}_{n+1} = e^{(n+1)\tau A}\widetilde{e}_0 + \tau \sum_{j=0}^{n} e^{(n-j)\tau A}\mathcal{J}\big(t_j, \widetilde{u}(t_j), u_j\big) - \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j. \tag{4.39}$$

In a first step we establish the bound

$$\Big\| \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j \Big\|_X \leq C_\delta \tau^2 \tag{4.40}$$

with a constant C_δ being independent of τ and n. In the second step we close the proof with the bound in (4.16) and the application of a discrete Gronwall lemma.

(i) The proof is done by induction on n in order to control the X-norm of the numerical approximations. For $n = 0$, the statement is obviously true. Hence we assume that for all $0 \leq k \leq n$ it holds

$$\|u_k\|_X \leq r_X, \qquad \|u_k - \widetilde{u}(t_k)\|_X \leq C_{\mathrm{e}}\, \tau^2, \qquad C_{\mathrm{e}} \coloneqq C_\delta\, e^{C_{\mathcal{J}}(\widetilde{K}, r_X)t_{\mathrm{end}}}.$$

By Assumption 4.20, the defect is split into three parts, which motivates to write

$$\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j = \widetilde{e}_{n+1}^{(1)} + \widetilde{e}_{n+1}^{(2)} + \widetilde{e}_{n+1}^{(D)}, \tag{4.41}$$

where

$$\widetilde{e}_{n+1}^{(\ell)} = \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(\ell)}, \quad \ell = 1, 2, \qquad \widetilde{e}_{n+1}^{(D)} = \sum_{j=0}^{n} e^{(n-j)\tau A} D_j.$$

Since $\|D_j\|_X \leq C\tau^3$ and $n\tau \leq t_{\mathrm{end}}$ we easily see

$$\Big\|\widetilde{e}_{n+1}^{(D)}\Big\|_X = \Big\| \sum_{j=0}^{n} e^{(n-j)\tau A} D_j \Big\|_X \leq C\tau^2.$$

To bound $\widetilde{e}_{n+1}^{(\ell)}$, $\ell = 1, 2$, we define the sums

$$E_n = \sum_{j=0}^{n} e^{j\tau A} \qquad \text{and} \qquad F_n = \sum_{j=0}^{n} \widetilde{u}(t_j). \tag{4.42}$$

We employ the summation by parts formula (4.38) with $a_j = e^{j\tau A}$ and $b_j = \delta_j^{(1)}$, use the representation of the defect in Assumption 4.20, and property (OF4) with $\chi = \Psi$ to obtain

$$\begin{aligned}
\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(1)} &= E_n\delta_0^{(1)} + \sum_{j=0}^{n-1} E_{n-j-1}\big(\delta_{j+1}^{(1)} - \delta_j^{(1)}\big) \\
&= \tau^3 E_n A\Psi w_0 \ + \tau^3 \sum_{j=0}^{n-1} E_{n-j-1} A\Psi\big(w_{j+1} - w_j\big) \\
&= \tau^2 E_n(e^{\tau A} - I)\Theta_\Psi w_0 \\
&\quad + \tau^2\Big(\tau \sum_{j=0}^{n-1} E_{n-j-1}(e^{\tau A} - I)\Theta_\Psi \frac{1}{\tau}\big(w_{j+1} - w_j\big)\Big).
\end{aligned} \tag{4.43}$$

We note that estimating E_n in a naive way leads to a factor n and hence one loses one order of convergence. However, we can do better if we bound $E_j(e^{\tau A} - I)$ together. We exploit a telescopic sum to get

$$\big\|E_j(e^{\tau A} - I)\big\|_X = \Big\|\sum_{k=0}^{j} e^{k\tau A}(e^{\tau A} - I)\Big\|_X = \Big\|e^{(j+1)\tau A} - I\Big\|_X \le 2. \tag{4.44}$$

Together with (4.27a) and (OF1) this yields (4.40) for $\delta_j^{(1)}$ instead of δ_j.

We proceed similarly for the term $\widetilde{e}_{n+1}^{(2)}$. We use the representation in (4.28), apply the summation by parts formula (4.37) with $a_j = e^{(n-j)\tau A}W_j$ and $b_j = A\Phi A\widetilde{u}_j$, and (OF4) with $\chi = \Phi$ to get

$$\begin{aligned}
\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(2)} &= \tau^3 W_n A\Phi A F_n + \tau^3 \sum_{j=0}^{n-1} e^{(n-j)\tau A}\big(W_j - e^{-\tau A}W_{j+1}\big)A\Phi A F_j \\
&= \tau^2 W_n \Theta_\Phi(e^{\tau A} - I)AF_n \\
&\quad + \tau^2\Big(\tau \sum_{j=0}^{n-1} e^{(n-j)\tau A}\frac{1}{\tau}\big(W_j - e^{-\tau A}W_{j+1}\big)\Theta_\Phi(e^{\tau A} - I)AF_j\Big).
\end{aligned} \tag{4.45}$$

In order to obtain second-order error bounds we bound the terms separately. If we expand the term

$$\frac{1}{\tau}\big(W_j - e^{-\tau A}W_{j+1}\big) = \frac{1}{\tau}e^{-\tau A}\big(W_j - W_{j+1}\big) - \frac{1}{\tau}(e^{-\tau A} - I)W_j\,,$$

we may use the bounds (4.27b) and (4.27c) to derive

$$\begin{aligned}
\Big\|\frac{1}{\tau}e^{-\tau A}\big(W_j - W_{j+1}\big)\Big\|_{X\leftarrow X} &= \Big\|\frac{1}{\tau}\big(W_j - W_{j+1}\big)\Big\|_{X\leftarrow X} \le C, \\
\Big\|\frac{1}{\tau}(e^{-\tau A} - I)W_j\Big\|_{X\leftarrow X} &= \Big\|\varphi_1(-\tau A)AW_j\Big\|_{X\leftarrow X} \le C,
\end{aligned}$$

since $|\varphi_1(z)| \le 1$ for $z \in i\mathbb{R}$.

As in (4.44), we estimate the term $(e^{\tau A} - I)AF_j$ for $j \le n$ since the term F_j alone does not have the right order. After adding the exact solution we apply the variation-of-constants formula, (A3'),

and (4.13), which gives

$$\begin{aligned}\left\|(e^{\tau A}-I)AF_j\right\|_X &= \left\|A\sum_{k=0}^{j}(e^{\tau A}\widetilde{u}(t_k)-\widetilde{u}(t_k+\tau))+A\sum_{k=0}^{j}(\widetilde{u}(t_k+\tau)-\widetilde{u}(t_k))\right\|_X\\ &= \Big\|-\sum_{k=0}^{j}\int_0^\tau e^{(\tau-s)A}A\widetilde{f}(\widetilde{u}(t_k+s))\,ds+A(\widetilde{u}(t_{j+1})-\widetilde{u}_0)\Big\|_X\\ &\le t_{\mathrm{end}}C(\widetilde{K})+2\widetilde{K}.\end{aligned}$$

This yields (4.40) for $\delta_j^{(2)}$ instead of δ_j and together with the results above proves (4.40).

(ii) Finally, turning back to (4.39), we plug in $\widetilde{e}_0=0$, the bounds on the defects (4.40) and the stability in (4.16) and arrive at

$$\begin{aligned}\|\widetilde{e}_{n+1}\|_X &= \Big\|\tau\sum_{j=0}^{n}e^{(n-j)\tau A}\mathcal{J}\big(t_j,\widetilde{u}(t_j),u_j\big)-\sum_{j=0}^{n}e^{(n-j)\tau A}\delta_j\Big\|_X\\ &\le C_\delta\tau^2+\tau\sum_{j=1}^{n}C_{\mathcal{J}}(\widetilde{K},r_X)\,\|\widetilde{e}_j\|_X\,.\end{aligned}$$

A discrete Gronwall lemma thus yields

$$\begin{aligned}&\|\widetilde{e}_{n+1}\|_X \le \tau^2\,C_\delta\,e^{C_{\mathcal{J}}(\widetilde{K},r_X)t_{\mathrm{end}}}=C_{\mathrm{e}}\tau^2,\\ &\|u_{n+1}\|_X \le \|\widetilde{u}(t_{n+1})\|_X+\|\widetilde{e}_{n+1}\|_X \le \tfrac{3}{4}r_X+C_{\mathrm{e}}\tau^2\le r_X\end{aligned}$$

for $\tau\le\tau_0\le\frac{1}{2}\big(\frac{r_X}{C_{\mathrm{e}}}\big)^{1/2}$ and the induction is closed. □

From all these preparations we now easily conclude our main result for averaged exponential one-step methods.

Theorem 4.24 ([9, Thm. 6.2])**.** *Let Assumptions 4.1, 4.2, and 4.4 be fulfilled. Further, let $(u_n)_n$ be the numerical approximations of a scheme that satisfies Assumption 4.17.*

(a) *If the method also satisfies Assumption 4.19 and is applied to the original equation (2.4), then there is a $\tau_0>0$ and a constant $C_1>0$ such that for all $\tau\le\tau_0$*

$$\|u_n-u(t_n)\|_X\le C_1\tau,\qquad 0\le t_n=n\tau\le t_{\mathrm{end}}\,.$$

(b) *Let ϕ,ψ such that Assumption 4.20 is satisfied. Then there is a $\tau_0>0$ and a constant $C_2>0$ such that for all $\tau\le\tau_0$*

$$\|u_n-u(t_n)\|_X\le C_2\tau^2,\qquad 0\le t_n=n\tau\le t_{\mathrm{end}},$$

if the method is applied to the averaged equation (4.1).

The constants C_1,C_2 and τ_0 depend on u_0, t_{end}, the generalized finite-energy K from Proposition 4.9, the filter functions, and the embedding constant C_{emb}, but are independent of τ and n.

Proof. Part (a) follows directly from Assumption 4.19 and equation (4.39). For part (b), we simply combine Theorem 4.14 and Theorem 4.23 by the triangle inequality (4.3). □

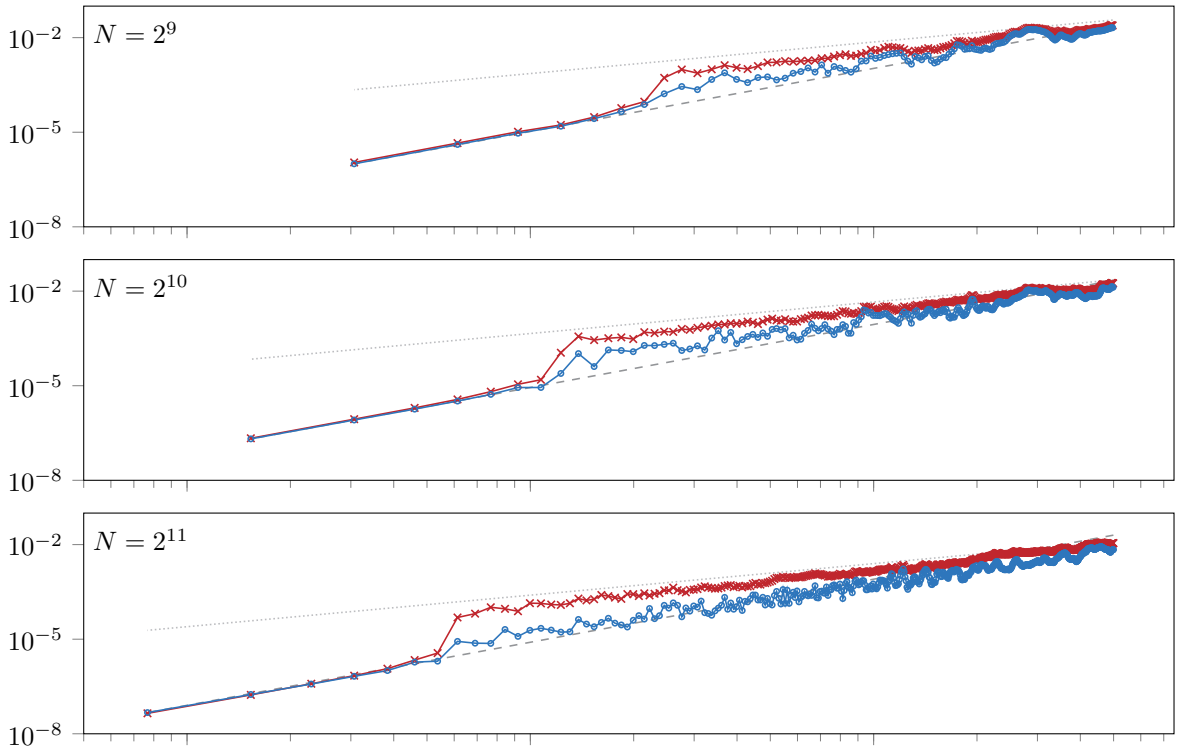

Figure 4.3: Discrete $L^\infty\Big([0,1], L^2(\mathbb{T}) \times H^{-1}(\mathbb{T})\Big)$ error (on the y-axis) of the numerical solution of (4.4) plotted against the stepsize τ (on the x-axis) without filters (red, crosses) and with outer filter only (blue, dots) with N grid points. The gray lines indicate order one (dotted) and two (dashed).

4.5.2 On the necessity of the inner filter

Gauckler [22] proves that in the setting of a one-dimensional wave equation with periodic boundary condition it is not necessary to use an inner filter ϕ in order to obtain second-order error bounds. Hence, we comment on how this insight is present in our more general framework. Numerical experiments indicate that in certain examples the inner filter cannot be neglected. We used an example similar to that of Section 4.1.4 and only describe the differences. We changed the right-hand side to

$$G(q) = m_a \sin(m_i\, q)\, q, \tag{4.46}$$

and made the choice $\phi = 1$ and ψ as in (4.5) for the numerical scheme (blue, dots), see Figure 4.3. The code to reproduce the plots is available on https://doi.org/10.5445/IR/1000130189. We still observe an improvement of the numerical scheme with outer filter compared to not using any filter (red, crosses), but the order reduction cannot be prevented.

Going back into the error analysis, we see that the inner filter Φ is important in the defect that stems from the part of the quadrature error which is considered in (4.25). In more detail we examine this term in (4.32), and it is obvious that this term was not problematic if we could simply conclude

$$\left\| f''_{n+\xi s} \right\|_X \leq C\,. \tag{4.47}$$

For the following considerations we restrict ourselves to the case $g(t,x,q) = g(x,q)$ and hence (4.47) is

equivalent to

$$\left\|g''_{n+\xi s}\right\|_{\mathcal{H}} = \left\|g_{yy}(q_{n+\xi s})(q'_{n+\xi s})^2 + g_y(q_{n+\xi s})q''_{n+\xi s}\right\|_{\mathcal{H}} \leq C\,.$$

which is precisely the conclusion in [22, Proposition 3.3]. In all the examples mentioned above we have at least, cf. Appendix A,

$$\left\|g_{yy}(q_{n+\xi s})(q'_{n+\xi s})^2\right\|_V \leq C\big(\|q_{n+\xi s}\|_{\mathcal{D}(L)}\,, \left\|q'_{n+\xi s}\right\|_V\big) \leq C(K)$$

and hence (4.47) is equivalent to

$$\left\|g_y(q_{n+\xi s})q''_{n+\xi s}\right\|_{\mathcal{H}} \leq C\,. \tag{4.48}$$

Since by the generalized finite-energy condition (2.13) it holds $\left\|q''_{n+\xi s}\right\|_{\mathcal{H}} \leq K$ a sufficient condition for (4.48) is

$$\|g_y(q_{n+\xi s})\|_{\mathcal{H}\leftarrow\mathcal{H}} \leq C\,.$$

However, from (A2) we can in general only conclude that it is bounded from V to $\mathcal{H}$, and we would need a bound on $\left\|q''_{n+\xi s}\right\|_V$ which is not covered by the generalized finite-energy condition.

So, the question to answer is when the multiplication by $g_y(q_{n+\xi s})$ is a bounded operator from $\mathcal{H}$ to $\mathcal{H}$. We give an exemplary overview on different scenarios.

(a) $\mathcal{H} = H^{-1}(\Omega)$ **and** $d = 1$

Since in our framework $g'(q_{n+\xi s})$ acts as a multiplication operator, it is sufficient to check whether a multiplication is continuous as an operator from $H^{-1}(\Omega)$ into itself. We compute for $a \in H^1(\Omega)$ and $v \in L^2(\Omega)$

$$\|av\|_{H^{-1}} = \sup_{\|w\|_{H^1}=1} \langle av, w\rangle_{L^2} = \sup_{\|w\|_{H^1}=1} \langle v, aw\rangle_{L^2} \leq \sup_{\|w\|_{H^1}=1} \|v\|_{H^{-1}} \|aw\|_{H^1} \leq C\,\|v\|_{H^{-1}}\,\|a\|_{H^1}\,.$$

Hence, we can extend this to a bounded linear operator $m\colon H^{-1}(\Omega) \to H^{-1}(\Omega)$ if $a \in H^1(\Omega)$. Therefore, a sufficient condition for (4.48) to hold, is $g_y(q_{n+\xi s}) \in H^1(\Omega)$ for $q_{n+\xi s} \in H^1(\Omega)$.

Smooth coefficients In [22] g is a polynomial and since $H^1(\Omega)$ is an algebra for $d = 1$, this directly implies $g_y(q_{n+\xi s}) \in H^1(\Omega)$. More general, we can assure this condition if we assume $(x, y) \mapsto g_y(x, y)$ to be weakly differentiable in x and continuously differentiable in y since then

$$x \mapsto g_y(x, q_{n+\xi s}(t, x)) \in H^1(\Omega)$$

for all $t \in [0, t_{\text{end}}]$ by the standard arguments.

Irregular coefficients If we use a right-hand side as in (4.46) the above considerations do not apply. For example, take the linear case with $g(x, y) = m(x)y$, $m \in L^\infty(\Omega)$. We then have $g_y(x, y) = m(x)$, which is in general not a map from $H^{-1}(\Omega)$ into itself. This explains the behavior in Figure 4.3.

(b) $\mathcal{H} = L^2(\Omega)$

For any spatial dimension $d \in \{1, 2, 3\}$ a sufficient condition for (4.48) is

$$g_y(q_{n+\xi s}) \in L^\infty(\Omega)\,, \tag{4.49}$$

since we only need a bounded multiplication operator from $L^2(\Omega)$ to $L^2(\Omega)$. Surprisingly, this is far less restrictive than in the case $\mathcal{H} = H^{-1}(\Omega)$. In addition, since $q_{n+\xi s} \in H^2(\Omega) \hookrightarrow L^\infty(\Omega)$ holds, the assumptions in Table 4.1 directly imply (4.49) and no inner filter is needed.

(c) $\mathcal{H} = H_0^1(\Omega)$

The situation is not so much different to the $L^2(\Omega)$-case since (4.48) is implied by

$$g_y(q_{n+\xi s}) \in W^{1,\infty}(\Omega)\,. \tag{4.50}$$

Since $q_{n+\xi s} \in H^3(\Omega) \hookrightarrow W^{1,\infty}(\Omega)$ holds for the exact solution, the regularity of g is sufficient for (4.50) to be valid and also no inner filter is needed.

4.5.3 On the necessity of the outer filter

After the discussion above, naturally the question arises whether one needs an outer filter and if one can characterize the scenarios where it is necessary. Again checking the proof of Proposition 4.22, we observe that the outer filter Ψ only enters in the term (4.34). In particular, if we could establish the bound

$$\left\|A^2 f(t_n, u(t_n)\right\|_X \leq C\,, \tag{4.51}$$

with C independent of τ, then one can take $\psi = 1$. Note that (4.51) is equivalent to

$$\|G(t_n, q(t_n)\|_{\mathcal{D}(L)} = \|L\, G(t_n, q(t_n)\|_{\mathcal{H}} \leq C\,. \tag{4.52}$$

For simplicity we only consider $L = -\Delta$ and check the different scenarios.

(a) $\mathcal{H} = H^{-1}(\Omega)$

In this case we have chosen $\alpha = (2, 0, 2)$, but (4.52) is given by

$$\|g(t_n, \cdot, q(t_n, \cdot))\|_{H^1} \leq C\,,$$

which is not defined under the assumed smoothness of g. For example, the right-hand side in (4.46) is not weakly differentiable in x due to the terms $m_a, m_i \in L^\infty(\Omega)$. However, since $q \in C^1([0,T], H_0^1(\Omega))$ holds, differentiability in the x-component of g and the assumption $g(t, \cdot, 0) = 0$ on $\partial\Omega$ imply (4.52) by the chain rule. In particular in the framework of [22], this section together with the observations in the previous one yield second-order error bounds without any filter.

(b) $\mathcal{H} = L^2(\Omega)$ **and** $\mathcal{H} = H_0^1(\Omega)$

As in the previous case, it is easily seen that the bound in (4.52) can be achieved requiring more regularity of g in the spatial variable x and possibly adding more compatibility conditions on the boundary.

4.5.4 Bounds in the graph norm

In this section we adapt the technique previously used to obtain error bounds also in the stronger graph norm. It is no surprise that the order is decreased to one for the filtered scheme. We emphasize that the usage of the inner filter Φ is redundant. However, it does also not deteriorate the result. We first present the slightly different stability assumption for the method compared to Assumption 4.17.

Assumption 4.25 (Stability)**.** *The method applied to* (4.1) *is stable in the sense that for all* $v, w \in \mathcal{D}(A)$*,* $t \geq 0$*,*

$$S_\tau(t, v) - S_\tau(t, w) = e^{\tau A}\,(v - w) + \tau \mathcal{J}\,(t, v, w)\,,$$

where $\mathcal{J} : \mathbb{R} \times \mathcal{D}(A) \times \mathcal{D}(A) \to \mathcal{D}(A)$ *is bounded by*

$$\|\mathcal{J}(t, v, w)\|_{\mathcal{D}(A)} \leq C_{\mathcal{J}}\left(\|v\|_{\mathcal{D}(A)}, \|w\|_{\mathcal{D}(A)}\right)\|v - w\|_{\mathcal{D}(A)}, \qquad t \in [0, t_{\text{end}}]. \tag{4.53}$$

The following proposition shows that the assumption on the stability is satisfied by all the one-step methods considered.

Proposition 4.26. *Let Assumptions 4.1, 4.2, and 4.4 be satisfied.*

(a) *The general explicit exponential class (3.6) satisfies the stability Assumption 4.25.*

(b) *There is some* $\tau_0 > 0$ *such that the general implicit exponential class (3.7) satisfies the stability Assumption 4.25 for all* $\tau \leq \tau_0$.

(c) *The second-order variant of the Lie splitting (3.14) applied to the averaged equation (4.1) satisfies Assumption 4.25.*

Proof. The proof is very similar to the one of Proposition 4.18 and in particular the operator $\mathcal{J}$ remains the same. We only need to prove the additional bounds. We consider the three cases separately.

(a) As before it suffices to consider the part of $\mathcal{J}$ given in (4.18). With the property given in (A5b'), we hence may conclude the assertion if we establish for $\|v\|_{\mathcal{D}(A)} \leq r_A$ a bound of the form

$$\|s_\tau^{\text{ex}}(t_n, v)\|_{\mathcal{D}(A)} \leq C(r_A),$$

where s_τ^{ex} denotes the flow of the inner stage defined in (4.17). This is obtained by

$$\|s_\tau^{\text{ex}}(t_n, v)\|_{\mathcal{D}(A)} \leq \|v\|_{\mathcal{D}(A)} + C\tau\|f(t_n, v)\|_{\mathcal{D}(A)} \leq C(r_A)$$

using the bound in (A3').

(b) As in part (a) it is sufficient to prove a bound for the solution of (4.19) in $\|\cdot\|_{\mathcal{D}(A)}$. To achieve this, we consider (4.19) and repeat the fixed-point argument in the stronger norm. Let again $\|v\|_{\mathcal{D}(A)} \leq r_A$ and $\|U\|_{\mathcal{D}(A)} \leq 2r_A$ and compute

$$\|T_{v,t_n}(U)\|_{\mathcal{D}(A)} \leq r_A + C\tau\|f(t_{n+c_2}, U)\|_{\mathcal{D}(A)} \leq r_A + \tau C(r_A) \leq 2r_A$$

for $\tau \leq \tau_0 \leq \frac{r_A}{C(r_A)}$. By the same means we also obtain the contractivity and conclude

$$\left\|s_\tau^{\text{im}}(v)\right\|_{\mathcal{D}(A)} \leq 2r_A$$

for sufficiently small $\tau \leq \tau_0$.

(c) We use the decomposition of (4.21) and obtain by (OF1), (OF4), and (A5b') for

$$\|\mathcal{J}_1\|_{\mathcal{D}(A)} + \|\mathcal{J}_3\|_{\mathcal{D}(A)} \leq C(\|v\|_{\mathcal{D}(A)}, \|w\|_{\mathcal{D}(A)})\|v - w\|_{\mathcal{D}(A)}. \tag{4.54}$$

For $\mathcal{J}_2$ we use the expansion

$$\mathcal{J}_2 = \tfrac{1}{2}\Big(\Psi J_f\big(t, \Phi v\big) - \Psi J_f\big(t, \Phi w\big)\Big)\begin{pmatrix} 0 \\ (\tau A\Phi)v \end{pmatrix} + \tfrac{1}{2}\Psi J_f\big(t, \Phi w\big)\begin{pmatrix} 0 \\ (\tau A\Phi)\big(v - w\big) \end{pmatrix}.$$

Again (OF4), the bound (A-CLS-2') for the first term and (A4b') for the second term yield

$$\|\mathcal{J}_2\|_X \leq C(\|v\|_{\mathcal{D}(A)}, \|w\|_{\mathcal{D}(A)})\|v - w\|_{\mathcal{D}(A)} + C(\|w\|_{\mathcal{D}(A)})\|v - w\|_{\mathcal{D}(A)}. \tag{4.55}$$

Combining (4.54) and (4.55) we have shown the assertion of Assumption 4.25. □

We further need an assumption on the consistency of the methods. Since we only prove an error bound of order one, the structure of the defect takes a far simpler form as in Assumption 4.20 and only one part is left that needs to be taken special care of.

Assumption 4.27 (Structure of defects)**.** *The defect δ_n defined in (4.14) of a numerical method applied to the averaged equation (4.1) is of the form*

$$\delta_n = \delta_n^{(1)} + D_n$$

with $\|D_n\|_{\mathcal{D}(A)} \leq C\tau^2$, where the constant $C > 0$ is independent of τ and n. In addition, one of the following sets of conditions is satisfied:

(a) *If ψ is a filter of order 1, then there exists $w_n \in X$ which satisfies*

$$\|w_n\|_X \leq C, \qquad \left\|\frac{1}{\tau}\big(w_{n+1} - w_n\big)\right\|_X \leq C, \tag{4.56}$$

with a constant C which is independent of τ and n such that $\delta_n^{(1)}$ can be written as

$$A\,\delta_n^{(1)} = \tau\,\big(\tau A\Psi\big)w_n\,. \tag{4.57}$$

(b) *If $\psi = 1$, then (4.56) and (4.57) hold with $w_n = 0$ for all n.*

Remark 4.28. *We emphasize that a method that satisfies condition (b) of Assumptions 4.27 actually does not need any filter. These methods are characterized by $\rho = \widehat{\rho} = 0$ in (3.8).*

In the following proposition we prove that all the methods we consider allow for this special structure of the defect.

Proposition 4.29. *Let Assumptions 4.1, 4.2, and 4.4 be satisfied.*

(a) *The general explicit and implicit exponential class (3.6) and (3.7) applied to the averaged equation (4.1) satisfy Assumption 4.27 (a).*

(b) *If the coefficients in (3.8) are chosen such that $\rho = \widehat{\rho} = 0$, then (3.6) and (3.7) applied to the averaged equation (4.1) also satisfy Assumption 4.27 (b).*

(c) *The second-order variant of the Lie splitting (3.14) applied to the averaged equation (4.1) satisfies Assumption 4.27 (a).*

Proof. The proof is very similar to the one of Proposition 4.22. We first proof part (a) and explain how (b) follows from this. We recall the decomposition $I = \pi_1 + \pi_2$ which led to the defects $\delta_{n,i} = \pi_i \delta_n$ and that it is sufficient to bound one of them. We further use (4.30) where we further decomposed the defect as

$$\begin{aligned}
\delta_{n,2} &= \tau\pi_2 B_2(\tau A)\Big(\widetilde{f}\big(t_{n+c_2}, s_\tau\big(t_n, \widetilde{u}_n\big)\big) - \widetilde{f}_{n+c_2}\Big) \\
&\quad + \tau\pi_2\Big(B_1(\tau A)\widetilde{f}_n + B_2(\tau A)\widetilde{f}_{n+c_2} - \int_0^1 e^{(1-\xi)\tau A}\widetilde{f}_{n+\xi}\,d\xi\Big) \\
&= \tau\pi_2\widehat{I}_1 + \tau\pi_2\widehat{I}_2\,.
\end{aligned}$$

The first term gives with (A5b')

$$\big\|A\tau\pi_2\widehat{I}_1\big\|_X = \tau\big\|\pi_1 A\widehat{I}_1\big\|_X \leq \tau C(\widetilde{K})\,\big\|s_\tau\big(t_n, \widetilde{u}_n\big) - \widetilde{u}_{n+c_2}\big\|_X \leq C\tau^3\,,$$

where we used (4.29). Note that this lead to a part of $\delta_{n,2}$ which is of order $\mathcal{O}(\tau^3)$, however, this fact cannot be used. The second part requires a more careful treatment. We compute

$$\begin{aligned}
A\tau\pi_2\widehat{I}_2 &= \tau\pi_1\Big(B_1(\tau A)A\widetilde{f}_n + B_2(\tau A)A\widetilde{f}_{n+c_2} - \int_0^1 e^{(1-\xi)\tau A}A\widetilde{f}_{n+\xi}\,d\xi\Big)\\
&= \tau\pi_1\big(B_1(\tau A) + B_2(\tau A) - \varphi_1(\tau A)\big)A\widetilde{f}_n\\
&\quad + \tau\pi_1 B_2(\tau A)A\big(\widetilde{f}_{n+c_2} - \widetilde{f}_n\big) - \tau\pi_1\int_0^1 e^{(1-\xi)\tau A}A\big(\widetilde{f}_{n+\xi} - \widetilde{f}_n\big)\,d\xi\\
&= \widehat{I}^A_{2,1} + \widehat{I}^A_{2,2}
\end{aligned}$$

and obtain by (A5b') directly $\big\|\widehat{I}^A_{2,2}\big\|_X \le C\tau^2$. If $\rho = \widehat{\rho} = 0$ holds the conditions in (3.8) imply $\widehat{I}_{2,1} = 0$ and part (b) is shown.

For the other cases we use (4.35) and (4.36) to write

$$\begin{aligned}
\widehat{I}^A_{2,1} &= \tau\pi_1\big(B_1(\tau A) + B_2(\tau A) - \varphi_1(\tau A)\big)A\widetilde{f}_n\\
&= \tau\big(\tau A\Psi\big)\pi_2\widetilde{\rho}(\tau A)Af(t_n,\Phi\widetilde{u}_n)\\
&= \tau\big(\tau A\Psi\big)w_n\,,
\end{aligned}$$

which gives us (4.57). It remains to prove the properties of w_n in (4.56). From (A3') we obtain

$$\|w_n\|_X \le \|\widetilde{\rho}(\tau A)Af(t_n,\Phi\widetilde{u}_n)\|_X \le C(\widetilde{K})$$

and with (A5b') it follows

$$\|w_{n+1} - w_n\|_X \le \big\|\widetilde{\rho}(\tau A)A\big(f(t_{n+1},\Phi\widetilde{u}_{n+1}) - f(t_n,\Phi\widetilde{u}_n)\big)\big\|_X \le C(\widetilde{K})\tau$$

such that part (a) and (b) are proven.

In order to show the assertion in (c) we consider the defect

$$\begin{aligned}
A\delta_n &= Ae^{\tau A}\Big(\widetilde{u}_n + \tau\widetilde{f}\big(t_{n+1/2},\widetilde{u}_n\big) + \tfrac{\tau^2}{2}r_{\widetilde{f}}\big(t_{n+1/2},\widetilde{u}_n\big)\Big) - A\widetilde{u}_{n+1}\\
&= \tau\Big(A\widetilde{f}\big(t_{n+1/2},\widetilde{u}_n\big) - \int_0^1 e^{(1-\xi)\tau A}A\widetilde{f}_{n+\xi}\,d\xi\Big) + \tfrac{\tau^2}{2}e^{\tau A}Ar_{\widetilde{f}}\big(t_{n+1/2},\widetilde{u}_n\big)\\
&= \widehat{I}^A_3 + \widehat{I}^A_4\,.
\end{aligned}$$

We expand the first term

$$\begin{aligned}
\widehat{I}^A_3 &= \tau\Big(A\widetilde{f}\big(t_{n+1/2},\widetilde{u}_n\big) - A\widetilde{f}_{n+1/2} + A\widetilde{f}_{n+1/2} - A\varphi_1(\tau A)\widetilde{f}_{n+1/2} - \int_0^1 e^{(1-\xi)\tau A}A\big(\widetilde{f}_{n+\xi} - \widetilde{f}_{n+1/2}\big)\,d\xi\Big)\\
&= \tau\Big(A\widetilde{f}_{n+1/2} - A\varphi_1(\tau A)\widetilde{f}_{n+1/2}\Big)\\
&\quad + \tau\Big(A\widetilde{f}\big(t_{n+1/2},\widetilde{u}_n\big) - A\widetilde{f}_{n+1/2} - \int_0^1 e^{(1-\xi)\tau A}A\big(\widetilde{f}_{n+\xi} - \widetilde{f}_{n+1/2}\big)\,d\xi\Big)\\
&= \widehat{I}^A_{3,1} + \widehat{I}^A_{3,2}
\end{aligned}$$

and obtain from (A5b') and (A1') the bound $\|\widehat{I}^A_{3,2}\|_X \le C\tau^2$. Concerning $\widehat{I}^A_{3,1}$ we get from (3.2)

$$\widehat{I}^A_{3,1} = \tau(\tau A\Psi)\big(-\varphi_2(\tau A)Af(t_{n+1/2}, \Phi\widetilde{u}_{n+1/2})\big) = \tau(\tau A\Psi)w_{n,1}$$

where $w_{n,1}$ satisfies the properties in (4.56) by the same arguments as in part (a).

We finally study the term $\widehat{I}^A_4$

$$\begin{aligned}\widehat{I}^A_4 &= \tfrac{\tau^2}{2}e^{\tau A}Ar(t_{n+1/2}, \widetilde{u}_n)\\ &= \tfrac{\tau^2}{2}e^{\tau A}\Psi AJ_f(t_{n+1/2}, \Phi\widetilde{u}_n)\begin{pmatrix}0\\ A\Phi\widetilde{u}_n\end{pmatrix} + \tau(\tau A\Psi)\big(-\tfrac{1}{2}e^{\tau A}Af(t_{n+1/2}, \Phi\widetilde{u}_n)\big)\\ &= \widehat{I}^A_{4,1} + \tau(\tau A\Psi)w_{n,2}\end{aligned}$$

with $\|\widehat{I}^A_{4,1}\|_X \le C\tau^2$ by (A4b') and $w_{n,2}$ also satisfies (4.56). Setting $w_n = w_{n,1} + w_{n,2}$ gives (4.57) and part (c) is proved. □

Theorem 4.30. *(Global error of the averaged problem) Let Assumptions 4.1, 4.2, and 4.4 be fulfilled. Moreover, let $(u_n)_n$ be the numerical approximations of a scheme applied to the averaged equation (4.1) that satisfies Assumptions 4.25 and 4.27. Then there is a $\tau_0 > 0$ and a constant $C_e > 0$ such that for all $\tau \le \tau_0$*

$$\|u_n - \widetilde{u}(t_n)\|_{\mathcal{D}(A)} \le C\tau, \qquad 0 \le t_n = n\tau \le t_{\text{end}},$$

The constant C_e and τ_0 depend on u_0, t_{end}, the generalized finite-energy K from Proposition 4.9, the filter functions, and the embedding constant C_{emb}, but are independent of τ and n.

Proof. We proceed as in the proof of Theorem 4.23. Using Assumption 4.25, we get from (4.39) by multiplying with A

$$A\widetilde{e}_{n+1} = e^{(n+1)\tau A}A\widetilde{e}_0 + \tau\sum_{j=0}^{n} e^{(n-j)\tau A}A\mathcal{J}\left(t_j, \widetilde{u}(t_j), u_j\right) - \sum_{j=0}^{n} e^{(n-j)\tau A}A\delta_j. \tag{4.58}$$

In a first step we again establish the bound

$$\Big\|\sum_{j=0}^{n} e^{(n-j)\tau A}A\delta_j\Big\|_X \le C_\delta\tau \tag{4.59}$$

with a constant C_δ being independent of τ and n. Similarly, in the second step we close the proof with the bound in (4.53) and the application of a discrete Gronwall lemma.

(i) The proof is done by induction on n. For $n = 0$, the statement is obviously true. Hence we assume that for all $0 \le k \le n$ it holds

$$\|u_k\|_{\mathcal{D}(A)} \le 2\widetilde{K}, \qquad \|u_k - \widetilde{u}(t_k)\|_{\mathcal{D}(A)} \le C_{\mathrm{e}}\,\tau, \qquad C_{\mathrm{e}} := C_\delta\, e^{C_{\mathcal{J}}(2\widetilde{K})t_{\text{end}}}.$$

By Assumption 4.27, the defect is split into two parts

$$\sum_{j=0}^{n} e^{(n-j)\tau A}A\delta_j = \widetilde{e}^{(1)}_{n+1} + \widetilde{e}^{(D)}_{n+1},$$

analogously to (4.41). Since $\|D_j\|_{\mathcal{D}(A)} \le C\tau^2$ and $n\tau \le t_{\text{end}}$ we immediately obtain

$$\big\|\widetilde{e}^{(D)}_{n+1}\big\|_X = \Big\|\sum_{j=0}^{n} e^{(n-j)\tau A}AD_j\Big\|_X \le C\tau\,.$$

As in (4.43) we use the integration by parts formula (4.37) and obtain

$$\widetilde{e}_{n+1}^{(1)} = \sum_{j=0}^{n} e^{(n-j)\tau A} A\delta_j^{(1)} = \tau E_n(e^{\tau A} - I)\Theta_\Psi w_0 + \tau\Big(\tau \sum_{j=0}^{n-1} E_{n-j-1}(e^{\tau A} - I)\Theta_\Psi \frac{1}{\tau}\big(w_{j+1} - w_j\big)\Big)$$

and estimate it as before to arrive at (4.59).

(ii) Taking norms in (4.58), using (4.59) and $\widetilde{e}_0 = 0$ we have

$$\|\widetilde{e}_{n+1}\|_{\mathcal{D}(A)} \le C_\delta\, \tau + \tau \sum_{j=1}^{n} C_{\mathcal{J}}(2\widetilde{K})\, \|\widetilde{e}_j\|_{\mathcal{D}(A)}\,.$$

A discrete Gronwall Lemma thus yields

$$\begin{aligned}
\|\widetilde{e}_{n+1}\|_{\mathcal{D}(A)} &\le \tau\, C_\delta\, e^{C_{\mathcal{J}}(2\widetilde{K})t_{\mathrm{end}}} = C_{\mathrm{e}}\tau,\\
\|u_{n+1}\|_{\mathcal{D}(A)} &\le \|\widetilde{u}(t_{n+1})\|_{\mathcal{D}(A)} + \|\widetilde{e}_{n+1}\|_{\mathcal{D}(A)} \le \widetilde{K} + C_{\mathrm{e}}\tau \le 2\widetilde{K}
\end{aligned}$$

for $\tau \le \tau_0 \le \frac{\widetilde{K}}{C_{\mathrm{e}}}$ and the induction is closed. □

This leads to the desired error bound in the graph norm.

Theorem 4.31. *Let Assumptions 4.1, 4.2, and 4.4 be fulfilled. Further let $(u_n)_n$ be the numerical approximations of a scheme that satisfies Assumptions 4.25 and 4.27. Then there is a $\tau_0 > 0$ and a constant $C > 0$ such that for all $\tau \le \tau_0$*

$$\|u_n - u(t_n)\|_{\mathcal{D}(A)} \le C\tau, \qquad 0 \le t_n = n\tau \le t_{\mathrm{end}},$$

if the method is applied to the averaged equation (4.1). The constants C and τ_0 depend on u_0, t_{end}, the generalized finite-energy K from Proposition 4.9, the filter functions, and the embedding constant C_{emb}, but are independent of τ and n.

Proof. We simply combine Lemma 4.15 and Theorem 4.30 to conclude

$$\|u(t_n) - u_n\|_{\mathcal{D}(A)} \le \|u(t_n) - \widetilde{u}(t_n)\|_{\mathcal{D}(A)} + \|\widetilde{u}(t_n) - u_n\|_{\mathcal{D}(A)} \le C\tau$$

for $0 \le t_n = n\tau \le t_{\mathrm{end}}$. □

4.6 Error bounds for exponential multistep methods

We briefly indicate how to extend the developed theory to the exponential multistep methods of Section 3.2.2. In order to get a useful representation for the defects we give a proof for the generalization of a known result on quadrature errors.

4.6.1 Peano kernels and defects

Let X be some Hilbert or Banach space and consider a weight $w\colon [0,1] \to \mathcal{B}(X)$. For a sufficiently smooth function $f\colon [0,1] \to X$ we consider

$$T(f) = \sum_{i=1}^{m} \alpha_i f(c_i), \qquad I(f) = \int_0^1 w(s) f(s)\, ds\,, \tag{4.60}$$

where T is a quadrature formula with nodes $c_i \in \mathbb{R}$ and weights $\alpha_i \in \mathcal{B}(X)$, $i = 1, \dots, m$. We recall the notion of polynomials in Banach spaces X from [4, Section IV.3] where any polynomial p is given for some $N \in \mathbb{N}$ and coefficients $k_i \in X$, $i = 0, \dots, N$, as

$$p\colon [0,1] \to X, \quad s \mapsto p(s) = \sum_{i=0}^{N} k_i s^i \,.$$

We assume that the quadrature formula is of degree q, meaning that polynomials in the above sense of degree $q-1$ are integrated exactly. In the case where all nodes c_i lie in $[0,1]$ it is well known that the error functional defined by

$$E(f) = T(f) - I(f)$$

allows for a representation of the form

$$E(f) = \int_0^1 K(s) f^{(q)}(s)\, ds \,,$$

with some bounded $K\colon [0,1] \to \mathcal{B}(X)$, often called the Peano kernel. We will now slightly generalize this result since for multistep methods nodes will also lie outside this interval. We note that the proof is straightforward, however, we could not find any reference. Thus, we give the proof in detail here.

Lemma 4.32. *Consider T and I from (4.60) with degree q and let $[0,1] \subseteq [x_0, x_1]$ such that $c_i \in [x_0, x_1]$, $i = 1, \dots, m$. Then there exists a bounded $K\colon [x_0, x_1] \to \mathcal{B}(X)$ such that*

$$E(f) = \int_{x_0}^{x_1} K(s)\, f^{(q)}(s)\, ds$$

for all q-times differentiable $f\colon [x_0, x_1] \to X$.

Proof. Consider the Taylor expansion of f for $s \in [x_0, x_1]$ by

$$\begin{aligned} f(s) &= f(0) + sf'(0) + \cdots + \tfrac{s^{q-1}}{(q-1)!} f^{(q-1)}(0) + \tfrac{1}{(q-1)!} \int_0^s (s-t)^{(q-1)} f^{(q)}(t)\, dt \\ &= p_f(s) + r(s) \,, \end{aligned}$$

where p_f is a polynomial of degree $q-1$ and r can be written as

$$r(s) = \tfrac{1}{(q-1)!} \int_{x_0}^{x_1} (s-t)^{q-1} f^{(q)}(t) \chi(s,t)\, dt, \qquad \chi(s,t) = \begin{cases} \mathbb{1}_{[0,s]}(t) \,, & 0 \le s \le x_1 \,, \\ -\mathbb{1}_{[s,0]}(t) \,, & x_0 \le s < 0 \,. \end{cases}$$

Since $E(p_f) = 0$ holds, we get

$$\begin{aligned} (q-1)!\, E(f) &= (q-1)!\, E(r) \\ &= \sum_{i=1}^{n} \alpha_i \int_{x_0}^{x_1} (c_i - t)^{q-1} f^{(q)}(t) \chi(c_i, t)\, dt - \int_0^1 \int_{x_0}^{x_1} w(s)(s-t)^{q-1} f^{(q)}(t) \chi(s,t)\, dt\, ds \\ &= \int_{x_0}^{x_1} f^{(q)}(t) \Big(\sum_{i=1}^{n} \alpha_i (c_i - t)^{q-1} \chi(c_i, t) - \int_0^1 w(s)(s-t)^{q-1} \chi(s,t)\, ds \Big)\, dt \\ &= \int_{x_0}^{x_1} f^{(q)}(t)\; E\left((\cdot - t)^{q-1} \chi(\cdot, t) \right)\, dt. \end{aligned}$$

Hence, we obtain the assertion if we define

$$K(t) \coloneqq \tfrac{1}{(q-1)!} E\left((\cdot - t)^{q-1}\chi(\cdot, t)\right)$$

and the boundedness is clear if the kernel w and the weights α_i are bounded. □

4.6.2 Bounds in the X- and the graph norm

In this section we prove first- and second-order error bounds in the X-norm and first-order error bounds in the $\mathcal{D}(A)$-norm. Since the first step is performed by an exponential Euler step or a Lie splitting step, we only mention that in any case the error of the first step is given by

$$\|e_1\|_{\mathcal{D}(A)} \leq C\tau^2 \tag{4.61}$$

by simply adapting the proofs for the inner stages in Propositions 4.22 and 4.26. For the three schemes (3.15), (3.16), and (3.17) Assumption 4.17 needs to be modified.

Exponential multistep method of Adams-type For method (3.15), we denote the numerical flow by $S_\tau(t, v_n, v_{n-1})$ and obtain

$$S_\tau(t, v_n, v_{n-1}) - S_\tau(t, w_n, w_{n-1}) = e^{\tau A}\big(v_n - w_n\big) + \tau \mathcal{J}_n, \tag{4.62}$$

where $\mathcal{J}_n = \mathcal{J}\big(t, v_n, v_{n-1}, w_n, w_{n-1}\big)$ satisfies by (A5a') similar to (4.18) the bound

$$\begin{aligned} \|\mathcal{J}_n\|_X \leq{}& C_{\mathcal{J}}\big(\|v_n\|_X, \|w_n\|_X\big)\, \|v_n - w_n\|_X \\ &+ C_{\mathcal{J}}\big(\|v_{n-1}\|_X, \|w_{n-1}\|_X\big)\, \|v_{n-1} - w_{n-1}\|_X, \qquad t \in [0, t_{\text{end}}], \end{aligned} \tag{4.63}$$

and also by (A5b') in the graph norm

$$\begin{aligned} \|\mathcal{J}_n\|_{\mathcal{D}(A)} \leq{}& C_{\mathcal{J}}\big(\|v_n\|_{\mathcal{D}(A)}, \|w_n\|_{\mathcal{D}(A)}\big)\, \|v_n - w_n\|_{\mathcal{D}(A)} \\ &+ C_{\mathcal{J}}\big(\|v_{n-1}\|_{\mathcal{D}(A)}, \|w_{n-1}\|_{\mathcal{D}(A)}\big)\, \|v_{n-1} - w_{n-1}\|_{\mathcal{D}(A)}, \qquad t \in [0, t_{\text{end}}]. \end{aligned} \tag{4.64}$$

This yields the following error bound.

Corollary 4.33 ([9, Cor. 7.1])**.** *Let Assumptions 4.1, 4.2, and 4.4 be valid and consider the numerical approximations $(u_n)_n$ from (3.15). Then the following error bounds hold:*

(a) *If the method is applied to the original equation (2.4), then there is a $\tau_0 > 0$ and a constant $C > 0$ such that for all $\tau \leq \tau_0$*

$$\|u(t_n) - u_n\|_X + \|u(t_n) - u_n\|_{\mathcal{D}(A)} \leq C_1\tau, \qquad 0 \leq t_n = n\tau \leq t_{\text{end}}.$$

(b) *If the method is applied to the averaged equation (4.1) with $\psi = 1$ and a filter ϕ of order 2, then there is a $\tau_0 > 0$ and a constant $C > 0$ such that for all $\tau \leq \tau_0$*

$$\|u(t_n) - u_n\|_X \leq C_2\tau^2, \qquad 0 \leq t_n = n\tau \leq t_{\text{end}}.$$

Here, C_1, C_2, and τ_0 depend on u_0, t_{end}, the generalized finite-energy K from Proposition 4.9, the embedding constant C_{emb}, C_2, and in addition on the filter functions, but are independent of τ and n.

Proof. (a) For the first part is is sufficient to prove that the defects are of order 2 since we can then use (4.64) conclude by the standard arguments. As in the proof of [40, Thm. 4.3] the defect stems from a quadrature error that can be represented by Lemma 4.32 as

$$\delta_n = S_\tau\big(t_n, \widetilde{u}(t_n), \widetilde{u}(t_{n-1})\big) - \widetilde{u}(t_{n+1}) = \tau^2 \int_{-1}^{1} K_1(s)\, \widetilde{f}'_{n+s}\, ds\,,$$

where we use the notation of (4.31) for the derivatives of $\widetilde{f}_{n+s}$. The integral term can be bounded uniformly in both norms by Assumption (A1'). We obtain by (4.62) the error recursion

$$\begin{aligned}\widetilde{e}_{n+1} &= S_\tau\big(t_n, \widetilde{u}_n, \widetilde{u}_{n-1}\big) - S_\tau\big(t_n, u_n, u_{n-1}\big) - \delta_n \\ &= e^{\tau A}\widetilde{e}_n + \tau \mathcal{J}_n - \delta_n\,,\end{aligned}$$

which is resolved by

$$\widetilde{e}_{n+1} = e^{n\tau A}\widetilde{e}_1 + \tau \sum_{j=1}^{n} e^{(n-j)\tau A}\mathcal{J}_j - \sum_{j=1}^{n} e^{(n-j)\tau A}\delta_j. \tag{4.65}$$

Since the last term is bounded by

$$\Big\| \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j \Big\|_{\mathcal{D}(A)} \le C\tau\,,$$

and (4.61) holds, the assertion is easily derived by a Gronwall lemma. The bound in the X-norm follows from Lemma 2.8.

(b) In order to prove the second statement, we first employ Theorem 4.14 and Lemma 4.15, so again it remains to prove the error in approximating the filtered solution. We obtain the similar representation

$$\delta_n = \tau^3 \int_{-1}^{1} K_2(s)\, \widetilde{f}''_{n+s}\, ds\,,$$

which yields the dominant terms as in (4.32). As above, it also satisfies the conditions on W_n in Assumption 4.20. We note that the error recursion in (4.65) is still valid and with (4.63) we may close the proof by the lines of the one of Theorem 4.23.

□

Exponential multistep methods of Nyström-type For the methods (3.16) and (3.17) we have

$$S_\tau(t, v_n, v_{n-1}) - S_\tau(t, w_n, w_{n-1}) = e^{2\tau A}\big(v_{n-1} - w_{n-1}\big) + \tau \mathcal{J}_n\,, \tag{4.66}$$

where $\mathcal{J}_n = \mathcal{J}\,(t, v_n, w_n)$ is bounded with (A5a') by

$$\|\mathcal{J}_n\|_X \le C_{\mathcal{J}}\big(\|v_n\|_X\,, \|w_n\|_X\big)\, \|v_n - w_n\|_X\,, \qquad t \in [0, t_{\mathrm{end}}]\,, \tag{4.67}$$

and in the stronger norm with (A5b') by

$$\|\mathcal{J}_n\|_{\mathcal{D}(A)} \le C_{\mathcal{J}}\big(\|v_n\|_{\mathcal{D}(A)}\,, \|w_n\|_{\mathcal{D}(A)}\big)\, \|v_n - w_n\|_{\mathcal{D}(A)}\,, \qquad t \in [0, t_{\mathrm{end}}]\,. \tag{4.68}$$

In order to apply the techniques from above we define the modification

$$\chi_2 : \mathcal{C}_b\big(i\mathbb{R}\big) \to \mathcal{C}_b\big(i\mathbb{R}\big), \quad \chi(\cdot) \mapsto \chi(2\cdot)\,,$$

and can state the following result.

Corollary 4.34 ([9, Cor. 7.2])**.** *Let Assumptions 4.1, 4.2, and 4.4 be valid and u be the classical solution of (2.4). Consider the numerical approximations $(u_n)_n$ from (3.16).*

(a) *If the method is applied to the original equation (2.4), then there is a $\tau_0 > 0$ and a constant $C > 0$ such that for all $\tau \le \tau_0$ it holds*

$$\|u(t_n) - u_n\|_X \le C\tau, \qquad 0 \le t_n = n\tau \le t_{\mathrm{end}}\,.$$

(b) *If the method is applied to the averaged equation (4.1) with filters $\chi_2\psi, \chi_2\phi$, where ψ, ϕ are filters of order 2, then there is a $\tau_0 > 0$ and a constant $C > 0$ such that for all $\tau \le \tau_0$*

$$\|u(t_n) - u_n\|_X \le C\tau^2, \qquad 0 \le t_n = n\tau \le t_{\mathrm{end}}.$$

(c) *If the method is applied to the averaged equation (4.1) with $\phi = 1$ and the filter $\chi_2\psi$, where ψ is a filter of order 1, then there is a $\tau_0 > 0$ and a constant $C > 0$ such that for all $\tau \le \tau_0$*

$$\|u(t_n) - u_n\|_{\mathcal{D}(A)} \le C\tau, \qquad 0 \le t_n = n\tau \le t_{\mathrm{end}}.$$

Here, C and τ_0 depend on u_0, t_{end}, the generalized finite-energy K from Proposition 4.9, the filter functions, and the embedding constant C_{emb}, but are independent of τ and n.

Proof. (a) Since the method stems from a midpoint rule applied to the variation-of-constants formula, the defect can again be written as

$$\begin{aligned}\delta_n &= \tau \int_{-1}^{1} K_1(s)\, \tfrac{d}{ds}\big(e^{(1-s)\tau A}\widetilde{f}_{n+s}\big)\, ds \\ &= \tau^2 \int_{-1}^{1} K_1(s)\, e^{(1-s)\tau A}\big(\widetilde{f}'_{n+s} - A\widetilde{f}_{n+s}\big)\, ds\,,\end{aligned}$$

and we may bound the integral term uniformly by (A1') and (A3'). We use (4.66) to obtain the error recursion

$$\begin{aligned}\widetilde{e}_{n+1} &= S_\tau\big(t_n, \widetilde{u}_n, \widetilde{u}_{n-1}\big) - S_\tau\big(t_n, u_n, u_{n-1}\big) - \delta_n \\ &= e^{2\tau A}\widetilde{e}_{n-1} + \tau\mathcal{J}_n - \delta_n\,,\end{aligned}$$

which is resolved and bounded by

$$\|e_{n+1}\|_X \le \|e_1\|_X + \tau \sum_{0\le j\le \frac{n}{2}} \|\mathcal{J}_{n-2j}\|_X + \Big\| \sum_{0\le j\le \frac{n}{2}} e^{2j\tau A}\delta_{n-2j}\Big\|_X\,. \tag{4.69}$$

Since the last term is bounded by $\mathcal{O}(\tau)$ and (4.61) holds, the assertion directly follows from (4.67).

(b) As the representation in (4.69) is still valid we only have to bound the last term by $\mathcal{O}(\tau^2)$. The defect can be written as

$$\delta_n = \tau \int_{-1}^{1} K_2(s)\, \tfrac{d^2}{ds^2}\big(e^{(1-s)\tau A}\widetilde{f}_{n+s}\big)\, ds\,,$$

and we obtain dominant terms similar to (4.26) and (4.32). They also satisfy Assumption 4.20 if we replace the following properties in (4.27) by

$$\Big\|\frac{1}{2\tau}(w_{n+2}-w_n)\Big\|_X \le C\,, \qquad \Big\|\frac{1}{2\tau}(W_{n+2}-W_n)\Big\|_X \le C\,. \tag{4.70}$$

As e^z in (F3) is replaced by e^{2z}, this can be combined to conclude the assertion similar to the proof of Theorem 4.23.

(c) The last part easily follows by the arguments of part (*b*), the bounds derived in Proposition 4.29, the stability in (4.68), and the ideas of the proof of Theorem 4.30. □

Corollary 4.35. *Let Assumptions 4.1, 4.2, and 4.4 be valid and u be the classical solution of (2.4). Consider the numerical approximations $(u_n)_n$ from (3.17).*

(a) *If the method is applied to the original equation (2.4), then there is a $\tau_0 > 0$ and a constant $C > 0$ such that for all $\tau \le \tau_0$ it holds*

$$\|u(t_n) - u_n\|_X + \|u(t_n) - u_n\|_{\mathcal{D}(A)} \le C\tau, \qquad 0 \le t_n = n\tau \le t_{\mathrm{end}}\,.$$

(b) *If the method is applied to the averaged equation (4.1) with filter $\chi_2\phi$, where ϕ is a filter of order 2, then there is a $\tau_0 > 0$ and a constant $C > 0$ such that for all $\tau \le \tau_0$*

$$\|u(t_n) - u_n\|_X \le C\tau^2, \qquad 0 \le t_n = n\tau \le t_{\mathrm{end}},$$

Here, C and τ_0 depend on u_0, t_{end}, the generalized finite-energy K from Proposition 4.9, the filter function, and the embedding constant C_{emb}, but are independent of τ and n.

Proof. The proof combines the ideas of Proposition 4.22 and Corollary 4.34, and it only remains to investigate the defect. The bound in the X-norm in part (a) again follows from Lemma 2.8.

(a) We use the definition of φ_1 to compute

$$\begin{aligned} A\delta_n &= 2\tau\Big(\varphi_1(2\tau A)A\widetilde{f}_n - \int_0^1 e^{(1-\xi)2\tau A}A\widetilde{f}_{n-1+2\xi}\, d\xi\Big) \\ &= 2\tau\int_0^1 e^{(1-\xi)2\tau A}A\big(\widetilde{f}_n - \widetilde{f}_{n+2\xi-1}\big)\, d\xi\,, \end{aligned}$$

which directly gives $\|\delta_n\|_{\mathcal{D}(A)} \le C\tau^2$ by (A1').

(b) For second-order we compute for the defect expanding with Taylor as in (4.31)

$$\begin{aligned} \delta_n &= 2\tau\Big(\varphi_1(2\tau A)\widetilde{f}_n - \int_0^1 e^{(1-\xi)2\tau A}\widetilde{f}_{n+2\xi-1}\, d\xi\Big) \\ &= -2\tau^2\int_0^1 e^{(1-\xi)2\tau A}(2\xi-1)\widetilde{f}'_n\, d\xi - 2\tau^3\int_0^1(2\xi-1)^2\int_0^1(1-s)\widetilde{f}''_{n+(2\xi-1)s}\, ds\, d\xi \\ &= \widehat{I}_1 + \widehat{I}_2\,. \end{aligned}$$

For the first term we use integration by parts and obtain with (A1')

$$\left\|\widehat{I}_1\right\|_X = 4\tau^3 \left\| \int_0^1 e^{(1-\xi)2\tau A}(\xi^2-\xi)A\widetilde{f}_n' \, d\xi \right\|_X \le C\tau^3 .$$

Along the lines of (4.33) we deduce from $\widehat{I}_2$ the map W_n which also satisfies the second part of (4.70). □

4.7 Error bounds for first-order methods with mild solutions

In this last section we consider the linear version of equation (2.4)

$$u'(t) = Au(t) + Fu(t), \qquad u(0) = u_0 \in X, \tag{4.71}$$

which is discretized by the first-order schemes presented in (3.4) and (3.5). The error analysis is performed under the following assumption on the linear term F.

Assumption 4.36. *The linear operator $F\colon X \to X$ satisfies the bounds*

$$\|Fx\|_X, \ \|Fx\|_{\mathcal{D}(A)} \le L_F \|x\|_X, \qquad x \in X .$$

Remark 4.37. *In the second-order formulation (2.3), consider a linear operator G satisfying*

$$G\colon V \to V, \qquad \|Gq\|_V \le C \|q\|_V .$$

If we define the operator

$$Fu = \begin{pmatrix} 0 \\ Gq \end{pmatrix}, \qquad u = \begin{pmatrix} q \\ v \end{pmatrix},$$

then F satisfies Assumption 4.36.

From Theorem 2.17 we immediately obtain a mild solution $u \in C([0,t_{\mathrm{end}}],X)$ and it holds the variation-of-constants formula

$$u(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A}Fu(s)\,ds$$

for all $t \in [0,t_{\mathrm{end}}]$. We emphasize that the initial value only satisfies $u_0 \in X$ and hence there is no hope for a classical solution of (4.71). Nevertheless, we are able to prove an error bound of order one under this regularity. We start with an explicit bound on the norm of the solution u.

Lemma 4.38. *Let Assumption 4.36 be satisfied. Then the mild solution $u \in C([0,t_{\mathrm{end}}],X)$ of (4.71) satisfies the bound*

$$\|u(t)\|_X \le \|u_0\|_X e^{L_F t} \le \|u_0\|_X e^{L_F t_{\mathrm{end}}} =: r_X .$$

Proof. We simply compute

$$\|u(t)\|_X \le \left\|e^{tA}u_0\right\|_X + \left\| \int_0^t e^{(t-s)A}Fu(s)\,ds \right\|_X \le \|u_0\|_X + \int_0^t L_F \|u(s)\|_X \, ds ,$$

and a Gronwall lemma yields the assertion. □

Let ϕ be a filter of order 1 and define $\widetilde{F}\widetilde{u} = F\Phi\widetilde{u}$. We then consider the linear version of equation (4.1)

$$\widetilde{u}'(t) = A\widetilde{u}(t) + \widetilde{F}\widetilde{u}(t), \qquad \widetilde{u}(0) = u_0 \in X\,, \tag{4.72}$$

and directly obtain the following bound.

Corollary 4.39. *Let Assumption 4.36 be satisfied. Then the mild solution $\widetilde{u} \in C([0,t_{\mathrm{end}}],X)$ of (4.72) satisfies the bound*

$$\sup_{t\in[0,t_{\mathrm{end}}]} \|\widetilde{u}(t)\|_X \le r_X\,.$$

Proof. We only use that by (OF1) it holds

$$\left\|\widetilde{F}x\right\|_X,\ \left\|\widetilde{F}x\right\|_{\mathcal{D}(A)} \le L_F\,\|x\|_X$$

and conclude by the lines of Lemma 4.38. □

In the next step we bound the difference of the original and the averaged solution. The idea is the same as before, but we need to take care of the lack of regularity when deriving the error terms.

Theorem 4.40. *Let Assumption 4.36 be valid and consider the averaged nonlinearity $\widetilde{F}$ with a first-order filter. Then there is a constant $C_{av} > 0$ such that for all $\tau > 0$*

$$\|u(t) - \widetilde{u}(t)\|_X \le C_{av}\tau, \qquad 0 \le t \le t_{\mathrm{end}}\,.$$

The constant C_{av} depends on L_F, u_0, t_{end}, the filter functions, and the embedding constant C_{emb}, but not on τ.

Proof. We employ the variation-of-constants formula to write

$$\begin{aligned} u(t) - \widetilde{u}(t) &= \int_0^t e^{(t-s)A}\left(Fu(s) - \widetilde{F}\widetilde{u}(s)\right)\,ds \\ &= \int_0^t e^{(t-s)A}F\big(I-\Phi\big)u(s)\,ds + \int_0^t e^{(t-s)A}\widetilde{F}\big(u(s)-\widetilde{u}(s)\big)\,ds \\ &= I_1(t) + I_2(t)\,. \end{aligned} \tag{4.73}$$

By (OF1) and Assumption 4.36 the second term in (4.73) is bounded by

$$\|I_2(t)\|_X = \left\|\int_0^t e^{(t-s)A}\widetilde{F}\big(u(s)-\widetilde{u}(s)\big)\,ds\right\|_X \le L_F\int_0^t \|u(s)-\widetilde{u}(s)\|_X\,ds.$$

It remains to prove

$$\|I_1(t)\|_X \le C\tau\,, \tag{4.74}$$

since this bound is sufficient to apply a Gronwall lemma. To bound I_1 we use the variation-of-constants formula to obtain

$$\begin{aligned} I_1(t) &= \int_0^t e^{(t-s)A}F\big(I-\Phi\big)e^{sA}u_0\,ds, \\ &\qquad + \int_0^t e^{(t-s)A}F\big(I-\Phi\big)\int_0^s e^{(s-\theta)A}Fu(\theta)\,d\theta\,ds, \\ &= I_{1,1}(t) + I_{1,2}(t) \end{aligned}$$

For the first term we again use (OF3) and integration by parts to obtain

$$\begin{aligned} I_{1,1}(t) &= \tau \int_0^t e^{(t-s)A} F e^{sA} A\vartheta(\tau A) u_0 \, ds, \\ &= \tau \Big(\big[e^{(t-s)A} F e^{sA} \vartheta(\tau A) u_0 \big]_0^t + \int_0^t e^{(t-s)A} A F e^{sA} \vartheta(\tau A) u_0 \, ds \Big) , \end{aligned}$$

which gives the first part of (4.74). The second part follows by the estimate

$$\|I_{1,2}\|_X (t) = \tau \left\| \int_0^t e^{(t-s)A} F \vartheta(\tau A) \int_0^s e^{(s-\theta)A} A F u(\theta) \, d\theta \, ds \right\|_X \leq C\tau ,$$

and the assertion is proved. □

In order to stick to the established framework of the preceding sections we formulate the properties of the two first-order schemes as abstract assumptions.

Assumption 4.41 (Stability)**.** *The method applied to (4.72) is stable in the sense that for all $v, w \in X$, $t \geq 0$,*

$$S_\tau(v) - S_\tau(w) = e^{\tau A} (v - w) + \tau \mathcal{J} (v, w) ,$$

where $\mathcal{J} : X \times X \to X$ is bounded by

$$\|\mathcal{J} (v, w)\|_X \leq C_{\mathcal{J}} \|v - w\|_X . \tag{4.75}$$

As we have already seen for the error bound in the graph norm, the structure of the defect becomes simpler if one only wants to prove bounds of order 1. We think it is worth mentioning that in comparison to Assumption 4.27 we now have $\delta_n^{(2)}$ instead of $\delta_n^{(1)}$ whereas both appeared originally in Assumption 4.20.

Assumption 4.42 (Structure of defects)**.** *The defect δ_n defined in (4.14) of a numerical method applied to the averaged equation (4.72) is of the form*

$$\delta_n = \delta_n^{(2)} + D_n$$

with $\|D_n\|_X \leq C\tau^2$, where the constant $C > 0$ is independent of τ and n. In addition, there exists a linear map $W \colon X \to \mathcal{D}(A)$ which satisfies

$$\|W\|_{X \leftarrow X} \leq C, \qquad \|AW\|_{X \leftarrow X} \leq C, \tag{4.76}$$

with a constant C which is independent of τ and n such that $\delta_n^{(2)}$ can be written as

$$\delta_n^{(2)} = \tau W\big(\tau A \Phi\big) \widetilde{u}_n . \tag{4.77}$$

Proposition 4.43. *Let Assumption 4.36 be satisfied. The exponential Euler method (3.4) and the Lie splitting (3.5) applied to the averaged equation (4.72) satisfy Assumptions 4.41 and 4.42.*

Proof. (a) We first investigate the Lie splitting (3.5). Concerning the stability, we note that Assumption 4.41 is fulfilled with

$$J(v, w) = e^{\tau A} \widetilde{F}\big(v - w\big) .$$

which clearly satisfies (4.75). By the variation-of-constants formula we expand the defect as

$$\begin{aligned}
\delta_n &= e^{\tau A}\big(I+\tau\widetilde{F}\big)\widetilde{u}_n - \widetilde{u}_{n+1}\\
&= e^{\tau A}\tau\widetilde{F}\widetilde{u}_n - \tau\int_0^1 e^{(1-s)\tau A}\widetilde{F}\widetilde{u}_{n+s}\,ds\\
&= \tau e^{\tau A}\widetilde{F}\widetilde{u}_n - \tau\int_0^1 e^{(1-s)\tau A}\widetilde{F}e^{s\tau A}\widetilde{u}_n\,ds - \tau^2\int_0^1\int_0^s e^{(1-s)\tau A}\widetilde{F}e^{(s-\sigma)\tau A}\widetilde{F}\widetilde{u}_{n+\sigma}\,d\sigma\,ds\\
&= \tau\int_0^1 \big(e^{\tau A}\widetilde{F}\widetilde{u}_n - e^{(1-s)\tau A}\widetilde{F}e^{s\tau A}\widetilde{u}_n\big)\,ds - \tau^2\int_0^1\int_0^s e^{(1-s)\tau A}\widetilde{F}e^{(s-\sigma)\tau A}\widetilde{F}\widetilde{u}_{n+\sigma}\,d\sigma\,ds\\
&= \tau\int_0^1 \big(e^{\tau A}F - e^{(1-s)\tau A}Fe^{s\tau A}\big)\Phi\widetilde{u}_n\,ds + D_n^1\,.
\end{aligned}$$

By (OF2) we have $\Phi\widetilde{u}_n$ in $D(A)$ and hence we may differentiate the semigroup in the following computation

$$\begin{aligned}
\delta_n &= \tau\int_0^1\int_{1-s}^1 \tfrac{d}{d\sigma}\left(e^{\sigma\tau A}Fe^{(1-\sigma)\tau A}\right)\Phi\widetilde{u}_n\,d\sigma\,ds + D_n^1\\
&= \tau^2\int_0^1\int_{1-s}^1 e^{\sigma\tau A}\big(AF-FA\big)e^{(1-\sigma)\tau A}\Phi\widetilde{u}_n\,d\sigma\,ds + D_n^1\\
&= -\tau^2\int_0^1\int_{1-s}^1 e^{\sigma\tau A}Fe^{(1-\sigma)\tau A}A\Phi\widetilde{u}_n\,d\sigma\,ds + \tau^2\int_0^1\int_{1-s}^1 e^{\sigma\tau A}AFe^{(1-\sigma)\tau A}\Phi\widetilde{u}_n\,d\sigma\,ds + D_n^1\\
&= \tau^2 WA\Phi\widetilde{u}_n + D_n^2 + D_n^1
\end{aligned}$$

with

$$Wx = -\int_0^1\int_{1-s}^1 e^{\sigma\tau A}Fe^{(1-\sigma)\tau A}x\,d\sigma\,ds\,.$$

Hence, (4.76) and (4.77) are satisfied by Assumption 4.36. We further set $D_n = D_n^1 + D_n^2$ and obtain the bound $\|D_n\|_X \le C\tau^2$ which yields the claim of Assumption 4.42.

(b) We similarly proceed for the exponential Euler method (3.4) and note that Assumption 4.41 is fulfilled with

$$J(v,w) = \varphi_1(\tau A)\widetilde{F}\big(v-w\big)\,.$$

which satisfies (4.75). By the variation-of-constants formula we expand the defect as

$$\begin{aligned}
\delta_n &= e^{\tau A}\widetilde{u}_n + \tau\varphi_1(\tau A)\widetilde{F}\widetilde{u}_n - \widetilde{u}_{n+1}\\
&= \tau\int_0^1 e^{(1-s)\tau A}\widetilde{F}\widetilde{u}_n\,ds - \tau\int_0^1 e^{(1-s)\tau A}\widetilde{F}\widetilde{u}_{n+s}\,ds\\
&= \tau\int_0^1 e^{(1-s)\tau A}F\big(I-e^{s\tau A}\big)\Phi\widetilde{u}_n\,ds - \tau^2\int_0^1\int_0^s e^{(1-s)\tau A}\widetilde{F}e^{(s-\sigma)\tau A}\widetilde{F}\widetilde{u}_{n+\sigma}\,d\sigma\,ds\\
&= \tau^2 WA\Phi\widetilde{u}_n + D_n
\end{aligned}$$

with

$$Wx = -\int_0^1 \int_{1-s}^1 s e^{\sigma\tau A} F\varphi_1(s\tau A)x \, d\sigma \, ds\,.$$

As before, (4.76) and (4.77) are satisfied, and it holds $\|D_n\|_X \le C\tau^2$. □

Theorem 4.44. *(Global error of the averaged problem) Let Assumption 4.36 be fulfilled. Moreover, let $(u_n)_n$ be the numerical approximations of a scheme applied to the averaged equation (4.72) that satisfies Assumptions 4.41 and 4.42. Then there is a constant $C_e > 0$ such that for all $\tau \ge 0$*

$$\|u_n - \widetilde{u}(t_n)\|_X \le C_e\tau, \qquad 0 \le t_n = n\tau \le t_{\mathrm{end}},$$

The constant C_e depends on u_0, t_{end}, the radius L_F, the filter functions, and the embedding constant C_{emb}, but is independent of τ and n.

Proof. We proceed as in the proof of Theorem 4.23 and expand the global error by Assumption 4.41

$$\widetilde{e}_{n+1} = e^{(n+1)\tau A}\widetilde{e}_0 + \tau\sum_{j=0}^{n} e^{(n-j)\tau A}\mathcal{J}\big(\widetilde{u}(t_j), u_j\big) - \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j.$$

Once we established the bound

$$\Big\|\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j\Big\|_X \le C_\delta\tau \tag{4.78}$$

with a constant C_δ being independent of τ and n. The proof is closed by a discrete Gronwall lemma which then yields

$$\|\widetilde{e}_{n+1}\|_X \le \tau\, C_\delta\, e^{C_{\mathcal{J}} t_{\mathrm{end}}}\,.$$

By Assumption 4.42, the defect is split into two parts, which motivates to write

$$\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j = \widetilde{e}_{n+1}^{(2)} + \widetilde{e}_{n+1}^{(D)},$$

where

$$\widetilde{e}_{n+1}^{(2)} = \sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(2)}, \qquad \widetilde{e}_{n+1}^{(D)} = \sum_{j=0}^{n} e^{(n-j)\tau A}D_j.$$

Since $\|D_j\|_X \le C\tau^2$ and $n\tau \le t_{\mathrm{end}}$ we easily see

$$\Big\|\widetilde{e}_{n+1}^{(D)}\Big\|_X = \Big\|\sum_{j=0}^{n} e^{(n-j)\tau A}D_j\Big\|_X \le C\tau\,.$$

Next we consider $\widetilde{e}_{n+1}^{(2)}$. Recall F_n from (4.42) and, as in (4.45), we arrive at

$$\begin{aligned}\sum_{j=0}^{n} e^{(n-j)\tau A}\delta_j^{(2)} &= \tau W\Theta_\Phi(e^{\tau A} - I)F_n \\ &\quad + \tau\Big(\tau\sum_{j=0}^{n-1} e^{(n-j)\tau A}\frac{1}{\tau}\big(I - e^{-\tau A}\big)W\Theta_\Phi(e^{\tau A} - I)F_j\Big).\end{aligned}$$

We estimate by (4.76) the difference

$$\Big\|\frac{1}{\tau}(e^{-\tau A} - I)W\Big\|_{X\leftarrow X} = \Big\|\varphi_1(-\tau A)AW\Big\|_{X\leftarrow X} \le C,$$

since $|\varphi_1(z)| \leq 1$ for $z \in i\mathbb{R}$.

Next we consider $(e^{\tau A} - I)F_j$ for $j \leq n$. After adding the exact solution we apply the variation-of-constants formula, which gives

$$\begin{aligned}\left\|(e^{\tau A} - I)F_j\right\|_X &= \Big\| \sum_{k=0}^{j} (e^{\tau A}\widetilde{u}(t_k) - \widetilde{u}(t_k + \tau)) + \sum_{k=0}^{j} (\widetilde{u}(t_k + \tau) - \widetilde{u}(t_k)) \Big\|_X \\ &= \Big\| \sum_{k=0}^{j} \int_0^\tau e^{(\tau - s)A} \widetilde{F}\widetilde{u}(t_k + s)\, ds + (\widetilde{u}(t_{j+1}) - \widetilde{u}_0) \Big\|_X \\ &\leq t_{\mathrm{end}} L_F r_X + 2r_X \,.\end{aligned}$$

This yields (4.78) and thus the assertion. □

From this we may conclude the final error bound of this section.

Theorem 4.45. *Let Assumption 4.36 be fulfilled. Further let $(u_n)_n$ be the numerical approximations of the exponential Euler method (3.4) or the Lie splitting (3.5) applied to the averaged equation (4.72). Then there is a constant $C > 0$ such that for all $\tau \geq 0$*

$$\|u_n - u(t_n)\|_X \leq C\tau, \qquad 0 \leq t_n = n\tau \leq t_{\mathrm{end}} \,.$$

The constant C depends on u_0, t_{end}, the radius L_F, the filter functions, and the embedding constant C_{emb}, but is independent of τ and n.

Proof. We simply combine Theorem 4.40 and Theorem 4.44 using the triangle inequality. □

APPENDIX A

Semilinear examples

This appendix is devoted to the verification of the assumptions made in Sections 2.1.1 and 4.2.1. We show that Example 2.5 with its specification made in Table 4.1 is fully covered by the analysis presented in Part I. Hence, we check every column of Table 4.1 in the following sections.

A.1 Basic estimates

Throughout we need estimates related to Sobolev spaces and products of functions lying in them. We collect them in this section. Several times we employ for a bounded Lipschitz domain $\emptyset \neq \Omega \subseteq \mathbb{R}^d$, $d \in \{1,2,3\}$ the continuous embeddings [1, Theorem 4.12]

$$\begin{aligned} &H^1(\Omega) \hookrightarrow L^\infty(\Omega), && d = 1, \\ &H^1(\Omega) \hookrightarrow L^q(\Omega), && d = 2, \quad q \in [1,\infty), \\ &H^1(\Omega) \hookrightarrow L^6(\Omega), && d = 3, \\ &H^2(\Omega) \hookrightarrow L^\infty(\Omega), && d = 2,3. \end{aligned} \tag{A.1}$$

Throughout, we consider the norms

$$\begin{aligned} \|q\|_{H_0^1}^2 &= \|\nabla q\|_{L^2}^2 , \\ \|q\|_{H^1}^2 &= \|q\|_{L^2}^2 + \|\nabla q\|_{L^2}^2 , \\ \|q\|_{H^2}^2 &= \|q\|_{L^2}^2 + \|\nabla q\|_{L^2}^2 + \sum_{i,j=1}^{d} \left\|\partial_{x_i}\partial_{x_j} q\right\|_{L^2}^2 . \end{aligned} \tag{A.2}$$

In the next lemma we collect some estimates that remain valid for the dimensions $d = 1,2,3$. They either directly follow from (A.1) or are extensions of the computations in [24, Section 7.4].

Lemma A.1. *Let $\Omega \subseteq \mathbb{R}^d$ be bounded with $d = 1, 2, 3$. Then the following estimates hold for functions f, g in the respective spaces:*

$$\|fg\|_{L^2} \leq C \|f\|_{H^1} \|g\|_{H^1}, \tag{A.3}$$
$$\|fg\|_{L^2} \leq C \|f\|_{L^2} \|g\|_{H^2}, \tag{A.4}$$
$$\|fg\|_{H^1} \leq C \|f\|_{H^1} \|g\|_{H^2}, \tag{A.5}$$
$$\|fg\|_{H^2} \leq C \|f\|_{H^2} \|g\|_{H^2}. \tag{A.6}$$

Using the notation of Example 2.5, assume that for $\psi\colon (t, x, y) \mapsto \mathbb{R}$ all partial derivatives $\partial^\beta \psi$, $\beta \leq \alpha$, exist, are continuous in t and y and bounded in x. Then for $p \in [2, \infty)$ it holds

$$\alpha = (0,0,0): \quad \|\psi(t,\cdot,f)\|_{L^2} \leq C\big(\|f\|_{L^\infty}\big), \tag{A.7}$$
$$\alpha = (1,0,1): \quad \|\psi(t,\cdot,f) - \psi(s,\cdot,g)\|_{L^p} \leq C\big(\|f\|_{L^\infty}, \|g\|_{L^\infty}\big)\big(|t-s| + \|f-g\|_{L^p}\big), \tag{A.8}$$
$$\alpha = (0,1,1): \quad \|\psi(t,\cdot,f)\|_{H^1} \leq C\big(\|f\|_{L^\infty}\big) \|f\|_{H^1}, \tag{A.9}$$
$$\alpha = (1,1,2): \quad \|\psi(t,\cdot,f) - \psi(s,\cdot,g)\|_{H^1} \leq C\big(\|f\|_{H^2}, \|g\|_{H^2}\big)\big(|t-s| + \|f-g\|_{H^1}\big), \tag{A.10}$$
$$\alpha = (0,2,2): \quad \|\psi(t,\cdot,f)\|_{H^2} \leq C\big(\|f\|_{H^2}\big), \tag{A.11}$$
$$\alpha = (1,2,3): \quad \|\psi(t,\cdot,f) - \psi(s,\cdot,g)\|_{H^2} \leq C\big(\|f\|_{H^2}, \|g\|_{H^2}\big)\big(|t-s| + \|f-g\|_{H^2}\big), \tag{A.12}$$

We further denote the evaluation of a function G at a function $q\colon [0,T] \mapsto \mathcal{X}$, where $\mathcal{X}$ is some Banach space by

$$\widehat{G}(t) \coloneqq G(t, q(t)), \qquad \widehat{G}_t(t) \coloneqq \partial_t G(t,y)|_{y=q(t)}$$

and for higher derivatives analogously.

A.2 $\mathcal{H} = H^{-1}(\Omega)$

We start with the first column of Table 4.1 with $d = 1$ and $\Omega \subseteq \mathbb{R}$ some finite interval. The operator L is defined on $\mathcal{H} = H^{-1}(\Omega)$ by

$$\langle Lq, \phi \rangle_{H^{-1} \times H_0^1} = \langle \mathbf{A} \nabla q, \nabla \phi \rangle_{L^2}, \quad q, \phi \in H_0^1(\Omega),$$

for some uniformly positive $\mathbf{A} \in L^\infty(\Omega)$, i.e., $\mathbf{A} \geq \delta$ almost everywhere. The additional spaces are given by

$$V = L^2(\Omega), \qquad \mathcal{D}(L) = H_0^1(\Omega),$$

but we need to be careful with the choice of the norms. Usually, $H_0^1(\Omega)$ is equipped with the inner product $\langle u, v \rangle_{H_0^1} = \langle \nabla u, \nabla v \rangle_{L^2}$ which induces a norm by the Friedrich's inequality. Further, its dual comes with the operator norm

$$\|f\|_{H^{-1}} = \sup_{\|u\|_{H_0^1} = 1} \langle f, u \rangle_{H^{-1} \times H_0^1}. \tag{A.13}$$

However, we need to work with equivalent norms below. For the nonlinearity g we assume $\alpha = (2,0,2)$ and the growth bounds

$$\begin{aligned} |g(t,x,y)|, |\partial_t g(t,x,y)| &\leq C_{\mathrm{g}}\big(1 + |y|^2\big), \\ |\partial_y g(t,x,y)| &\leq C_{\mathrm{g}}\big(1 + |y|\big). \end{aligned} \tag{A.14}$$

For the corrected Lie Splitting (3.14) we assume in addition

$$|\partial_{yy} g(t,x,y)| \leq C_{\mathrm{g}}\big(1+|y|\big)\,. \tag{A.15}$$

We first consider the operator theoretic assumptions from Section 2.1.1.

Lemma A.2. *The operator $L\colon \mathcal{D}(L) \subseteq \mathcal{H} \to \mathcal{H}$ is strictly positive and self adjoint with respect to $\langle\cdot,\cdot\rangle_{H^{-1},\mathbf{A}}$ defined in (A.19). Further, (2.1) and (2.2) hold and the embeddings are compact.*

Proof. (a) We first consider wellposedness and the spectral bounds. By the Lax-Milgram Lemma we obtain for some $\epsilon > 0$ that for any $f \in H^{-1}(\Omega)$ there is a unique solution $q \in H_0^1(\Omega)$ of the problem

$$\begin{aligned} &Lq - \lambda q = f \quad \text{ in } H^{-1}(\Omega) \\ \iff &\langle \mathbf{A}\nabla q, \nabla\phi\rangle_{L^2} - \lambda\langle q,\phi\rangle_{L^2} = \langle f\,,\phi\rangle_{H^{-1}\times H_0^1} \quad \text{for all } \phi \in H_0^1(\Omega) \end{aligned} \tag{A.16}$$

for all λ with $\operatorname{Re}\lambda < \epsilon$. Hence, the spectrum of L is part of the right half plane and for $\lambda = 0$, we can define $L^{-1}\colon H^{-1}(\Omega) \to H_0^1(\Omega)$ via

$$\langle \mathbf{A}\nabla(L^{-1}f), \nabla\phi\rangle_{L^2} = \langle f\,,\phi\rangle_{H^{-1}\times H_0^1}\,. \tag{A.17}$$

Let $q = L^{-1}f \in H_0^1(\Omega)$ be the solution of (A.16), then the boundedness of L^{-1} follows from

$$\big\|L^{-1}f\big\|_{H_0^1}^2 = \|\nabla q\|_{L^2}^2 \leq \frac{1}{\delta}\langle \mathbf{A}\nabla q, \nabla q\rangle_{L^2} = \frac{1}{\delta}\langle f\,,q\rangle_{H^{-1}\times H_0^1} \leq \frac{1}{\delta}\|\nabla q\|_{L^2}\|f\|_{H^{-1}}$$

such that $\big\|L^{-1}f\big\|_{H_0^1} \leq \delta^{-1}\|f\|_{H^{-1}}$.

(b) We now prove that L is self adjoint. To this end we introduce the scalar products

$$\langle f,g\rangle_{\mathbf{A}} \coloneqq \langle \mathbf{A}\nabla f, \nabla g\rangle_{L^2}\,, \tag{A.18}$$

$$\langle f,g\rangle_{H^{-1},\mathbf{A}} \coloneqq \langle L^{-1}f, L^{-1}g\rangle_{\mathbf{A}}\,. \tag{A.19}$$

We prove their equivalence to the standard inner product and then show that L is self adjoint with respect to $\langle\cdot,\cdot\rangle_{H^{-1},\mathbf{A}}$.

(1) Obviously, (A.18) is equivalent to the standard inner product by the properties of $\mathbf{A}$ with

$$\delta\,\|q\|_{H_0^1}^2 \leq \|q\|_{\mathbf{A}}^2 \leq \|\mathbf{A}\|_\infty\,\|q\|_{H_0^1}^2$$

and we have to check that the norm induced by (A.19) is equivalent to (A.13). It holds with the definition of L^{-1} in (A.17) with $\lambda = 0$

$$\begin{aligned} \|f\|_{H^{-1}} &= \sup_{\|u\|_{H_0^1}=1} \langle f\,,u\rangle_{H^{-1}\times H_0^1} = \sup_{\|u\|_{H_0^1}=1} \langle L^{-1}f, u\rangle_{\mathbf{A}} \leq \sup_{\|u\|_{H_0^1}=1} \big\|L^{-1}f\big\|_{\mathbf{A}}\,\|u\|_{\mathbf{A}} \\ &\leq \|\mathbf{A}\|_\infty^{1/2}\,\big\|L^{-1}f\big\|_{\mathbf{A}} \end{aligned}$$

and choosing $u_0 = \big\|L^{-1}f\big\|_{H_0^1}^{-1}\; L^{-1}f$ yields

$$\|u_0\|_{H_0^1} = 1, \qquad \langle f\,,u_0\rangle_{H^{-1}\times H_0^1} = \big\|L^{-1}f\big\|_{\mathbf{A}}^2\,\big\|L^{-1}f\big\|_{H_0^1}^{-1} \geq \delta^{1/2}\,\big\|L^{-1}f\big\|_{\mathbf{A}}\,,$$

and hence the equivalence is shown.

(2) By (A.19) and the definition of L^{-1} in (A.17) we obtain symmetry for $f, g \in H_0^1(\Omega)$ by

$$\langle Lf, g\rangle_{H^{-1},\mathbf{A}} = \langle f, L^{-1}g\rangle_{\mathbf{A}} = \langle L^{-1}g, f\rangle_{\mathbf{A}} = \langle g\,, f\rangle_{H^{-1}\times H_0^1} = \langle g, f\rangle_{L^2}\,,$$
$$\langle f, Lg\rangle_{H^{-1},\mathbf{A}} = \langle L^{-1}f, g\rangle_{\mathbf{A}} = \langle f\,, g\rangle_{H^{-1}\times H_0^1} = \langle f, g\rangle_{L^2}\,.$$

As in (A.16) we can also solve $(\pm iI + L)q = f$ for every $f \in H^{-1}(\Omega)$ and we may conclude self-adjointness.

(3) By [1, Theorem 6.3] we have the compact embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$. Now let $\|f_n\|_{L^2} \leq C$. Using the strictly positive square root of L, and the relations

$$\left\|L^{1/2}\phi\right\|_{H^{-1},\mathbf{A}} = \|\phi\|_{L^2} = \left\|L^{-1/2}\phi\right\|_{\mathbf{A}}\,,$$

we define $g_n \coloneqq L^{-1/2}f_n$ with $\|g_n\|_{\mathbf{A}} \leq C$. Hence, there is a converging subsequence $(g_{n_j})_j$ in $L^2(\Omega)$ and we obtain

$$\left\|f_{n_j} - f_{n_k}\right\|_{H^{-1},\mathbf{A}} = \left\|g_{n_j} - g_{n_k}\right\|_{L^2} \to 0$$

for $j, k \to \infty$ such that also $L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ is compact. □

We then turn to the assumptions from Section 4.2.1 and verify the necessary Fréchet-differentiability.

Lemma A.3. *In the case $\mathcal{H} = H^{-1}(\Omega)$ and the framework recalled above Assumption 4.1 is valid.*

Proof. We start with a more general calculation. Since we have the fundamental theorem of calculus for almost every $x \in \Omega$ we get for functions $q, p \in V$ and $t, t+s \in [0, t_{\mathrm{def}}]$

$$\begin{aligned} D_g(t,s,q,p)(x) &\coloneqq g(t+s,x,q+p) - g(t,x,q) - (\partial_t g(t,x,q)s + \partial_y g(t,x,q)p) \\ &= \int_0^1 \tfrac{d}{d\sigma}\left(g(t+\sigma s, x, q+\sigma p)\right) - (\partial_t g(t,x,q)s + \partial_y g(t,x,q)p)\; d\sigma \qquad \text{(A.20)} \\ &= \int_0^1 \left[\partial_t g(t+\sigma s, x, q+\sigma p) - \partial_t g(t,x,q)\right] s + \left[\partial_y g(t+\sigma s, x, q+\sigma p) - \partial_y g(t,x,q)\right] p\, d\sigma \end{aligned}$$

First note that the embedding (A.1) by [1, Result 3.13] also implies the embedding $L^1(\Omega) \hookrightarrow H^{-1}(\Omega)$. Hence, we will use the L^1-norm instead of the H^{-1}-norm. For $p, q \in L^2(\Omega)$, taking the L^1-norm of (A.20) and recalling $G(t,q)(x) \coloneqq g(t,x,q(x))$ we obtain by Hölder's inequality

$$\begin{aligned} \|D_g(t,s,q,p)\|_{L^1} \leq \int_0^1 &\|\partial_t G(t+\sigma s, q+\sigma p) - \partial_t G(t,q)\|_{L^1}\,|s| \\ &+ \|\partial_y G(t+\sigma s, q+\sigma p) - \partial_y G(t,q)\|_{L^2}\,\|p\|_{L^2}\; d\sigma \end{aligned}$$

and the growth bounds in (A.14) guarantee by the dominated convergence theorem

$$\frac{1}{|s| + \|p\|_{L^2}}\,\|D_g(t,s,q,p)\|_{L^1} \to 0, \quad s, p \to 0\,.$$

This yields the Fréchet derivative for $h \in \mathbb{R}$ and $p \in V$

$$J_G(t,q)\begin{pmatrix} h \\ p \end{pmatrix} = \partial_t G(t,q)h + \partial_y G(t,q)p \qquad \text{(A.21)}$$

and by the same computation as above we check that it is continuous in $\mathcal{L}\big([0,t_{\mathrm{def}}] \times L^2(\Omega), H^{-1}(\Omega)\big)$. □

In the next lemma we consider the differentiability of the right-hand side evaluated at a smooth function.

Lemma A.4. *In the case $\mathcal{H} = H^{-1}(\Omega)$ and the framework recalled above Assumption 4.2 is valid.*

Proof. (A1) For $J_G(t,q)$ defined in (A.21) we compute

$$\begin{aligned}\widehat{D}_G(t,s) &:= \tfrac{1}{s}\Big(\widehat{G}(t+s) - \widehat{G}(t)\Big) - J_G(t,q(t))\begin{pmatrix}1\\ q'(t)\end{pmatrix}\\ &= \int_0^1 \tfrac{d}{d\sigma}\tfrac{1}{s}\Big(\widehat{G}(t+\sigma s)\Big) - J_G(t,q(t))\begin{pmatrix}1\\ q'(t)\end{pmatrix} d\sigma\\ &= \int_0^1 \widehat{G}_t(t+\sigma s) - \widehat{G}_t(t) + \widehat{G}_y(t+\sigma s)q'(t+\sigma s) - \widehat{G}_y(t)q'(t)\, d\sigma\end{aligned} \tag{A.22}$$

For $q \in C([0,T], H^1(\Omega)) \cap C^1([0,T], L^2(\Omega))$ we get

$$\Big\|\widehat{D}_G(t,s)\Big\|_{L^2} \le \int_0^1 \Big\|\widehat{G}_t(t+\sigma s) - \widehat{G}_t(t)\Big\|_{L^2} + \Big\|\widehat{G}_y(t+\sigma s)q'(t+\sigma s) - \widehat{G}_y(t)q'(t)\Big\|_{L^2}\, d\sigma$$

which goes to zero for $s \to 0$ by $t \mapsto \widehat{G}_z(t) \in C([0,T], L^\infty(\Omega))$ for $z \in \{t, y\}$ due to (A.1). By the same argument we get the continuity of the derivative.

(A2) To shorten notation, we define for $h \in \mathbb{R}$ and $p \in V$

$$\widehat{D}^2_G(t) := \widehat{G}_{tt}(t) + \widehat{G}_{ty}(t)\big(hq'(t) + p\big) + \widehat{G}_{yy}(t)q'(t)\cdot p \tag{A.23}$$

and compute

$$\begin{aligned}&\tfrac{1}{s}\left(J_G(t+s, q(t+s))\begin{pmatrix}h\\ p\end{pmatrix} - J_G(t,q(t))\begin{pmatrix}h\\ p\end{pmatrix}\right) - \widehat{D}^2_G(t)\\ &= \int_0^1 \tfrac{d}{d\sigma}\tfrac{1}{s}\left(J_G(t+\sigma s, q(t+\sigma s))\begin{pmatrix}h\\ p\end{pmatrix} - \widehat{D}^2_G(t)\right) d\sigma\\ &= \int_0^1 \widehat{D}^2_G(t+\sigma s) - \widehat{D}^2_G(t)\, d\sigma\,.\end{aligned} \tag{A.24}$$

For $p \in L^2$, $h \in \mathbb{R}$ and $q \in C([0,T], H^1(\Omega)) \cap C^1([0,T], L^2(\Omega)$ we get $q'p \in C([0,T], L^1(\Omega))$ and hence

$$\begin{aligned}\Big\|\int_0^1 \widehat{D}^2_G(t+\sigma s) - \widehat{D}^2_G(t)\, d\sigma\Big\|_{L^1} \le \int_0^1 &\Big\|\widehat{G}_{tt}(t+\sigma s) - \widehat{G}_{tt}(t)\Big\|_{L^1}\\ &+ \Big\|\widehat{G}_{ty}(t+\sigma s)\,(hq'(t+\sigma s)+p) - \widehat{G}_{ty}(t)\,(hq'(t)+p)\Big\|_{L^1}\\ &+ \Big\|\widehat{G}_{yy}(t+\sigma s)q'(t+\sigma s)\;p - \widehat{G}_{yy}(t)q'(t)\;p\Big\|_{L^1}\, d\sigma\,.\end{aligned}$$

Since $t \mapsto \widehat{G}_z(t) \in C([0,T], L^\infty(\Omega))$ holds for $z \in \{tt, ty, yy\}$, the expression tends to 0 uniformly in $h, p \to 0$. □

In the final lemma we consider different bounds of the nonlinearity.

Lemma A.5. *In the case $\mathcal{H} = H^{-1}(\Omega)$ and the framework recalled above Assumption 4.4 is valid.*

Proof. (A3) For $q \in H^1(\Omega)$, $t \in [0, t_{\text{def}}]$ we get by (A.14) and (A.3) the estimate

$$\|G(t,q)\|_{L^2} \leq C\left(1 + \left\|q^2\right\|_{L^2}\right) \leq C\left(\|q\|_{H^1}\right).$$

(A4a) For $q, p \in L^2(\Omega)$, $t, s \in [0, t_{\text{def}}]$ we get by (A.14) and Hölder's inequality

$$\begin{aligned}\left\|J_G(t,q)\begin{pmatrix} s \\ p \end{pmatrix}\right\|_{L^1} &= \|\partial_t G(t,q)s + \partial_t G(t,q)p\|_{L^1} \\ &\leq C\left(1 + \left\|q^2\right\|_{L^1}\right)|s| + C\left(1 + \|q\|_{L^2}\right)\|p\|_{L^2} \\ &\leq C\left(\|q\|_{L^2}\right)\left(|s| + \|p\|_{L^2}\right).\end{aligned}$$

(A4b) For $q \in H^1(\Omega)$, $p \in L^2(\Omega)$, $t, s \in [0, t_{\text{def}}]$ we get by (A.14), (A.3), (A.7), and (A.1)

$$\begin{aligned}\left\|J_G(t,q)\begin{pmatrix} s \\ p \end{pmatrix}\right\|_{L^2} &= \|\partial_t G(t,q)s + \partial_y G(t,q)p\|_{L^2} \\ &\leq C\left(1 + \left\|q^2\right\|_{L^2}\right)|s| + C\left(\|q\|_{L^\infty}\right)\|p\|_{L^2} \\ &\leq C\left(\|q\|_{H^1}\right)\left(|s| + \|p\|_{L^2}\right).\end{aligned}$$

(A-CLS-1) For $p_i \in L^2(\Omega)$, $i = 1, 2$, $q \in H_0^1(\Omega)$ and $t \in [0, t_{\text{def}}]$ we obtain by (A.1) and the Hölder's inequality and (A.15)

$$\begin{aligned}\left\|\left(J_G(t,p_1) - J_G(t,p_2)\right)\begin{pmatrix} 0 \\ q \end{pmatrix}\right\|_{L^1} &= \left\|\left(\partial_y G(t,p_1) - \partial_y G(t,p_2)\right)q\right\|_{L^1} \\ &\leq C\left(\|q\|_{H^1}\right)\|\partial_y G(t,p_1) - \partial_y G(t,p_2)\|_{L^1} \\ &\leq C\left(\|q\|_{H^1}\right)\Big(\sup_{s\in[0,1]}\|\partial_{yy} G(t, sp_1 + (1-s)p_2)\|_{L^2}\Big)\|p_1 - p_2\|_{L^2} \\ &\leq C\left(\|q\|_{H^1}, \|p_1\|_{L^2}, \|p_2\|_{L^2}\right)\|p_1 - p_2\|_{L^2}.\end{aligned}$$

(A-CLS-2) For $p_i, q \in H_0^1(\Omega)$, $i = 1, 2$, and $t \in [0, t_{\text{def}}]$ we obtain by (A.1)

$$\begin{aligned}\left\|\left(J_G(t,p_1) - J_G(t,p_2)\right)\begin{pmatrix} 0 \\ q \end{pmatrix}\right\|_{L^2} &= \left\|\left(\partial_y G(t,p_1) - \partial_y G(t,p_2)\right)q\right\|_{L^2} \\ &\leq C\left(\|q\|_{L^\infty}\right)\|\partial_y G(t,p_1) - \partial_y G(t,p_2)\|_{L^2} \\ &\leq C\left(\|q\|_{H^1}\right)\Big(\sup_{s\in[0,1]}\|\partial_{yy} G(t, sp_1 + (1-s)p_2)\|_{L^2}\Big)\|p_1 - p_2\|_{H^1} \\ &\leq C\left(\|q\|_{H^1}, \|p_1\|_{H^1}, \|p_2\|_{H^1}\right)\|p_1 - p_2\|_{H^1}.\end{aligned}$$

□

A.3 $\mathcal{H} = L^2(\Omega)$

Next, we consider the second column of Table 4.1. For $d = 1, 2, 3$ and a convex Lipschitz domain $\Omega \subseteq \mathbb{R}^d$, the operator L is defined on $\mathcal{H} = L^2(\Omega)$ by

$$Lq = -\operatorname{div}\left(\mathbf{A}\nabla q\right)$$

for some symmetric, uniformly positive matrix $\mathbf{A} \in W^{1,\infty}(\Omega)^{d\times d}$ with lower bound $\delta > 0$ and the additional spaces are given by

$$V = H_0^1(\Omega), \qquad \mathcal{D}(L) = H^2(\Omega) \cap H_0^1(\Omega)\,.$$

with their standard norms. For the nonlinearity g we assume $\alpha = (2, 1, 3)$ and the growth bounds

$$\begin{aligned} |g(t,x,y)|, |\partial_t g(t,x,y)| &\le C_{\mathrm{g}}\big(1 + |y|^{\gamma}\big)\,, \\ |\partial_y g(t,x,y)| &\le C_{\mathrm{g}}\big(1 + |y|^{\gamma-1}\big)\,. \end{aligned} \tag{A.25}$$

and for the corrected Lie Splitting (3.14) we assume in addition

$$|\partial_{yy} g(t,x,y)| \le C_{\mathrm{g}}\big(1 + |y|^{\gamma-1}\big)\,. \tag{A.26}$$

For $d = 2$ we may choose $\gamma > 1$ arbitrarily large and for $d = 3$ we need $\gamma \le 3$. In the case $d = 1$, we note that (A.25) and (A.26) are not necessary.

We first consider the operator theoretic assumptions from Section 2.1.1.

Lemma A.6. *The operator $L\colon \mathcal{D}(L) \subseteq \mathcal{H} \to \mathcal{H}$ is positive and self adjoint. Further, (2.1) and (2.2) hold and the embeddings are compact.*

Proof. (a) We first consider wellposedness and the spectral bounds. We compute as in (A.16)

$$Lq - \lambda q = f \quad \text{in } L^2(\Omega) \iff \langle \mathbf{A}\nabla q, \nabla\phi\rangle_{L^2} - \lambda\langle q, \phi\rangle_{L^2} = \langle f\,, \phi\rangle_{L^2} \quad \forall \phi \in H_0^1(\Omega) \tag{A.27}$$

and obtain by the Lax-Milgram Lemma for some $\epsilon > 0$ that there is a unique solution $q \in H_0^1(\Omega)$ for all λ with $\operatorname{Re}\lambda < \epsilon$. By the convexity of Ω the result [32, Theorem 3.2.1.2] further yields $q \in H^2(\Omega)$. Hence, the spectrum of L is part of the right half plane.

(b) By [1, Theorem 6.3] we have the compact embeddings

$$H^2(\Omega) \hookrightarrow H^1(\Omega) \hookrightarrow L^2(\Omega)\,.$$

(c) We finally prove that L is self adjoint on the L^2-scalar product. Symmetry directly follows from (A.27) as well as the solvability of $(\pm iI + L)q = f$ in $L^2(\Omega)$ which gives the assertion. □

We now turn to the assumptions made in Section 4.2.1 and verify the necessary Fréchet-differentiability.

Lemma A.7. *In the case $\mathcal{H} = L^2(\Omega)$ and the framework recalled above Assumption 4.1 is valid.*

Proof. We only prove the case $d = 2$ and $d = 3$, as the case $d = 1$ is even easier. By the choice of γ, for $p, q \in H^1(\Omega) \hookrightarrow L^{2\gamma}(\Omega)$, taking the L^2-norm of (A.20) we obtain for $\rho = \frac{2\gamma}{\gamma-1}$ with Hölder's inequality

$$\begin{aligned} \|D_g(t,s,q,p)\|_{L^2} \le \int_0^1 &\|\partial_t G(t+\sigma s, q+\sigma p) - \partial_t G(t,q)\|_{L^2}\,|s| \\ &+ \|\partial_y G(t+\sigma s, q+\sigma p) - \partial_y G(t,q)\|_{L^\rho}\,\|p\|_{L^{2\gamma}}\,d\sigma\,. \end{aligned}$$

By the growth bounds in (A.25) we estimate with (A.1)

$$\begin{aligned} \|\partial_t G(t,q)\|_{L^2}^2 &\le C\big(1 + \|q\|_{L^{2\gamma}}^{2\gamma}\big) \le C\big(\|q\|_{H^1}\big)\,, \\ \|\partial_y G(t,q)\|_{L^\rho}^\rho &\le C\big(1 + \|q\|_{L^{2\gamma}}^{2\gamma}\big) \le C\big(\|q\|_{H^1}\big)\,, \end{aligned}$$

which leads to convergence of

$$\frac{1}{|s| + \|p\|_{H^1}} \|D_g(t,s,q,p)\|_{L^2} \to 0, \quad s,p \to 0\,,$$

by the standard arguments as above. □

In the next lemma we consider the differentiability of the right-hand side evaluated at a smooth function.

Lemma A.8. *In the case $\mathcal{H} = L^2(\Omega)$ and the framework recalled above Assumption 4.2 is valid.*

Proof. (A1) For $q \in C([0,T], H^2(\Omega)) \cap C^1([0,T], H^1(\Omega))$ we get by (A.1) and (A.3)

$$q \in C([0,T], L^\infty(\Omega)), \quad \nabla q \in C([0,T], H^1(\Omega)), \quad \nabla q\ q' \in C([0,T], L^2(\Omega)), \quad \nabla q' \in C([0,T], L^2(\Omega))$$

and hence for $\widehat{D}_G$ defined in (A.22)

$$\begin{aligned}
\left\|\widehat{D}_G(t,s)\right\|_{H^1} &\le \int_0^1 \left\|\widehat{G}_t(t+\sigma s) - \widehat{G}_t(t)\right\|_{H^1} + \left\|\widehat{G}_y(t+\sigma s)q'(t+\sigma s) - \widehat{G}_y(t)q'(t)\right\|_{H^1} d\sigma \\
&\le \int_0^1 \left\|\widehat{G}_{tx}(t+\sigma s) - \widehat{G}_{tx}(t)\right\|_{L^2} \\
&\quad + \left\|\widehat{G}_{ty}(t+\sigma s)\nabla q(t+\sigma s) - \widehat{G}_{ty}(t)\nabla q(t)\right\|_{L^2} \\
&\quad + \left\|\widehat{G}_{xy}(t+\sigma s)q'(t+\sigma s) - \widehat{G}_{xy}(t)q'(t)\right\|_{L^2} \\
&\quad + \left\|\widehat{G}_{yy}(t+\sigma s)\nabla q(t+\sigma s)\ q'(t+\sigma s) - \widehat{G}_{yy}(t)\nabla q(t)\ q'(t)\right\|_{L^2} \\
&\quad + \left\|\widehat{G}_y(t+\sigma s)\nabla q'(t+\sigma s) - \widehat{G}_y(t)\nabla q'(t)\right\|_{L^2} d\sigma
\end{aligned}$$

goes to zero for $s \to 0$ since $t \mapsto \widehat{G}_z(t) \in C([0,T], L^\infty(\Omega))$ holds for $z \in \{tx, ty, xy, yy, y\}$.

(A2) For $p \in H^1(\Omega)$, $h \in \mathbb{R}$ and $q \in C([0,T], H^2(\Omega)) \cap C^1([0,T], H^1(\Omega))$ we get $q'\ p \in C([0,T], L^2(\Omega))$ and hence for $\widehat{D}_G^2$ defined in (A.23)

$$\begin{aligned}
\Big\| \int_0^1 \widehat{D}_G^2(t+\sigma s) - \widehat{D}_G^2(t)\, d\sigma \Big\|_{L^2} &\le \int_0^1 \left\|\widehat{G}_{tt}(t+\sigma s)h - \widehat{G}_{tt}(t)h\right\|_{L^2} \\
&\quad + \left\|\widehat{G}_{ty}(t+\sigma s)\,(hq'(t+\sigma s)+p) - \widehat{G}_{ty}(t)\,(hq'(t)+p)\right\|_{L^2} \\
&\quad + \left\|\widehat{G}_{yy}(t+\sigma s)q'(t+\sigma s)\ p - \widehat{G}_{yy}(t)q'(t)\ p\right\|_{L^2} d\sigma
\end{aligned}$$

which goes to zero for $s \to 0$ since $t \mapsto \widehat{G}_z(t) \in C([0,T], L^\infty(\Omega))$ holds for $z \in \{tt, ty, yy\}$. □

In the final lemma of this section we consider different bounds of the nonlinearity.

Lemma A.9. *In the case $\mathcal{H} = L^2(\Omega)$ and the framework recalled above Assumption 4.4 is valid.*

Proof. (A3) For $q \in H^2(\Omega)$, $t \in [0, t_{\mathrm{def}}]$ we get by (A.9)

$$\|g(t,q)\|_{H^1} \le C\left(\|q\|_{H^1}\right).$$

(A4a) For $q, p \in H^1(\Omega) \hookrightarrow L^{2\gamma}$ and γ from (A.25) with $\rho = \frac{2\gamma}{\gamma-1}$ and $t, s \in [0, t_{\text{def}}]$ we get by Hölder's inequality and (A.1)

$$\begin{aligned}\|\partial_t G(t,q)s + \partial_y G(t,q)p\|_{L^2} &\leq C\left(1 + \||q|^\gamma\|_{L^2}\right)|s| + C\left(1 + \left\||q|^{\gamma-1}\right\|_{L^\rho}\right)\|p\|_{L^{2\gamma}} \\ &\leq C\left(\|q\|_{H^1}\right)\left(|s| + \|p\|_{H^1}\right).\end{aligned}$$

(A4b) For $q \in H^2(\Omega)$, $p \in H^1(\Omega)$ and $t, s \in [0, t_{\text{def}}]$ we get by (A.5) and (A.11)

$$\begin{aligned}\|\partial_t G(t,q)s + \partial_y G(t,q)p\|_{H^1} &\leq C\left(\|q\|_{H^2}\right)|s| + C\,\|\partial_y G(t,q)\|_{H^2}\,\|p\|_{H^1} \\ &\leq C\left(\|q\|_{H^2}\right)\left(|s| + \|p\|_{H^1}\right).\end{aligned}$$

(A-CLS-1) For $p_i \in H^1(\Omega) \hookrightarrow L^{2\gamma}(\Omega)$, $i = 1, 2$, and γ from (A.25) with $\rho = \frac{2\gamma}{\gamma-1}$, $q \in H^2(\Omega)$ and $t \in [0, t_{\text{def}}]$ we obtain by the Hölder's inequality and (A.26)

$$\begin{aligned}\left\|\left(\partial_y G(t,p_1) - \partial_y G(t,p_2)\right)q\right\|_{L^2} &\leq C\left(\|q\|_{H^2}\right)\|\partial_y G(t,p_1) - \partial_y G(t,p_2)\|_{L^2} \\ &\leq C\left(\|q\|_{H^2}\right)\Big(\sup_{s\in[0,1]}\|\partial_{yy} G(t, sp_1 + (1-s)p_2)\|_{L^\rho}\Big)\|p_1 - p_2\|_{L^{2\gamma}} \\ &\leq C\left(\|q\|_{H^2}, \|p_1\|_{H^1}, \|p_2\|_{H^1}\right)\|p_1 - p_2\|_{H^1}.\end{aligned}$$

(A-CLS-2) For $p_i, q \in H^2(\Omega)$, $i = 1, 2$, $t \in [0, t_{\text{def}}]$ we obtain by (A.5), (A.10) and (A.9)

$$\begin{aligned}\left\|\left(\partial_y G(t,p_1) - \partial_y G(t,p_2)\right)q\right\|_{H^1} &\leq C\left(\|q\|_{H^2}\right)\|\partial_y G(t,p_1) - \partial_y G(t,p_2)\|_{H^1} \\ &\leq C\left(\|q\|_{H^2}\right)\Big(\sup_{s\in[0,1]}\|\partial_{yy} G(t, sp_1 + (1-s)p_2)\|_{H^1}\Big)\|p_1 - p_2\|_{H^2} \\ &\leq C\left(\|q\|_{H^2}, \|p_1\|_{H^2}, \|p_2\|_{H^2}\right).\|p_1 - p_2\|_{H^2}.\end{aligned}$$

□

A.4 $\mathcal{H} = H_0^1(\Omega)$

For the last column of Table 4.1 let $d = 1, 2, 3$ and consider a convex domain $\Omega \subseteq \mathbb{R}^d$ with boundary of class C^3. The operator L is defined on $\mathcal{H} = H_0^1(\Omega)$, equipped with $\langle\cdot,\cdot\rangle_{\mathbf{A}}$ defined in (A.18), by

$$Lq = -\operatorname{div}\left(\mathbf{A}\nabla q\right)$$

for some $\mathbf{A} \in C^{1,1}(\Omega)^{d\times d} \cap W^{2,\infty}(\Omega)^{d\times d}$ or $\mathbf{A} \in H^4(\Omega)^{d\times d}$ and the additional spaces are given by

$$V = H^2(\Omega) \cap H_0^1(\Omega), \qquad \mathcal{D}(L) = \{p \in H^3(\Omega) \cap H_0^1(\Omega) \mid Lp \in H_0^1(\Omega)\}.$$

We note that for the nonlinearity g we assume $\alpha = (3, 2, 3)$, but no bounds of the form (A.14) or (A.25) are necessary. We first consider the operator theoretic assumptions from Section 2.1.1.

Lemma A.10. *The operator $L\colon \mathcal{D}(L) \subseteq \mathcal{H} \to \mathcal{H}$ is positive and self adjoint. Further, (2.1) and (2.2) hold and the embeddings are compact.*

Proof. (a) We first consider wellposedness and the spectral bounds. We want to solve the equation

$$Lq - \lambda q = f \in H_0^1(\Omega). \tag{A.28}$$

Let $\phi \in C_c^\infty(\Omega)$ and take the inner product to get

$$\begin{aligned} \langle Lq, \phi\rangle_{\mathbf{A}} - \lambda\langle q, \phi\rangle_{\mathbf{A}} = \langle f, \phi\rangle_{\mathbf{A}} &\iff \langle -\mathbf{A}\nabla \operatorname{div} \mathbf{A}\nabla q, \nabla\phi\rangle_{L^2} - \lambda\langle \mathbf{A}\nabla q, \nabla\phi\rangle_{L^2} = \langle f, \phi\rangle_{\mathbf{A}} \\ &\iff \langle \operatorname{div} \mathbf{A}\nabla q, \operatorname{div} \mathbf{A}\nabla\phi\rangle_{L^2} - \lambda\langle \mathbf{A}\nabla q, \nabla\phi\rangle_{L^2} = \langle f, \phi\rangle_{\mathbf{A}}\,. \end{aligned} \tag{A.29}$$

This equation remains valid for all $\phi \in H^2(\Omega)\cap H_0^1(\Omega)$. By [32, Theorem 3.1.3.1] there is some $\epsilon > 0$ such that the bilinear form

$$a\colon H^2(\Omega)\cap H_0^1(\Omega)\times H^2(\Omega)\cap H_0^1(\Omega) \to \mathbb{R}, \quad a(\phi,\psi) = \langle \operatorname{div} \mathbf{A}\nabla\phi, \operatorname{div} \mathbf{A}\nabla\psi\rangle_{L^2} - \lambda\langle \mathbf{A}\nabla\phi, \nabla\psi\rangle_{L^2}$$

is bounded and coercive for all $\operatorname{Re}\lambda > \epsilon$ and, hence, by Lax-Milgram we obtain the unique solution $q \in H^2(\Omega)\cap H_0^1(\Omega)$ of (A.28) in $L^2(\Omega)$. With [32, Theorem 2.5.1.1] the smoothness of the boundary and the coefficients $\mathbf{A}$ further imply $q \in H^3(\Omega)$, i.e., $q \in \mathcal{D}(L)$.

(b) By [1, Theorem 6.3] we have the compact embeddings

$$H^3(\Omega) \hookrightarrow H^2(\Omega) \hookrightarrow H^1(\Omega)\,.$$

(c) We finally prove that L is self adjoint in $\big(H_0^1(\Omega), \langle\cdot,\cdot\rangle_{\mathbf{A}}\big)$. The symmetry directly follows from (A.29) as well as the solvability of $(\pm iI + L)q = f$ in $H_0^1(\Omega)$ which gives the assertion. □

We now turn to the assumptions made in Section 4.2.1 and verify the necessary Fréchet-differentiability.

Lemma A.11. *In the case $\mathcal{H} = H_0^1(\Omega)$ and the framework recalled above Assumption 4.1 is valid.*

Proof. For $p, q \in H^2(\Omega)$ we obtain by (A.5) and (A.10)

$$\begin{aligned} \|D_g(t,s,q,p)\|_{H^1} \le \int_0^1 &\|\partial_t G(t+\sigma s, q+\sigma p) - \partial_t G(t,q)\|_{H^1}\,|s| \\ &+ \|\partial_y G(t+\sigma s, q+\sigma p) - \partial_y G(t,q)\|_{H^1}\,\|p\|_{H^2}\,d\sigma \\ &\le C\big(\|q\|_{H^2}, \|p\|_{H^2}\big)\big(|s| + \|p\|_{H^1}\big)\,|s| \\ &+ C\big(\|q\|_{H^2}, \|p\|_{H^2}\big)\big(|s| + \|p\|_{H^1}\big)\,\|p\|_{H^2}\,d\sigma\,, \end{aligned}$$

which leads to convergence of

$$\frac{1}{|s| + \|p\|_{H^2}}\,\|D_g(t,s,q,p)\|_{H^1} \to 0, \quad s,p \to 0\,.$$ □

In the next lemma we consider the differentiability of the right-hand side evaluated at a smooth function.

Lemma A.12. *In the case $\mathcal{H} = H_0^1(\Omega)$ and the framework recalled above Assumption 4.2 is valid.*

Proof. (A1) For $q \in C([0,T], H^3(\Omega))\cap C^1([0,T], H^2(\Omega))$ and $\widehat{D}_G$ defined in (A.22) we get

$$\Big\|\widehat{D}_G(t,s)\Big\|_{H^2} \le \int_0^1 \Big\|\widehat{G}_t(t+\sigma s) - \widehat{G}_t(t)\Big\|_{H^2} + \Big\|\widehat{G}_y(t+\sigma s)q'(t+\sigma s) - \widehat{G}_y(t)q'(t)\Big\|_{H^2}\,d\sigma\,,$$

which goes to zero for $s \to 0$ by $t \mapsto \widehat{G}_z(t) \in C([0,T], H^2(\Omega))$ for $z \in \{t, y\}$. By the same argument we get the continuity of the derivative.

(A2) For $p \in H^2(\Omega)$ and $q \in C([0,T], H^3(\Omega)) \cap C^1([0,T], H^2(\Omega))$ we get $q'\ p \in C([0,T], H^2(\Omega))$ and hence for $\widehat{D}_G^2$ defined in (A.23)

$$\begin{aligned}\Big\| \int_0^1 \widehat{D}_G^2(t+\sigma s) - \widehat{D}_G^2(t)\, d\sigma \Big\|_{H^1} \le \int_0^1 & \Big\| \widehat{G}_{tt}(t+\sigma s) - \widehat{G}_{tt}(t) \Big\|_{H^1} \\ & + \Big\| \widehat{G}_{ty}(t+\sigma s)\,(hq'(t+\sigma s)+p) - \widehat{G}_{ty}(t)\,(hq'(t)+p) \Big\|_{H^1} \\ & + \Big\| \widehat{G}_{yy}(t+\sigma s) q'(t+\sigma s)\ p - \widehat{G}_{yy}(t) q'(t)\ p \Big\|_{H^1}\ d\sigma\,,\end{aligned}$$

which goes to zero for $s \to 0$ by $t \mapsto \widehat{G}_z(t) \in C([0,T], H^2(\Omega))$ for $z \in \{tt, ty, yy\}$. □

In the final lemma we consider different bounds of the nonlinearity.

Lemma A.13. *In the case $\mathcal{H} = H_0^1(\Omega)$ and the framework recalled above Assumption 4.4 is valid.*

Proof. (A3) For $q \in H^3(\Omega)$, $t \in [0, t_{\text{def}}]$ we get by (A.11)

$$\|g(t,q)\|_{H^2} \le C\left(\|q\|_{H^2}\right).$$

(A4a) For $q, p \in H^2(\Omega)$, $t, s \in [0, t_{\text{def}}]$ we get by (A.5), (A.9), and (A.11)

$$\begin{aligned}\|\partial_t G(t,q)s + \partial_y G(t,q)p\|_{H^1} &\le C\left(\|q\|_{H^2}\right)|s| + C\,\|\partial_y G(t,q)\|_{H^1}\,\|p\|_{H^2} \\ &\le C\left(\|q\|_{H^2}\right)\left(|s| + \|p\|_{H^2}\right).\end{aligned}$$

(A4b) For $q \in H^3(\Omega)$, $p \in H^2(\Omega)$, $t, s \in [0, t_{\text{def}}]$ we get by (A.6) and (A.11)

$$\|\partial_t G(t,q)s + \partial_y G(t,q)p\|_{H^2} \le C\left(\|q\|_{H^2}\right)\left(|s| + \|p\|_{H^2}\right).$$

(A-CLS-1) For $p_i \in H^2(\Omega)$, $i = 1, 2$, $q \in H^3(\Omega)$ and $t \in [0, t_{\text{def}}]$ we obtain by (A.5), (A.9) and (A.10)

$$\begin{aligned}\big\|\big(\partial_y G(t,p_1) - \partial_y G(t,p_2)\big)q\big\|_{H^1} &\le C\left(\|q\|_{H^2}\right)\|\partial_y G(t,p_1) - \partial_y G(t,p_2)\|_{H^1} \\ &\le C\left(\|q\|_{H^2}\right)\Big(\sup_{s\in[0,1]}\|\partial_{yy} G(t, sp_1 + (1-s)p_2)\|_{H^1}\Big)\,\|p_1 - p_2\|_{H^2} \\ &\le C\left(\|q\|_{H^2}, \|p_1\|_{H^2}, \|p_2\|_{H^2}\right).\|p_1 - p_2\|_{H^2}\,.\end{aligned}$$

(A-CLS-2) For $p_i, q \in H^3(\Omega)$, $i = 1, 2$, $t \in [0, t_{\text{def}}]$ we obtain by (A.5), (A.9) and (A.12)

$$\begin{aligned}\big\|\big(\partial_y G(t,p_1) - \partial_y G(t,p_2)\big)q\big\|_{H^2} &\le C\left(\|q\|_{H^2}\right)\|\partial_y G(t,p_1) - \partial_y G(t,p_2)\|_{H^2} \\ &\le C\left(\|q\|_{H^2}, \|p_1\|_{H^2}, \|p_2\|_{H^2}\right).\|p_1 - p_2\|_{H^2}\,.\end{aligned}$$

□

Part II

Exponential integrators for quasilinear wave-type equations

CHAPTER 5

Analytical framework - quasilinear problems

In this chapter we introduce the analytical framework necessary to treat quasilinear evolution equations of the form

$$\Lambda(u(t))u'(t) = Au(t) + g(t, u(t)), \qquad u(0) = u_0. \tag{5.1}$$

We recall the results from [61], explain the examples which fit into the framework and extend the well-posedness result from the literature. We introduce the three nested Hilbert spaces

$$Z \hookrightarrow Y \hookrightarrow X$$

which are continuously and densely embedded. The space Y is an interpolation space between Z and X, see [55] for details on interpolation spaces. The linear operator A is skew adjoint on $\mathcal{D}(A)$ where $Y \hookrightarrow \mathcal{D}(A) \hookrightarrow X$ with

$$\|A\|_{X \leftarrow Y} \leq \alpha_{XY}, \qquad \|A\|_{Y \leftarrow Z} \leq \alpha_{YZ},$$

holds. We reformulate (5.1) as

$$u'(t) = \mathbf{A}(u(t))u(t) + f(t, u(t)) \tag{5.2}$$

where we use the notation

$$\mathbf{A}(u) = \mathbf{A}_u = \Lambda^{-1}(u)A, \qquad f(t, u) = \Lambda^{-1}(u)g(t, u). \tag{5.3}$$

The situation of semilinear problems is recovered for constant Λ such that this framework extends the one of Chapter 2. Before going into details of the framework, we discuss the examples which are covered by the presented error analysis.

5.1 Prototypical examples

The two classes of examples are the quasilinear wave equation on a bounded domain and the Maxwell's equations on a domain or the full space. We discuss these examples separately and prove the assumptions in Appendix B.

5.1.1 Wave equation

Consider the quasilinear wave equation from [18] on a bounded domain $\Omega \subseteq \mathbb{R}^d$, $d = 1, 2, 3$, with a C^3-boundary $\partial\Omega$ of the form

$$\begin{aligned} \partial_{tt} q(t) + \partial_{tt} K\big(q(t)\big) &= \Delta q(t) + r(t, q(t), q'(t)), \quad \text{in } \Omega, \quad t \geq 0\,, \\ q(t) &= 0, \qquad\qquad\qquad\qquad\quad \text{on } \partial\Omega, \quad t \geq 0 \end{aligned} \tag{5.4}$$

with

$$K \in C^5(\mathbb{R}), \qquad 1 + K'(0) > 0, \qquad r \in C^3(\mathbb{R} \times \overline{\Omega} \times \mathbb{R} \times \mathbb{R})\,, \tag{5.5}$$

and $r(t, \cdot, 0, 0) = 0$ on $\partial\Omega$ for $t \geq 0$. We note that in [18] the term f was not present, but is covered by our extension of the wellposedness result. This equation fits into the framework of (5.1) by rewriting it in first-order with $u = \big(\, q, q' \,\big)^T$ and the operators

$$\Lambda(u) = \begin{pmatrix} 1 & 0 \\ 0 & 1 + K'(q) \end{pmatrix}, \quad A = \begin{pmatrix} 0 & I \\ \Delta & 0 \end{pmatrix}, \quad g(t, u) = \begin{pmatrix} 0 \\ -K''(q)\big(q'\big)^2 + r(t, q, q') \end{pmatrix}. \tag{5.6}$$

The Hilbert spaces in this example are

$$\begin{aligned} X &\coloneqq H_0^1(\Omega) \times L^2(\Omega), \qquad Y \coloneqq \big(H^2(\Omega) \cap H_0^1(\Omega)\big) \times H_0^1(\Omega), \\ Z &\coloneqq \{q \in H^3(\Omega) \cap H_0^1(\Omega) : \Delta q \in H_0^1(\Omega)\} \times \big(H^2(\Omega) \cap H_0^1(\Omega)\big)\,. \end{aligned} \tag{5.7}$$

An important step throughout the wellposedness theory and error analysis is to ensure that the operator Λ is invertible. This was used above to rewrite (5.1) into the formulation (5.2). In addition, we note that the equation (5.1) degenerates in the case that Λ is not invertible and the whole theory is not applicable.

In the model above, a typical choice is the Kerr-type nonlinearity

$$K(z) = \chi z^3, \quad \chi \in \mathbb{R}\,, \tag{5.8}$$

see for example [12, 59, 64]. In this case, one needs to ensure that for the solution q it holds

$$1 + K'(q) = 1 + 3\chi q^2 > 0\,, \tag{5.9}$$

which is always satisfied for $\chi \geq 0$. Since we consider $d \leq 3$, there is a continuous embedding $H^2(\Omega) \hookrightarrow L^\infty(\Omega)$ with constant C_{emb}, and we may estimate

$$\|q\|_{L^\infty} \leq C_{\text{emb}} \|q\|_{H^2} \leq C_{\text{emb}} \|u\|_Y\,.$$

Hence, we can guarantee (5.9) also for $\chi < 0$ if we control $\|u\|_Y$ by some radius R satisfying

$$R^2 < \frac{1}{C_{\text{emb}}^2 3\,|\chi|}\,.$$

This radius R then ensures that equation (5.1) does not degenerate, and we will use it for the wellposedness and the error analysis in this part of the thesis. From now on we consider R as a given quantity of the problem that might have an a priori bound as in the case $\chi < 0$.

Further, we need another radius r with

$$\|u\|_Z \leq r\,,$$

to obtain uniform bounds in the later appearing constants. However, this parameter can be chosen arbitrarily.

5.1.2 Maxwell's equations

Another example are the quasilinear Maxwell's equations. They were for example considered in [61] where a detailed framework is provided and most of the assumption made in this section are verified. This framework was amended in [41]. The Maxwell's equations are given by the system of evolution equations

$$\begin{aligned} \partial_t D(t) &= \operatorname{curl} H(t) - \sigma(E(t))E(t), && \text{in } \Omega, \quad t \geq 0\,, \\ \partial_t B(t) &= -\operatorname{curl} E(t), && \text{in } \Omega, \quad t \geq 0\,, \\ \operatorname{div} D(t) &= 0, && \text{in } \Omega, \quad t \geq 0\,, \\ \operatorname{div} B(t) &= 0, && \text{in } \Omega, \quad t \geq 0\,, \end{aligned} \tag{5.10}$$

with the nonlinear material laws

$$D = E + P(E), \qquad B = H + M(H)\,.$$

This equation fits in the framework of (5.1) by rewriting it in first order with $u = \big(E, H\big)^T$ and the operators

$$\Lambda(u) = \begin{pmatrix} I + P'(E) & 0 \\ 0 & I + M'(H) \end{pmatrix}, \quad A = \begin{pmatrix} 0 & \operatorname{curl} \\ -\operatorname{curl} & 0 \end{pmatrix}, \quad g(t,u) = \begin{pmatrix} -\sigma(E)E \\ 0 \end{pmatrix}. \tag{5.11}$$

For the coefficients we assume

$$\sigma \in C^4(\mathbb{R}^3, \mathbb{R}^{3,3}), \qquad P, M \in C^4(\mathbb{R}^3, \mathbb{R}^{3,3})\,, \tag{5.12}$$

where $P'(x)$ and $M'(x)$ are symmetric for all $x \in \Omega$. Further, $I + P'(0)$ and $I + M'(0)$ are assumed to be positive definite. The most prominent example is again the Kerr-type nonlinearity

$$P(E) = \chi |E|^2 E, \quad \chi \in \mathbb{R}, \qquad M = 0\,,$$

see for example [2, 12, 64].

On the full space For $\Omega = \mathbb{R}^3$ we use the Hilbert spaces

$$X \coloneqq L^2(\mathbb{R}^3)^6, \qquad Y \coloneqq H^2(\mathbb{R}^3)^6, \qquad Z \coloneqq H^3(\mathbb{R}^3)^6\,. \tag{5.13}$$

By the embedding $H^2(\mathbb{R}^3)^6 \hookrightarrow L^\infty(\mathbb{R}^3)^6$ similar arguments as for the wave equation guarantee that there is some $R > 0$ such that $I + P'(x)$ and $I + M'(x)$ are positive definite in a ball $\mathcal{B}_Y(R)$. One could also replace the Hilbert spaces Y and Z by

$$Y = H^s(\mathbb{R}^3)^6, \qquad Z = H^{s+1}(\mathbb{R}^3)^6\,, \tag{5.14}$$

for $s > \frac{3}{2}$ since also $H^s(\mathbb{R}^3)^6 \hookrightarrow L^\infty(\mathbb{R}^3)^6$ holds. However, we only consider the choice (5.13) for the verification of the examples, but we expect that everything can be transferred to the situation (5.14).

On a bounded domain Let Ω be a domain with a boundary $\partial\Omega$ of class C^4. The framework then also covers homogeneous Dirichlet boundary conditions with the spaces

$$X \coloneqq L^2(\mathbb{R}^3)^6, \qquad Y \coloneqq H^2(\Omega)^6 \cap H_0^1(\Omega)^6, \qquad Z \coloneqq \{q \in H^4(\Omega)^3 \cap H_0^1(\Omega)^3 : \Delta q \in H_0^1(\Omega)^3\}^2\,. \tag{5.15}$$

5.2 Assumptions

Recall that the radius $R < \infty$ is given by the problem, and there might be an a priori bound as in (5.9). The radius $r < \infty$, however, can always be chosen arbitrarily large. We drop the dependency of the constants on R and r for the sake of readability, i.e., we always abbreviate $C = C(R, r)$, where C is any constant appearing in the following.

Assumption 5.1 (properties of Λ)**.** *The set $\{\Lambda(y)\colon y \in \mathcal{B}_Y(R)\}$ forms a family of invertible self-adjoint operators in $\mathcal{L}(X)$ such that the ranges $Ran(I \mp \Lambda^{-1}(y)A)$ are dense in X and the inverses $\Lambda^{-1}(y)$ also belong to $\mathcal{L}(Y)$. Moreover, for all $x \in X$ and $y, \widetilde{y} \in \mathcal{B}_Y(R)$, we have*

$$\|\Lambda(y)\|_{X \leftarrow X} \leq \lambda_X \tag{5.16a}$$

$$\langle x, \Lambda(y)x\rangle_X \geq \nu_X^{-1} \|x\|_X^2 \tag{5.16b}$$

$$\|\Lambda(y) - \Lambda(\widetilde{y})\|_{X \leftarrow X} \leq \ell \|y - \widetilde{y}\|_Y \tag{5.16c}$$

and there are constants ℓ_X, ℓ_Y, ℓ_Z such that for $\phi, \widetilde{\phi} \in \mathcal{B}$:

$$\left\|\Lambda^{-1}(\phi) - \Lambda^{-1}(\widetilde{\phi})\right\|_{V \leftarrow W} \leq \ell_V \left\|\phi - \widetilde{\phi}\right\|_V , \tag{5.16d}$$

with the triples

$$(V, W, \mathcal{B}) \in \Big\{\big(X, Y, \mathcal{B}_Y(R)\big), \big(Y, Y, \mathcal{B}_Y(R)\big), \big(Z, Z, \mathcal{B}_Z(r)\big)\Big\} .$$

As a direct consequence of the previous assumption we obtain with ν_X from (5.16d) and constants ν_Y, ν_Z that for $\phi \in \mathcal{B}$ it holds:

$$\left\|\Lambda^{-1}(\phi)\right\|_{V \leftarrow V} \leq \nu_V , \tag{5.17}$$

with the tuples

$$(V, \mathcal{B}) \in \Big\{\big(X, \mathcal{B}_Y(R)\big), \big(Y, \mathcal{B}_Y(R)\big), \big(Z, \mathcal{B}_Z(r)\big)\Big\} .$$

In the following we make frequent use of the state dependent inner product

$$\langle x, y\rangle_\phi = \langle \Lambda(\phi)x, y\rangle_X$$

which is defined for $\phi \in \mathcal{B}_Y(R)$ by (5.16a) and (5.16b). We state two important properties which can be found in the Appendix of [41].

Lemma 5.2 (relation of norms)**.** *Let Assumption 5.1 hold.*

(a) *For $\phi \in \mathcal{B}_Y(R)$*

$$\lambda_X^{-1} \|u\|_\phi^2 \leq \|u\|_X^2 \leq \nu_X \|u\|_\phi^2 . \tag{5.18}$$

(b) *For $\phi, \psi \in \mathcal{B}_Y(R)$ and $\tau > 0$*

$$\|u\|_\phi \leq e^{k_1 \tau} \|u\|_\psi , \qquad \text{for} \quad \|\phi - \psi\|_Y \leq \gamma\tau , \tag{5.19}$$

where $k_1 = k_1(\gamma) = \frac{1}{2}\nu_X \ell\, \gamma$.

With the bounds on Λ in Assumption 5.1, we establish several properties of the composed differential operator $\mathbf{A}_\phi$ in the following lemma.

Lemma 5.3 (properties of $\mathbf{A}_\phi$)**.** *Let Assumption 5.1 hold. Then for $\phi \in \mathcal{B}_Y(R)$*

$$\|\mathbf{A}_\phi\|_{X\leftarrow Y} \leq \nu_X \alpha_{XY} \tag{5.20a}$$

and for $\phi, \psi \in \mathcal{B}_Y(R) \cap \mathcal{B}_Z(r)$

$$\|\mathbf{A}_\phi\|_{Y\leftarrow Z} \leq \nu_Y \alpha_{YZ}\,, \tag{5.20b}$$

$$\|\mathbf{A}_\phi - \mathbf{A}_\psi\|_{X\leftarrow Z} \leq L_X \|\phi - \psi\|_X\,, \tag{5.20c}$$

$$\|\mathbf{A}_\phi - \mathbf{A}_\psi\|_{Y\leftarrow Z} \leq L_Y \|\phi - \psi\|_Y\,. \tag{5.20d}$$

Proof. Equation (5.20a) is easily verified by Assumption 5.1 and the other statements are proved in [41, Lemma 3.6]. □

In the papers of Kato, a key assumption is given by the following commutator condition. It is used in his proofs for the wellposedness, but is also employed in the error analysis of this thesis. It guarantees that the quasilinear operator can be lifted to the stronger space Z while only taking a small perturbation, in form of a bounded linear operator in the space X, into account.

Assumption 5.4 (commutator condition)**.** *We assume that there is an continuous isomorphism $S\colon Z \to X$ such that for $z \in \mathcal{B}_Y(R) \cap \mathcal{B}_Z(r)$ it holds*

$$\mathbf{A}_z^S = S\mathbf{A}_z S^{-1} = \mathbf{A}_z + B(z)$$

with

$$\|B(z)\|_{X\leftarrow X} \leq \beta\,.$$

In order to make the assumptions easily verifiable for the semilinear term, we pose the assumptions on the original term g in (5.1).

Assumption 5.5 (properties of g)**.** *For $V \in \{X, Y, Z\}$ there are constants $L_{g,V}$ such that for $\phi_1, \phi_2 \in \mathcal{B}_Z(r)$ and $t, s \in [0, T]$ it holds*

$$\|g(t, \phi_1) - g(s, \phi_2)\|_V \leq L_{g,V}\big(|t - s| + \|\phi_1 - \phi_2\|_V\big)\,. \tag{5.21}$$

From this we can deduce the properties of f which will be the ones used in the wellposedness theory and the error analysis.

Lemma 5.6 (properties of f)**.** *Let Assumptions 5.1 and 5.5 hold.*

(a) *The Lipschitz bound (5.21) also holds for f with constants $L_{f,V}$.*

(b) *For $V \in \{Y, Z\}$ there are constants $C_{f,V,\infty}$ such that for $\phi \in \mathcal{B}_Z(r)$ and $t \in [0, T]$*

$$\|f(t, \phi)\|_V \leq C_{f,V,\infty}\,.$$

Proof. The properties are simply deduced by combining Assumption 5.1 with the properties (5.16d), (5.17), and (5.21). □

Notation

We briefly collect some relevant constant used in the error analysis later and introduce a shorthand notation. In the following, $\gamma > 0$ denotes a given parameter, which will be determined later.

$$k_0 = (\nu_X \lambda_X)^{1/2} \geq 1 \qquad k_1 = k_1(\gamma) = \tfrac{1}{2}\,\nu_X \ell \gamma, \tag{5.22a}$$

$$c_0 = \|S\|_{X \leftarrow Z} \left\|S^{-1}\right\|_{Z \leftarrow X} k_0 \geq 1, \qquad c_1 = c_0 \nu_Y \alpha_{YZ} \tag{5.22b}$$

We further use for a Hilbert space V and a function $v \in C([0,T], V)$

$$\|v\|_{V,\infty} \coloneqq \max_{t \in [0,T]} \|v(t)\|_V \,.$$

5.3 Wellposedness

The aim of this section is to provide a wellposedness result for the equation (5.2). The standard approach is to use the Banach fixed-point theorem. This can be done by choosing a complete metric space E and considering for fixed $\phi \in E$ the linear, non-autonomous evolution equation

$$\begin{aligned} u'(t) &= \mathbf{A}(\phi(t))u(t) + f(t, \phi(t)) \\ &= \mathbf{A}_\phi(t)u(t) + f_\phi(t)\,, \end{aligned} \tag{5.23}$$

with initial value $u(0) = u_0$. For this equation, wellposedness and a priori bounds need to be established. In the next step the solution map

$$\mathcal{S} \colon \phi \mapsto u = u_\phi$$

is studied. Obviously, a fixed point of $\mathcal{S}$ is a solution of (5.2). Hence, the main task lies in the construction of a suitable space E which allows $\mathcal{S}$ to be a contractive self-map.

This has been successfully done in [61, Thm. 3.41] for the slightly simpler right-hand side

$$f(t, \phi) = \Lambda^{-1}(\phi) Q(\phi) \phi\,.$$

We mention that in [61] the special structure of f was used in order to define the operator $\mathbf{A}(\phi)$ differently by

$$\widetilde{\mathbf{A}}(\phi) = \Lambda^{-1}(\phi)\big(A + Q(\phi)\big)\,,$$

and to set f in (5.2) to zero. Our contribution is the generalization of this result, and we may apply the results of [61] by setting $Q = 0$, but therefore have to treat the inhomogeneous term f with additional technical effort.

For the fixed-point argument we use the same (complete) metric space as in [61]

$$E(T, r, \gamma) \coloneqq \{\phi \in C([0,T], Z) \mid \|\phi(t)\|_Y \leq R, \quad \|\phi(t)\|_Z \leq r, \quad [\phi]_{Lip([0,T],Y)} \leq \gamma\}\,, \tag{5.24}$$

for positive parameters T, R, r, γ chosen later, equipped with the metric

$$d\big(\phi, \psi\big) \coloneqq \max_{t \in [0,T]} \|\phi(t) - \psi(t)\|_Y \,.$$

As explained above, we fix a function $\phi \in E(T, r, \gamma)$ and study equation (5.23).

5.3.1 A priori bounds for the non-autonomous evolution equation

In the following we recall known results for non-autonomous evolution equations since they are needed for the analysis of (5.23). The first ingredient are so-called evolution families that generalize the concept of one-parameter semigroups discussed in Chapter 2. They arise in solving the problem

$$u'(t) = A(t)u(t) \tag{5.25}$$

where $A(t)$ depends on time. One can show under suitable assumptions on $A(t)$ that there is a family of operators U depending on two variables such that u given by

$$u(t) = U(t,s)u_0$$

is the solution of (5.25) with initial value $u(s) = u_0$. This family of operators is often called evolution family. For example if $A(t) = A$ is constant, the standard semigroup theory applies and U is simply given by

$$U(t,s) = e^{(t-s)A}\,.$$

We put this together in the following definition, see, e.g., [49].

Definition 5.7 (evolution family)**.** *Let $J = [a,b]$ be an interval and define $\Delta_J \coloneqq \{(t,s) \in J \times J \colon s \le t\}$. Further, consider a Hilbert space Y. The family of operators $U : \Delta_J \to \mathcal{L}(Y)$ is called an evolution family on Y if it satisfies the following properties for $a \le s \le r \le t \le b$.*

(a) *For any $y \in Y$, the map $(t,s) \mapsto U(t,s)y$ is continuous in Y with $U(t,t) = I$ and there are constants $M \ge 1$ and $\omega \in \mathbb{R}$ such that $\|U(t,s)\|_{Y \leftarrow Y} \le Me^{\omega(t-s)}$ holds.*

(b) *It holds $U(t,s) = U(t,r)U(r,s)$.*

Evolution families are a useful tool in the representation of the solution of non-autonomous evolution equations. This also applies for equation (5.23) where there is an additional inhomogeneity present.

Theorem 5.8. *Let $\phi \in E(T,r,\gamma)$. Then there exists an evolution family U_ϕ on Y with $J = [0,T]$ such that (5.23) with initial value $u(0) = u_0$ has a unique solution given by*

$$u_\phi(t) = U_\phi(t,0)u_0 + \int_0^t U_\phi(t,\sigma)f_\phi(\sigma)\,d\sigma\,. \tag{5.26}$$

In addition, the evolution family has the following properties:

(a) *For any $z \in Z$ the following derivatives exist in Y for $0 \le s \le t \le T$:*

$$\begin{aligned} \partial_t U_\phi(t,s)z &= \mathbf{A}_\phi(t)U_\phi(t,s)z\,,\\ \partial_s U_\phi(t,s)z &= -U_\phi(t,s)\mathbf{A}_\phi(s)z\,. \end{aligned} \tag{5.27}$$

(b) *The evolution family satisfies for $0 \le s \le t \le T$ the bounds*

$$\begin{aligned} \|U_\phi(t,s)\|_{Y \leftarrow Y} &\le c_0 e^{k_1 T} e^{k_0 \beta(t-s)}\,,\\ \|U_\phi(t,s)\|_{Z \leftarrow Z} &\le c_0 e^{k_1 T} e^{k_0 \beta(t-s)}\,. \end{aligned} \tag{5.28}$$

Proof. The representation (5.26) is given in [61, Thm.3.13]. The additional properties are verified in [61, Thm. 3.35] and in the proof of [61, Thm.3.41] where (5.28) is explicitly stated in [61, (3.10)]. □

In order to derive the a priori bounds, we decompose the solution by

$$u_\phi = \mathscr{L}_\phi + \mathscr{C}_\phi, \qquad \mathscr{L}_\phi(t) \coloneqq U_\phi(t,0)u_0, \qquad \mathscr{C}_\phi(t) \coloneqq \int_0^t U_\phi(t,\sigma) f_\phi(\sigma)\, d\sigma\,,$$

since the bounds on the linear part $\mathscr{L}_\phi$ and the convolution part $\mathscr{C}_\phi$ are derived separately. Before we estimate the two terms, we first guarantee the desired regularity of the solution u_ϕ. The following lemma is the firs step towards this.

Lemma 5.9. *Let $\phi \in E(T,r,\gamma)$. Then $f_\phi \in C([0,T],Z)$ holds and the maps*

$$\begin{aligned} \Delta_J \to Z, &\qquad (t,s) \mapsto U_\phi(t,s) f_\phi(s) \\ \Delta_J \to Y, &\qquad (t,s) \mapsto \mathbf{A}_\phi(t) U_\phi(t,s) f_\phi(s) \end{aligned}$$

are jointly continuous in both variables.

Proof. By Lemma 5.6 and $u \in C([0,T],Z)$ we immediately obtain $f_\phi \in C([0,T],Z)$. Since we have a constant isomorphism S, the second assertion can be deduced from [48, Thm. 6.1] where the continuity of

$$\Delta_J \to Z, \qquad (t,s) \mapsto U_\phi(t,s)z$$

for $z \in Z$ is shown. From this and the continuity of A_ϕ the last claim is easily derived. □

This immediately implies that a fixed-point of (5.26) is a classical solution of (5.2), and it suffices to find u_ϕ in the metric space E.

Theorem 5.10. *Let $u_0 \in Z$ and $\phi \in E(T,r,\gamma)$. Then the function u_ϕ defined in (5.26) satisfies*

$$u_\phi \in C([0,T],Z) \cap C^1([0,T],Y)\,.$$

Proof. For the linear part $\mathscr{L}_\phi$ we use the differentiability in (5.27) and obtain the same result for $\mathscr{C}_\phi$ if we combine Lemma 5.9 with the proof of [48, Theorem 7.1] replacing the space X by Y. □

We now turn to the a priori bounds. For the linear part $\mathscr{L}_\phi$ they were already derived in [61], and we only state the bounds. We remark that the constants have been adjusted to the notation in this thesis and introduce the constants

$$\omega_2 = \omega_2(\gamma) = k_1(\gamma) + k_0\beta, \qquad \gamma = \gamma(r) \coloneqq \tfrac{c_1}{c_0} r + 2c_0 C_{f,Y,\infty}\,. \tag{5.29}$$

Proposition 5.11. *For $\phi \in E(T,r,\gamma)$ the following bounds hold:*

$$\begin{aligned} \|\mathscr{L}_\phi(t)\|_Y &\le c_0 e^{\omega_2 t} \|u_0\|_Y\,, \\ \|\mathscr{L}_\phi(t)\|_Z &\le c_0 e^{\omega_2 t} \|u_0\|_Z\,, \\ [\mathscr{L}_\phi]_{Lip(Y,[0,T])} &\le c_1 e^{\omega_2 T} \|u_0\|_Z\,, \\ \|\mathscr{L}_\phi - \mathscr{L}_\psi\|_{Y,\infty} &\le T\big(c_0^2 L_Y e^{k_1 T}\big) e^{\omega_2 T} \|u_0\|_Z \|\phi - \psi\|_{Y,\infty}\,. \end{aligned}$$

Proof. The results can be found in the proof of [61, Thm. 3.41] in step 4. □

The a priori estimates for $\mathscr{C}_\phi$ are derived by similar arguments. We employ the bounds on the evolution family in Theorem 5.8 and on f given in Lemma 5.6.

Proposition 5.12. *For $\phi, \psi \in E(T, r, \gamma)$ the following bounds hold:*

$$\begin{aligned}
\|\mathscr{C}_\phi(t)\|_Y &\le c_0 e^{\omega_2 T} T C_{f,Y,\infty}\,,\\
\|\mathscr{C}_\phi(t)\|_Z &\le c_0 e^{\omega_2 T} T C_{f,Z,\infty}\,,\\
[\mathscr{C}_\phi]_{Lip(Y,[0,T])} &\le \big(c_0 C_{f,Y,\infty} + T c_1 C_{f,Z,\infty}\big) e^{\omega_2 T}\,,\\
\|\mathscr{C}_\phi - \mathscr{C}_\psi\|_{Y,\infty} &\le \big(c_0 T e^{\omega_2 T} L_{f,Y} + T^2 c_0^2 e^{k_1 T} e^{\omega_2 T} L_Y C_{f,Z,\infty}\big)\, \|\phi - \psi\|_{Y,\infty}\,.
\end{aligned}$$

Proof. (a) We first provide the bounds in Y and Z. By Lemma 5.6 and (5.28) we have

$$\begin{aligned}
\|\mathscr{C}_\phi(t)\|_Y &\le \int_0^t c_0 e^{k_1 T} e^{k_0 \beta (t-\sigma)} C_{f,Y,\infty}\, d\sigma \le c_0 T e^{\omega_2 T} C_{f,Y,\infty}\,,\\
\|\mathscr{C}_\phi(t)\|_Z &\le \int_0^t c_0 e^{k_1 T} e^{k_0 \beta (t-\sigma)} C_{f,Z,\infty}\, d\sigma \le c_0 T e^{\omega_2 T} C_{f,Z,\infty}\,.
\end{aligned}$$

(b) The Lipschitz-continuity in Y for $0 \le s \le t \le T$ is obtained by

$$\begin{aligned}
\|\mathscr{C}_\phi(t) - \mathscr{C}_\phi(s)\|_Y &\le \Big\| \int_s^t U_\phi(t,\sigma) f_\phi(\sigma)\, d\sigma \Big\|_Y + \Big\| \int_0^s \big(U_\phi(t,\sigma) - U_\phi(s,\sigma)\big) f_\phi(\sigma)\, d\sigma \Big\|_Y\\
&\le c_0 e^{\omega_2 T} C_{f,Y,\infty}\, |t-s| + \Big\| \int_0^s \int_s^t A_\phi(r) U_\phi(r,\sigma) f_\phi(\sigma)\, dr\, d\sigma \Big\|_Y\\
&\le \big(c_0 C_{f,Y,\infty} + T c_1 C_{f,Z,\infty}\big) e^{\omega_2 T}\, |t-s|\,,
\end{aligned}$$

where we used (5.27) in the second step. This implies

$$[\mathscr{C}_\phi]_{\mathrm{Lip}(Y,[0,T])} \le \big(c_0 C_{f,Y,\infty} + T c_1 C_{f,Z,\infty}\big) e^{\omega_2 T}\,.$$

(c) We finally estimate the Lipschitz constant for the contraction. To this end we compute for $\phi, \psi \in E(T, r, \gamma)$ and $0 \le t \le T$

$$\begin{aligned}
\|\mathscr{C}_\phi(t) - \mathscr{C}_\psi(t)\|_Y &\le \Big\| \int_0^t \big(U_\phi(t,\sigma) f_\phi(\sigma) - U_\psi(t,\sigma) f_\psi(\sigma)\big)\, d\sigma \Big\|_Y\\
&\le \Big\| \int_0^t U_\phi(t,\sigma) \big(f_\phi(\sigma) - f_\psi(\sigma)\big)\, d\sigma \Big\|_Y + \Big\| \int_0^t \big(U_\phi(t,\sigma) - U_\psi(t,\sigma)\big) f_\psi(\sigma)\, d\sigma \Big\|_Y\\
&= \mathscr{C}^1 + \mathscr{C}^2\,.
\end{aligned}$$

We estimate separately by Lemma 5.6 and (5.28)

$$\mathscr{C}^1 \le c_0 T e^{\omega_2 T} L_{f,Y}\, \|\phi - \psi\|_{Y,\infty}$$

and as in [61, p.75 bottom]

$$\begin{aligned}\mathscr{C}^2 &= \Big\| \int_0^t \big(U_\phi(t,\sigma) - U_\psi(t,\sigma)\big) f_\psi(\sigma)\, d\sigma \Big\|_Y \\ &\le \int_0^t \big\| \big(U_\phi(t,\sigma) - U_\psi(t,\sigma)\big) f_\psi(\sigma) \big\|_Y \, d\sigma \\ &\le \int_0^t c_0^2 e^{2k_1 T} e^{k_0 \beta t} \int_\sigma^t \|A_\phi(s) - A_\psi(s)\|_{Y \leftarrow Z} \, ds \, \|f_\psi(\sigma)\|_Z \, d\sigma \\ &\le T^2 c_0^2 e^{k_1 T} e^{\omega_2 T} L_Y C_{f,Z,\infty} \|\phi - \psi\|_{Y,\infty} \,,\end{aligned}$$

and conclude

$$\|\mathscr{C}_\phi - \mathscr{C}_\psi\|_{Y,\infty} \le \big(c_0 T e^{\omega_2 T} L_{f,Y} + T^2 c_0^2 e^{k_1 T} e^{\omega_2 T} L_Y C_{f,Z,\infty}\big) \|\phi - \psi\|_{Y,\infty} \,. \qquad \square$$

We finally arrive at the a priori bound for the solution u_ϕ by combining the results of Propositions 5.11 and 5.12.

Corollary 5.13. *For $\phi, \psi \in E(T, r, \gamma)$ it holds*

$$\begin{aligned}\|u_\phi(t)\|_Y &\le c_0 e^{\omega_2 T} \Big(\|u_0\|_Y + T C_{f,Y,\infty} \Big) \,, \\ \|u_\phi(t)\|_Z &\le c_0 e^{\omega_2 T} \Big(\|u_0\|_Z + T C_{f,Z,\infty} \Big) \,, \\ [u_\phi]_{Lip(Y,[0,T])} &\le c_1 e^{\omega_2 T} \big(\|u_0\|_Z + T C_{f,Z,\infty} \big) + c_0 e^{\omega_2 T} C_{f,Y,\infty} \,, \\ \|u_\phi - u_\psi\|_{Y,\infty} &\le \big(T c_0^2 L_Y e^{k_1 T} e^{\omega_2 T} \|u_0\|_Z + \big(c_0 T e^{\omega_2 T} L_{f,Y} + T^2 c_0^2 e^{k_1 T} e^{\omega_2 T} L_Y C_{f,Z,\infty} \big) \big) \|\phi - \psi\|_{Y,\infty} \,.\end{aligned}$$

5.3.2 Quasilinear evolution equation

With this preparation we are now in the position to close the proof of the fixed-point argument.

Theorem 5.14. *Let Assumptions 5.1, 5.4, and 5.5 be satisfied. For an initial value*

$$\|u_0\|_Y \le R_0 := \tfrac{1}{4c_0} R, \qquad \|u_0\|_Z \le r_0 := \tfrac{1}{4c_0} r,$$

define the time

$$T := \min\Big\{ \frac{\ln 2}{\omega_2}, \frac{R}{4c_0 C_{f,Y,\infty}}, \frac{r}{4c_0 C_{f,Z,\infty}}, \frac{1}{4c_0 \big(L_Y r + L_{f,Y}\big)} \Big\}, \tag{5.30}$$

where ω_2 and γ are given in (5.29). Then there is a unique solution u of (5.2) with

$$u \in C([0,T], Z) \cap C^1([0,T], Y) \,,$$

satisfying

$$\|u(t)\|_Y \le R, \qquad \|u(t)\|_Z \le r$$

on the interval $[0,T]$.

Proof. With the definition of T, R_0 and r_0 we obtain

$$e^{k_1 T} \leq e^{\omega_2 T} \leq 2, \qquad c_0 T C_{f,Y,\infty} \leq \tfrac{1}{4} R, \qquad c_0 T C_{f,Z,\infty} \leq \tfrac{1}{4} r,$$

and hence with Corollary 5.13 directly

$$\begin{aligned} \|u_\phi(t)\|_Y &\leq R\,, \\ \|u_\phi(t)\|_Z &\leq r\,, \\ [u_\phi]_{\mathrm{Lip}(Y,[0,T])} &\leq \gamma\,. \end{aligned}$$

It remains to prove the contraction bound

$$\|u_\phi - u_\psi\|_{Y,\infty} \leq \tfrac{1}{2} \|\phi - \psi\|_{Y,\infty}\,. \tag{5.31}$$

With this one can apply Banach fixed-point theorem and close the proof by the same arguments as in [61, Thm. 3.41], in particular using Theorem 5.10. We rewrite the last constant of Corollary 5.13

$$C_{\mathrm{Lip}} = c_0 T \Big(L_Y e^{k_1 T} \Big(c_0 e^{\omega_2 T} \big(\|u_0\|_Z + T C_{f,Z,\infty} \big) \Big) + e^{\omega_2 T} L_{f,Y} \Big)\,.$$

and as above we obtain with the definition of T

$$C_{\mathrm{Lip}} \leq 2 c_0 T \big(L_Y r + L_{f,Y} \big) \leq \tfrac{1}{2}$$

such that (5.31) follows. □

One can also obtain additional differentiability of the solution u in the weaker space X if we assume more differentiability of the data.

Theorem 5.15. *Let the assumptions of Theorem 5.14 be satisfied and let u be the solution of (5.2) with $u \in C([0,T], Z) \cap C^1([0,T], Y)$. Further, assume for $y \in Y$ the following differentiability*

$$\begin{aligned} t &\mapsto f(t, u(t)) \in C^1([0,T], X)\,, \\ t &\mapsto \mathbf{A}(u(t)) y \in C^1([0,T], X)\,. \end{aligned}$$

Then the solution u of (5.2) satisfies in addition

$$u \in C^2([0,T], X)\,.$$

Proof. The assumptions basically guarantee that we may differentiate u' in X using equation (5.2). □

CHAPTER 6

Review on time integration of quasilinear evolution equations

In this chapter we give an overview on the results obtained for time integration of quasilinear evolution equations. We mainly focus on wave-type equation except the Magnus-type integrators which were analyzed for parabolic problems. The approaches in Section 6.1 and 6.2 are the main motivation for our methods proposed in the next Chapter. In Section 6.3 we present an alternative approach for the time integration of quasilinear wave equations by trigonometric integrators and in Section 6.4 we briefly discuss a numerical comparison of exponential integrators for quasilinear Maxwell's equations.

6.1 Implicit Runge–Kutta methods for quasilinear hyperbolic systems

We start with implicit Runge–Kutta methods that were analyzed in the same framework as the analysis in this part of the thesis. We remark that in Hochbruck, Pažur and Schnaubelt [41, 44] the problem was of the form

$$u'(t) = \mathbf{A}(u(t))u(t), \qquad u(0) = u_0. \tag{6.1}$$

We start with explaining the first- and second-order methods that gave rise to the methods proposed later and afterwards we briefly show how higher-order was achieved.

Euler method

In [41] problem (6.1) was discretized in time by the Euler method. Applying the well-known implicit Euler rule (6.1) with the notation of (5.3) results in

$$u_{n+1} = u_n + \tau \mathbf{A}_{u_{n+1}} u_{n+1}, \quad n \geq 0\,, \tag{6.2}$$

where in each step a fully nonlinear problem has to be solved. On can linearize, this resulting in the so-called semi-implicit Euler method given by

$$u_{n+1} = u_n + \tau \mathbf{A}_{u_n} u_{n+1}, \quad n \geq 0\,. \tag{6.3}$$

Here, in each step only a linear system has to be solved which is computationally far more attractive. The idea of (6.3) is later employed for the exponential Euler method. Both methods are of order 1 as can be seen in the following theorems.

In the first one, error bounds in the X-norm are proven under regularity assumptions that follow from Theorem 5.15

Theorem 6.1 ([41, Thm. 4.3])**.** *Let u be the classical solution of Theorem 5.14 and assume in addition $u'' \in L^2([0,T],X)$. Further, let u_n be the numerical approximation obtained from (6.2) or (6.3). Under certain assumptions on the data there is $\tau_0 > 0$ such that for all $\tau \leq \tau_0$ it holds*

$$\|u(t_n) - u_n\|_X \leq C\tau \left(\int_0^T \|u''(t)\|_X^2 + \|u'(t)\|_X^2 \, dt \right)^{1/2}$$

with a constant $C > 0$ independent of τ and n.

Under additional assumptions on the data and the regularity of the solution, first-order error bounds are also shown in the stronger Z-norm.

Theorem 6.2 ([41, Thm. 4.5])**.** *Let u be the classical solution of Theorem 5.14 and assume in addition $Au \in L^\infty([0,T],Z)$ and $u', u'' \in L^2([0,T],X)$. Further, let u_n be the numerical approximation obtain from (6.2) or (6.3). Under certain assumptions on the data there is $\tau_0 > 0$ such that for all $\tau \leq \tau_0$ it holds*

$$\|u(t_n) - u_n\|_Z \leq C\tau \left(\int_0^T \|u''(t)\|_Z^2 + \|u'(t)\|_Z^2 \, dt \right)^{1/2}$$

with a constant $C > 0$ independent of τ and n.

Midpoint rule

Next we bring the attention to two second-order methods proposed by Kovács and Lubich [53]. In this work the authors considered equation (5.2) in a slightly different framework. For simplicity, we omit the additional nonlinearity f and consider only problem (6.1). They consider the implicit midpoint rule

$$u_{n+1} = u_n + \tfrac{\tau}{2} \mathbf{A}_{u_{n+1/2}} \big(u_n + u_{n+1} \big) \tag{6.4}$$

with two different choices of $u_{n+1/2}$ which are given by

$$u_{n+1/2} = \tfrac{1}{2} \big(u_{n+1} + u_n \big)\,, \qquad n \geq 0\,, \tag{6.4, FI}$$

$$u_{n+1/2} = u_n + \tfrac{1}{2} \big(u_n - u_{n-1} \big), \qquad n \geq 1, \qquad u_{1/2} = u_0\,. \tag{6.4, LI}$$

Similar to (6.2) the method (6.4, FI) is fully nonlinear whereas (6.4, LI) is only linearly implicit as this was the case in (6.3). The idea of (6.4, LI) is later employed for the exponential midpoint rule. For both schemes the following error bound was derived.

Theorem 6.3 ([53, Thm. 3.1]). *Let u be a sufficiently regular solution of (5.2) and let u_n be the numerical approximation obtain from (6.4, FI) or (6.4, LI). Under certain assumptions on the data there is $\tau_0 > 0$ such that for all $\tau \leq \tau_0$ it holds*

$$\|u(t_n) - u_n\|_Z \leq C\tau^2$$

with a constant $C > 0$ independent of τ and n.

Higher-order methods

Even though this thesis is not concerned with methods of order higher than two, we briefly sketch the results which have their basis in the theory of the preceding two sections. This might be a starting point for future research.

Despite slightly different analytical frameworks in [41, 53], the papers both considered implicit Runge–Kutta methods that are coercive [33, Def. IV.14.1] and algebraically stable [11], [33, Def. IV.12.5]. Since we will not further work with these concepts we only refer to the given literature and state the Gauss and Radau IIA methods as the main examples, [33, Thm. IV.12.9]. For nodes $c_i \in [0,1]$, coefficients a_{ij} and positive weights $b_i > 0$, they are given by

$$\begin{aligned} \dot{U}_{ni} &= \mathbf{A}_{U_{ni}} U_{ni}, && i = 1, \dots, s, \\ U_{ni} &= u_n + \tau \sum_{j=1}^{s} a_{ij} \dot{U}_{nj}, && i = 1, \dots, s, \\ u_{n+1} &= u_n + \tau \sum_{i=1}^{s} b_i \dot{U}_{ni}, \end{aligned} \tag{6.5}$$

where $u_n \approx u(t_n)$ approximates the exact solution u at time $t_n = n\tau$ and the internal stages satisfy $U_{ni} \approx u(t_n + c_i\tau)$.

In both papers we find results on the convergence of the schemes with stage order q. Without being precise about assumptions and frameworks for completeness we state the following result which combines error bounds in different norms.

Theorem 6.4 ([41, Thm. 5.3 & Thm. 6.3],[53, Thm. 4.1]). *Let u be a sufficiently regular solution of (5.2) and let u_n be the numerical approximation obtained from a method of type (6.5) with stage order q. For $V \in \{X, Y, Z\}$ under certain assumptions on the data there is $\tau_0 > 0$ such that for all $\tau \leq \tau_0$ it holds*

$$\|u(t_n) - u_n\|_V \leq C\tau^{q+1}$$

with a constant $C > 0$ independent of τ and n.

We remark that under stronger assumptions on the commutator compared to Assumption 5.4, Kovács and Lubich also proved an error bound of the classical order p, see [53, Thm. 4.2].

6.2 Magnus-type integrators for quasilinear parabolic problems

The motivation to use exponential integrators for (5.2) comes from the paper [26] by González and Thalhammer. With the ideas developed by González, Thalhammer and Ostermann [29] for non-autonomous

parabolic problems, the quasilinear equation

$$u'(t) = \mathbf{L}(u(t))u(t) + b(t), \qquad u(0) = u_0 \tag{6.6}$$

is considered in some Banach space X. For sufficiently regular u the operator $L(u) : \mathcal{D} \to X$ is of elliptic type. The main example is given by an elliptic operator with solution dependent coefficients in some L^p space over a domain Ω.

The method they considered in [26] is constructed in the following way. They freeze the argument of $\mathbf{L}$ in (6.6) at some midpoint $U_{n+1/2}$ and b at $t_{n+1/2}$, such that they arrive at a linear equation with a constant inhomogeneity. The outer stage then simply is the exact solution of this equation. To obtain $U_{n+1/2}$ they use an exponential Euler step with stepsize $\frac{\tau}{2}$ which yields the following method

$$\begin{aligned} U_{n+1/2} &= e^{\tau/2\mathbf{L}_{u_n}} u_n + \tfrac{\tau}{2}\varphi_1(\tfrac{\tau}{2}\mathbf{L}_{u_n})b(t_n)\,, \\ u_{n+1} &= e^{\tau\mathbf{L}_{U_{n+1/2}}} u_n + \tau\varphi_1(\tau\mathbf{L}_{U_{n+1/2}})b(t_{n+1/2})\,. \end{aligned} \tag{6.7}$$

On expects this method to be of order two, but this is not true in general as can be seen from the theory and numerical experiments in [26]. In fact, they prove for some interpolation space of X and $\mathcal{D}(L)$ that the method converges with order sightly less than 2 depending on certain parameters, in particular the exponent p of the L^p-space. We state their main theorem in this sloppy way.

Theorem 6.5 ([26, Thm. 5.1])**.** *Let u be a sufficiently regular solution of (6.6) and let u_n be the numerical approximation obtained from (6.7). Further, let X^β be some interpolation space between $\mathcal{D}(L)$ and X. Under certain assumptions on the data, there is $\tau_0 > 0$ and $\epsilon > 0$ such that all $\tau \le \tau_0$ it holds*

$$\|u(t_n) - u_n\|_{X^\beta} \le C\tau^{2-\epsilon}$$

with a constant $C > 0$ independent of τ and n.

In [27, 28], González and Thalhammer extended these results to higher order methods and proved error bounds for a larger class of methods, but we omit the details here.

6.3 Trigonometric integrators for quasilinear wave equations

Gauckler et al. [23] considered a quasilinear wave equation in one space dimension. It is given in the form

$$\partial_{tt}q(t) = \partial_{xx}q(t) - q(t) + \kappa a(q(t))\partial_{xx}q(t) + \kappa r(q(t), \partial_x q(t)) \quad \text{on } \mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})\,,$$

with smooth and real-valued functions a, r. Similar to (5.6), the equation is rewritten in first order as

$$u'(t) = Au(t) + g(t, u(t), \nabla u(t))$$

with the positive, self-adjoint operator $L = -\partial_{xx} + I \colon H^{s+2}(\mathbb{T}) \to H^s(\mathbb{T})$ and

$$A = \begin{pmatrix} 0 & I \\ -L & 0 \end{pmatrix}, \qquad g(t, u(t), \nabla u(t)) = \begin{pmatrix} 0 \\ \kappa a(q)\partial_{xx}q + \kappa r(q, \partial_x q) \end{pmatrix}.$$

The basis of the numerical method is the Strang splitting (3.12b) applied to the first-order formulation with some modified $\widetilde{f}$ similar to what was analyzed in the first part of this thesis. However, the properties

of the filters proposed in Definition 4.10 are in general not sufficient for quasilinear problems. In the error analysis, solutions with

$$u(t) = \big(q(t), q'(t)\big) \in H^5(\mathbb{T}) \times H^4(\mathbb{T}), \qquad t \in [0, T]\,,$$

are considered. One of the main results of the paper is an error bound for the semi-discretization in time [23, Thm 3.2] which allows for the estimate

$$\|u_n - u(t_n)\|_{H^2 \times H^1} \leq C\tau^2\,.$$

In addition, the authors provide for the full discretization with Fourier spectral methods in space [23, Thm 3.4] a similar error bound.

We emphasize that we are not able to properly compare the trigonometric integrator to our later proposed methods. This is due to the fact that we work in the framework of Chapter 5, which mainly treats the dimensions $d = 2$ and $d = 3$. Even though, the case $d = 1$ can also be handled, our results would be by far not optimal. To give an example, in $d = 1$ we have the embedding $H^1(\mathbb{T}) \hookrightarrow L^\infty(\mathbb{T})$ such that the condition (5.9) can already be guaranteed with the H^1- instead of the H^2-norm.

6.4 Numerical comparison of exponential integrators for quasilinear Maxwell's equations

We conclude this chapter with some comments on the reference by Pototschnig et al. [64]. In this paper, two exponential integration schemes are proposed for the time integration of quasilinear Maxwell's equations of a form closely related to (5.10) and are compared to classical integration schemes. The spatial discretization is given by a staggered Yee-grid [74] for all methods. However, no error analysis is provided, and we are not aware of it published elsewhere.

For the classical scheme, the authors choose the Leapfrog method where the nonlinear part is solved by a Newton solver and the classical Runge–Kutta method (RK4) of classical order 4. This first exponential integrator they consider is given by a Lawson method which is a fourth-order variant of (3.10). The underlying Runge–Kutta method is the RK4 from above. The equation is split in a linear and nonlinear part where the linear part is integrated exactly. Further, the fourth-order exponential Runge–Kutta method proposed in [42] is used. In comparison to the Lawson method, in each time step the exact Jacobian is used as linear part which is integrated exactly. The evaluation of the matrix exponential applied to a vector is approximated by Krylov subspace methods.

In the one-dimensional test case a performance comparison is carried out where the computational time is plotted against the relative error. The authors observe a very nice behavior of the exponential methods and can even outperform the classical methods. These numerical findings clearly indicate that exponential integrators for quasilinear Maxwell's equations can be very efficient, and it might also be interesting to compare our newly proposed methods with the ones from [64].

CHAPTER 7

Exponential integrators for quasilinear hyperbolic systems and main results

After the preparations in the previous chapters, we now propose and analyze the exponential integrators used to solve (5.2). In the first section we derive the new methods, state the main results in Theorem 7.1 and Theorem 7.7, and compare them to the results from Section 6.1. In the following two sections 7.2 and 7.3, we prove the main results and show some numerical experiments in Section 7.4. We conclude the chapter in Section 7.5 with some further results concerning error bounds in stronger norms. However, in contrast to the main results discussed in Section 7.1, we have to assume additional regularity of the solution which cannot be deduced from the wellposedness result in Chapter 5.

Recall the stepsize $\tau > 0$ and, given a numerical approximation $u_n \approx u(t_n)$ and the time $t_n = n\tau$, we define the operators

$$\mathbf{A}_n = \mathbf{A}(u_n), \qquad \mathbf{f}_n = f(t_n, u_n)\,. \tag{7.1}$$

Similarly, let $u(t)$ be the solution of Theorem 5.14 and define $\widehat{u}_{n+\sigma} = u(t_n + \tau\sigma)$. We introduce the notation

$$\begin{aligned} \widehat{\mathbf{f}}(t) &= f(t, u(t))\,, & \widehat{\mathbf{f}}_{n+\sigma} &= f(t_n + \tau\sigma, \widehat{u}_{n+\sigma})\,, \\ \widehat{\mathbf{A}}(t) &= \mathbf{A}(u(t))\,, & \widehat{\mathbf{A}}_{n+\sigma} &= \mathbf{A}(\widehat{u}_{n+\sigma})\,. \end{aligned} \tag{7.2}$$

Throughout the chapter the assumptions of Chapter 5 are valid. In particular, we do not state the precise dependence of the appearing constants on the bounds assumed in Chapter 5. We will only be precise about the regularity of the solution u and the dependence on τ, n and t_n.

7.1 Overview of methods and main results

In this section we propose the new exponential integrators for the time discretization of (5.2) and explain how they are connected to the methods explained in Chapter 6. The common feature goes back to the idea explained in Section 6.2 where we freeze the argument of the differential operator and the semilinear term in (5.2) and use the exact representation of the solution of the resulting linear equation. This results in the following two methods.

Exponential Euler method

If we freeze at the last approximation u_n, we obtain the exponential Euler scheme

$$\begin{aligned} u_{n+1} &= e^{\tau \mathbf{A}_n} u_n + \tau \varphi_1(\tau \mathbf{A}_n) \mathbf{f}_n \\ &= u_n + \tau \varphi_1(\tau \mathbf{A}_n) \big(\mathbf{A}_n u_n + \mathbf{f}_n \big) , \end{aligned} \tag{7.3}$$

where we used the notation introduced in (7.1). We note that this can also be seen as a variant of (6.3) where the resolvent $\big(I - \tau A_{u_n}\big)^{-1}$ is replaced by the exponential. In fact, we see in the latter stability analysis a very similar behavior of the two methods. For the method (7.3), the first main result in the second part of this thesis are the following error bounds.

Theorem 7.1. *Let u be the solution of (5.1) obtained by Theorem 5.14 and u_n the approximation obtained from (7.3). If Assumptions 5.1, 5.4, and 5.5 are satisfied, we obtain for $V \in \{X, Y\}$ the error bounds*

$$\|u(t_n) - u_n\|_V \leq t_n e^{c_V \, t_n} C_V \, \tau, \quad 0 \leq n\tau = t_n \leq T,$$

with constants $C_V, c_V > 0$ that only depend on $\|u'\|_{V,\infty}$ and $\|u\|_{Z,\infty}$, but are independent of τ, n and t_n.

We note that compared to Theorem 6.1 for the bound in the X-norm, we also need $\|u\|_{Z,\infty}$, but in our theorem only $\|u'\|_{X,\infty}$ enters compared to the L^2-norm

$$\int_0^T \|u''(t)\|_X^2 \, dt \, ,$$

which shows a slight advantage of the exponential integrator in terms of regularity assumptions.

Exponential midpoint rule

We also study a second-order method, inspired by the exponential ansatz in (6.7). We could directly study this method in our framework and we expect that second-order would also be achieved. However, we do not want to compute another exponential Euler step as an inner stage.

So we combine this scheme with the ideas of [53] where the midpoint is computed by (6.4, FI) or (6.4, LI). Classically, one would like to use the average of u_n and u_{n+1} as in (6.4, FI), but this would make the method implicit in the unbounded operator and thus computationally very expensive. Hence, as in (6.4, LI) we replace the average by the extrapolation using the last two approximations and arrive at the following scheme

$$\begin{aligned} u_{1/2} &= u_0, \\ u_{n+1/2} &= \tfrac{1}{2}\big(3u_n - u_{n-1}\big), \quad n \geq 1, \\ u_{n+1} &= e^{\tau \mathbf{A}_{n+1/2}} u_n + \tau \varphi_1(\tau \mathbf{A}_{n+1/2}) \mathbf{f}_{n+1/2} \, , \end{aligned} \tag{7.4}$$

which we call the exponential midpoint rule.

In order to derive error bounds of second-order for the scheme (7.4), the Lipschitz bounds from the previous chapter are not sufficient. Indeed, we have to apply Taylor expansion not only to the exact solution u, but also to the terms on the right-hand side of (5.2). Otherwise, we can only achieve bounds under the same regularity as in Theorem 6.3 and there is no gain in an exponential method. The necessary differentiability is formulated as assumptions to ensure readability of the chapter. We provide the detailed computations to verify the assumptions in Appendix B. We begin with the differentiability of the semilinear term g.

Assumption 7.2 (additional properties of g)**.** *Let $u \in C^1([0,T],Y) \cap C([0,T],Z)$ and consider the map*

$$t \mapsto \widehat{\mathbf{g}}(t) = g(t,u(t))\,. \tag{7.5}$$

Then there is a constant $C_{g',Y,\infty}$ with

(a) $t \mapsto \widehat{\mathbf{g}}(t) \in C^1([0,T],Y)$, $\quad \|\widehat{\mathbf{g}}'(t)\|_Y \le C_{g',Y,\infty}$,

and, if in addition, $u \in C^2([0,T],X)$ holds, then there is $C_{g'',X,\infty}$ such that

(b) $t \mapsto \widehat{\mathbf{g}}(t) \in C^2([0,T],X)$, $\quad \|\widehat{\mathbf{g}}''(t)\|_X \le C_{g'',X,\infty}$,

with constants only depending on $\|u''\|_{X,\infty}$, $\|u'\|_{Y,\infty}$, $\|u\|_{Z,\infty}$.

Whereas similar conditions to those in Assumption 7.2 are known from the analysis of semilinear evolution equations, we need an additional assumption in order to treat the differential operator and the composition of Λ^{-1} and g.

Assumption 7.3 (additional properties of Λ)**.** *Let $u \in C^1([0,T],Y) \cap C([0,T],Z)$ and consider the map*

$$t \mapsto \mathbf{\Lambda^{-1}}(t) \coloneqq \Lambda^{-1}(u(t))\,.$$

For $V \in \{X,Y\}$ and $v \in V$ it holds

(a) $t \mapsto \mathbf{\Lambda^{-1}}(t)v \in C^1([0,T],V)$, $\quad \left\|\left(\mathbf{\Lambda^{-1}}\right)'(t)\right\|_{V \leftarrow V} \le C_{VV}$,

and, if in addition, $u \in C^2([0,T],X)$, it further holds for $y \in Y$

(b) $t \mapsto \mathbf{\Lambda^{-1}}(t)y \in C^2([0,T],X)$, $\quad \left\|\left(\mathbf{\Lambda^{-1}}\right)''(t)\right\|_{X \leftarrow Y} \le C_{XY}$,

with constants C_{XX}, C_{XY}, C_{YY} only depending on $\|u''\|_{X,\infty}$, $\|u'\|_{Y,\infty}$, $\|u\|_{Z,\infty}$.

With the two preceding assumptions, we can conclude differentiability of the right-hand side in (5.2). We first consider the semilinear term f.

Lemma 7.4. *Let $u \in C^2([0,T],X) \cap C^1([0,T],Y) \cap C([0,T],Z)$ and consider the map*

$$t \mapsto \widehat{\mathbf{f}}(t) = f(t,u(t))\,.$$

If Assumptions 7.2 and 7.3 hold, then $\widehat{\mathbf{f}}$ satisfies Assumption 7.2 with constants $C_{f',Y,\infty}, C_{f'',X,\infty}$ only depending on $\|u''\|_{X,\infty}$, $\|u'\|_{Y,\infty}$, $\|u\|_{Z,\infty}$.

Proof. The assertion directly follows from the product rule. Note however, that part (a) holds already true for $u \in C^1([0,T],Y) \cap C([0,T],Z)$, since we only employ part (a) of Assumptions 7.2 and 7.3. □

By the structure of $\mathbf{A}(u)$, we directly conclude the following lemma which gives differentiability of the differential operator evaluated at a smooth function.

Lemma 7.5. *Let $u \in C^1([0,T],Y) \cap C([0,T],Z)$ and consider the map*

$$t \mapsto \widehat{\mathbf{A}}(t) = \mathbf{\Lambda^{-1}}(t)A\,.$$

If Assumption 5.1 and 7.3 are satisfied, then for $y \in Y$ and $z \in Z$ it holds

(a) $t \mapsto \widehat{\mathbf{A}}(t)y$ *is* $C^1([0,T],X)$, $\quad \|\widehat{\mathbf{A}}'(t)\|_{X \leftarrow Y} \le C^A_{XY}$,

(b) $t \mapsto \widehat{\mathbf{A}}(t)z$ *is* $C^1([0,T],Y)$, $\quad \|\widehat{\mathbf{A}}'(t)\|_{Y \leftarrow Z} \le C^A_{YZ}$,

and, if in addition, $u \in C^2([0,T],X)$, it further holds

(c) $t \mapsto \widehat{\mathbf{A}}(t)z$ *is* $C^2([0,T],X)$, $\qquad \big\|\widehat{\mathbf{A}}''(t)\big\|_{X\leftarrow Z} \le C^A_{XZ}$,
with constants $C^A_{XY}, C^A_{YZ}, C^A_{XZ}$ *only depending on* $\|u''\|_{X,\infty}$, $\|u'\|_{Y,\infty}$, $\|u\|_{Z,\infty}$.

We remark that Lemmas 7.4 and 7.5 are a key ingredient in the error analysis. Additionally, they allow us to derive further regularity of the solution u.

Corollary 7.6. *Let* $u \in C^1([0,T],Y)\cap C([0,T],Z)$ *be the solution obtained in Theorem 5.14. If Assumptions 7.2 (a) and 7.3 (a) hold, then*

$$u \in C^2([0,T],X)\cap C^1([0,T],Y)\cap C([0,T],Z)\,.$$

Proof. This is a direct consequence of Theorem 5.15 using Lemmas 7.4 and 7.5 (a). □

We are almost in the position to state the error bound for the exponential midpoint rule. However, we need to take care of the bounds on the extrapolated approximations $u_{n+1/2}$. Since this is not a convex combination of previous approximation, the bounds of u_n and u_{n-1} do not hold. To overcome this, we choose some radius $\widehat{R} > R$ such that Assumption 5.1 on $\Lambda(y)$ is still valid. If we can guarantee that the numerical approximations stay in the slightly larger ball $\mathcal{B}_Y(\widehat{R})$, the scheme remains stable. This enters later as a mild stepsize restriction $\tau \le \tau_0$ with

$$\tfrac{\widehat{\gamma}\tau_0}{2} \le \widehat{R} - R\,, \tag{7.6}$$

where $\widehat{\gamma}$ is chosen below in (7.8). Due to similar arguments we also have to replace the radius r by $\widehat{r} = 2r$.

All assumptions in Chapter 5 have been posed for the radii R and r. For the analysis of the exponential midpoint rule, we have to assume that they also hold for the new radii $\widehat{R}$ and $\widehat{r}$. We denote the constants by the same name but with an additional hat, e.g., we replace

$$C_{f,X,\infty} = C_{f,X,\infty}(R,r) \quad \text{by} \quad \widehat{C}_{f,X,\infty} = \widehat{C}_{f,X,\infty}(\widehat{R},\widehat{r})\,.$$

Without loss of generality we may also assume a monotone growth of the constants in the radii such that, e.g., $C_{f,X,\infty} \le \widehat{C}_{f,X,\infty}$ holds. Due to the possibly larger constants we can only simulate up to the time

$$\widehat{T}_{\text{mid}} := \min\Big\{\frac{\ln 2}{\widehat{\omega}_2}, \frac{R}{4\widehat{c}_0\widehat{C}_{f,Y,\infty}}, \frac{r}{4\widehat{c}_0\widehat{C}_{f,Z,\infty}}\Big\}, \tag{7.7}$$

where

$$\widehat{\omega}_2 = 2\widehat{k}_1(\widehat{\gamma}) + \widehat{k}_0\widehat{\beta}, \qquad \widehat{\gamma} := \frac{\widehat{c}_1}{\widehat{c}_0}r + 2\widehat{c}_0\widehat{C}_{f,Y,\infty}\,. \tag{7.8}$$

If we compare (7.7) to the end time T given in (5.30), then in general the three terms appearing here are smaller than the corresponding ones in (5.30) and thus also their minimum is smaller. However, we do not know in general how the fourth term in (5.30) relates to these quantities and hence, in general we can not decide which time is larger. Hence, we prove the following error bound in the X- and Y-norm on the intersection of both time intervals.

Theorem 7.7. *Let* u *be the solution of* (5.1) *obtained by Corollary 7.6 and* u_n *the approximation obtained from* (7.4). *If Assumptions 5.1, 5.4, and 5.5, are satisfied, and in addition Assumptions 7.2 and 7.3 hold true, and* τ_0 *is given by* (7.6), *then for all* $\tau \le \tau_0$ *the error is bounded by*

$$\|u(t_n) - u_n\|_X + \tau\,\|u(t_n) - u_n\|_Y \le t_n e^{c\,t_n} C\tau^2, \quad 0 \le n\tau = t_n \le \min\{T, \widehat{T}_{mid}\}\,,$$

with constants $C, c > 0$ *that only depend on* $\|u''\|_{X,\infty}$, $\|u'\|_{Y,\infty}$, $\|u\|_{Z,\infty}$, *but are independent of* τ, n *and* t_n.

For the second-order error bound in the X-norm in Theorem 6.4, see [41, Thm. 5.3] for the precise statement, a bound on

$$\int_0^T \left\|Au''(t)\right\|_X^2 + \left\|u^{(3)}(t)\right\|_X^2 \, dt$$

is required. This is, roughly speaking, one scale of regularity more than used in the above Theorem 7.7.

7.2 Error analysis of the exponential Euler method

This section is devoted to the proof of Theorem 7.1 and it is divided into three steps. We first establish stability of the numerical approximations in the stronger Y- and Z-norms in order to use the numerical flow for the error propagation. The analysis closely follows [41]. In the next step, we derive an error recursion for the global error and prove bounds on the defect. Lastly, we solve the error recursion and conclude a bound on the global error.

7.2.1 Stability

The first observation is a variant of [41, Lemma 3.7]. In this lemma we use a space that contains all numerical approximations. For $N \in \mathbb{N}$ and $\xi > 0$ we define the space

$$\begin{aligned} E\big(N, R, r, \xi\big) \coloneqq \{\phi = \big(\phi_0, \dots, \phi_N\big) \in Z^{N+1} \mid \\ \|\phi_k\|_Y \leq R, \|\phi_k\|_Z \leq r, \quad k = 0, \dots, N, \\ \|\phi_k - \phi_{k-1}\|_Y \leq \xi, \quad k = 1, \dots, N\}, \end{aligned} \tag{7.9}$$

which can be seen as a discrete analogue of the space (5.24). It is constructed in such a way that starting with approximations of this space for some $N \geq 1$, and inserting them in the numerical scheme, yields that the following approximation, say ϕ_{N+1}, together with the preceding approximations then lies in $E\big(N+1, R, r, \xi\big)$. The proof is done by induction in Lemma 7.11 and needs the following auxiliary results.

Lemma 7.8. *Let Assumptions 5.1 and 5.4 hold. Further, let $\phi = \big(\phi_0, \dots, \phi_N\big) \in E\big(N, R, r, \tau\gamma\big)$ and $0 \leq j \leq k \leq N$ for $j, k \in \mathbb{N}$. Then:*

$$\begin{aligned} \left\| e^{\tau \mathbf{A}_{\phi_k}} e^{\tau \mathbf{A}_{\phi_{k-1}}} \dots e^{\tau \mathbf{A}_{\phi_j}} \right\|_{X \leftarrow X} &\leq k_0 e^{\omega_1 (k-j+1)\tau}, \\ \left\| e^{\tau \mathbf{A}_{\phi_k}} e^{\tau \mathbf{A}_{\phi_{k-1}}} \dots e^{\tau \mathbf{A}_{\phi_j}} \right\|_{Y \leftarrow Y} &\leq c_0 e^{\omega_2 (k-j+1)\tau}, \\ \left\| e^{\tau \mathbf{A}_{\phi_k}} e^{\tau \mathbf{A}_{\phi_{k-1}}} \dots e^{\tau \mathbf{A}_{\phi_j}} \right\|_{Z \leftarrow Z} &\leq c_0 e^{\omega_2 (k-j+1)\tau}, \end{aligned}$$

with $\omega_1 = \omega_1(\gamma) = k_1(\gamma)$ and ω_2 given in (5.29).

Proof. The proof can be found the in Appendix of [41]. However, since we need an extension of this result, we give the proof in detail here. In a first step we prove the bound in the X-norm and then adapt it to the Z-norm. The last bound is then obtained by interpolation and $k_0 \leq c_0$ due to (5.22).

(a) Let $x \in X$ be arbitrary. By Assumption 5.1 we have that $\mathbf{A}_\phi$ generates a C_0-group with $\left\|e^{t\mathbf{A}_\phi}x\right\|_\phi = \|x\|_\phi$. Using this with (5.18) and (5.19), we compute

$$\begin{aligned}
\left\|e^{\tau\mathbf{A}_{\phi_k}}e^{\tau\mathbf{A}_{\phi_{k-1}}}\dots e^{\tau\mathbf{A}_{\phi_j}}x\right\|_X &\le \nu_X^{1/2}\left\|e^{\tau\mathbf{A}_{\phi_k}}e^{\tau\mathbf{A}_{\phi_{k-1}}}\dots e^{\tau\mathbf{A}_{\phi_j}}x\right\|_{\phi_k} \\
&= \nu_X^{1/2}\left\|e^{\tau\mathbf{A}_{\phi_{k-1}}}\dots e^{\tau\mathbf{A}_{\phi_j}}x\right\|_{\phi_k} \\
&\le \nu_X^{1/2}e^{k_1\tau}\left\|e^{\tau\mathbf{A}_{\phi_{k-1}}}\dots e^{\tau\mathbf{A}_{\phi_j}}x\right\|_{\phi_{k-1}} \\
&\le \dots \\
&\le \nu_X^{1/2}e^{k_1(k-j)\tau}\|x\|_{\phi_j} \\
&\le k_0 e^{k_1(k-j)\tau}\|x\|_X ,
\end{aligned}$$

by the definition of k_0 in (5.22a), which gives the assertion.

(b) Let $z \in Z$ be arbitrary. By Assumption 5.4 we obtain for

$$\mathbf{A}_{\phi_\ell}^S = S\mathbf{A}_{\phi_\ell}S^{-1} = \mathbf{A}_{\phi_\ell} + B(\phi_\ell), \qquad \|B(\phi_\ell)x\|_{\phi_\ell} \le k_0\beta\|x\|_{\phi_\ell}$$

that the semigroups satisfies $\left\|e^{t\mathbf{A}_\phi^S}x\right\|_\phi \le e^{k_0\beta t}\|x\|_\phi$. From this we conclude

$$\begin{aligned}
\left\|e^{\tau\mathbf{A}_{\phi_k}}e^{\tau\mathbf{A}_{\phi_{k-1}}}\dots e^{\tau\mathbf{A}_{\phi_j}}x\right\|_Z &= \left\|S^{-1}e^{\tau\mathbf{A}_{\phi_k}^S}e^{\tau\mathbf{A}_{\phi_{k-1}}^S}\dots e^{\tau\mathbf{A}_{\phi_j}^S}Sx\right\|_Z \\
&\le \left\|S^{-1}\right\|_{X\leftarrow Z}\nu_X^{1/2}\left\|e^{\tau\mathbf{A}_{\phi_k}^S}e^{\tau\mathbf{A}_{\phi_{k-1}}^S}\dots e^{\tau\mathbf{A}_{\phi_j}^S}Sx\right\|_{\phi_k} \\
&\le \left\|S^{-1}\right\|_{X\leftarrow Z}\nu_X^{1/2}e^{k_0\beta\tau}\left\|e^{\tau\mathbf{A}_{\phi_{k-1}}^S}\dots e^{\tau\mathbf{A}_{\phi_j}^S}Sx\right\|_{\phi_k} \\
&\le \left\|S^{-1}\right\|_{X\leftarrow Z}\nu_X^{1/2}e^{(k_1+k_0\beta)\tau}\left\|e^{\tau\mathbf{A}_{\phi_{k-1}}^S}\dots e^{\tau\mathbf{A}_{\phi_j}^S}x\right\|_{\phi_{k-1}} \\
&\le \dots \\
&\le \left\|S^{-1}\right\|_{X\leftarrow Z}k_0 e^{\omega_2(k-j+1)\tau}\|Sx\|_X \\
&\le c_0 e^{\omega_2(k-j+1)\tau}\|x\|_X ,
\end{aligned}$$

by the definition of c_0 in (5.22b) and of ω_2 in (5.29). □

Remark 7.9. *From the proof we can see that in the X-norm one could replace $k-j+1$ by $k-j$, but this is not possible in the Z-norm. However, we cannot gain anything from this in the latter error analysis, and we hence stay with this suboptimal bound for the sake of consistency with the error bounds in the stronger norms.*

Corollary 7.10. *The bounds in Lemma 7.8 hold true, if we replace $e^{\tau\mathbf{A}_{\phi_\ell}}$ by $\varphi_1(\tau\mathbf{A}_{\phi_\ell})$ for some $\ell \in \{0,\dots,N\}$.*

Proof. For $x \in X$, this simply follows from the bounds

$$\begin{aligned}
\|\varphi_1(\tau\mathbf{A}_{\phi_\ell})x\|_{\phi_\ell} &\le \int_0^1 \left\|e^{(1-s)\tau\mathbf{A}_{\phi_\ell}}x\right\|_{\phi_\ell} ds \le \|x\|_{\phi_\ell} , \\
\left\|\varphi_1(\tau\mathbf{A}_{\phi_\ell}^S)x\right\|_{\phi_\ell} &\le \int_0^1 \left\|e^{(1-s)\tau\mathbf{A}_{\phi_\ell}^S}x\right\|_{\phi_\ell} ds \le e^{k_0\beta\tau}\|x\|_{\phi_\ell} ,
\end{aligned}$$

which are the same as for $e^{\tau\mathbf{A}_{\phi_\ell}}$ and $e^{\tau\mathbf{A}_{\phi_\ell}^S}$. □

In order to use the result of Lemma 7.8 in the error analysis, we have to guarantee that the numerical approximations stay in the space E from (7.9). The following lemma is an extension of [41, Theorem 4.1] and establishes this at least as long as the lower bound on the existence time T of the exact solution.

Lemma 7.11. *Let Assumptions 5.1, 5.4, and 5.5 hold. For T defined in (5.30) and initial values*

$$\|u_0\|_Y \leq R_0 \coloneqq \tfrac{1}{4c_0} R, \qquad \|u_0\|_Z \leq r_0 \coloneqq \tfrac{1}{4c_0} r,$$

the numerical approximations given by (7.3) satisfy for $N\tau \leq T$

$$(u_0, \dots, u_N) \in E\big(N, R, r, \tau\gamma\big)\,, \tag{7.10}$$

for E defined in (7.9) and γ in (5.29).

Proof. We first introduce an abbreviation for the product of several semigroups

$$\mathbf{S}_i^k \coloneqq \begin{cases} e^{\tau \mathbf{A}_k} \dots e^{\tau \mathbf{A}_i}, & i \leq k, \\ I, & i > k \end{cases}$$

and with this it holds

$$\begin{aligned} u_{n+1} &= e^{\tau \mathbf{A}_n} u_n + \tau \varphi_1(\tau \mathbf{A}_n) \mathbf{f}_n \\ &= e^{\tau \mathbf{A}_n} \Big(e^{\tau \mathbf{A}_{n-1}} u_{n-1} + \tau \varphi_1(\tau \mathbf{A}_{n-1}) \mathbf{f}_{n-1} \Big) + \tau \varphi_1(\tau \mathbf{A}_n) \mathbf{f}_n \\ &= \mathbf{S}_0^n u_0 + \tau \sum_{j=0}^{n} \mathbf{S}_{j+1}^n \varphi_1(\tau \mathbf{A}_j) \mathbf{f}_j\,. \end{aligned} \tag{7.11}$$

We prove (7.10) by induction on n. Hence, let $n \leq N-1$ and assume $(u_0, \dots, u_n) \in E\big(n, R, r, \gamma\tau\big)$. Then by Lemma 7.8 and Corollary 7.10 we estimate for $j \leq n$

$$\big\|\mathbf{S}_{j+1}^n \varphi_1(\tau \mathbf{A}_j)\big\|_{Y \leftarrow Y}\,,\ \big\|\mathbf{S}_{j+1}^n \varphi_1(\tau \mathbf{A}_j)\big\|_{Z \leftarrow Z} \leq c_0 e^{\omega_2 (n-j+1)\tau}\,. \tag{7.12}$$

Taking the Y-norm in (7.11) gives with the bounds in Lemma 5.6

$$\begin{aligned} \|u_{n+1}\|_Y &\leq c_0 e^{\omega_2 t_{n+1}} \|u_0\|_Y + c_0 \tau \sum_{j=0}^{n} e^{\omega_2 (n-j+1)\tau} \|\mathbf{f}_j\|_Y \\ &\leq c_0 e^{\omega_2 t_{n+1}} \big(\|u_0\|_Y + T\, C_{f,Y,\infty} \big) \\ &\leq 2c_0 \big(R_0 + \tfrac{1}{4c_0} R \big) = R\,, \end{aligned} \tag{7.13}$$

since $t_{n+1} \leq T$ and (5.30) hold, where we used the induction hypothesis to bound $\mathbf{f}_j$. In the same way, we get with Lemma 5.6

$$\begin{aligned} \|u_{n+1}\|_Z &\leq c_0 e^{\omega_2 t_{n+1}} \|u_0\|_Z + c_0 \tau \sum_{j=0}^{n} e^{\omega_2 (n-j+1)\tau} \|\mathbf{f}_j\|_Z \\ &\leq c_0 e^{\omega_2 T} \big(\|u_0\|_Z + T C_{f,Z,\infty} \big) \\ &\leq 2\big(\tfrac{r}{4} + \tfrac{r}{4} \big) = r\,. \end{aligned} \tag{7.14}$$

We close the induction estimating with (5.20b) the term

$$\begin{aligned} \|u_{n+1} - u_n\|_Y &\leq \big\| \big(e^{\tau \mathbf{A}_n} - I \big) u_n \big\|_Y + \tau \|\varphi_1(\tau \mathbf{A}_n) \mathbf{f}_n\|_Y \\ &= \tau \|\mathbf{A}_n \varphi_1(\tau \mathbf{A}_n) u_n\|_Y + \tau \|\varphi_1(\tau \mathbf{A}_n) \mathbf{f}_n\|_Y \\ &\leq \tau \nu_Y \alpha_{YZ} \|\varphi_1(\tau \mathbf{A}_n) u_n\|_Z + \tau \|\varphi_1(\tau \mathbf{A}_n) \mathbf{f}_n\|_Y\,. \end{aligned} \tag{7.15}$$

If we use the representation in (7.11) for u_n,

$$\varphi_1(\tau\mathbf{A}_n)u_n = \varphi_1(\tau\mathbf{A}_n)\mathbf{S}_0^{n-1}u_0 + \tau\sum_{j=0}^{n-1}\varphi_1(\tau\mathbf{A}_n)\mathbf{S}_{j+1}^{n-1}\varphi_1(\tau\mathbf{A}_j)\mathbf{f}_j\,,$$

then Corollary 7.10 and the same computations as in (7.14) yield

$$\|\varphi_1(\tau\mathbf{A}_n)u_n\|_Z \le c_0 e^{\omega_2 T}\big(\|u_0\|_Z + TC_{f,Z,\infty}\big) \le r\,. \tag{7.16}$$

With Lemma 5.6 and (7.12) we further get

$$\|\varphi_1(\tau\mathbf{A}_n)\mathbf{f}_n\|_Y \le c_0 e^{\tau\omega_2}\, C_{f,Y,\infty} \le 2c_0\, C_{f,Y,\infty}\,, \tag{7.17}$$

where we used $\tau\omega_2 \le \ln 2$. From (7.15), together with the definition (5.22b), we arrive at

$$\|u_{n+1} - u_n\|_Y \le \tau\big(\tfrac{c_1}{c_0}r + 2c_0\, C_{f,Y,\infty}\big) = \gamma\tau\,, \tag{7.18}$$

which finally yields $(u_0,\dots,u_{n+1}) \in E\big(n+1,R,r,\gamma\tau\big)$ and the induction is closed. □

7.2.2 Defect

In this step we present a recursion for the global error given by

$$e_n \coloneqq u(t_n) - u_n\,.$$

In order to make $u(t_n)$ and u_n comparable, we use (5.2), replace $\mathbf{A}(u(t))$ by $\mathbf{A}(u_n)$ and treat the remainder as an inhomogeneity. Then the error propagation is driven by the semigroups studied in Lemma 7.8, and it remains to bound the defects. This is the main task in the following proposition.

Proposition 7.12. *Let Assumptions 5.1, 5.4, and 5.5 hold and consider the solution u given by Theorem 5.14 and numerical approximations $(u_n)_n$ given by (7.3). Then the global error satisfies the error recursion*

$$e_{n+1} = e^{\tau\mathbf{A}_n}e_n + \delta_n\,, \tag{7.19}$$

where the defect is bounded by

$$\|\delta_n\|_X \le \big(C_{\sigma,X}\,\tau\,\|e_n\|_X + C_{\delta,X}\,\tau^2\big)\,e^{\tau\omega_1}\,,$$

with constants $C_{\sigma,X}, C_{\delta,X} > 0$ that only depend on $\|u'\|_{X,\infty}$, $\|u\|_{Z,\infty}$.

Proof. We obtain from equation (5.2), plugging in the last approximation u_n and using the notation in (7.1) and (7.2), the differential equation

$$\begin{aligned} u'(t) &= \widehat{\mathbf{A}}(t)u(t) + \widehat{\mathbf{f}}(t) \\ &= \mathbf{A}_n u(t) + \mathbf{f}_n \\ &\qquad + \big(\widehat{\mathbf{A}}_n - \mathbf{A}_n\big)u(t) + \big(\widehat{\mathbf{f}}_n - \mathbf{f}_n\big) + \big(\widehat{\mathbf{A}}(t) - \widehat{\mathbf{A}}_n\big)u(t) + \big(\widehat{\mathbf{f}}(t) - \widehat{\mathbf{f}}_n\big) \\ &=: \mathbf{A}_n u(t) + \mathbf{f}_n + \sum_{i=1}^{4}\widetilde{\delta}_{n,i}(t)\,. \end{aligned}$$

The variation-of-constants formula enables us to solve this equation by

$$u(t_{n+1}) = e^{\tau \mathbf{A}_n} u(t_n) + \tau \varphi_1(\tau \mathbf{A}_n)\mathbf{f}_n + \sum_{i=1}^{4} \delta_{n,i} \tag{7.20}$$

where

$$\delta_{n,i} = \int_0^\tau e^{(\tau-s)\mathbf{A}_n} \widetilde{\delta}_{n,i}(t_n + s)\, ds, \quad i = 1, \dots, 4\,.$$

The four terms are estimated separately. By (5.20c) and Lemma 7.8 it holds

$$\begin{aligned}
\|\delta_{n,1}\|_X &= \tau \Big\| \int_0^1 e^{(1-s)\tau \mathbf{A}_n} \big(\widehat{\mathbf{A}}_n - \mathbf{A}_n\big) \widehat{u}_{n+s}\, ds \Big\|_X \\
&\le \tau \int_0^1 \Big\| e^{(1-s)\tau \mathbf{A}_n} \big(\widehat{\mathbf{A}}_n - \mathbf{A}_n\big) \widehat{u}_{n+s} \Big\|_X\, ds \\
&\le \tau k_0 \int_0^1 e^{(1-s)\tau\omega_1} \Big\| \big(\widehat{\mathbf{A}}_n - \mathbf{A}_n\big) \widehat{u}_{n+s} \Big\|_X\, ds \\
&\le \tau k_0 L_X \|e_n\|_X \int_0^1 e^{(1-s)\tau\omega_1} \|\widehat{u}_{n+s}\|_Z\, ds \\
&\le \tau k_0 L_X e^{\tau\omega_1} \|e_n\|_X \|u\|_{Z,\infty}
\end{aligned} \tag{7.21}$$

and in the same manner with Lemma 5.6

$$\begin{aligned}
\|\delta_{n,2}\|_X &= \tau \Big\| \int_0^1 e^{(1-s)\tau \mathbf{A}_n} \big(\widehat{\mathbf{f}}_n - \mathbf{f}_n\big)\, ds \Big\|_X \\
&\le \tau k_0 \int_0^1 e^{(1-s)\tau\omega_1} \Big\| \widehat{\mathbf{f}}_n - \mathbf{f}_n \Big\|_X\, ds \\
&\le \tau k_0 e^{\tau\omega_1} L_{f,X} \|e_n\|_X\,.
\end{aligned} \tag{7.22}$$

The other defects can be bounded with (5.20c) by

$$\begin{aligned}
\|\delta_{n,3}\|_X &= \tau \Big\| \int_0^1 e^{(1-s)\tau \mathbf{A}_n} \big(\widehat{\mathbf{A}}_{n+s} - \widehat{\mathbf{A}}_n\big) \widehat{u}_{n+s}\, ds \Big\|_X \\
&\le \tau k_0 \int_0^1 e^{(1-s)\tau\omega_1} \Big\| \big(\widehat{\mathbf{A}}_{n+s} - \widehat{\mathbf{A}}_n\big) \widehat{u}_{n+s} \Big\|_X\, ds \\
&\le \tau k_0 L_X \int_0^1 e^{(1-s)\tau\omega_1} \|\widehat{u}_{n+s} - \widehat{u}_n\|_X \|\widehat{u}_{n+s}\|_Z\, ds \\
&\le \tau^2 k_0 L_X e^{\tau\omega_1} \|u'\|_{X,\infty} \|u\|_{Z,\infty}\,,
\end{aligned} \tag{7.23}$$

and similarly by Lemma 5.6

$$\begin{aligned}\|\delta_{n,4}\|_X &= \tau \Big\| \int_0^1 e^{(1-s)\tau \mathbf{A}_n} \big(\widehat{\mathbf{f}}_{n+s} - \widehat{\mathbf{f}}_n\big)\, ds \Big\|_X \\ &\le \tau k_0 \int_0^1 e^{(1-s)\tau\omega_1} \Big\| \widehat{\mathbf{f}}_{n+s} - \widehat{\mathbf{f}}_n \Big\|_X \, ds \\ &\le \tau^2 k_0 e^{\tau\omega_1} L_{f,X} \big(1 + \|u'\|_{X,\infty}\big)\,.\end{aligned} \tag{7.24}$$

The assertion follows by setting

$$\delta_n := \sum_{i=1}^4 \delta_{n,i}$$

and subtracting u_{n+1} given in (7.3) from (7.20). □

Very similar computations lead to bounds in the stronger Y-norm, where we employ the additional regularity $u \in C^1([0,T], Y)$.

Corollary 7.13. *The defect in (7.19) can also be bounded by*

$$\|\delta_n\|_Y \le \big(C_{\sigma,Y}\, \tau \,\|e_n\|_Y + C_{\delta,Y}\, \tau^2\big)\, e^{\tau\omega_2}\,,$$

with constants $C_{\sigma,Y}, C_{\delta,Y} > 0$ that only depend on $\|u'\|_{Y,\infty}$, $\|u\|_{Z,\infty}$.

Proof. We only need to establish the bounds on the defects to verify the assertion.

By Lemma 7.8 and (5.20d) it holds

$$\begin{aligned}\|\delta_{n,1}\|_Y &\le \tau c_0 \int_0^1 e^{(1-s)\tau\omega_2} \Big\| \big(\widehat{\mathbf{A}}_n - \mathbf{A}_n\big)\widehat{u}_{n+s} \Big\|_Y \, ds \\ &\le \tau c_0 L_Y \,\|e_n\|_Y \int_0^1 e^{(1-s)\tau\omega_2} \,\|\widehat{u}_{n+s}\|_Z \, ds \\ &\le \tau c_0 L_Y e^{\tau\omega_2} \,\|e_n\|_Y \,\|u\|_{Z,\infty}\end{aligned} \tag{7.25}$$

and in the same manner with Lemma 5.6

$$\|\delta_{n,2}\|_Y \le \tau c_0 \int_0^1 e^{(\tau-s)\omega_2} \Big\| \widehat{\mathbf{f}}_n - \mathbf{f}_n \Big\|_Y \, ds \le \tau c_0 e^{\tau\omega_2} L_{f,Y} \,\|e_n\|_Y\,. \tag{7.26}$$

The other defects can be bounded with (5.20d) by

$$\begin{aligned}\|\delta_{n,3}\|_Y &\le \tau c_0 \int_0^1 e^{(1-s)\tau\omega_2} \Big\| \big(\widehat{\mathbf{A}}_{n+s} - \widehat{\mathbf{A}}_n\big)\widehat{u}_{n+s} \Big\|_Y \, ds \\ &\le \tau c_0 L_Y \int_0^1 e^{(1-s)\tau\omega_2} \,\|\widehat{u}_{n+s} - \widehat{u}_n\|_Y \,\|\widehat{u}_{n+s}\|_Z \, ds \\ &\le \tau^2 c_0 L_Y e^{\tau\omega_2} \,\|u'\|_{Y,\infty} \,\|u\|_{Z,\infty}\,,\end{aligned} \tag{7.27}$$

and similarly by Lemma 5.6

$$\|\delta_{n,4}\|_Y \le \tau c_0 \int_0^1 e^{(1-s)\tau\omega_2} \left\|\widehat{\mathbf{f}}_{n+s} - \widehat{\mathbf{f}}_n\right\|_Y ds \le \tau^2 c_0 e^{\tau\omega_2} L_{f,Y}\left(1 + \|u'\|_{Y,\infty}\right), \tag{7.28}$$

which yields the required bound on the defect. □

7.2.3 Global error

A combination of the stability bounds and the defects yields the global error result.

Proof of Theorem 7.1. We note that the assumptions of the theorem allow us to apply all results of Sections 7.2.1 and 7.2.2.

(a) We first prove the bound in the X-norm. Using the error recursion in (7.19) and recalling the product $\mathbf{S}_i^k = e^{\tau\mathbf{A}_k} \dots e^{\tau\mathbf{A}_i}$ for $k \ge i$, we obtain by a discrete version of the variation-of-constants formula

$$e_{n+1} = e^{\tau\mathbf{A}_n} e_n + \delta_n = \mathbf{S}_0^n e_0 + \sum_{j=0}^n \mathbf{S}_{j+1}^n \delta_j \,. \tag{7.29}$$

As it holds $e_0 = 0$, with Lemma 7.8 and Proposition 7.12 we get as in (7.12)

$$\begin{aligned}\|e_{n+1}\|_X &\le \sum_{j=0}^n \left\|\mathbf{S}_{j+1}^n\right\|_{X\leftarrow X} \|\delta_j\|_X \\ &\le k_0\tau \sum_{j=0}^n e^{\omega_1(n+1-j)\tau} C_\sigma \|e_j\|_X + k_0\tau \sum_{j=0}^n e^{\omega_1(n+1-j)\tau} C_\delta \tau\end{aligned}$$

which is equivalent to

$$e^{-\omega_1(n+1)\tau} \|e_{n+1}\|_X \le C_\sigma k_0 \tau \sum_{j=0}^n e^{-\omega_1 j\tau} \|e_j\|_X + k_0\tau \sum_{j=0}^n e^{-\omega_1 j\tau} C_\delta \tau \,.$$

A Gronwall argument yields with $t_{n+1} = (n+1)\tau$

$$e^{-\omega_1 t_{n+1}} \|e_{n+1}\|_X \le t_{n+1} e^{C_\sigma k_0 t_{n+1}} k_0 C_\delta \, \tau$$

and hence

$$\|e_{n+1}\|_X \le t_{n+1} e^{(\omega_1 + C_\sigma k_0) t_{n+1}} k_0 C_\delta \, \tau \,,$$

which completes the proof.

(b) The error bound in the Y-norm is easily derived replacing Proposition 7.12 by Corollary 7.13 and ω_1 by ω_2 which yields

$$\|e_{n+1}\|_Y \le c_0\tau \sum_{j=0}^n e^{\omega_2(n+1-j)\tau} C_\sigma \|e_j\|_Y + c_0\tau \sum_{j=0}^n e^{\omega_2(n+1-j)\tau} C_\delta \tau$$

and again bringing $e^{\omega_2(n+1)\tau}$ to the other side and a Gronwall argument yield the assertion. □

7.3 Error analysis of the exponential midpoint rule

The proof of Theorem 7.7 has a very similar structure to the one of Theorem 7.1, but we need to take special care of the extrapolations of previous approximations. This induces some technical difficulties in the stability analysis. Next, we derive the error propagation similar to Proposition 7.12 and bound the appearing defects. Several terms are treated in the standard way as the exponential Euler method, whereas for the remaining terms the additionally required differentiability come into play.

7.3.1 Stability

Since the numerical method is now driven by the exponential evaluated at the extrapolated midpoints, we again derive a result for bounds on the composition of these linear flows. The choice of the larger constants in the space E becomes clearer in Lemma 7.15 when we derive the bounds on the midpoints.

Lemma 7.14. *Let Assumptions 5.1 and 5.4 hold. Further, let*

$$\phi = \left(\phi_{1/2}, \phi_{3/2} \ldots, \phi_{N+1/2}\right) \in E\left(N, \widehat{R}, \widehat{r}, 2\tau\widehat{\gamma}\right).$$

We obtain the stability bounds as in (7.8) for $j \leq k$ and $j,k \in \{\frac{1}{2}, \frac{3}{2}, \ldots, N+\frac{1}{2}\}$ with k_0, c_0, ω_1 and ω_2 replaced by $\widehat{k}_0$, $\widehat{c}_0$, $\widehat{\omega}_1$ and $\widehat{\omega}_2$, respectively, where $\widehat{\omega}_1 \coloneqq 2\widehat{k}_1(\widehat{\gamma})$ and $\widehat{\omega}_2$ is given in (7.8).

Proof. The proof is similar to the one of Lemma 7.8 and can be found in the Appendix of [41]. □

This enables us to prove bounds on the numerical approximations very similar to the bounds provided in Lemma 7.11. The only difference lies in the time $\widehat{T}_{\text{mid}}$ defined in (7.7), which is necessary to obtain uniform bounds in the numerical approximations.

Lemma 7.15. *Let Assumptions 5.1, 5.4, and 5.5 hold. For $\widehat{T}_{mid}$ defined in (7.7), $\tau \leq \tau_0$ with τ_0 given in (7.6) and initial values*

$$\|u_0\|_Y \leq R_0 \coloneqq \tfrac{1}{4\widehat{c}_0} R, \qquad \|u_0\|_Z \leq r_0 \coloneqq \tfrac{1}{4\widehat{c}_0} r,$$

the numerical approximations given in (7.4) satisfy for $N\tau \leq \widehat{T}_{mid}$

$$(u_0, \ldots, u_N) \in E\big(N, R, r, \tau\widehat{\gamma}\big), \qquad (u_{1/2}, \ldots, u_{N-1/2}) \in E\left(N-1, \widehat{R}, \widehat{r}, 2\tau\widehat{\gamma}\right). \tag{7.30}$$

Proof. We proof the assertion by induction on n and assume (7.30) is true for some $1 \leq n \leq N-1$, i.e.,

$$(u_0, \ldots, u_n) \in E\big(n, R, r, \gamma\tau\big), \qquad (u_{1/2}, \ldots, u_{n-1/2}) \in E\left(n-1, \widehat{R}, \widehat{r}, 2\tau\widehat{\gamma}\right).$$

Note that the base case $n = 1$ is the same as for the exponential Euler due to the choice $u_{1/2} \coloneqq u_0$ and hence true by Lemma 7.11.

(a) By the induction hypothesis and $\tau \leq \tau_0$ given in (7.6), we obtain for the extrapolated midpoint

$$\left\|u_{n+1/2}\right\|_Y \leq \|u_n\|_Y + \tfrac{1}{2}\|u_n - u_{n-1}\|_Y \leq R + \tfrac{\widehat{\gamma}\tau}{2} \leq \widehat{R},$$

as well as

$$\left\|u_{n+1/2}\right\|_Z \leq \tfrac{3}{2}\|u_n\|_Z + \tfrac{1}{2}\|u_{n-1}\|_Z \leq 2r = \widehat{r}.$$

Defining $u_{-1} \coloneqq u_0$, it holds $u_{1/2} = \frac{3}{2}u_0 - \frac{1}{2}u_{-1}$ and we estimate for $n \geq 1$

$$\left\|u_{n+1/2} - u_{n-1/2}\right\|_Y \leq \tfrac{3}{2}\left\|u_n - u_{n-1}\right\|_Y + \tfrac{1}{2}\left\|u_{n-1} - u_{n-2}\right\|_Y \leq 2\widehat{\gamma}\tau\,.$$

Hence, it holds $(u_{1/2}, \dots, u_{n+1/2}) \in E\left(n, \widehat{R}, \widehat{r}, 2\tau\widehat{\gamma}\right)$.

(b) By part (a), we can apply Lemmas 5.6 and 7.14 and, together with the induction hypothesis, we obtain as in (7.13)

$$\|u_{n+1}\|_Y \leq \widehat{c}_0 e^{\widehat{\omega}_2 t_{n+1}}\left(\|u_0\|_Y + \widehat{T}_{\mathrm{mid}}\widehat{C}_{f,Y,\infty}\right) \leq R$$

and in the same way

$$\|u_{n+1}\|_Z \leq \widehat{c}_0 e^{\widehat{\omega}_2 t_{n+1}}\left(\|u_0\|_Z + \widehat{T}_{\mathrm{mid}}\widehat{C}_{f,Z,\infty}\right) \leq r\,.$$

Finally, along the lines of (7.15), (7.16), (7.17), and (7.18) we establish

$$\begin{aligned}\|u_{n+1} - u_n\|_Y &= \left\|\left(e^{\tau\mathbf{A}_{n+1/2}} - I\right)u_n + \tau\varphi_1(\tau\mathbf{A}_{n+1/2})\mathbf{f}_{n+1/2}\right\|_Y \\ &\leq \tau\left\|\mathbf{A}_{n+1/2}\varphi_1(\tau\mathbf{A}_{n+1/2})u_n\right\|_Y + \tau\widehat{c}_0 e^{\widehat{\omega}_2\tau}\widehat{C}_{f,Y,\infty} \\ &\leq \widehat{\gamma}\tau\,,\end{aligned}$$

which gives $(u_0, \dots, u_{n+1}) \in E\big(n+1, R, r, \gamma\tau\big)$, so the induction is closed. □

7.3.2 Defects and global error

In order to increase the readability of the proof we define analogously to $u_{n+1/2}$ in (7.4) the extrapolation of the exact solution and the corresponding operator by

$$\underline{\widehat{u}}_{n+1/2} = \tfrac{1}{2}\left(3\widehat{u}_n - \widehat{u}_{n-1}\right), \quad \underline{\widehat{\mathbf{A}}}_{n+1/2} = \mathbf{A}(\underline{\widehat{u}}_{n+1/2}), \quad \underline{\widehat{\mathbf{f}}}_{n+1/2} = f(t_{n+1/2}, \underline{\widehat{u}}_{n+1/2})$$

with $\underline{\widehat{u}}_{1/2} = u_0$. By the proof of Theorem 5.14, we have

$$\left\|\underline{\widehat{u}}_{n+1/2}\right\|_Y \leq \|\widehat{u}_n\|_Y + \frac{1}{2}\|\widehat{u}_n - \widehat{u}_{n-1}\|_Y \leq R + \tfrac{\gamma\tau}{2} \leq \widehat{R}, \qquad \left\|\underline{\widehat{u}}_{n+1/2}\right\|_Z \leq 2r = \widehat{r}$$

and thus we can use the same bounds as for $u_{n+1/2}$. Further, we consider the extrapolated error

$$e_{n+1/2} = \underline{\widehat{u}}_{n+1/2} - u_{n+1/2}\,.$$

We emphasize that one does not necessarily need to introduce this extrapolated error. However, it makes the following computations a bit shorter, and we get rid of this term at the very end in the error accumulation.

Proposition 7.16. *Let Assumptions 5.1, 5.4, 5.5, 7.2, and 7.3 be satisfied and consider the solution u given by Corollary 7.6 and numerical approximations $(u_n)_n$ given by (7.4). Then the global error satisfies the error recursion*

$$e_{n+1} = e^{\tau\mathbf{A}_{n+1/2}}e_n + \delta_n\,, \tag{7.31}$$

where the defect is bounded by

$$\begin{aligned}\|\delta_0\|_X &\leq C_\delta\,\tau^2\,e^{\tau\widehat{\omega}_1}\,, \\ \|\delta_n\|_X &\leq \left(C_{\sigma,X}\,\tau\left\|e_{n+1/2}\right\|_X + C_{\delta,X}\,\tau^3\right)e^{\tau\widehat{\omega}_1}\,, \qquad n \geq 1\,,\end{aligned}$$

with constants $C_{\sigma,X}, C_{\delta,X}$ only depending on $\|u''\|_{X,\infty}$, $\|u'\|_{Y,\infty}$, $\|u\|_{Z,\infty}$.

Proof. We proceed as in Proposition 7.12 plugging in $u_{n+1/2}$ to obtain

$$\begin{aligned}
u'(t) &= \widehat{\mathbf{A}}(t)u(t) + \widehat{\mathbf{f}}(t) \\
&= \mathbf{A}_{n+1/2}u(t) + \mathbf{f}_{n+1/2} \\
&\quad + \left(\underline{\widehat{\mathbf{A}}}_{n+1/2} - \mathbf{A}_{n+1/2}\right) u(t) + \left(\underline{\widehat{\mathbf{f}}}_{n+1/2} - \mathbf{f}_{n+1/2}\right) \\
&\quad + \left(\widehat{\mathbf{A}}_{n+1/2} - \underline{\widehat{\mathbf{A}}}_{n+1/2}\right) u(t) + \left(\widehat{\mathbf{f}}_{n+1/2} - \underline{\widehat{\mathbf{f}}}_{n+1/2}\right) \\
&\quad + \left(\widehat{\mathbf{A}}(t) - \widehat{\mathbf{A}}_{n+1/2}\right) u(t) \qquad + \left(\widehat{\mathbf{f}}(t) - \widehat{\mathbf{f}}_{n+1/2}\right) \\
&=: \mathbf{A}_{n+1/2}u(t) + \mathbf{f}_{n+1/2} + \sum_{i=1}^{6} \widetilde{\delta}_{n,i}(t)\,.
\end{aligned}$$

Applying the variation-of-constants formula as above yields the terms

$$u(t_{n+1}) = e^{\tau \mathbf{A}_{n+1/2}} u(t_n) + \tau\varphi_1(\tau\mathbf{A}_{n+1/2})\mathbf{f}_{n+1/2} + \sum_{i=1}^{6} \delta_{n,i}\,. \tag{7.32}$$

We split the proof into four parts. We first bound the four terms that have appeared similarly in the proof of Proposition 7.12. Here, we need to distinguish the defect of the first step from the others. In the third and forth part the assumptions on the differentiability enter.

(a) By definition we have $u_{1/2} = \underline{\widehat{u}}_{1/2}$ and hence $\delta_{0,1} = \delta_{0,2} = 0$. Since the first step is given by an exponential Euler step, (7.23) and (7.24) yield

$$\|\delta_{0,3}\|_X + \|\delta_{0,4}\|_X \le C\tau^2\,.$$

(b) We now turn to the case $n \ge 1$. The same computation as in (7.21) gives

$$\begin{aligned}
\|\delta_{n,1}\|_X &\le \tau\widehat{k}_0 \int_0^1 e^{(1-s)\tau\widehat{\omega}_1} \left\| \left(\underline{\widehat{\mathbf{A}}}_{n+1/2} - \mathbf{A}_{n+1/2}\right) \widehat{u}_{n+s} \right\|_X ds \\
&\le \tau\widehat{k}_0\widehat{L}_X e^{\tau\widehat{\omega}_1} \left\|e_{n+1/2}\right\|_X \|u\|_{Z,\infty}
\end{aligned}$$

as well as in (7.22)

$$\|\delta_{n,2}\|_X \le \tau\widehat{k}_0 e^{\tau\widehat{\omega}_1} \widehat{L}_{f,X} \left\|e_{n+1/2}\right\|_X\,.$$

The defect $\delta_{n,3}$ can be bounded by Lemma 7.14 and (5.20c)

$$\begin{aligned}
\|\delta_{n,3}\|_X &\le \tau\widehat{k}_0 \int_0^1 e^{(1-s)\tau\widehat{\omega}_1} \left\| \left(\widehat{\mathbf{A}}_{n+1/2} - \underline{\widehat{\mathbf{A}}}_{n+1/2}\right) \widehat{u}_{n+s} \right\|_X ds \\
&\le \tau\widehat{k}_0\widehat{L}_X \int_0^1 e^{(1-s)\tau\widehat{\omega}_1} \left\| u(t_{n+1/2}) - \underline{\widehat{u}}_{n+1/2} \right\|_X \|\widehat{u}_{n+s}\|_Z \, ds \\
&\le \tau^3\widehat{k}_0\widehat{L}_X e^{\tau\widehat{\omega}_1} \tfrac{3}{8} \|u''\|_{X,\infty} \|u\|_{Z,\infty}\,,
\end{aligned}$$

as well as $\delta_{n,4}$ with Lemma 5.6 by

$$\|\delta_{n,4}\|_X \le \tau^3\widehat{k}_0 e^{\tau\widehat{\omega}_1} \widehat{L}_{f,X} \tfrac{3}{8} \|u''\|_{X,\infty}\,,$$

where we used Taylor expansion on $u(t_{n+1/2})$ for both defects, see Lemma B.11.

(c) The last two defects are considered for $n \geq 0$. We first prove the statement for $\delta_{n,5}$. We have

$$\delta_{n,5} = \int_0^\tau e^{(\tau-s)\mathbf{A}_{n+1/2}} d_n(t_n + s)\, ds \tag{7.33}$$

with the function

$$d_n(t) \coloneqq \big(\widehat{\mathbf{A}}(t) - \widehat{\mathbf{A}}_{n+1/2}\big)u(t), \qquad d_n(t_{n+1/2}) = 0\,.$$

To expand this we first need

$$\begin{aligned} d_n'(t) &= \widehat{\mathbf{A}}'(t)u(t) &&+ \big(\widehat{\mathbf{A}}(t) - \widehat{\mathbf{A}}_{n+1/2}\big)u'(t) \\ &=: \dot{d}_{n,1}(t) &&+ \dot{d}_{n,2}(t) \end{aligned}$$

and hence

$$d_n'(t_{n+1/2}) = \dot{d}_{n,1}(t_{n+1/2}) = \widehat{\mathbf{A}}'(t_{n+1/2})\widehat{u}_{n+1/2}\,.$$

We also obtain

$$\ddot{d}_{n,1}(t) \coloneqq \tfrac{d}{dt}\dot{d}_{n,1}(t) = \widehat{\mathbf{A}}''(t)u(t) + \widehat{\mathbf{A}}'(t)u'(t)\,.$$

Lemma 7.5 implies the following bounds

$$\big\|\dot{d}_{n,1}(t_{n+1/2})\big\|_Y = \big\|\widehat{\mathbf{A}}'(t_{n+1/2})\widehat{u}_{n+1/2}\big\|_Y \leq C_{YZ}^A \,\|u\|_{Z,\infty}\,, \tag{7.34}$$

and

$$\begin{aligned} \big\|\ddot{d}_{n,1}(t)\big\|_X &\leq \big\|\widehat{\mathbf{A}}''(t)u(t)\big\|_X + \big\|\widehat{\mathbf{A}}'(t)u'(t)\big\|_X \\ &\leq C_{XZ}^A \,\|u\|_{Z,\infty} + C_{XY}^A \,\|u'\|_{Y,\infty}\,, \end{aligned} \tag{7.35}$$

as well as

$$\big\|\dot{d}_{n,2}(t)\big\|_X = \big\|\big(\widehat{\mathbf{A}}(t) - \widehat{\mathbf{A}}_{n+1/2}\big)u'(t)\big\|_X \leq \tfrac{\tau}{2} C_{XY}^A \,\|u'\|_{Y,\infty}\,. \tag{7.36}$$

Using $d_n(t_{n+1/2}) = 0$ and integration by parts, we expand

$$\begin{aligned} d_n(t_n + s) &= \int_0^{s-\tau/2} \dot{d}_{n,1}(t_{n+1/2} + \sigma)\, d\sigma + \int_0^{s-\tau/2} \dot{d}_{n,2}(t_{n+1/2} + \sigma)\, d\sigma \\ &= \big(s - \tfrac{\tau}{2}\big)\, \dot{d}_{n,1}(t_{n+1/2}) + \int_0^{s-\tau/2} \big(s - \tfrac{\tau}{2} - \sigma\big)\, \ddot{d}_{n,1}(t_{n+1/2} + \sigma)\, d\sigma \\ &\quad + \int_0^{s-\tau/2} \dot{d}_{n,2}(t_{n+1/2} + \sigma)\, d\sigma\,. \end{aligned}$$

Plugging this in (7.33) gives

$$\begin{aligned}
\delta_{n,5} &= \int_0^\tau e^{(\tau-s)\mathbf{A}_{n+1/2}} d_n(t_n+s)\,ds \\
&= \left(\int_0^\tau e^{(\tau-s)\mathbf{A}_{n+1/2}} \left(s-\tfrac{\tau}{2}\right)\,ds \right) \dot d_{n,1}(t_{n+1/2}) \\
&\quad + \int_0^\tau e^{(\tau-s)\mathbf{A}_{n+1/2}} \int_0^{s-\tau/2} \left(s-\tfrac{\tau}{2}-\sigma\right) \ddot d_{n,1}(t_{n+1/2}+\sigma)\,d\sigma\,ds \\
&\quad + \int_0^\tau e^{(\tau-s)\mathbf{A}_{n+1/2}} \int_0^{s-\tau/2} \dot d_{n,2}(t_{n+1/2}+\sigma)\,d\sigma\,ds \\
&= \delta_{n,5}^1 + \delta_{n,5}^2 + \delta_{n,5}^3
\end{aligned}$$

We estimate these terms separately. By integration by parts we obtain

$$\begin{aligned}
\delta_{n,5}^1 &= \left(\int_0^\tau e^{(\tau-s)\mathbf{A}_{n+1/2}} \left(s-\tfrac{\tau}{2}\right)\,ds \right) \dot d_{n,1}(t_{n+1/2}) \\
&= \left(\frac{1}{2} \int_0^\tau e^{(\tau-s)\mathbf{A}_{n+1/2}} \left(s^2-\tau s\right)\,ds \right) \mathbf{A}_{n+1/2} \dot d_{n,1}(t_{n+1/2})
\end{aligned}$$

and estimate by Lemma 7.14, (5.20a), and (7.34)

$$\begin{aligned}
\left\|\delta_{n,5}^1\right\|_X &\le \tfrac{1}{12}\widehat{k}_0 \tau^3 e^{\tau\widehat{\omega}_1} \left\|\mathbf{A}_{n+1/2}\,\dot d_{n,1}(t_{n+1/2})\right\|_X \\
&\le \tfrac{1}{12}\widehat{k}_0 \widehat{\nu}_X \widehat{\alpha}_{XY} \tau^3 e^{\tau\widehat{\omega}_1} \left\|\dot d_{n,1}(t_{n+1/2})\right\|_Y \\
&\le \left(\tfrac{1}{12}\widehat{k}_0 \widehat{\nu}_X \widehat{\alpha}_{XY} C_{YZ}^A \|u\|_{Z,\infty}\right) \tau^3 e^{\tau\widehat{\omega}_1}\,.
\end{aligned}$$

We further obtain by (7.35)

$$\begin{aligned}
\left\|\delta_{n,5}^2\right\|_X &\le \tfrac{1}{24}\widehat{k}_0 \tau^3 e^{\tau\widehat{\omega}_1} \left\|\ddot d_{n,1}\right\|_{X,\infty} \\
&\le \tfrac{1}{24}\widehat{k}_0 \Big(C_{XZ}^A \|u\|_{Z,\infty} + C_{XY}^A \|u'\|_{Y,\infty} \Big) \tau^3 e^{\tau\widehat{\omega}_1}\,,
\end{aligned}$$

and at last by (7.36)

$$\begin{aligned}
\left\|\delta_{n,5}^3\right\|_X &\le \tfrac{1}{4}\widehat{k}_0 \tau^2 e^{\tau\widehat{\omega}_1} \left\|\dot d_{n,2}\right\|_{X,\infty} \\
&\le \Big(\tfrac{1}{8}\widehat{k}_0 C_{XY}^A \|u'\|_{Y,\infty} \Big) \tau^3 e^{\tau\widehat{\omega}_1}\,.
\end{aligned}$$

This gives the assertion for $\delta_{n,5}$.

(d) The proof for $\delta_{n,6}$ is very similar. We have the representation (7.33) with d_n replaced by

$$D_n(t) \coloneqq \widehat{\mathbf{f}}(t) - \widehat{\mathbf{f}}_{n+1/2}, \qquad D_n(t_{n+1/2}) = 0\,.$$

Computing the derivatives, we obtain with Lemma 7.4 similar to (7.34) and (7.35)

$$\begin{aligned}
\left\|\dot D_{n,1}\right\|_Y &= \left\|\widehat{\mathbf{f}}'(t)\right\|_Y \le C_{f',Y,\infty} \\
\left\|\ddot D_{n,1}\right\|_X &= \left\|\widehat{\mathbf{f}}''(t)\right\|_X \le C_{f'',X,\infty}
\end{aligned} \tag{7.37}$$

and in particular the term corresponding to $\dot d_{n,2}$ does not appear. Hence, we proceed as in part (c) which yields the desired bound for $\delta_{n,6}$.

Finally, setting

$$\delta_n \coloneqq \sum_{i=1}^{6} \delta_{n,i}$$

and subtracting u_{n+1} given in (7.4) from (7.32) closes the proof. □

Corollary 7.17. *The defect in* (7.31) *can also be bounded by*

$$\|\delta_n\|_Y \leq \left(C_{\sigma,Y}\,\tau \left\|e_{j+1/2}\right\|_Y + C_{\delta,Y}\,\tau^2\right) e^{\tau\widehat{\omega}_2},$$

with constants $C_{\sigma,Y}, C_{\delta,Y} > 0$ *that only depend on* $\|u'\|_{Y,\infty}$, $\|u\|_{Z,\infty}$.

Proof. We proceed analogously to Corollary 7.13. By Lemma 7.14, Lemma 5.6 and (5.20d) it holds

$$\begin{aligned}
\|\delta_{n,1}\|_Y &\leq \tau\widehat{c}_0 e^{\tau\widehat{\omega}_2}\widehat{L}_Y \left\|e_{n+1/2}\right\|_Y \|u\|_{Z,\infty}\,,\\
\|\delta_{n,2}\|_Y &\leq \tau\widehat{c}_0 e^{\tau\widehat{\omega}_2}\widehat{L}_{f,Y} \left\|e_{n+1/2}\right\|_Y\,.
\end{aligned}$$

Using Taylor expansion only up to order 1, see Lemma B.11, with (5.20d) and Lemma 5.6 we bound

$$\begin{aligned}
\|\delta_{n,3}\|_Y &\leq \tau^2\widehat{c}_0 e^{\tau\widehat{\omega}_2}\widehat{L}_Y \|u'\|_{Y,\infty}\|u\|_{Z,\infty}\,,\\
\|\delta_{n,4}\|_Y &\leq \tau^2\widehat{c}_0 e^{\tau\widehat{\omega}_2}\widehat{L}_{f,Y} \|u'\|_{Y,\infty}\,.
\end{aligned}$$

Since we only aim for defects of order 2, we estimate in the exact same way

$$\begin{aligned}
\|\delta_{n,5}\|_Y &\leq \tau\widehat{c}_0 \int_0^1 e^{(1-s)\tau\widehat{\omega}_2} \left\|\left(\widehat{\mathbf{A}}_{n+s} - \widehat{\mathbf{A}}_{n+1/2}\right)\widehat{u}_{n+s}\right\|_Y ds\\
&\leq \tau^2\widehat{c}_0 e^{\tau\widehat{\omega}_2}\widehat{L}_Y \tfrac{1}{2}\|u'\|_{Y,\infty}\|u\|_{Z,\infty}\,,
\end{aligned}$$

and also

$$\|\delta_{n,6}\|_Y \leq \tau^2\widehat{c}_0 e^{\tau\widehat{\omega}_2}\widehat{L}_{f,Y}\tfrac{1}{2}\left(1 + \|u'\|_{Y,\infty}\right),$$

which gives the assertion. □

We can finally give the proof of the error bound of the exponential midpoint rule.

Proof of Theorem 7.7. We note that the assumptions of the theorem allow us to apply the results of Section 7.3.1 as well as Proposition 7.16 and Corollary 7.17.

By (7.31) we resolve the error recursion as in (7.29) and use the bounds provided in Lemma 7.14, Proposition 7.16, and Corollary 7.17. With the observation

$$\sum_{j=1}^{n} \left\|e_{j+1/2}\right\|_V \leq 2\sum_{j=1}^{n} \|e_j\|_V\,,$$

for $V \in \{X, Y\}$, the bound in the X- and the Y-norm is derived analogously to Theorem 7.1. □

7.4 Numerical experiments

In order to illustrate the theoretical findings in Theorems 7.1 and 7.7, we consider the quasilinear wave equation (5.4) rewritten in the form

$$\lambda(q)q'' = \Delta q + r(t, q, q') \tag{7.38}$$

obtained from the Kerr-type nonlinearity (5.8) with $\chi = -\frac{1}{30}$ and coefficients

$$\lambda(q) = 1 - \tfrac{1}{10}q^2, \qquad r(t, q, q') = \tfrac{1}{5}q \cdot (q')^2 - \tfrac{1}{5}\sin(q) + f(t)\,,$$

on the unit disc $\Omega \subseteq \mathbb{R}^2$ subject to homogeneous Dirichlet boundary conditions. With $x = \big(\, x_1, x_2 \,\big)$, we chose the source term f by

$$f(t, x) = \cos^2(t)\sin\Big((1+t)\big(1 - |x|^2\big)^3\Big)\,.$$

To illustrate the sufficiency of our regularity assumptions, we chose the initial position

$$q_0(x) = -\frac{1}{4}|x|^2\ln(-\ln(\rho|x|^2)) + C_1(|x|^2 - 1) + C_2$$

with $\rho = \frac{2}{5}$ and constants C_1 and C_2 such that $q_0 = \Delta q_0 = 0$ holds on $\partial\Omega$. A straightforward calculation shows that $q_0 \in H^3(\Omega)$, see Lemma B.13, and hence it satisfies the conditions on the first component of the product space Z defined in (5.7).

Note however, that there is no $\varepsilon > 0$ such that $q_0 \in H^{3+\varepsilon}(\Omega)$ holds. Indeed, computing the second derivatives, we are left with nice terms, that are in $H^1(\Omega) \cap L^\infty(\Omega)$, but also the critical term

$$p_0 \colon x \mapsto \ln(-\ln(\rho|x|^2)) \in H^1(\Omega) \setminus L^\infty(\Omega)\,, \tag{7.39}$$

which is a well-known function to prove the sharpness of the Sobolev's embedding theorem. We have provided more details in Appendix B.

For the initial value in the second component we take

$$q_0'(x) = -\big(1 - |x|^2\big)^2\,,$$

which is a smooth function, but Δq_0 does not satisfy the homogeneous Dirichlet boundary conditions. In particular, the initial value $u_0 = \big(\, q_0, q_0' \,\big)^T$ is an element of Z.

Space discretization

We performed the space discretization by linear Lagrange finite elements and used the open source Python tool FEniCS [3, version 2018.1.0]. This gives the ansatz space $V_h \subseteq H_0^1(\Omega_h)$, with $\Omega_h \subseteq \Omega$ and we then seek for $q_h(t) \in V_h$ which solves

$$\langle\lambda(q_h(t))q_h''(t), \phi\rangle_{L^2(\Omega_h)} = -\langle q_h(t), \phi\rangle_{H_0^1(\Omega_h)} + \langle\lambda(q_h(t))\mathcal{I}_h\big(\lambda(q_h(t))^{-1}r(t, q_h(t), q_h'(t))\big), \phi\rangle_{L^2(\Omega_h)}$$

for all $\phi \in V_h$, where $\mathcal{I}_h$ denotes the interpolation onto V_h. Testing against a basis, leads to the system of ordinary differential equations

$$M_h(q_h(t))q_h''(t) = -L_h q_h(t) + M_h(q_h(t))g_h(t, q_h(t), q_h'(t)) \tag{7.40}$$

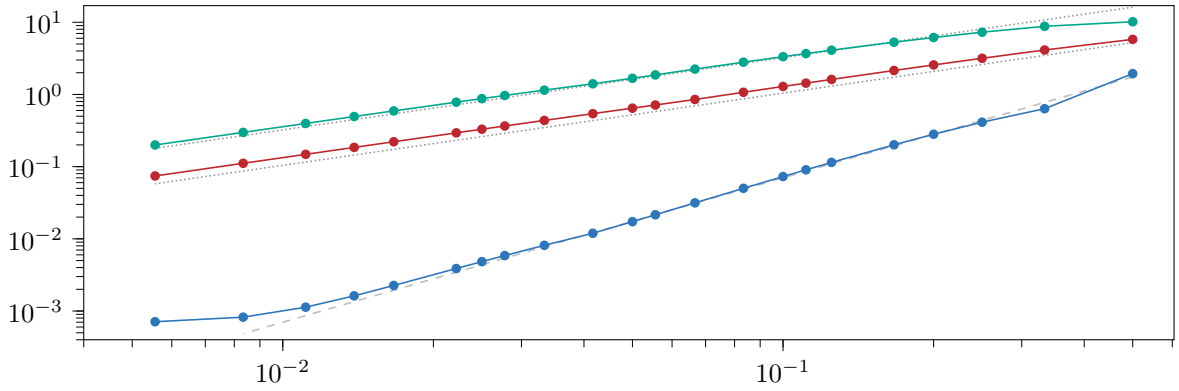

Figure 7.1: Discrete $L^\infty\Big([0,1], H_0^1(\Omega) \times L^2(\Omega)\Big)$ error (on the y-axis) of the numerical solution of (7.38) computed with (7.41) (middle line, red) and (7.42) (lower line, blue) plotted against the stepsize τ (on the x-axis). Further, the discrete $L^\infty\Big([0,1], H_0^1(\Omega)\Big)$ error in the velocity q' computed with (7.41) is shown (upper line, green). The gray lines indicate order one (dotted) and two (dashed).

with the mass and stiffness matrix

$$\big(M_h(q_h(t))\big)_{i,j} = \langle \lambda(q_h(t))\phi_i, \phi_j \rangle_{L^2(\Omega_h)}, \qquad \big(L_h\big)_{i,j} = \langle \nabla\phi_i, \nabla\phi_j \rangle_{L^2(\Omega_h)},$$

and discretized nonlinearity

$$g_h(t, q_h(t)) = \mathcal{I}_h\big(\Lambda(q_h(t))^{-1} r(t, q_h(t), q_h'(t))\big).$$

Time discretization

Recalling the construction of the method, we freeze the argument of the differential operator and the semilinear term in (7.40) either on the last approximation or on the extrapolation to the midpoint. Denoting the fully discrete approximation by $q_h^n \approx q(t_n)$ and $v_h^n \approx q'(t_n)$, we compute the exponential Euler step by solving the linearized version of (7.40)

$$M_h(q_h^n)q_h''(t) = -L_h q_h(t) + M_h(q_h^n)g_h(t_n, q_h^n, v_h^n), \qquad t \in [t_n, t_n + \tau], \tag{7.41}$$

exactly to obtain q_h^{n+1} and v_h^{n+1}, where q_h^n is given from the previous step. We note that this is equivalent to first rewriting (7.40) as a first-order system and then applying the exponential Euler method.

Similarly, we define the extrapolation term $q_h^{n+1/2} = \frac{3}{2}q_h^n - \frac{1}{2}q_h^{n-1}$ and $v_h^{n+1/2} = \frac{3}{2}v_h^n - \frac{1}{2}v_h^{n-1}$ and a step of the exponential midpoint rule is given by the solution of

$$M_h(q_h^{n+1/2})q_h''(t) = -L_h q_h(t) + M_h(q_h^{n+1/2})g_h(t_{n+1/2}, q_h^{n+1/2}, v_h^{n+1/2}). \tag{7.42}$$

The exact solution of these equations is approximated using rational Krylov methods to evaluate the trigonometric matrix functions as it was suggested in [31] and [43]. The code to reproduce the plots is available on https://doi.org/10.5445/IR/1000130189.

Numerical results

Unfortunately, there is no exact solution available to this problem. However, we know by the well-posedness result that the solution is sufficiently regular in order to apply our theorems. We thus used

the midpoint rule on a fine grid with maximal diameter $h_{\text{ref}} = 6 \cdot 10^{-3}$ and stepsize $\tau_{\text{ref}} = \frac{1}{360}$ to obtain a reference solution. On a coarser mesh with maximal diameter $h_{\max} = 10^{-2}$, we computed the approximated solutions of (7.41) and (7.42). The stepsizes τ were chosen such that the quotient $\frac{\tau}{\tau_{\text{ref}}}$ is an integer and hence the reference solution at this time is available.

In Figure 7.1 we depicted the error between the projection of the reference solution and the numerical approximations in the different norms. To compute the X-norm we used the discrete $H_0^1(\Omega) \times L^2(\Omega)$ norm obtained by the mass and stiffness matrix. However, Lagrange finite elements are not contained in $H^2(\Omega)$ such that the full Y-norm cannot be computed. We thus only provide the error in the velocity q' in the $H_0^1(\Omega)$-norm. We included lines that indicate order one and two, and we observe a good alignment with the error bounds shown in Theorems 7.1 and 7.7. The deviation of the last two or three points of the midpoint rule can be explained by the error induced by the space discretization which is only relevant in the regime below 10^{-3}.

7.5 Error bounds in stronger norms

In this final section, we explain how error bounds for the exponential Euler (7.3) and the exponential midpoint rule (7.4) in stronger norms compared to Section 7.1 can be achieved. However, we have used all the regularity provided in Theorem 5.14 and 5.15, and hence need to assume additional regularity of the solution. Note that this cannot be guaranteed by the wellposedness theory considered in this thesis. To this end we introduce the space $Z^A := \{z \in Z \colon Az \in Z\}$ with norm

$$\|z\|_{Z^A}^2 = \|z\|_Z^2 + \|Az\|_Z^2$$

and continuous embedding $Z^A \hookrightarrow Z$. We further assume that the solution of (5.2) satisfies

$$u \in C([0,T], Z^A) \cap C^1([0,T], Z) \tag{7.43}$$

and discuss how this helps to extend our so far obtained results.

Exponential Euler method We first extend Lemma 5.3 by a Lipschitz bound that uses the new space Z^A, see [41, Lemma 3.6].

Lemma 7.18. *Let Assumption 5.1 hold. Then for $\phi, \psi \in \mathcal{B}_Y(R) \cap \mathcal{B}_Z(r)$*

$$\|\mathbf{A}_\phi - \mathbf{A}_\psi\|_{Z \leftarrow Z^A} \leq L_Z \|\phi - \psi\|_Z$$

Proof. This directly follows from (5.16d). □

With this additional Lemma we can immediately prove a bound on the defect in the Z-norm.

Corollary 7.19. *Let Assumptions 5.1, 5.4, and 5.5 hold. Further, consider the solution u which satisfies (7.43) and the numerical approximations $(u_n)_n$ given by (7.3). Then the defect in (7.19) can also be bounded by*

$$\|\delta_n\|_Z \leq \left(C_{\sigma,Z}\, \tau \|e_n\|_Z + C_{\delta,Z}\, \tau^2\right) e^{\tau \omega_2},$$

with constants $C_{\sigma,Z}, C_{\delta,Z} > 0$ that only depend on $\|u'\|_{Z,\infty}$ and $\|u\|_{Z^A,\infty}$.

Proof. As in the proof of Corollary 7.13, we only establish the bounds on the defects. By Lemma 7.8 and 7.18 it holds

$$\|\delta_{n,1}\|_Z \leq \tau c_0 L_Z e^{\tau\omega_2} \|e_n\|_Z \|u\|_{Z^A,\infty} ,$$

and in the same manner with Lemma 5.6

$$\|\delta_{n,2}\|_Z \leq \tau c_0 e^{\tau\omega_2} L_{f,Z} \|e_n\|_Z .$$

The other defects can be bounded with Lemma 7.18 by

$$\|\delta_{n,3}\|_Z \leq \tau^2 c_0 L_Z e^{\tau\omega_2} \|u'\|_{Z,\infty} \|u\|_{Z^A,\infty} ,$$

and similarly by Lemma 5.6

$$\|\delta_{n,4}\|_Z \leq \tau^2 c_0 e^{\tau\omega_2} L_{f,Z} \big(1 + \|u'\|_{Z,\infty}\big) . \qquad \square$$

Analogously to Theorem 7.1 we can derive the first-order error bound in the Z-norm.

Theorem 7.20. *Let u be the solution of (5.1) and u_n the approximation obtained from (7.3). Further, assume that $u \in C([0,T], Z^A) \cap C^1([0,T], Z)$ holds. If Assumptions 5.1, 5.4, and 5.5 are satisfied, we obtain the error bound*

$$\|u(t_n) - u_n\|_Z \leq t_n e^{c_Z t_n} C_Z \tau, \quad 0 \leq n\tau = t_n \leq T$$

with constants $C_Z, c_Z > 0$ that only depend on $\|u'\|_{Z,\infty}$ and $\|u\|_{Z^A,\infty}$, but are independent of τ, n and t_n.

Similar to the observations for Theorem 7.1, comparing our result with Theorem 6.2, we could improve the regularity assumptions using only $\|u'\|_{Z,\infty}$ instead of the L^2-norm

$$\int_0^T \|u''(t)\|_Z^2 \, dt .$$

Exponential midpoint rule In Theorem 7.7 we have shown a second-order error bound in the X-norm and an first-order error bound in the Y-norm. We now improve the result in the Y-norm and study the additional regularity that can be deduced from (7.43).

Corollary 7.21. *If $u \in C^1([0,T], Z) \cap C([0,T], Z^A)$ and in addition Assumptions 7.2 (a) and 7.3 (a) hold, then the solution u of (5.2) satisfies*

$$u \in C^2([0,T], Y) \cap C^1([0,T], Z) \cap C([0,T], Z^A) .$$

Nevertheless, we have to assume additional regularity of the data as well. We again formulate this in assumptions and give the detailed proofs in Appendix B.

Assumption 7.22 (additional properties of g)**.** *Let $u \in C^1([0,T], Z) \cap C([0,T], Z^A)$ and consider the map*

$$t \mapsto \widehat{\mathbf{g}}(t) = g(t, u(t)) .$$

Then there is a constant $C_{g',Z,\infty}$ with

(a) $t \mapsto \widehat{\mathbf{g}}(t) \in C^1([0,T],Z)$, $\quad \|\widehat{\mathbf{g}}'(t)\|_Z \le C_{g',Z,\infty}$,
and, if in addition, $u \in C^2([0,T],Y)$ *holds, then there is* $C_{g'',Y,\infty}$ *such that*
(b) $t \mapsto \widehat{\mathbf{g}}(t) \in C^2([0,T],Y)$, $\quad \|\widehat{\mathbf{g}}''(t)\|_Y \le C_{g'',Y,\infty}$,
with constants only depending on $\|u''\|_{Y,\infty}$, $\|u'\|_{Z,\infty}$, $\|u\|_{Z^A,\infty}$.

We further add assumptions on the differentiability of Λ in some stronger norms.

Assumption 7.23 (additional properties of Λ)**.** *Let* $u \in C^1([0,T],Z) \cap C([0,T],Z^A)$ *and consider the map*

$$t \mapsto \mathbf{\Lambda^{-1}}(t) \coloneqq \Lambda^{-1}(u(t))\,.$$

For $z \in Z$ *it holds*
(a) $t \mapsto \mathbf{\Lambda^{-1}}(t)z \in C^1([0,T],Z)$, $\quad \left\|\left(\mathbf{\Lambda^{-1}}\right)'(t)\right\|_{Z\leftarrow Z} \le C_{ZZ}$,
and, if in addition, $u \in C^2([0,T],Y)$*, it further holds*
(b) $t \mapsto \mathbf{\Lambda^{-1}}(t)z \in C^2([0,T],Y)$, $\quad \left\|\left(\mathbf{\Lambda^{-1}}\right)''(t)\right\|_{Y\leftarrow Z} \le C_{YZ}$
with constants C_{YZ}, C_{ZZ} *only depending on* $\|u''\|_{Y,\infty}$, $\|u'\|_{Z,\infty}$, $\|u\|_{Z^A,\infty}$.

Combining the two preceding assumptions gives us the following stronger differentiability and extends Lemma 7.4.

Lemma 7.24. *Let* $u \in C^1([0,T],Y) \cap C([0,T],Z)$ *and consider the map*

$$t \mapsto \widehat{\mathbf{f}}(t) = f(t,u(t))\,.$$

If Assumptions 7.22 and 7.23 hold, then $\widehat{\mathbf{f}}$ *satisfies Assumption 7.22 with constants* $C_{f',Z,\infty}, C_{f'',Y,\infty}$ *only depending on* $\|u''\|_{Y,\infty}$, $\|u'\|_{Z,\infty}$, $\|u\|_{Z^A,\infty}$.

Further, we easily obtain together with Assumption 5.1 (a) the following differentiability of the differential operator evaluated at a smooth function.

Lemma 7.25. *Let* $u \in C^1([0,T],Z) \cap C([0,T],Z^A)$ *and consider the map*

$$t \mapsto \widehat{\mathbf{A}}(t) = \mathbf{\Lambda^{-1}}(t)A\,.$$

If Assumptions 5.1, 7.3, and 7.23 are satisfied, then for $w \in Z^A$ *it holds*
(a) $t \mapsto \widehat{\mathbf{A}}(t)w$ *is* $C^1([0,T],Z)$, $\quad \left\|\widehat{\mathbf{A}}'(t)\right\|_{Z\leftarrow Z^A} \le C^A_{ZZ}$
and, if in addition, $u \in C^2([0,T],Y)$*, it further holds*
(b) $t \mapsto \widehat{\mathbf{A}}(t)w$ *is* $C^2([0,T],Y)$, $\quad \left\|\widehat{\mathbf{A}}''(t)\right\|_{Y\leftarrow Z^A} \le C^A_{YZ^A}$
with constants $C^A_{ZZ}, C^A_{YZ^A}$ *only depending on* $\|u''\|_{Y,\infty}$, $\|u'\|_{Z,\infty}$, $\|u\|_{Z^A,\infty}$.

With these preparations we can bound the defect of the exponential midpoint rule (7.4) in the Y-norm which will lead to the desired second-order error bound.

Corollary 7.26. *Let Assumptions 5.1, 5.4, 5.5, 7.2, 7.3, 7.22, and 7.23 be satisfied and consider the solution* u *given by Corollary 7.21 and numerical approximations* $(u_n)_n$ *given by (7.4). Then the global error satisfies the error recursion (7.31) where the defect is bounded by*

$$\|\delta_0\|_Y \le C_{\delta,Y}\,\tau^2\,e^{\tau\widehat{\omega}_2}\,,$$
$$\|\delta_n\|_Y \le \left(C_{\sigma,Y}\,\tau\left\|e_{n+1/2}\right\|_Y + C_{\delta,Y}\,\tau^3\right)e^{\tau\widehat{\omega}_2}\,, \qquad n \ge 1\,,$$

with constants $C_{\sigma,Y}, C_{\delta,Y}$ *only depending on* $\|u''\|_Y$, $\|u'\|_Z$, $\|u\|_{Z^A}$.

Proof. (a) By definition we have $u_{1/2} = \widehat{u}_{1/2}$ and hence $\delta_{0,1} = \delta_{0,2} = 0$. Since the first step is given by an exponential Euler step, (7.27) and (7.28) yield

$$\|\delta_{0,3}\|_Y + \|\delta_{0,4}\|_Y \leq C\tau^2 .$$

(b) We now turn to the case $n \geq 1$. The same computation as in (7.25) gives

$$\|\delta_{n,1}\|_Y \leq \tau \widehat{c}_0 \widehat{L}_Y e^{\tau\widehat{\omega}_2} \left\|e_{n+1/2}\right\|_Y \|u\|_{Z,\infty}$$

as well as in (7.26)

$$\|\delta_{n,2}\|_Y \leq \tau \widehat{c}_0 e^{\tau\widehat{\omega}_2} \widehat{L}_{f,Y} \left\|e_{n+1/2}\right\|_Y .$$

The defect $\delta_{n,3}$ can be bounded by Lemma 7.14 and 7.18

$$\|\delta_{n,3}\|_Y \leq \tau^3 \widehat{c}_0 \widehat{L}_Y e^{\tau\widehat{\omega}_2} \tfrac{3}{8} \|u''\|_{Y,\infty} \|u\|_{Z,\infty} ,$$

as well as $\delta_{n,4}$ with Lemma 5.6 by

$$\|\delta_{n,4}\|_Y \leq \tau^3 \widehat{c}_0 e^{\tau\widehat{\omega}_2} \widehat{L}_{f,Y} \tfrac{3}{8} \|u''\|_{Y,\infty} ,$$

where we used Taylor expansion on $u(t_{n+1/2})$ for both defects, see Lemma B.11.

(c) Since the representation in (7.33) is still valid and Lemma 7.5 (b) and 7.25 implies the bounds

$$\left\|\dot{d}_{n,1}(t_{n+1/2})\right\|_Z = \left\|\widehat{\mathbf{A}}'(t_{n+1/2})\widehat{u}_{n+1/2}\right\|_Z \leq C^A_{ZZ} \|u\|_{Z^A,\infty}$$

and

$$\begin{aligned}\left\|\ddot{d}_{n,1}(t)\right\|_Y &\leq \left\|\widehat{\mathbf{A}}''(t)u(t)\right\|_Y + \left\|\widehat{\mathbf{A}}'(t)u'(t)\right\|_Y \\ &\leq C^A_{YZ^A} \|u\|_{Z^A,\infty} + C^A_{YZ} \|u'\|_{Z,\infty} ,\end{aligned}$$

as well as

$$\left\|\dot{d}_{n,2}(t)\right\|_Y = \left\|\big(\widehat{\mathbf{A}}(t) - \widehat{\mathbf{A}}_{n+1/2}\big)u'(t)\right\|_Y \leq \tfrac{\tau}{2} C^A_{YZ} \|u'\|_{Z,\infty} ,$$

the bound for $\|\delta_{n,5}\|_Y$ is derived as before.

(d) Analogously to (7.37) we can establish with Lemma 7.24

$$\begin{aligned}\left\|\dot{D}_{n,1}\right\|_Z &= \left\|\widehat{\mathbf{f}}'(t)\right\|_Z \leq C_{f',Z,\infty} , \\ \left\|\ddot{D}_{n,1}\right\|_Y &= \left\|\widehat{\mathbf{f}}''(t)\right\|_Y \leq C_{f'',Y,\infty} ,\end{aligned}$$

which then provides the bound for $\|\delta_{n,6}\|_Y$. □

Along the lines of Theorem 7.7 we deduce the global error in the Y-norm.

Theorem 7.27. *Let u be the solution of (5.1) and u_n the approximation obtained from (7.4). Further, assume that $u \in C([0,T], Z^A) \cap C^1([0,T], Z)$. If Assumptions 5.1, 5.4, and 5.5, are satisfied, and in addition Assumptions 7.2, 7.3, 7.22, and 7.23 hold true, and τ_0 is given by (7.6), then for all $\tau \leq \tau_0$ the error is bounded by*

$$\|u(t_n) - u_n\|_Y \leq t_n e^{c_Y t_n} C_Y \tau^2, \quad 0 \leq n\tau = t_n \leq \min\{T, \widehat{T}_{mid}\} ,$$

with constants $C_Y, c_Y > 0$ that only depend on $\|u''\|_Y$, $\|u'\|_Z$, $\|u\|_{Z^A}$, but are independent of τ, n and t_n.

APPENDIX B

Quasilinear examples

In this part of the Appendix we will comment on the assumptions made for the error analysis. Some of them have been used and verified before in [41, 44, 61] and we will give the references. This concerns in particular the assumptions on the operator Λ and A. Moreover, we only check the assumptions for Maxwell's equations in the full space case (5.13) since the case (5.15) is fully analogous.

B.1 Assumptions on Λ

We first comment on Assumption 5.1. The detailed computation in order to verify the assumption are given in the proof of [61, Thm. 4.6 & 4.9] for Maxwell's equations and in [61, Thm. 4.12] for the wave equation. In particular, properties (5.16a), (5.16c), (5.16b) and (5.16d) for the triple $\big(Y, Y, \mathcal{B}_Y(R)\big)$ are proven. The remaining two cases are derived fully analogously and we omit the details.

B.2 Kato's commutator condition

A crucial tool for the wellposedness is Assumption 5.4. The discussion of this property is beyond the scope of this thesis. The assertions are verified in the proof of [61, Thm. 4.6 & 4.9] for Maxwell's equations and in [61, Thm. 4.12] for the wave equation.

B.3 Lipschitz assumptions on the semilinear term

We now turn to Assumption 5.5. Since this term has not appeared in the previous works in this framework we prove all details here. Throughout we use the norms defined in (A.2). We consider the two examples separately.

Wave equation

We recall from (5.6) that g is given by

$$g(t,u) = \begin{pmatrix} 0 \\ \gamma_1(q,q') + \gamma_2(t,q,q') \end{pmatrix}, \qquad \gamma_1(q,q') = -K''(q)\big(q'\big)^2, \qquad \gamma_2(t,q,q') = r(t,q,q'), \tag{B.1}$$

and state the following lemma.

Lemma B.1. *For g given in (B.1) with the regularity given in (5.5) Assumption 5.5 is satisfied.*

Proof. It is sufficient to prove for $u_1 = \big(q_1, q_1'\big)^T$, $u_2 = \big(q_2, q_2'\big)^T$ there are constants C such that for $u_1, u_2 \in \mathcal{B}_Z(r)$, i.e., $q_i \in H^3(\Omega)$, $q_i' \in H^2(\Omega)$, and $t, s \in [0,T]$:

$$\|\gamma_i(t,q_1,q_1') - \gamma_i(t,q_2,q_2')\|_V \le L_{g,V}\big(|t-s| + \|u_1 - u_2\|_W\big)\,,$$

with the tuples

$$(V,W) \in \Big\{\big(L^2,X\big), \big(H^1(\Omega),Y\big), \big(H^2(\Omega),Z\big),\Big\}$$

$\underline{i=1}$: We write

$$\begin{aligned}\gamma_1(q_1,q_1') - \gamma_1(q_2,q_2') &= K''(q_2)\big(q_2'\big)^2 - K''(q_1)\big(q_1'\big)^2 \\ &= \big(K''(q_2) - K''(q_1)\big)\big(q_2'\big)^2 - K''(q_1)\big(q_2' + q_1'\big)\big(q_2' - q_1'\big)\,.\end{aligned}$$

(a) In the L^2-norm we have by (A.4)

$$\|\gamma_1(q_1,q_1') - \gamma_1(q_2,q_2')\|_{L^2} \le \|K'''\|_{L^\infty}\,\|q_1' - q_2'\|_{L^2}\,\|q_2'\|_{H^2} + \|K''\|_{L^\infty}\,\|q_1' + q_2'\|_{H^2}\,\|q_1' - q_2'\|_{L^2}\,.$$

(b) In the H^1-norm we have by (A.9) and (A.5)

$$\begin{aligned}\|\gamma_1(q_1,q_1') - \gamma_1(q_2,q_2')\|_{H^1} &\le \Big\|\big(K''(q_2) - K''(q_1)\big)\big(q_2'\big)^2\Big\|_{H^1} + \big\|K''(q_1)\big(q_2' + q_1'\big)\big(q_2' - q_1'\big)\big\|_{H^1} \\ &\le C\big(\|q_1\|_{H^2}, \|q_2\|_{H^2}\big)\,\|q_2 - q_1\|_{H^1}\,\|q_2'\|_{H^2}^2 \\ &\quad + C\big(\|q_1\|_{H^2}, \|q_2\|_{H^2}\big)\,\|q_2' - q_1'\|_{H^1}\,.\end{aligned}$$

(c) In the H^2-norm we have with (A.11) and (A.12)

$$\begin{aligned}\|\gamma_1(q_1,q_1') - \gamma_1(q_2,q_2')\|_{H^2} &\le C\big(\|q_1\|_{H^2}, \|q_2\|_{H^2}\big)\,\|q_1' - q_2'\|_{H^2}\,\|q_2'\|_{H^2}^2 \\ &\quad + C\big(\|q_1\|_{H^2}\big)\,\|q_1' + q_2'\|_{H^2}\,\|q_1' - q_2'\|_{H^2}\,.\end{aligned}$$

$\underline{i=2}$: We write

$$\begin{aligned}\gamma_2(t,q_1,q_1') - \gamma_2(s,q_2,q_2') &= r(t,q_1,q_1') - r(s,q_2,q_2') \\ &= r(t,q_1,q_1') - r(s,q_1,q_1') + r(s,q_1,q_1') - r(s,q_2,q_1') + r(t,q_2,q_1') - r(s,q_2,q_2') \\ &= \Delta_{r,1} + \Delta_{r,2} + \Delta_{r,3}\end{aligned}$$

(a) In the L^2-norm we have by (A.8)

$$\begin{aligned}\|\Delta_{r,1}\|_{L^2} &\le C\big(\|q_1\|_{L^\infty}, \|q_1'\|_{L^\infty}\big)|t-s|\,, \\ \|\Delta_{r,2}\|_{L^2} &\le C\big(\|q_1\|_{L^\infty}, \|q_2\|_{L^\infty}, \|q_1'\|_{L^\infty}\big)\,\|q_1 - q_2\|_{L^2}\,, \\ \|\Delta_{r,3}\|_{L^2} &\le C\big(\|q_1\|_{L^\infty}, \|q_1'\|_{L^\infty}, \|q_2'\|_{L^\infty}\big)\,\|q_1' - q_2'\|_{L^2}\,.\end{aligned}$$

(b) In the H^1-norm we have with (A.9) and (A.10)

$$\begin{aligned}
\|\Delta_{r,1}\|_{H^1} &\leq C\big(\|q_1\|_{H^2}, \|q_1'\|_{H^2}\big)|t-s|\,,\\
\|\Delta_{r,2}\|_{H^1} &\leq C\big(\|q_1\|_{H^2}, \|q_2\|_{H^2}, \|q_1'\|_{H^2}\big)\,\|q_1-q_2\|_{H^1}\,,\\
\|\Delta_{r,3}\|_{H^1} &\leq C\big(\|q_2\|_{H^2}, \|q_1'\|_{H^2}, \|q_2'\|_{H^2}\big)\,\|q_1'-q_2'\|_{H^1}\,.
\end{aligned}$$

(c) In the H^2-norm we have with (A.11) and (A.12)

$$\begin{aligned}
\|\Delta_{r,1}\|_{H^2} &\leq C\big(\|q_1\|_{H^2}, \|q_1'\|_{H^2}\big)|t-s|\,,\\
\|\Delta_{r,2}\|_{H^2} &\leq C\big(\|q_1\|_{H^2}, \|q_2\|_{H^2}, \|q_1'\|_{H^2}\big)\,\|q_1-q_2\|_{H^2}\,,\\
\|\Delta_{r,3}\|_{H^2} &\leq C\big(\|q_2\|_{H^2}, \|q_1'\|_{H^2}, \|q_2'\|_{H^2}\big)\,\|q_1'-q_2'\|_{H^1}\,.
\end{aligned}$$

□

Maxwell's equations

We recall from (5.11) that g is given by

$$g(t,u) = \begin{pmatrix} -\sigma(E)E \\ 0 \end{pmatrix} \tag{B.2}$$

and state the following lemma.

Lemma B.2. *For g given in (B.2) with the regularity given in (5.12) Assumption 5.5 is satisfied.*

Proof. It is sufficient to proof for $E_1, E_2 \in H^3(\mathbb{R}^3)$ that there are constants C such that

$$\|\sigma(E_1)E_1 - \sigma(E_2)E_2\|_V \leq C\,\|E_1-E_2\|_V\,,$$

with $V \in \Big\{L^2(\mathbb{R}^3), H^2(\mathbb{R}^3), H^3(\mathbb{R}^3)\Big\}$. We use the representation

$$\sigma(E_1)E_1 - \sigma(E_2)E_2 = \sigma(E_1)\big(E_1-E_2\big) + \big(\sigma(E_1)E - \sigma(E_2)\big)E_2\,.$$

(a) In the L^2-norm we have by (A.1)

$$\begin{aligned}
\|\sigma(E_1)E_1 - \sigma(E_2)E_2\|_{L^2} &\leq \|\sigma(E_1)\|_{L^\infty}\,\|E_1-E_2\|_{L^2} + \|\sigma(E_1)-\sigma(E_2)\|_{L^2}\,\|E_2\|_{L^\infty}\\
&\leq C\big(\|E_1\|_{H^2}, \|E_2\|_{H^2},\big)\,\|E_1-E_2\|_{L^2}\,.
\end{aligned}$$

(b) In the H^2-norm we have by (A.6), (A.11), and (A.12)

$$\begin{aligned}
\|\sigma(E_1)E_1 - \sigma(E_2)E_2\|_{H^2} &\leq \|\sigma(E_1)\|_{H^2}\,\|E_1-E_2\|_{H^2} + \|\sigma(E_1)-\sigma(E_2)\|_{H^2}\,\|E_2\|_{H^2}\\
&\leq C\big(\|E_1\|_{H^2}, \|E_2\|_{H^2},\big)\,\|E_1-E_2\|_{H^2}\,.
\end{aligned}$$

(c) In the H^3-norm we have

$$\begin{aligned}
\|\sigma(E_1)E_1 - \sigma(E_2)E_2\|_{H^3} &\leq \|\sigma(E_1)\|_{H^3}\,\|E_1-E_2\|_{H^3} + \|\sigma(E_1)-\sigma(E_2)\|_{H^3}\,\|E_2\|_{H^3}\\
&\leq C\big(\|E_1\|_{H^3}, \|E_2\|_{H^3},\big)\,\|E_1-E_2\|_{H^3}\,.
\end{aligned}$$

□

B.4 Differentiability of the semilinear term

In this section we discuss Assumptions 7.2 and 7.22. We restrict ourselves to the boundedness of the formally obtained derivatives. However, we note that continuity is achieved by the precise same computations and in (A.20), (A.22), and (A.24) we have shown how to conclude differentiability from this.

Further we introduce the notation

$$C^m(H^k) \coloneqq C^m([0,T], H^k(\Omega)), \qquad k, m \geq 0\,.$$

Wave equation

We recall g, γ_1 and γ_2 from (B.1), $\widehat{\mathbf{g}}$ from (7.5) and set

$$\widehat{\mathbf{g}}(t) = \begin{pmatrix} 0 \\ \widehat{\gamma}_1(t) + \widehat{\gamma}_2(t) \end{pmatrix}, \qquad \widehat{\gamma}_1(t) = \gamma_1(q(t), q'(t)), \quad \widehat{\gamma}_2(t) = \gamma_2(t, q(t), q'(t))\,. \tag{B.3}$$

Lemma B.3. *For $\widehat{\mathbf{g}}$ given in (B.3) with the regularity given in (5.5) Assumption 7.2 is satisfied.*

Proof. It is sufficient to prove for $q \in C^2(H^1) \cap C^1(H^2) \cap C(H^3)$ that for $i = 1, 2$

(a) $t \mapsto \widehat{\gamma}_i(t)$ is $C^1([0,T], H^1(\Omega))$,

(b) $t \mapsto \widehat{\gamma}_i(t)$ is $C^2([0,T], L^2(\Omega))$.

$\underline{i = 1:}$ We first compute

$$\begin{aligned} -\partial_t \widehat{\gamma}_1(t) &= K'''(q)\big(q'\big)^3 + 2K''(q)q'q'' \\ -\partial_t^2 \widehat{\gamma}_1(t) &= K^{(4)}(q)\big(q'\big)^4 + K'''(q)3\big(q'\big)^2 q'' \\ &\quad + 2K'''(q)\big(q'\big)^2 q'' + 2K''(q)\big(q''\big)^2 + 2K''(q)q'q''' \end{aligned} \tag{B.4}$$

(a) We have for $q \in C^2(H^1) \cap C^1(H^2) \cap C(H^3)$ by (A.5)

$$\begin{aligned} \big\|\partial_t \widehat{\gamma}_1(t)\big\|_{H^1} &\leq \big\|K'''(q)\big(q'\big)^3\big\|_{H^1} + \|2K''(q)q'q''\|_{H^1} \\ &\leq C\,\|K'''(q)\|_{H^1}\,\|q'\|_{H^2}^3 + C\,\|2K''(q)q'\|_{H^2}\,\|q''\|_{H^1} \\ &\leq C\big(\|q\|_{H^2}\big)\big(\|q'\|_{H^2}^3 + \|q'\|_{H^2}\,\|q''\|_{H^1}\big) \end{aligned}$$

(b) If in addition $q \in C^3(L^2)$, it holds by (A.1)

$$\begin{aligned} \big\|\partial_t^2 \widehat{\gamma}_1(t)\big\|_{L^2} &\leq \big\|K^{(4)}(q)\big(q'\big)^4\big\|_{L^2} + 5\big\|K'''(q)\big(q'\big)^2 q''\big\|_{L^2} \\ &\quad + \big\|2K''(q)\big(q''\big)^2\big\|_{L^2} + \big\|2K''(q)q'q'''\big\|_{L^2} \\ &\leq C\big(\|q\|_{H^2}\big)\big(\|q'\|_{H^2}^4 + \|q'\|_{H^2}^2\,\|q''\|_{H^1} + \|q''\|_{H^1}^2 + \|q'\|_{H^2}\,\|q'''\|_{L^2}\big) \end{aligned}$$

$\underline{i = 2:}$ We first compute

$$\begin{aligned} \partial_t \widehat{\gamma}_2(t) &= \partial_t r(\cdot) + \partial_q r(\cdot)q' + \partial_{q'} r(\cdot)q'' \\ \partial_t^2 \widehat{\gamma}_2(t) &= \partial_{t,t} r(\cdot) + \partial_q r(\cdot)q'' + \partial_{q'} r(\cdot)q''' \\ &\quad + 2\partial_{t,q} r(\cdot)q' + 2\partial_{t,q'} r(\cdot)q'' \\ &\quad + \partial_{q,q} r(\cdot)\big(q'\big)^2 + 2\partial_{q,q'} r(\cdot)q'q'' + \partial_{q',q'} r(\cdot)\big(q''\big)^2 \end{aligned} \tag{B.5}$$

(a) We have for $q \in C^2(H^1) \cap C^1(H^2) \cap C(H^3)$ by (A.5), (A.9), and (A.11)

$$\begin{aligned}\|\partial_t \widehat{\gamma}_2(t)\|_{H^1} &\le \|\partial_t r(\cdot)\|_{H^1} + \|\partial_q r(\cdot) q'\|_{H^1} + \|\partial_{q'} r(\cdot) q''\|_{H^1} \\ &\le C\big(\|q\|_{H^2}, \|q'\|_{H^2}\big)\big(1 + \|q'\|_{H^2} + \|q''\|_{H^1}\big)\,.\end{aligned}$$

(b) If in addition $q \in C^3(L^2)$, it holds by (A.3), (A.7), (A.9), and (A.11)

$$\begin{aligned}\big\|\partial_t^2 \widehat{\gamma}_2(t)\big\|_{L^2} &\le \|\partial_{t,t} r(\cdot) + \partial_q r(\cdot) q'' + \partial_{q'} r(\cdot) q'''\|_{L^2} \\ &\quad + 2\,\|\partial_{t,q} r(\cdot) q' + \partial_{t,q'} r(\cdot) q''\|_{L^2} \\ &\quad + \big\|\partial_{q,q} r(\cdot)\big(q'\big)^2 + 2\partial_{q,q'} r(\cdot) q' q'' + \partial_{q',q'} r(\cdot)\big(q''\big)^2\big\|_{L^2} \\ &\le C\big(\|q\|_{H^2}, \|q'\|_{H^2}\big)\big(1 + \|q''\|_{H^1} + \|q'''\|_{L^2} + \|q'\|_{L^2} + \|q''\|_{L^2} \\ &\qquad + \|q'\|_{H^2}^2 + \|q''\|_{H^1}\,\|q'\|_{H^2} + \|q''\|_{H^1}^2\big)\,.\end{aligned}$$

□

Lemma B.4. *For $\widehat{\mathbf{g}}$ given in (B.3) with the regularity given in (5.5) and with $K \in C^6(\mathbb{R})$, Assumption 7.22 is satisfied.*

Proof. It is sufficient to prove for $q \in C^2(H^2)$ that

(a) $t \mapsto \widehat{\gamma}_i(t)$ is $C^1([0,T], H^2(\Omega))$,

(b) $t \mapsto \widehat{\gamma}_i(t)$ is $C^2([0,T], H^1(\Omega))$.

$\underline{i = 1:}$ We use the representation in (B.4) for the following computations.

(a) We have for $q \in C^2(H^2)$ by (A.6)

$$\begin{aligned}\|\partial_t \widehat{\gamma}_1(t)\|_{H^2} &\le \big\|K'''(q)\big(q'\big)^3\big\|_{H^2} + \|2K''(q) q' q''\|_{H^2} \\ &\le C\big(\|q\|_{H^2}, \|q'\|_{H^2}\big)\big(\|q'\|_{H^2}^3 + \|q'\|_{H^2}\,\|q''\|_{H^2}\big)\,.\end{aligned}$$

(b) If in addition $q \in C^3(H^1)$, it holds with (A.5)

$$\begin{aligned}\big\|\partial_t^2 \widehat{\gamma}_1(t)\big\|_{H^1} &\le \big\|K^{(4)}(q)\big(q'\big)^4\big\|_{H^1} + 5\big\|K'''(q)\big(q'\big)^2 q''\big\|_{H^1} + \big\|2K''(q)\big(q''\big)^2\big\|_{H^1} + \big\|2K''(q) q' q'''\big\|_{H^1} \\ &\le C\big(\|q\|_{H^2}, \|q'\|_{H^2}\big)\big(\|q'\|_{H^2}^4 + \|q'\|_{H^2}^2\,\|q''\|_{H^2} + \|q''\|_{H^2}^2 + \|q'\|_{H^2}\,\|q'''\|_{H^1}\big)\,.\end{aligned}$$

$\underline{i = 2:}$ We use the representation in (B.5) for the following computations.

(a) We have for $q \in C^2(H^2)$ by (A.6)

$$\begin{aligned}\|\partial_t \widehat{\gamma}_2(t)\|_{H^2} &\le \|\partial_t r(\cdot)\|_{H^2} + \|\partial_q r(\cdot) q'\|_{H^2} + \|\partial_{q'} r(\cdot) q''\|_{H^2} \\ &\le C\big(\|q\|_{H^2}, \|q'\|_{H^2}\big)\big(1 + \|q'\|_{H^2} + \|q''\|_{H^2}\big)\,.\end{aligned}$$

(b) If in addition $q \in C^3(H^1)$, it holds by (A.5)

$$\begin{aligned}\big\|\partial_t^2 \widehat{\gamma}_2(t)\big\|_{H^1} &\le \|\partial_{t,t} r(\cdot) + \partial_q r(\cdot) q'' + \partial_{q'} r(\cdot) q'''\|_{H^1} \\ &\quad + 2\,\|\partial_{t,q} r(\cdot) q' + \partial_{t,q'} r(\cdot) q''\|_{H^1} \\ &\quad + \big\|\partial_{q,q} r(\cdot)\big(q'\big)^2 + 2\partial_{q,q'} r(\cdot) q' q'' + \partial_{q',q'} r(\cdot)\big(q''\big)^2\big\|_{H^1} \\ &\le C\big(\|q\|_{H^2}, \|q'\|_{H^2}\big)\big(1 + \|q''\|_{H^2} + \|q'''\|_{H^1} + \|q'\|_{L^2} + \|q''\|_{H^2} \\ &\qquad + \|q'\|_{H^2}^2 + \|q''\|_{H^2}\,\|q'\|_{H^2} + \|q''\|_{H^2}^2\big)\,.\end{aligned}$$

□

Maxwell's equations

We recall g from (B.2), $\widehat{\mathbf{g}}$ from (7.5) and set

$$\widehat{\mathbf{g}}(t) = \begin{pmatrix} -\widehat{\sigma}(t) \\ 0 \end{pmatrix}, \qquad \widehat{\sigma}(t) = \sigma(E(t))E(t) \tag{B.6}$$

Lemma B.5. *For $\widehat{\mathbf{g}}$ given in (B.6) with the regularity given in (5.12) Assumption 7.2 is satisfied.*

Proof. It is sufficient to prove for $E \in C^1(H^2) \cap C(H^3)$

(a) $t \mapsto \widehat{\sigma}(t)$ is $C^1([0,T], H^2(\mathbb{R}^3))$,

(b) $t \mapsto \widehat{\sigma}(t)$ is $C^2([0,T], L^2(\mathbb{R}^3))$.

(a) We have for $E \in C^1(H^2) \cap C(H^3)$ by (A.6)

$$\begin{aligned} \|\widehat{\sigma}'(t)\|_{H^2} &\leq \|\sigma'(E(t))[E'(t), E(t)]\|_{H^2} + \|\sigma(E(t))E'(t)\|_{H^2} \\ &\leq C\big(\|E\|_{H^2}, \|E'\|_{H^2}\big). \end{aligned}$$

(b) If in addition $E \in C^2(L^2)$, it holds by (A.4)

$$\begin{aligned} \|\widehat{\sigma}''(t)\|_{L^2} &\leq \|\sigma''(E(t))[E'(t), E'(t), E(t)]\|_{L^2} + 2\,\|\sigma(E(t))[E'(t), E'(t)]\|_{L^2} + \|\sigma(E(t))E''(t)\|_{L^2} \\ &\leq C\big(\|E\|_{H^2}, \|E'\|_{H^2}, \|E''\|_{L^2}\big). \end{aligned}$$

□

Lemma B.6. *For $\widehat{\mathbf{g}}$ given in (B.6) with $\sigma \in C^4(\mathbb{R}^3, \mathbb{R}^{3,3})$ Assumption 7.22 is satisfied.*

Proof. It is sufficient to prove for $E \in C^1(H^3)$

(a) $t \mapsto \widehat{\sigma}(t)$ is $C^1([0,T], H^3(\mathbb{R}^3))$,

(b) $t \mapsto \widehat{\sigma}(t)$ is $C^2([0,T], H^2(\mathbb{R}^3))$.

(a) We have for $E \in C^1(H^3)$ by the algebra property of $H^3(\mathbb{R}^3)$

$$\begin{aligned} \|\widehat{\sigma}'(t)\|_{H^3} &\leq \|\sigma'(E(t))[E'(t), E(t)]\|_{H^3} + \|\sigma(E(t))E'(t)\|_{H^3} \\ &\leq C\big(\|E\|_{H^3}, \|E'\|_{H^3}\big). \end{aligned}$$

(b) If in addition $E \in C^2(H^2)$, it holds by (A.6)

$$\begin{aligned} \|\widehat{\sigma}''(t)\|_{H^2} &\leq \|\sigma''(E(t))[E'(t), E'(t), E(t)]\|_{H^2} + 2\,\|\sigma(E(t))[E'(t), E'(t)]\|_{H^2} + \|\sigma(E(t))E''(t)\|_{H^2} \\ &\leq C\big(\|E\|_{H^2}, \|E'\|_{H^2}, \|E''\|_{H^2}\big). \end{aligned}$$

□

B.5 Differentiability of the quasilinear term

In this section we discuss Assumptions 7.3 and 7.23 and again restrict ourselves to the boundedness of the derivatives.

Wave equation

We recall the map

$$t \mapsto \mathbf{\Lambda^{-1}}(t) = \Lambda^{-1}(u(t)), \quad u \in C^1([0,T],Y) \cap C([0,T],Z)$$

and define

$$\mathbf{\Lambda^{-1}}(t) = \begin{pmatrix} I & 0 \\ 0 & \frac{1}{1+K'(q(t))} \end{pmatrix} =: \begin{pmatrix} I & 0 \\ 0 & \lambda(q(t)) \end{pmatrix} \tag{B.7}$$

Lemma B.7. *For $\mathbf{\Lambda^{-1}}$ given in (B.7) with the regularity given in (5.5) Assumption 7.3 is satisfied.*

Proof. It suffices to prove for $x \in L^2(\Omega)$ and $y \in H^1(\Omega)$ and $q \in C^2(H^1) \cap C^1(H^2) \cap C(H^3)$ that

(a) $t \mapsto \lambda(q(t))x$ is $C^1([0,T],L^2(\Omega))$

(b) $t \mapsto \lambda(q(t))y$ is $C^1([0,T],H^1(\Omega))$

(c) $t \mapsto \lambda(q(t))y$ is $C^2([0,T],L^2(\Omega))$.

We obtain the the expressions

$$\begin{aligned} \partial_t \lambda(q(t))y &= \lambda'(q(t))q'(t)y \\ \partial_t^2 \lambda(q(t))x &= \lambda''(q(t))\big(q'(t)\big)^2 x + \lambda'(q(t))q''(t)x\,. \end{aligned} \tag{B.8}$$

(a) We obtain the bounds by (A.1)

$$\|\partial_t \lambda(q(t))x\|_{L^2} \leq \|\lambda'(q(t))q'(t)y\|_{L^2} \leq \|\lambda'(q(t))\|_{L^\infty} \|q'\|_{H^2} \|x\|_{L^2} \leq C\big(\|q\|_{H^2}\big) \|q'\|_{H^2} \|x\|_{L^2}\,.$$

(b) By (A.5) we have

$$\|\partial_t \lambda(q(t))y\|_{H^1} \leq \|\lambda'(q(t))q'(t)y\|_{H^1} \leq C\big(\|q\|_{H^2}\big) \|q'\|_{H^2} \|y\|_{H^1}\,.$$

(c) Further, by (A.3) and (A.5) it holds

$$\begin{aligned} \big\|\partial_t^2 \lambda(q(t))y\big\|_{L^2} &\leq \big\|\lambda''(q(t))\big(q'(t)\big)^2 y\big\|_{L^2} + \|\lambda'(q(t))q''(t)y\|_{L^2} \\ &\leq C\big(\|q\|_{H^2}\big)\big(\|q'\|_{H^2}^2 + \|q''\|_{H^1}\big) \|y\|_{H^1}\,. \end{aligned}$$

□

Lemma B.8. *For $\mathbf{\Lambda^{-1}}$ given in (B.7) with the regularity given in (5.5) Assumption 7.23 is satisfied.*

Proof. It suffices to prove for $z \in H^2(\Omega)$ and $q \in C^2(H^1) \cap C^1(H^2) \cap C(H^3)$ that

(a) $t \mapsto \lambda(q(t))z$ is $C^1([0,T],H^2(\Omega))$

(b) $t \mapsto \lambda(q(t))z$ is $C^2([0,T],H^1(\Omega))$.

(a) With the the expressions in (B.8) we compute using (A.6)

$$\|\partial_t \lambda(q(t))z\|_{H^2} \leq \|\lambda'(q(t))q'(t)y\|_{H^2} \leq C\big(\|q\|_{H^2}\big) \|q'\|_{H^2} \|z\|_{H^2}\,.$$

(b) By (A.5) and (A.11) it holds

$$\begin{aligned} \big\|\partial_t^2 \lambda(q(t))z\big\|_{H^1} &\leq \big\|\lambda''(q(t))\big(q'(t)\big)^2 y\big\|_{H^1} + \|\lambda'(q(t))q''(t)y\|_{H^1} \\ &\leq C\big(\|q\|_{H^2}\big)\big(\|q'\|_{H^2}^2 + \|q''\|_{H^1}\big) \|y\|_{H^2}\,. \end{aligned}$$

□

Maxwell's equations

First note that

$$\mathbf{\Lambda^{-1}}(t) = \begin{pmatrix} \frac{1}{1+P'(E(t))} & 0 \\ 0 & \frac{1}{1+M'(H(t))} \end{pmatrix} =: \begin{pmatrix} \lambda_E(E(t)) & 0 \\ 0 & \lambda_H(H(t)) \end{pmatrix} \tag{B.9}$$

Lemma B.9. *For $\mathbf{\Lambda^{-1}}$ given in (B.9) with the regularity given in (5.12) Assumption 7.3 is satisfied.*

Proof. It suffices to prove for $x \in L^2(\mathbb{R}^3)$ and $y \in H_2(\mathbb{R}^3)$ to show for $E \in C^1(H^2) \cap C(H^3)$

(a) $t \mapsto \lambda_E(E(t))x$ is $C^1([0,T], L^2(\mathbb{R}^3))$

(b) $t \mapsto \lambda_E(E(t))y$ is $C^1([0,T], H^2(\mathbb{R}^3))$

and, if in addition, $E \in C^2([0,T], L^2(\mathbb{R}^3))$ holds, then

(c) $t \mapsto \lambda_E(E(t))y$ is $C^2([0,T], L^2(\mathbb{R}^3))$.

We differentiate to get

$$\partial_t \lambda_E(E(t))y = \lambda_E'(E(t))\big[E'(t), y\big] \tag{B.10}$$

(a) We obtain by (A.4) the bound

$$\|\partial_t \lambda_E(E(t))x\|_{L^2} = \big\|\lambda_E'(E(t))\big[E'(t), x\big]\big\|_{L^2} \le C\big(\|E\|_{H^2}\big)\,\|E'\|_{H^2}\,\|x\|_{L^2} \ ,$$

(b) Similar by (A.6) we have

$$\|\partial_t \lambda_E(E(t))y\|_{H^2} \le \big\|\lambda_E'(E(t))\big[E'(t), x\big]\big\|_{H^2} \le C\big(\|E\|_{H^2}\big)\,\|E'\|_{H^2}\,\|y\|_{H^2} \ .$$

(c) If in addition, we assume $E \in C^2(L^2)$, we get

$$\partial_t^2 \lambda_E(E(t))y = \lambda_E''(E(t))\big[E'(t), E'(t), y\big] + \lambda_E'(E(t))\big[E''(t), y\big] \,. \tag{B.11}$$

and by (A.4)

$$\begin{aligned} \big\|\partial_t^2 \lambda_E(E(t))y\big\|_{L^2} &\le \big\|\lambda_E''(E(t))\big[E'(t), E'(t), y\big]\big\|_{L^2} + \big\|\lambda_E'(E(t))\big[E''(t), y\big]\big\|_{L^2} \\ &\le C\big(\|E\|_{H^2}\big)\big(\|E'\|_{H^2}^2 + \|E''\|_{L^2}\big)\,\|y\|_{H^2} \ . \end{aligned}$$

□

Lemma B.10. *For $\mathbf{\Lambda^{-1}}$ given in (B.9) with $P, M \in C^5(\mathbb{R}^3, \mathbb{R}^{3,3})$ Assumption 7.23 is satisfied.*

Proof. It suffices to prove for $z \in H^3(\mathbb{R}^3)$ to show for $E \in C^1(H^3)$

(a) $t \mapsto \lambda_E(E(t))z$ is $C^1([0,T], H^3(\mathbb{R}^3))$

and, if in addition, $E \in C^2([0,T], L^2(\mathbb{R}^3))$ holds, then

(b) $t \mapsto \lambda_E(E(t))y$ is $C^2([0,T], H^2(\mathbb{R}^3))$.

(a) With (B.10) we obtain the bound

$$\|\partial_t \lambda_E(E(t))y\|_{H^3} \le \big\|\lambda_E'(E(t))\big[E'(t), x\big]\big\|_{H^3} \le C\big(\|E\|_{H^3}\big)\,\|E'\|_{H^3}\,\|y\|_{H^3} \ .$$

(b) If we assume in addition $E \in C^2(H^2)$, we get with (B.11)

$$\begin{aligned} \big\|\partial_t^2 \lambda_E(E(t))y\big\|_{H^2} &\le \big\|\lambda_E''(E(t))\big[E'(t), E'(t), y\big]\big\|_{H^2} + \big\|\lambda_E'(E(t))\big[E''(t), y\big]\big\|_{H^2} \\ &\le C\big(\|E\|_{H^2}\big)\big(\|E'\|_{H^2}^2 + \|E''\|_{H^2}\big)\,\|y\|_{H^2} \ . \end{aligned}$$

□

B.6 Miscellaneous

Bounds on the extrapolation error

Lemma B.11. *The following estimates hold for $u \in C^2([0,T],V)$:*

$$\begin{aligned}\left\|u(t_{n+1/2}) - \tfrac{1}{2}\left(3u(t_n) - u(t_{n-1})\right)\right\|_V &\le \tau \left\|u'\right\|_{V,\infty} ,\\ \left\|u(t_{n+1/2}) - \tfrac{1}{2}\left(3u(t_n) - u(t_{n-1})\right)\right\|_V &\le \tfrac{3}{8}\tau^2 \left\|u''\right\|_{V,\infty} .\end{aligned}$$

Proof. (a) We compute by the first-order Taylor approximation

$$\begin{aligned}\left\|u(t_{n+1/2}) - \tfrac{1}{2}\left(3u(t_n) - u(t_{n-1})\right)\right\|_X &\le \left\|u(t_{n+1/2}) - u(t_n)\right\|_V + \tfrac{1}{2}\left\|u(t_n) - u(t_{n-1})\right\|_V\\ &\le \tfrac{1}{2}\tau \left\|u'\right\|_{V,\infty} + \tfrac{1}{2}\tau \left\|u'\right\|_{V,\infty}\\ &= \tau \left\|u'\right\|_{V,\infty} .\end{aligned}$$

(b) For the second-order bound we expand with Taylor

$$\begin{aligned}u(t_{n+1/2}) &= u(t_n) + \int_{t_n}^{t_{n+1/2}} u'(s)\,ds\\ &= u(t_n) + \left[\left(s - t_{n+1/2}\right)u'(s)\right]_{t_n}^{t_{n+1/2}} + \int_{t_n}^{t_{n+1/2}} \left(t_{n+1/2} - s\right)u''(s)\,ds\\ &= u(t_n) + \tfrac{\tau}{2}u'(t_n) + \int_0^{\frac{\tau}{2}} \left(\tfrac{\tau}{2} - s\right)u''(t_n + s)\,ds\\ &= u(t_n) + \tfrac{\tau}{2}u'(t_n) + R_1\end{aligned}$$

with

$$\left\|R_1\right\|_V = \left\| \int_0^{\frac{\tau}{2}} \left(\tfrac{\tau}{2} - s\right)u''(t_n + s)\,ds\right\|_V \le \left[\tfrac{\tau}{2}s - \tfrac{s^2}{2}\right]_0^{\tau/2} \left\|u''\right\|_{X,\infty} = \tfrac{\tau^2}{8}\left\|u''\right\|_{X,\infty} .$$

In the same manner we obtain

$$\begin{aligned}u(t_{n-1}) &= u(t_n) - \tau u'(t_n) + \int_0^{-\tau} \left(-\tau - s\right)u''(t_n + s)\,ds\\ &= u(t_n) - \tau u'(t_n) + R_2\end{aligned}$$

with

$$\left\|R_2\right\|_V = \left\| \int_0^{-\tau} \left(-\tau - s\right)u''(t_n + s)\,ds\right\|_V \le \left|\left[\tau s + \tfrac{s^2}{2}\right]_0^{-\tau}\right| \left\|u''\right\|_{V,\infty} = \tfrac{\tau^2}{2}\left\|u''\right\|_{V,\infty} .$$

This gives us the assertion by

$$\left\|u(t_{n+1/2}) - \tfrac{1}{2}\left(3u(t_n) - u(t_{n-1})\right)\right\|_V \le \left\|R_1 + \tfrac{1}{2}R_2\right\|_V = \tau^2 \tfrac{3}{8}\left\|u''\right\|_{V,\infty} .$$

□

Regularity of the initial data in Section 7.4

We briefly discuss the regularity of the initial datum with $x = \big(x_1, x_2 \big)$

$$q_0(x) = -\frac{1}{4}(x_1^2 + x_2^2)\ln(-\ln(\rho(x_1^2 + x_2^2))) + C_1(x_1^2 + x_2^2 - 1) + C_2\,,$$

which was defined for $\rho \in (0,1)$ on the ball $\mathcal{B}_1(0)$. We start with the partial derivatives of q_0. Since the function is symmetric in x_1 and x_2 we do not need to compute all of them. The first derivative is given by

$$\begin{aligned}\partial_{x_1} q_0(x) &= -\frac{1}{2}x_1\ln(-\ln(\rho(x_1^2 + x_2^2))) + \frac{1}{4}(x_1^2 + x_2^2)\frac{1}{\ln(\rho(x_1^2 + x_2^2))}\frac{1}{\rho(x_1^2 + x_2^2)}2\rho x_1 + 2C_1 x_1 \\ &= -\frac{1}{2}x_1\ln(-\ln(\rho(x_1^2 + x_2^2))) + \frac{1}{2}x_1\frac{1}{\ln(\rho(x_1^2 + x_2^2))} + 2C_1 x_1\,.\end{aligned}$$

For the second derivative we need

$$\begin{aligned}\partial_{x_1}^2 q_0(x) &= -\frac{1}{2}\ln(-\ln(\rho(x_1^2 + x_2^2))) + \frac{1}{2}\frac{1}{\ln(\rho(x_1^2 + x_2^2))} + 2C_1 \\ &\quad + \frac{1}{2}x_1\frac{1}{\ln(\rho(x_1^2 + x_2^2))}\frac{1}{\rho(x_1^2 + x_2^2)}2\rho x_1 + \frac{1}{2}x_1\frac{1}{\ln^2(\rho(x_1^2 + x_2^2))}\frac{1}{\rho(x_1^2 + x_2^2)}2\rho x_1 \\ &= -\frac{1}{2}\ln(-\ln(\rho(x_1^2 + x_2^2))) + \frac{1}{2}\frac{1}{\ln(\rho(x_1^2 + x_2^2))} + 2C_1 \\ &\quad + \frac{1}{\ln(\rho(x_1^2 + x_2^2))}\frac{x_1^2}{(x_1^2 + x_2^2)} + \frac{1}{\ln^2(\rho(x_1^2 + x_2^2))}\frac{x_1^2}{(x_1^2 + x_2^2)}\,,\end{aligned}$$

as well as

$$\begin{aligned}\partial_{x_2}\partial_{x_1} q_0(x) &= \frac{1}{2}x_1\frac{1}{\ln(\rho(x_1^2 + x_2^2))}\frac{1}{\rho(x_1^2 + x_2^2)}2\rho x_2 + \frac{1}{2}x_1\frac{1}{\ln^2(\rho(x_1^2 + x_2^2))}\frac{1}{\rho(x_1^2 + x_2^2)}2\rho x_2 \\ &= \frac{1}{\ln(\rho(x_1^2 + x_2^2))}\frac{x_1 x_2}{(x_1^2 + x_2^2)} + \frac{1}{\ln^2(\rho(x_1^2 + x_2^2))}\frac{x_1 x_2}{(x_1^2 + x_2^2)}\,.\end{aligned}$$

From this we directly obtain

$$\partial_{x_2}\partial_{x_1} q_0 \in L^\infty(\mathcal{B}_1(0)), \qquad \partial_{x_1}^2 q_0 + \frac{1}{2}\ln(-\ln(\rho|\cdot|^2)) \in L^\infty(\mathcal{B}_1(0)), \qquad \ln(-\ln(\rho|\cdot|^2)) \notin L^\infty(\mathcal{B}_1(0))\,,$$

where the last term is the one from (7.39).

In the next step we show that the second derivatives are still H^1-functions. We need the following auxiliary result.

Lemma B.12. *Let $\rho \in (0,1)$, $k \geq 1$, and define the function*

$$D : \mathcal{B}_1(0) \subset \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto \frac{1}{\ln^k(\rho|x|^2)|x|}\,.$$

Then it holds $D \in L^2(\mathcal{B}_1(0))$.

Proof. We use polar coordinates to obtain

$$\begin{aligned}\int_{\mathcal{B}_1(0)} |D(x)|^2\,dx &= \int_{\mathcal{B}_1(0)} \frac{1}{\ln^{2k}(\rho|x|^2)|x|^2}\,dx \\ &= 2\pi \int_0^1 \frac{1}{r^2}\frac{1}{\ln^{2k}(\rho r^2)} r\,dr \\ &= 2\pi \int_0^1 \frac{1}{r\ln^{2k}(\rho r^2)}\,dr \\ &= \pi \left[\frac{1}{(2k-1)\ln^{2k-1}(\rho r^2)}\right]_0^1 \\ &= \frac{\pi}{-(2k-1)\ln^{2k-1}(\rho)} < \infty\,,\end{aligned}$$

since $\rho < 1$ holds. □

From this we can conclude the following regularity of q_0.

Lemma B.13. *The function q_0 is in $H^3(\mathcal{B}_1(0))$ and in particular it holds*

$$\partial_{x_2}\partial_{x_1} q_0,\ \partial_{x_1}^2 q_0 + \frac{1}{2}\ln(-\ln(\rho|\cdot|^2)),\ \frac{1}{2}\ln(-\ln(\rho|\cdot|^2)) \in H^1(\mathcal{B}_1(0))\,.$$

Proof. Computing all the derivatives, we observe that their absolute value is always dominated by a multiple of the function D from Lemma B.12. □

To summarize this, we have shown that q_0 is in $H^3(\mathcal{B}_1(0))$ but *not* in $W^{2,\infty}(\mathcal{B}_1(0))$. Assuming that $q_0 \in H^{3+\epsilon}(\mathcal{B}_1(0))$, would however imply $q_0 \in W^{2,\infty}(\mathcal{B}_1(0))$ by the Sobolev embedding, see [1, Thm. 7.34], which is a contradiction.

Bibliography

[1] 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.

[2] G. P. Agrawal. *Nonlinear fiber optics*. Elsevier, Academic Press, Amsterdam, 5. ed. edition, 2013. ISBN 978-0-12397-023-7.

[3] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. The fenics project version 1.5. *Archive of Numerical Software*, 3(100), 2015. URL https://doi.org/10.11588/ans.2015.100.20553.

[4] H. Amann and J. Escher. *Analysis I*. Grundstudium Mathematik. Birkhäuser Verlag, Basel, third edition, 2006. ISBN 978-3-7643-7755-7. URL https://doi.org/10.1007/978-3-7643-7756-4.

[5] H. Amann and J. Escher. *Analysis II*. Grundstudium Mathematik. Birkhäuser Verlag, Basel, second edition, 2006. ISBN 978-3-7643-7105-0. URL https://doi.org/10.1007/3-7643-7402-0.

[6] P. F. Antonietti, I. Mazzieri, M. Muhr, V. Nikolić, and B. Wohlmuth. A high-order discontinuous Galerkin method for nonlinear sound waves. *J. Comput. Phys.*, 415:109484, 27, 2020. ISSN 0021-9991. URL https://doi.org/10.1016/j.jcp.2020.109484.

[7] S. Baumstark, E. Faou, and K. Schratz. Uniformly accurate exponential-type integrators for Klein-Gordon equations with asymptotic convergence to the classical NLS splitting. *Math. Comp.*, 87(311):1227–1254, 2018. ISSN 0025-5718. URL https://doi.org/10.1090/mcom/3263.

[8] S. Buchholz, L. Gauckler, V. Grimm, M. Hochbruck, and T. Jahnke. Closing the gap between trigonometric integrators and splitting methods for highly oscillatory differential equations. *IMA J. Numer. Anal.*, 38(1):57–74, 2018. URL https://doi.org/10.1093/imanum/drx007.

[9] S. Buchholz, B. Dörich, and M. Hochbruck. On averaged exponential integrators for semilinear wave equations with solutions of low-regularity. *SN Partial Differ. Equ. Appl.*, 2(2), 2021. ISSN 2662-2963. URL https://doi.org/10.1007/s42985-020-00045-9.

[10] S. F. Buchholz. *Fehleranalyse von auf trigonometrischen Integratoren basierenden Splittingverfahren für hochoszillatorische, semilineare Probleme*. PhD thesis, Karlsruher Institut für Technologie (KIT), 2019. URL https://doi.org/10.5445/IR/1000088935.

[11] K. Burrage and J. C. Butcher. Stability criteria for implicit Runge-Kutta methods. *SIAM J. Numer. Anal.*, 16(1):46–57, 1979. ISSN 0036-1429. URL https://doi.org/10.1137/0716004.

[12] K. Busch, G. von Freymann, S. Linden, S. F. Mingaleev, L. Tkeshelashvili, and M. Wegener. Periodic nanostructures for photonics. *Physics Reports*, 444(3–6):101–202, 2007. ISSN 0370-1573. URL https://doi.org/10.1016/j.physrep.2007.02.011.

[13] M. Caliari, L. Einkemmer, A. Moriggl, and A. Ostermann. An accurate and time-parallel rational exponential integrator for hyperbolic and oscillatory PDEs. *arXiv*, page arXiv:2008.11607, 2020. URL http://arxiv.org/abs/2008.11607.

[14] E. Casas and K. Chrysafinos. Numerical analysis of quasilinear parabolic equations under low regularity assumptions.

Numer. Math., 143(4):749–780, 2019. ISSN 0029-599X. URL https://doi.org/10.1007/s00211-019-01071-5.

[15] E. Celledoni, D. Cohen, and B. Owren. Symmetric exponential integrators with an application to the cubic Schrödinger equation. *Found. Comput. Math.*, 8(3):303–317, 2008. ISSN 1615-3375. URL https://doi.org/10.1007/s10208-007-9016-7.

[16] M. G. Crandall and P. E. Souganidis. Convergence of difference approximations of quasilinear evolution equations. *Nonlinear Anal.*, 10(5):425–445, 1986. ISSN 0362-546X. URL http://dx.doi.org/10.1016/0362-546X(86)90049-0.

[17] N. Crouseilles, L. Einkemmer, and J. Massot. Exponential methods for solving hyperbolic problems with application to collisionless kinetic equations. *J. Comput. Phys.*, 420:109688, 25, 2020. ISSN 0021-9991. URL https://doi.org/10.1016/j.jcp.2020.109688.

[18] W. Dörfler, H. Gerner, and R. Schnaubelt. Local well-posedness of a quasilinear wave equation. *Appl. Anal.*, 95(9):2110–2123, 2016. ISSN 0003-6811. URL https://doi.org/10.1080/00036811.2015.1089236.

[19] K.-J. Engel and R. Nagel. *One-parameter semigroups for linear evolution equations*, volume 194 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2000. ISBN 0-387-98463-1. with contributions by S. Brendle, M. Campiti, T. Hahn, G. Metafune, G. Nickel, D. Pallara, C. Perazzoli, A. Rhandi, S. Romanelli and R. Schnaubelt.

[20] U. Frisch, Z.-S. She, and O. Thual. Viscoelastic behaviour of cellular solutions to the Kuramoto-Sivashinsky model. *J. Fluid Mech.*, 168:221–240, 1986. ISSN 0022-1120. URL https://doi.org/10.1017/S0022112086000356.

[21] B. García-Archilla, J. M. Sanz-Serna, and R. D. Skeel. Long-time-step methods for oscillatory differential equations. *SIAM J. Sci. Comput.*, 20(3):930–963, 1999. ISSN 1064-8275. URL http://dx.doi.org/10.1137/S1064827596313851.

[22] L. Gauckler. Error analysis of trigonometric integrators for semilinear wave equations. *SIAM J. Numer. Anal.*, 53(2):1082–1106, 2015. ISSN 0036-1429. URL http://dx.doi.org/10.1137/140977217.

[23] L. Gauckler, J. Lu, J. L. Marzuola, F. Rousset, and K. Schratz. Trigonometric integrators for quasilinear wave equations. *Math. Comp.*, 88(316):717–749, 2019. ISSN 0025-5718. URL https://doi.org/10.1090/mcom/3339.

[24] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. ISBN 3-540-41160-7. Reprint of the 1998 edition.

[25] 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. URL https://doi.org/10.1002/mana.19921550110.

[26] C. González and M. Thalhammer. A second-order Magnus-type integrator for quasi-linear parabolic problems. *Math. Comp.*, 76(257):205–231, 2007. ISSN 0025-5718. URL https://doi.org/10.1090/S0025-5718-06-01883-7.

[27] C. González and M. Thalhammer. Higher-order exponential integrators for quasi-linear parabolic problems. Part I: Stability. *SIAM J. Numer. Anal.*, 53(2):701–719, 2016. ISSN 0036-1429. URL https://doi.org/10.1137/140961845.

[28] C. González and M. Thalhammer. Higher-order exponential integrators for quasi-linear parabolic problems. Part II: Convergence. *SIAM J. Numer. Anal.*, 54(5):2868–2888, 2016. ISSN 0036-1429. URL https://doi.org/10.1137/15M103384.

[29] C. González, A. Ostermann, and M. Thalhammer. A second-order Magnus-type integrator for nonautonomous parabolic problems. *J. Comput. Appl. Math.*, 189(1-2):142–156, 2006. ISSN 0377-0427. URL https://doi.org/10.1016/j.cam.2005.04.036.

[30] V. Grimm and M. Hochbruck. Error analysis of exponential integrators for oscillatory second-order differential equations. *J. Phys. A*, 39(19):5495–5507, 2006. ISSN 0305-4470. URL http://dx.doi.org/10.1088/0305-4470/39/19/S10.

[31] V. Grimm and M. Hochbruck. Rational approximation to trigonometric operators. *BIT*, 48(2):215–229, 2008. ISSN 0006-3835. URL http://dx.doi.org/10.1007/s10543-008-0185-9.

[32] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985. ISBN 0-273-08647-2.

[33] E. Hairer and G. Wanner. *Solving ordinary differential equations II: Stiff and differential-algebraic problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 1996. ISBN 3-540-60452-9. URL https://doi.org/10.1007/978-3-642-05221-7.

[34] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations I: Nonstiff problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 1993. ISBN 3-540-56670-8.

[35] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 2006. ISBN 3-540-30663-3; 978-3-540-30663-4.

[36] E. Hansen and A. Ostermann. Exponential splitting for unbounded operators. *Math. Comp.*, 78(267):1485–1496, 2009.

ISSN 0025-5718. URL http://dx.doi.org/10.1090/S0025-5718-09-02213-3.

[37] E. Hansen and A. Ostermann. High-order splitting schemes for semilinear evolution equations. *BIT*, 56(4):1303–1316, 2016. ISSN 0006-3835. URL https://doi.org/10.1007/s10543-016-0604-2.

[38] M. Hochbruck and C. Lubich. A Gautschi-type method for oscillatory second-order differential equations. *Numer. Math.*, 83(3):403–426, 1999. ISSN 0029-599X. URL http://dx.doi.org/10.1007/s002110050456.

[39] M. Hochbruck and A. Ostermann. Explicit exponential Runge-Kutta methods for semilinear parabolic problems. *SIAM J. Numer. Anal.*, 43(3):1069–1090, 2005. URL http://dx.doi.org/10.1137/040611434.

[40] M. Hochbruck and A. Ostermann. Exponential multistep methods of Adams-type. *BIT*, 51(4):889–908, 2011. ISSN 0006-3835. URL http://dx.doi.org/10.1007/s10543-011-0332-6.

[41] M. Hochbruck and T. Pažur. Error analysis of implicit Euler methods for quasilinear hyperbolic evolution equations. *Numer. Math.*, 135(2):547–569, 2017. ISSN 0945-3245. URL http://dx.doi.org/10.1007/s00211-016-0810-5.

[42] M. Hochbruck, C. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.*, 19(5):1552–1574, 1998. ISSN 1064-8275. URL https://doi.org/10.1137/S1064827595295337.

[43] M. Hochbruck, T. Pažur, A. Schulz, E. Thawinan, and C. Wieners. Efficient time integration for discontinuous Galerkin approximations of linear wave equations. *ZAMM*, 95(3):237–259, 2015. URL http://dx.doi.org/10.1002/zamm.201300306.

[44] 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. URL https://doi.org/10.1007/s00211-017-0914-6.

[45] M. Hochbruck, J. Leibold, and A. Ostermann. On the convergence of Lawson methods for semilinear stiff problems. *Numer. Math.*, 145(3):553–580, 2020. ISSN 0029-599X. URL https://doi.org/10.1007/s00211-020-01120-4.

[46] T. Jahnke and C. Lubich. Error bounds for exponential operator splittings. *BIT*, 40(4):735–744, 2000. ISSN 0006-3835. URL http://dx.doi.org/10.1023/A:1022396519656.

[47] S. Kanda. Convergence of difference approximations and nonlinear semigroups. *Proc. Amer. Math. Soc.*, 108(3):741–748, 1990. ISSN 0002-9939. URL http://dx.doi.org/10.2307/2047796.

[48] T. Kato. Linear evolution equations of "hyperbolic" type. *J. Fac. Sci. Univ. Tokyo Sect. I*, 17:241–258, 1970. ISSN 0368-2269.

[49] T. Kato. Linear evolution equations of "hyperbolic" type. II. *J. Math. Soc. Japan*, 25:648–666, 1973. ISSN 0025-5645. URL https://doi.org/10.2969/jmsj/02540648.

[50] T. Kato. Quasi-linear equations of evolution, with applications to partial differential equations. In *Spectral theory and differential equations (Proc. Sympos., Dundee, 1974; dedicated to Konrad Jörgens)*, pages 25–70. Lecture Notes in Math., Vol. 448, 1975.

[51] T. Kato. *Abstract differential equations and nonlinear mixed problems*. Lezioni Fermiane. [Fermi Lectures]. Scuola Normale Superiore, Pisa; Accademia Nazionale dei Lincei, Rome, 1985.

[52] Y. Kobayashi. Difference approximation of Cauchy problems for quasi-dissipative operators and generation of nonlinear semigroups. *J. Math. Soc. Japan*, 27(4):640–665, 1975. ISSN 0025-5645. URL http://dx.doi.org/10.2969/jmsj/02740640.

[53] B. Kovács and C. Lubich. Stability and convergence of time discretizations of quasi-linear evolution equations of Kato type. *Numer. Math.*, 138(2):365–388, 2018. ISSN 0029-599X. URL https://doi.org/10.1007/s00211-017-0909-3.

[54] C. Lubich. On splitting methods for Schrödinger-Poisson and cubic nonlinear Schrödinger equations. *Math. Comp.*, 77(264):2141–2153, 2008. ISSN 0025-5718. URL http://dx.doi.org/10.1090/S0025-5718-08-02101-7.

[55] A. Lunardi. *Interpolation theory*, volume 16 of *Appunti. Scuola Normale Superiore di Pisa (Nuova Serie) [Lecture Notes. Scuola Normale Superiore di Pisa (New Series)]*. Edizioni della Normale, Pisa, 2018. ISBN 978-88-7642-639-1; 978-88-7642-638-4. URL https://doi.org/10.1007/978-88-7642-638-4. Third edition [of MR2523200].

[56] B. Maier. *Error analysis for space and time discretizations of quasilinear wave-type equations*. PhD thesis, Karlsruhe Institute of Technology, may 2020. URL https://doi.org/10.5445/IR/1000120935.

[57] R. I. McLachlan and G. R. W. Quispel. Splitting methods. *Acta Numer.*, 11:341–434, 2002. ISSN 0962-4929. URL https://doi.org/10.1017/S0962492902000053.

[58] M. Miklavčič. *Applied functional analysis and partial differential equations*. World Scientific Publishing Co., Inc., River Edge, NJ, 1998. ISBN 981-02-3535-6. URL https://doi.org/10.1142/9789812796233.

[59] J. Moloney and A. Newell. *Nonlinear optics*. Westview Press. Advanced Book Program, Boulder, CO, 2004. ISBN

0-8133-4118-3.

[60] M. Muhr, V. Nikolić, and B. Wohlmuth. Self-adaptive absorbing boundary conditions for quasilinear acoustic wave propagation. *J. Comput. Phys.*, 388:279–299, 2019. ISSN 0021-9991. URL https://doi.org/10.1016/j.jcp.2019.03.025.

[61] D. Müller. *Well-posedness for a general class of quasilinear evolution equations - with applications to Maxwell's equations*. PhD thesis, Karlsruhe Institute of Technology, 2014. URL https://doi.org/10.5445/IR/1000042147.

[62] V. Nikolić and B. Wohlmuth. A priori error estimates for the finite element approximation of Westervelt's quasi-linear acoustic wave equation. *SIAM J. Numer. Anal.*, 57(4):1897–1918, 2019. ISSN 0036-1429. URL https://doi.org/10.1137/19M1240873.

[63] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*, volume 44 of *Applied Mathematical Sciences*. Springer, New York, 1983. ISBN 9780387908458. URL https://doi.org/10.1007/978-1-4612-5561-1.

[64] M. Pototschnig, J. Niegemann, L. Tkeshelashvili, and K. Busch. Time-domain simulations of the nonlinear Maxwell equations using operator-exponential methods. *IEEE Transactions on Antennas and Propagation*, 57(2):475–483, Feb. 2009. ISSN 0018-926X. URL https://doi.org/10.1109/TAP.2008.2011181.

[65] M. Renardy and R. C. Rogers. *An introduction to partial differential equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2004. ISBN 0-387-00444-0.

[66] J. M. Sanz-Serna. Mollified impulse methods for highly oscillatory differential equations. *SIAM J. Numer. Anal.*, 46 (2):1040–1059, 2008. ISSN 0036-1429. URL https://doi.org/10.1137/070681636.

[67] K. Schmüdgen. *Unbounded self-adjoint operators on Hilbert space*, volume 265 of *Graduate Texts in Mathematics*. Springer, Dordrecht, 2012. ISBN 978-94-007-4752-4. URL https://doi.org/10.1007/978-94-007-4753-1.

[68] M. Suzuki. Generalized Trotter's formula and systematic approximants of exponential operators and inner derivations with applications to many-body problems. *Comm. Math. Phys.*, 51(2):183–190, 1976. ISSN 0010-3616. URL https://doi.org/10.1007/BF01609348.

[69] T. Takahashi. Convergence of difference approximation of nonlinear evolution equations and generation of semigroups. *J. Math. Soc. Japan*, 28(1):96–113, 1976. ISSN 0025-5645. URL https://doi.org/10.2969/jmsj/02810096.

[70] A. F. Timan. *Theory of approximation of functions of a real variable*. Translated from the Russian by J. Berry. English translation edited and editorial preface by J. Cossar. International Series of Monographs in Pure and Applied Mathematics, Vol. 34. A Pergamon Press Book. The Macmillan Co., New York, 1963.

[71] B. Wang and X. Wu. Global error bounds of one-stage extended RKN integrators for semilinear wave equations. *Numer. Algorithms*, 81(4):1203–1218, 2019. ISSN 1017-1398. URL https://doi.org/10.1007/s11075-018-0585-0.

[72] B. Wang, X. Wu, and J. Xia. Error bounds for explicit ERKN integrators for systems of multi-frequency oscillatory second-order differential equations. *Appl. Numer. Math.*, 74:17–34, 2013. ISSN 0168-9274. URL https://doi.org/10.1016/j.apnum.2013.08.002.

[73] D. Werner. *Funktionalanalysis*. Springer-Lehrbuch. Springer, Berlin, 7., korrigierte und erw. aufl. edition, 2011. ISBN 978-3-642-21017-4. URL http://doi.org/10.1007/978-3-642-21017-4.

[74] K. S. Yee. Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media. *IEEE Transactions on Antennas and Propagation*, 14(3):302–307, May 1966. ISSN 0018-926X. URL https://doi.org/10.1109/TAP.1966.1138693.