Trigonometric integrators for quasilinear wave equations

Ludwig Gauckler [1] Jianfeng Lu [2] Jeremy L. Marzuola [3]
Frédéric Rousset [4] Katharina Schratz [5]

August 25, 2017

Abstract

Trigonometric time integrators are introduced as a class of explicit numerical methods for quasilinear wave equations. Second-order convergence for the semi-discretization in time with these integrators is shown for a sufficiently regular exact solution. The time integrators are also combined with a Fourier spectral method into a fully discrete scheme, for which error bounds are provided without requiring any CFL-type coupling of the discretization parameters. The proofs of the error bounds are based on energy techniques and on the semiclassical Gårding inequality.

Mathematics Subject Classification (2010): 65M15, 65P10, 65L70, 65M20.

Keywords: Quasilinear wave equation, trigonometric integrators, exponential integrators, error bounds, loss of derivatives, energy estimates.

1 Introduction

The topic of the present paper is the numerical analysis of *quasilinear wave equations*. Such wave equations show up in a variety of applications, ranging from elastodynamics to general relativity. While the (local-in-time) analysis of them is well-developed since the seventies, the papers [24] by Kato and [23] by Hughes, Kato & Marsden being major contributions to the local well-posedness theory, and has meanwhile found its way into classical monographs on partial differential equation, see, for instance, the monograph [31] by Taylor, as well as the books by Sogge [29] and Hörmander [22], the numerical analysis of quasilinear wave equations is much less developed. The main challenge is, of course, the numerical treatment of the quasilinear term in the equation.

[1] Institut für Mathematik, Freie Universität Berlin, Arnimallee 9, D-14195 Berlin, Germany (gauckler@math.fu-berlin.de).
[2] Departments of Mathematics, Physics, and Chemistry, Duke University, Box 90320, Durham, NC 27708, USA (jianfeng@math.duke.edu).
[3] Department of Mathematics, UNC-Chapel Hill, CB#3250 Phillips Hall, Chapel Hill, NC 27599, USA (marzuola@math.unc.edu).
[4] Laboratoire de Mathématiques d'Orsay (UMR 8628), Université Paris-Sud, 91405 Orsay Cedex, France and Institut Universitaire de France (frederic.rousset@math.u-psud.fr).
[5] Fakultät für Mathematik, Karlsruhe Institute of Technology, Englerstr. 2, D-76131 Karlsruhe, Germany (katharina.schratz@kit.edu).

In the present paper, we focus on quasilinear wave equations of the form

$$\partial_t^2 u = \partial_x^2 u - u + \kappa\, a(u)\, \partial_x^2 u + \kappa\, g\big(u, \partial_x u\big) \tag{1}$$

where g and a are smooth and real-valued functions such that $g(0,0) = a(0) = 0$. We consider real-valued solutions to (1) with 2π-periodic boundary conditions in one space dimension, $x \in \mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$, for initial values

$$u(\cdot,0) = u_0, \qquad \partial_t u(\cdot,0) = \dot{u}_0 \tag{2}$$

given at time $t = 0$. The real-valued parameter κ will be used to emphasize the strength of the nonlinearities, and we will be interested both in the regime where κ is small so that the nonlinearities are small and the regime where κ is of order one. Quasilinear wave equations of this form with small $|\kappa|$ have been extensively studied by Groves & Schneider [17], Chong & Schneider [5], Chirilus-Bruckner, Düll & Schneider [4] and Düll [9]: the equations from the class (1) are prototypes for models in nonlinear optics with a nonlinear Schrödinger equation as a modulation equation [17, Introduction]. Many examples from elasticity and fluid mechanics can also be reduced to quasilinear wave equations under the form (1). Relevant applications from elasticity and general relativity appeared in [23] for instance, though of course many of the most physically interesting examples occur in dimensions 2 and include potential dependence upon $\partial_x u$ in a and possible dependence upon $\partial_t u$. Using the techniques developed here to study smooth solutions in relevant higher dimensional models will be a topic for future work and would likely require some higher regularity assumptions on the solutions.

The principal difficulty in the numerical discretization of (1) is the quasilinear term $\kappa a(u)\partial_x^2 u$. For typical explicit methods for the discretization in time, in which the numerical approximation at a discrete time depends in an explicit way on this quasilinear term at some previous time step, there is a risk of *losing derivatives*, in the sense that a control of a certain number of spatial derivatives of the numerical solution requires the control of *more* derivatives of the numerical solution at previous time steps. This phenomenon is by far not restricted to quasilinear wave equations, and an established way to prevent a loss of derivatives is to resort to carefully chosen implicit methods. In the case of quasilinear wave equations, this route has been taken recently by Hochbruck & Pažur [21] and Kovács & Lubich [25], who propose and study implicit and semi-implicit methods of Runge–Kutta type for semi-discretization in time for a more general class of quasilinear evolution equations.

In the present paper, we take another route and show how a class of *explicit* time discretizations can be used to numerically solve the quasilinear wave equation (1) (and also how it can be combined with a Fourier spectral method in space). The considered class of methods is the class of *trigonometric integrators*, which is described in detail in Section 2. These methods are exponential integrators and have been originally developed for highly oscillatory ordinary differential equations, see, for instance, [10] or Chapter XIII of the monograph [19]. Meanwhile, they were recognized to work well also for wave equations in the semilinear case, see [1–3, 6, 7, 11, 13]. We show here, how these explicit methods can be put to use also in the quasilinear case. A careful choice of the filters in these integrators turns out to play a crucial role in avoiding the above-mentioned loss of derivatives.

For the considered and derived trigonometric integrators, we rigorously prove second-order convergence in time. We also prove convergence of a fully discrete method which is based on a combination with a spectral discretization in space,

without requiring any CFL-type coupling of the discretization parameters. See Section 3 for statements of the error bounds. The proofs of our convergence results are based on *energy techniques* as they are widely used in mathematical analysis to prove well-posedness of quasilinear equations and also well-established in the numerical analysis of quasilinear *parabolic* equations, see, e.g., [28]. Furthermore, they have been applied in the recent analysis of *implicit* methods for quasilinear hyperbolic equations in [21] for (semi-)implicit Euler methods and in [25] for (linearly) implicit midpoint methods and implicit Runge–Kutta methods.

Here, we are interested in the analysis of *explicit* trigonometric integrators for quasilinear wave equations in the form (1). Note that in contrast to the semilinear case studied in [1–3, 6, 7, 11, 13], the proof uses energy techniques with a nontrivial modified discrete energy to prove stability of the methods under appropriate assumptions on the filters. In addition, in the case of non-small κ, we also need tools from semiclassical pseudodifferential calculus to ensure that the modified discrete energy is positive.

We mention that recently in [14, 15] explicit exponential integrators for quasilinear *parabolic* problems in Banach spaces were considered. Analysis of quasilinear parabolic equations can generally be done using simpler techniques stemming from the regularization implicit in the diffusion operators, but quasilinear waves must be handled with more care given the lack of smoothing properties of the leading order operator. We also point out that the exponential integrators in [14, 15] are based on solving exactly a differential equation with the linearization of the quasilinear part on the right, whereas the trigonometric integrators considered in the present paper are based on solving exactly a differential equation with the pure linear part $\partial_x^2 u - u$ on the right, which is usually simpler from a computational point of view.

The methods and their analysis as presented in the present paper can be extended to higher spatial dimensions or to quasilinear wave equations (1) without Klein–Gordon term $-u$ on the right-hand side. Moreover, we could also only assume that a and g are smooth on an open subset and deal with smooth solutions that stay in this subset on the considered time interval. In this way, our scheme can be used to approximate the classical p-system of elasticity and gas dynamics (as long as the solutions are smooth and with no vacuum). See [26] for instance for a discussion of this model. It would be interesting to see if our methods and their analysis can be extended to quasilinear wave equations (1) with a semilinear term g that depends also on $\partial_t u$ (for example in order to handle the equations considered in [8]), to the more abstract classes of equations considered in [21, 25], or to other equations with a possible loss of derivatives in explicit numerical methods, such as quasilinear Schrödinger equations as considered in [27].

The article is organized as follows. In Section 2, we introduce the considered trigonometric integrators, for which we state global error bounds in Section 3. The proof of the error bound for the semi-discretization in time is given in Section 4, and the one for the full discretization in Section 5. The necessary tools from semiclassical pseudodifferential calculus are collected in the appendix.

Notation. By $H^s = H^s(\mathbb{T})$, $s \geq 0$, we denote the usual Sobolev space, equipped with the norm $\|\cdot\|_s$ given by

$$\|v\|_s^2 = \sum_{j\in\mathbb{Z}} \langle j\rangle^{2s} |\hat{v}_j|^2 \qquad \text{for} \qquad v(x) = \sum_{j\in\mathbb{Z}} \hat{v}_j \mathrm{e}^{\mathrm{i}jx}$$

with the weights

$$\langle j\rangle = \sqrt{j^2+1}, \qquad j \in \mathbb{Z}.$$

By $\langle \cdot, \cdot \rangle_s$, we denote the corresponding scalar product,

$$\langle v, w \rangle_s = \sum_{j \in \mathbb{Z}} \langle j \rangle^{2s} \bar{\hat{v}}_j \hat{w}_j \qquad \text{for} \qquad v(x) = \sum_{j \in \mathbb{Z}} \hat{v}_j \mathrm{e}^{\mathrm{i}jx}, \quad w(x) = \sum_{j \in \mathbb{Z}} \hat{w}_j \mathrm{e}^{\mathrm{i}jx}.$$

We study solutions $(u(\cdot,t), \partial_t u(\cdot,t))$ of the quasilinear wave equation (1) in product spaces $H^{s+1} \times H^s$, on which we use the norm

$$|||(u, \dot{u})|||_s = \big(\|u\|_{s+1}^2 + \|\dot{u}\|_s^2 \big)^{1/2}.$$

For $s > \frac{1}{2}$, we will make frequent use of the classical estimates in Sobolev spaces

$$\|uv\|_0 \le C_s \|u\|_0 \|v\|_s, \qquad \|uv\|_s \le C_s \|u\|_s \|v\|_s \tag{3a}$$

and

$$\|G(u)\|_s \le \Lambda_s(\|u\|_s)\|u\|_s, \qquad \|G(u) - G(v)\|_s \le \Lambda_s(\|u\|_s + \|v\|_s)\|u - v\|_s, \tag{3b}$$

where G is any smooth function such that $G(0) = 0$ and $\Lambda_s(\cdot)$ is a continuous non-decreasing function, see for instance [32], Chapter 13, Section 3.

2 Discretization of quasilinear wave equations

2.1 Trigonometric integrators for the discretization in time

The quasilinear wave equation (1) can be written in compact form as

$$\partial_t^2 u = -\Omega^2 u + \kappa f(u). \tag{4}$$

with the nonlinearity

$$f(u) = a(u)\partial_x^2 u + g\big(u, \partial_x u\big) \tag{5}$$

and the linear operator

$$\Omega = \sqrt{-\partial_x^2 + 1},$$

that is, the operator Ω acts on a function by multiplication of the jth Fourier coefficient with $\sqrt{j^2+1}$.

For the discretization in time of (4), we use *trigonometric integrators*, see [19, Section XIII.2.2]. We introduce them here as splitting integrators as in [10], since this interpretation is convenient for the error analysis to be presented in this paper. Written in first-order form $\partial_t(u, \dot{u}) = (\dot{u}, -\Omega^2 u + \kappa f(u))$, equation (4) is split into

$$\partial_t \begin{pmatrix} u \\ \dot{u} \end{pmatrix} = \begin{pmatrix} \dot{u} \\ -\Omega^2 u \end{pmatrix} \qquad \text{and} \qquad \partial_t \begin{pmatrix} u \\ \dot{u} \end{pmatrix} = \begin{pmatrix} 0 \\ \kappa f(u) \end{pmatrix},$$

and the usual Strang splitting is applied with time step-size τ. In addition, the nonlinearity $f(u)$ of (5) is replaced by

$$\widehat{f}(u) = \Psi_1 f(\Phi u), \tag{6}$$

where

$$\Psi_1 = \psi_1(\tau\Omega) \qquad \text{and} \qquad \Phi = \phi(\tau\Omega)$$

are filter operators computed from suitably chosen filter functions ψ_1 and ϕ. Throughout, we assume that the filter functions are bounded and continuously differentiable

with bounded derivative. Denoting by u_n and $\dot{u}_n$ approximations to $u(\cdot, t_n)$ and $\partial_t u(\cdot, t_n)$ at time $t_n = n\tau$, the numerical method thus reads

$$\begin{aligned} \dot{u}_n^+ &= \dot{u}_n + \tfrac{1}{2}\tau\kappa\widehat{f}(u_n), \\ \begin{pmatrix} \Omega u_{n+1} \\ \dot{u}_{n+1}^- \end{pmatrix} &= \begin{pmatrix} \cos(\tau\Omega) & \sin(\tau\Omega) \\ -\sin(\tau\Omega) & \cos(\tau\Omega) \end{pmatrix} \begin{pmatrix} \Omega u_n \\ \dot{u}_n^+ \end{pmatrix}, \\ \dot{u}_{n+1} &= \dot{u}_{n+1}^- + \tfrac{1}{2}\tau\kappa\widehat{f}(u_{n+1}). \end{aligned} \tag{7}$$

By eliminating the intermediate values $\dot{u}_n^+$ and $\dot{u}_{n+1}^-$, one time step of the method is seen to be given by

$$\begin{aligned} u_{n+1} &= \cos(\tau\Omega)u_n + \tau\,\mathrm{sinc}(\tau\Omega)\dot{u}_n + \tfrac{1}{2}\tau^2\,\mathrm{sinc}(\tau\Omega)\kappa\widehat{f}(u_n), \\ \dot{u}_{n+1} &= -\Omega\sin(\tau\Omega)u_n + \cos(\tau\Omega)\dot{u}_n + \tfrac{1}{2}\tau\cos(\tau\Omega)\kappa\widehat{f}(u_n) + \tfrac{1}{2}\tau\kappa\widehat{f}(u_{n+1}). \end{aligned} \tag{8}$$

The numerical flow of this method is denoted in the following by φ_τ, i.e.,

$$(u_{n+1}, \dot{u}_{n+1}) = \varphi_\tau(u_n, \dot{u}_n). \tag{9}$$

2.2 On the filter functions

We collect some assumptions on the filter functions ϕ and ψ_1 that are going to play an important role in the following.

Assumption 1. Already in the semilinear case, the bounds

$$|\phi(\xi)| \le 1 \qquad \text{and} \qquad |1 - \phi(\xi)| \le c_0\xi^2 \qquad \text{for all} \qquad \xi \ge 0, \tag{10a}$$

$$|\psi_1(\xi)| \le 1 \qquad \text{and} \qquad |1 - \psi_1(\xi)| \le c_0\xi^2 \qquad \text{for all} \qquad \xi \ge 0 \tag{10b}$$

with a constant c_0 are needed for finite-time error bounds, see [11].

Assumption 2. In the quasilinear case, we need in addition that the functions ϕ and ψ_1 are continuous in ξ and satisfy

$$\psi_1(\xi) = \mathrm{sinc}(\xi)\phi(\xi) \qquad \text{for all} \qquad \xi \ge 0. \tag{11}$$

The condition in Assumption 2 has been originally derived in a study on energy conservation properties of trigonometric integrators applied to linear oscillatory ordinary differential equations, see [18, Equation (2.12)]. Surprisingly, it shows up here again in the somehow unrelated context of finite-time error bounds for quasilinear wave equations.

Assumption 3. We finally need in addition to (10) and (11) that, for prescribed $0 < \delta < 1$ and $A_0 \ge 0$ related to the size of κ and a in (1) and the solution u to (1),

$$A_0 \sin(\tfrac{1}{2}\xi)^2\phi(\xi)^2 \le 1 - \delta \qquad \text{for all} \qquad \xi \ge 0. \tag{12}$$

Remark 2.1 (Small nonlinearity)**.** The parameters δ and A_0 in Assumption 3 will be chosen later such that $1 + \kappa\, a(u(x,t)) \ge \delta$ and $\kappa\, a(u(x,t)) \le A_0$ for all $x \in \mathbb{T}$ and all t from the time interval under consideration. In particular, δ is small. In the regime $|\kappa| \ll 1$ of a small nonlinearity in (1), also the value A_0 is small, and hence Assumption 3 is satisfied in this case for *all* bounded filter functions ϕ. The remaining Assumptions 1 and 2 are thus sufficient to prove error bounds for small $|\kappa|$

in (1). They are, for example, satisfied (with $c_0 = 1$) for the trigonometric integrator of Hairer & Lubich [18], where

$$\phi(\xi) = 1, \qquad \psi_1(\xi) = \operatorname{sinc}(\xi), \tag{13}$$

and the one of Grimm & Hochbruck [16], where

$$\phi(\xi) = \operatorname{sinc}(\xi), \qquad \psi_1(\xi) = \operatorname{sinc}(\xi)^2. \tag{14}$$

Remark 2.2 (Non-small nonlinearity)**.** For non-small $|\kappa|$ in (1), the coefficient A_0 in (12) is not small and (12) not always true. A new method that we propose here for this case is the trigonometric integrator (7) with filter functions

$$\phi(\xi) = \operatorname{sinc}(c\xi), \qquad \psi_1(\xi) = \operatorname{sinc}(\xi)\operatorname{sinc}(c\xi) \tag{15}$$

with

$$c \geq \frac{1}{2}\sqrt{\frac{A_0}{1-\delta}}.$$

Here, $0 < \delta < 1$ and $A_0 \geq 0$ are the numbers of Assumption 3. This choice of filter functions satisfies Assumptions 1 and 2 (with $c_0 = \max(1, (c^2+1)/6)$), but it also satisfies Assumption 3. The latter follows from

$$A_0 \sin(\tfrac{1}{2}\xi)^2 \phi(\xi)^2 = \frac{A_0}{4c^2} \operatorname{sinc}(\tfrac{1}{2}\xi)^2 \sin(c\xi)^2 \leq \frac{A_0}{4c^2} \leq 1 - \delta.$$

Note that a filter function $\operatorname{sinc}(c\xi)$ can be motivated as an averaging of fast forces over a time interval of length $c\tau$, see [10] and [19, Section XIII.1.4]. For $c = 1$, the new method (15) reduces to the method (14) of Grimm & Hochbruck.

2.3 A spectral Galerkin method for the discretization in space

For a full discretization of (4), we combine the trigonometric integrators of the previous section with a spectral Galerkin method in space.

We denote by

$$\mathcal{V}^K = \left\{ \sum_{j=-K}^{K} \hat{v}_j \mathrm{e}^{\mathrm{i}jx} : \hat{v}_j \in \mathbb{C} \right\}$$

the space of trigonometric polynomials of degree K and by

$$\mathcal{P}^K(v) = \sum_{j=-K}^{K} \hat{v}_j \mathrm{e}^{\mathrm{i}jx} \qquad \text{for} \qquad v = \sum_{j=-\infty}^{\infty} \hat{v}_j \mathrm{e}^{\mathrm{i}jx} \in L^2 \tag{16}$$

the L^2-orthogonal projection onto this ansatz space. In the semi-discretization in time (7) or (8), we then replace the nonlinearity $\widehat{f}$ of (6) by

$$\widehat{f}^K(u) = \mathcal{P}^K\big(\Psi_1 f^K(\Phi u)\big) \tag{17}$$

with

$$f^K(u) = a^K(u)\partial_x^2 u + g^K(u, \partial_x u), \qquad a^K = \mathcal{I}^K \circ a, \qquad g^K = \mathcal{I}^K \circ g, \tag{18}$$

where $\mathcal{I}^K$ denotes the trigonometric interpolation in the space $\mathcal{V}^K$ of trigonometric polynomials of degree K.

This gives the fully discrete method

$$\begin{aligned} u_{n+1}^K &= \cos(\tau\Omega)u_n^K + \tau\,\mathrm{sinc}(\tau\Omega)\dot{u}_n^K + \tfrac{1}{2}\tau^2\,\mathrm{sinc}(\tau\Omega)\kappa\widehat{f}^K(u_n^K), \\ \dot{u}_{n+1}^K &= -\Omega\sin(\tau\Omega)u_n^K + \cos(\tau\Omega)\dot{u}_n^K + \tfrac{1}{2}\tau\cos(\tau\Omega)\kappa\widehat{f}^K(u_n^K) + \tfrac{1}{2}\tau\kappa\widehat{f}^K(u_{n+1}^K), \end{aligned} \tag{19}$$

which computes approximations $u_n^K \in \mathcal{V}^K$ and $\dot{u}_n^K \in \mathcal{V}^K$ to $u(\cdot, t_n)$ and $\partial_t u(\cdot, t_n)$, respectively. In addition, we replace the initial values u_0 and $\dot{u}_0$ of (2) by some approximations $u_0^K \in \mathcal{V}^K$ and $\dot{u}_0^K \in \mathcal{V}^K$, computed by an L^2-orthogonal projection onto $\mathcal{V}^K$:

$$u_0^K = \mathcal{P}^K(u_0), \qquad \dot{u}_0^K = \mathcal{P}^K(\dot{u}_0).$$

We emphasize that the nonlinearity $\widehat{f}^K$ as appearing in (19) can be computed efficiently using fast Fourier techniques: The functions $a^K = \mathcal{I}^K \circ a$ and $g^K = \mathcal{I}^K \circ g$ can be computed as usual with the fast Fourier transform. The full nonlinearity $\widehat{f}^K$ of (17) can then also be computed with fast Fourier techniques, even though it is defined via projection instead of trigonometric interpolation. This is based on the observation that the argument of the projection $\mathcal{P}^K$ in (17) as appearing in (19) is a trigonometric polynomial of degree $2K$, and hence

$$\widehat{f}^K(v^K) = \mathcal{P}^K\Big(\mathcal{I}^{2K}\big(\Psi_1 f^K(\Phi v^K)\big)\Big)$$

with the trigonometric interpolation $\mathcal{I}^{2K}$ in the larger space $\mathcal{V}^{2K}$ of trigonometric polynomials of degree $2K$.

3 Statement of global error bounds

In this section, we state our error bounds for the trigonometric integrator (7) and its fully discrete version (19) when applied to the quasilinear wave equation (1).

We will universally require Assumptions 1–3 on the filter functions of the trigonometric integrator (7). In addition, we will require that the exact solution $u(x,t)$ to (1) satisfies the following assumption.

Assumption 4. Let $s \geq 0$. We assume that the exact solution $(u(\cdot,t), \partial_t u(\cdot,t))$ to (1) is in $H^{5+s} \times H^{4+s}$ with

$$\big|\!\big|\!\big| \big(u(\cdot,t), \partial_t u(\cdot,t)\big) \big|\!\big|\!\big|_{4+s} \leq M \qquad \text{for} \qquad 0 \leq t \leq T \tag{20}$$

such that

$$1 + \kappa\, a(u(\cdot,t)) \geq \delta > 0 \qquad \text{for} \qquad 0 \leq t \leq T \tag{21}$$

and

$$\kappa\, a(u(\cdot,t)) \leq A_0 \qquad \text{for} \qquad 0 \leq t \leq T \tag{22}$$

for some constants $0 < \delta < 1$, $M > 0$ and $A_0 \geq 0$.

Remark 3.1. The restriction (21) in Assumption 4 is a natural assumption coming from the analysis of the equation. It ensures the hyperbolic character of the equation. By local well-posedness theory, the regularity assumption (20) (which implies (22)) on the exact solution holds locally in time for initial values in $H^{5+s} \times H^{4+s}$, see [23, 24, 31]. The time-scale of our numerical analysis is the time-scale where a solution to (1) actually exists and stays bounded.

We are now ready to state the main result for the semi-discretization in time (7) whose proof is given in Section 4 below.

Theorem 3.2 (Error bound for the semi-discretization in time)**.** *Fix $M > 0$, $T > 0$, $c_0 \geq 0$, $0 < \delta < 1$ and $A_0 \geq 0$. Then, there exists a positive constant τ_0 such that, for all time step-sizes $\tau \leq \tau_0$, the following global error bound holds for the time-discrete numerical solution $(u_n, \dot{u}_n)$ of (7).*

If the exact solution $(u(\cdot,t), \partial_t u(\cdot,t))$ satisfies Assumption 4 for $s = 0$ with constants M, T, δ and A_0, and if the filter functions in (7) satisfy Assumption 1–3 with constants c_0, δ and A_0, then we have in $H^2 \times H^1$ the global error bound

$$\left|\!\left|\!\left| \big(u_n, \dot{u}_n\big) - \big(u(\cdot,t_n), \partial_t u(\cdot,t_n)\big) \right|\!\right|\!\right|_1 \leq C\tau^2 \qquad \text{for} \qquad 0 \leq t_n = n\tau \leq T.$$

The constant C is of the form $C = C' \mathrm{e}^{C'|\kappa|T}$ with C' depending on $\max(1, |\kappa|)$ with the coefficient κ in (1), the smooth functions a and g in (1), the constant c_0 of (10), the constants δ and A_0 of (12), (21) and (22) and M from (20), but it is independent of the time step-size τ, the final time T and the parameter κ in (1).

Remark 3.3. For small nonlinearities with $|\kappa| \ll 1$, we need only Assumptions 1 and 2 on the filter functions of the trigonometric integrator (7) to prove the global error bound, as explained in Remark 2.1. In this case, the necessary energy estimates can be proved and bounded in a simpler fashion and the underlying ellipticity of the second order operator is more easily proved on long time scales. We will give some indication of these simplifications in Section 4.

For the fully discrete trigonometric integrator (19), we will prove in Section 5 the following global error bound.

Theorem 3.4 (Error bound for the full discretization)**.** *Fix $M > 0$, $T > 0$, $c_0 \geq 0$, $0 < \delta < 1$, $A_0 \geq 0$ and $s \geq 0$. Then, there exists a positive constant τ_0 such that, for all time step-sizes $\tau \leq \tau_0$, the following global error bound holds for the fully discrete numerical solution $(u_n^K, \dot{u}_n^K)$ of (19).*

If the exact solution $(u(\cdot,t), \partial_t u(\cdot,t))$ satisfies Assumption 4 for the above s with constants M, T, δ and A_0, and if the filter functions in (7) satisfy Assumption 1–3 with constants c_0, δ and A_0, then we have in $H^2 \times H^1$ the global error bound

$$\left|\!\left|\!\left| \big(u_n^K, \dot{u}_n^K\big) - \big(u(\cdot,t_n), \partial_t u(\cdot,t_n)\big) \right|\!\right|\!\right|_1 \leq C\tau^2 + CK^{-2-s} \qquad \text{for} \qquad 0 \leq t_n = n\tau \leq T.$$

The constant C is of the same form as in Theorem 3.2 with C' depending in addition on s.

The convergence rate τ^2 in Theorem 3.4 for the discretization in time is optimal as will be shown in the following numerical examples. It is not clear whether the convergence rate K^{-2-s} for the discretization in space is also optimal under the given regularity assumption. In fact, numerical experiments suggest that the error behaves like K^{-3-s} almost uniformly in the time step-size.

In the following numerical examples, we consider the trigonometric integrator (7) (or (19)) with

- no additional filter functions, i.e., $\phi = \psi_1 = 1$, which is known as impulse method or method of Deuflhard,
- filter functions (13), which is the method of Hairer & Lubich and coincides with the new method (15) for $c = 0$,

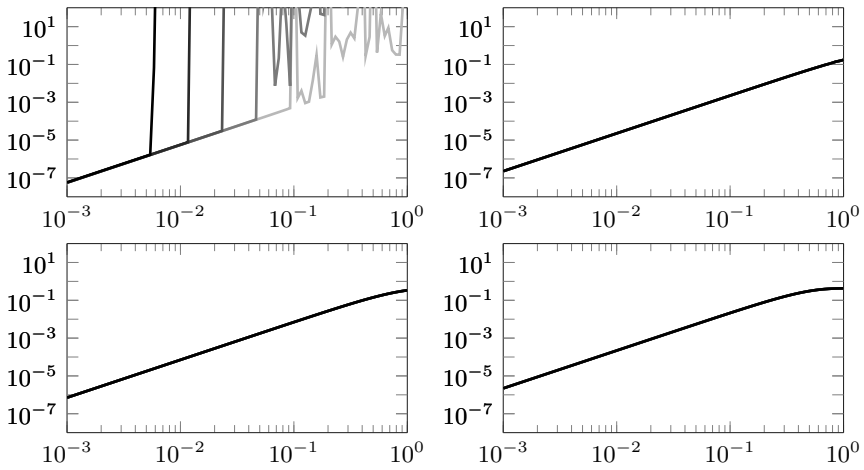

Figure 1: Error in $H^2 \times H^1$ at time 100 vs. time step-size τ for a small nonlinearity ($\kappa = 1/100$). The methods are the impulse method (top left), the method (13) of Hairer & Lubich (top right), the method (14) of Grimm & Hochbruck (bottom left) and our new method (15) with $c = 2$ (bottom right). Different lines correspond to different values of the discretization parameter $K = 2^5, 2^6, 2^7, 2^8, 2^9$, with darker lines for larger values of K.

- filter functions (14), which is the method of Grimm & Hochbruck and coincides with the new method (15) for $c = 1$,
- filter functions (15) with $c = 2$ and $c = 3$, which is the new method proposed in this paper for non-small nonlinearities.

The specific quasilinear wave equation that we consider is (1) with $a(u) = u$ and $g(u, \partial_x u) = (\partial_x u)^2 + \kappa u^3$:

$$\partial_t^2 u = \partial_x^2 u - u + \kappa u \partial_x^2 u + \kappa (\partial_x u)^2 + \kappa^2 u^3. \tag{23}$$

This is the model problem of [5]. As initial values we consider rather artificially

$$u(x,0) = \sum_{j\in\mathbb{Z}} \frac{1}{\sqrt{1+|j|^{11+1/50}}} \mathrm{e}^{\mathrm{i}jx}, \qquad \partial_t u(x,0) = \sum_{j\in\mathbb{Z}} \frac{1}{\sqrt{1+|j|^{9+1/50}}} \mathrm{e}^{\mathrm{i}jx}.$$

For this choice, the initial values are in $H^5 \times H^4$, but not in $H^{5+\sigma} \times H^{4+\sigma}$ for $\sigma \geq 1/100$, so that the initial values just don't fail to satisfy the regularity assumption (20) for $s = 0$.

Example 3.5 (Small nonlinearity)**.** We consider equation (23) with a small nonlinearity as in [5]. We choose $\kappa = 1/100$, and we consider correspondingly a long time interval of length $\kappa^{-1} = 100$. The error in $H^2 \times H^1$ of various trigonometric integrators at time $t = 100$ has been plotted in Figure 1. In the plots, only the temporal error has been taken into account by comparing the numerical solution to a reference solution with the same spatial discretization parameter.

For the method (13) of Hairer & Lubich, the method (14) of Grimm & Hochbruck and the new method (15) with $c = 2$, we observe second-order convergence in time uniformly in the spatial discretization parameter (the different lines corresponding to different spatial discretization parameters are all on top of each other). Note that the filter functions of these methods satisfy Assumptions 1 and 2 of Theorems 3.2

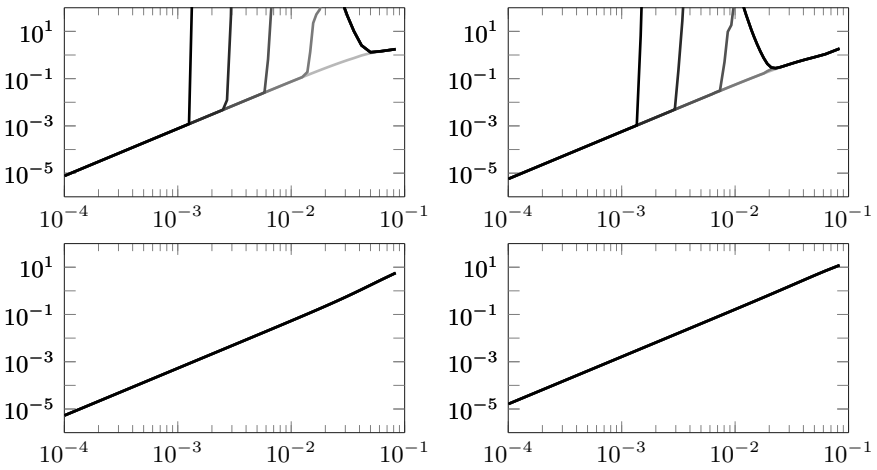

Figure 2: Error in $H^2 \times H^1$ at time $\frac{1}{4}$ vs. time step-size τ for a non-small nonlinearity ($\kappa = 1$). The methods are the method (13) of Hairer & Lubich (top left), the method (14) of Grimm & Hochbruck (top right) and our new method (15) with $c = 2$ (bottom left) and $c = 3$ (bottom right). Different lines correspond to different values of the discretization parameter $K = 2^5, 2^6, 2^7, 2^8, 2^9$, with darker lines for larger values of K.

and 3.4 and that Assumption 3 is an empty condition for small κ (see Remark 2.1). The observed convergence of these methods can thus be explained with Theorems 3.2 and 3.4.

For the impulse method, whose filter functions don't satisfy Assumption 2, we observe second-order convergence only for time step-sizes that are sufficiently small compared to the inverse of the spatial discretization parameter K. This observation is explained, for small $|\kappa|$, in a previous version of this paper [12, Section 5.2]. It is a quasilinear phenomenon which is not present in the semilinear case [11].

Example 3.6 (Non-small nonlinearity)**.** We consider again equation (23), but now with a non-small nonlinearity with $\kappa = 1$ on a time interval of length $\frac{1}{4}$. Numerical experiments suggest that the exact solution develops a singularity slightly beyond this time interval, and that $u = \kappa\, a(u)$ (which appears in assumption (22)) is bounded on this time interval by $A_0 = 13$. The error in $H^2 \times H^1$ at time $t = \frac{1}{4}$ of the methods has been plotted in Figure 2.

For the new method (15) with $c = 2$ and $c = 3$, we observe second-order convergence in time. These methods satisfy Assumption 1 and 2, but they also satisfy the additional Assumption 3 for non-small nonlinearities with the relevant value $A_0 = 13$ (this follows from Remark 2.2 since $2 > \frac{1}{2}\sqrt{A_0}$). The observed convergence of the new methods can thus be explained with Theorems 3.2 and 3.4. In practice, one will choose filter functions ϕ as in (15) with a value of c as small as possible (because the error constant deteriorates as c grows), and one will possibly adapt the value of c in the course of the computation (depending on the size of the numerical solution).

The methods (13) of Hairer & Lubich and (14) of Grimm & Hochbruck don't satisfy Assumption 3 with the necessary value $A_0 = 13$, although their filters are of the form (15) of Remark 2.2, but with a too small value $c = 0$ and $c = 1$, respectively. For these methods, the observed convergence is not uniform in the spatial discretization parameter K.

For additional numerical examples in connection with the questions studied in

[4, 5, 9, 17], we refer to a previous version of this paper [12, Section 3.5].

4 Proof of the error bound for the semi-discretization in time

In this section, we give the proof of Theorem 3.2 on the global error of the trigonometric integrator (7) applied to the quasilinear wave equation (1) without discretization in space. In the proof, we restrict to the case $g \equiv 0$ in (1), i.e.,

$$f(u) = a(u)\partial_x^2 u, \qquad \widehat{f}(u) = \Psi_1\big(a(\Phi u)(\Phi\partial_x^2 u)\big) \tag{24}$$

in the notation (5) and (6). Since the quasilinear term $a(u)\partial_x^2 u$ is the most critical part of the nonlinearity, the extension to nonzero g is rather straightforward and we will comment throughout on the necessary modifications to take $g \neq 0$.

Throughout the proof, we denote by C a generic constant that may depend on a, an upper bound $\max(1, |\kappa|)$ on the absolute value of the coefficient κ in (1) (but not on a lower bound), the order of the Sobolev space under consideration and on the constants c_0, δ and A_0 of (10) and (12). Additional dependencies of C are denoted by lower indices, e.g., C_M with M from (20).

4.1 Basic estimates

The estimates (3) and the smoothness of a imply the following fundamental properties of the nonlinearity f of (24): We have, for $s \geq 0$ and $u, v \in H^{s+2}$,

$$\|f(u)\|_s \leq \Lambda_s(\|u\|_\sigma)\|u\|_\sigma\|u\|_{s+2} \qquad \text{with} \qquad \sigma = \max(s, 1) \tag{25}$$

and the Lipschitz property

$$\|f(u) - f(v)\|_s \leq \Lambda_s\big(\|u\|_{s+2} + \|v\|_{s+2}\big)\big(\|u\|_{s+2} + \|v\|_{s+2}\big)\|u - v\|_{s+2}, \tag{26}$$

where $\Lambda_s(\cdot)$ is a continuous non-decreasing function.

Throughout the proof of Theorem 3.2, we make use of the fact that the numerical flow φ_τ given by (7) maps $H^2 \times H^1$ to itself and more generally $H^{s+1} \times H^s$ to itself for $s \geq 1$, as stated in the following lemma. This property of an explicit numerical method is in the quasilinear case by no means natural. It can be shown here using the smoothing properties of filter functions that satisfy (11).

Lemma 4.1 (Bounds for a single time step)**.** *Let $s \geq 1$, and let the filter functions satisfy Assumptions 1 and 2. For a numerical solution $(u_n, \dot{u}_n) \in H^{s+1} \times H^s$ with*

$$|||(u_n, \dot{u}_n)|||_s \leq M$$

we have $(u_{n+1}, \dot{u}_{n+1}) \in H^{s+1} \times H^s$ with

$$|||(u_{n+1}, \dot{u}_{n+1})|||_s \leq C_M.$$

Proof. We consider the method in the form (8). In this formulation, we use that

$$\tau\|\mathrm{sinc}(\tau\Omega)u\|_{s+1} = \|\Omega^{-1}\sin(\tau\Omega)u\|_{s+1} = \|\sin(\tau\Omega)u\|_s \leq \|u\|_s. \tag{27}$$

and that

$$\tau\|\widehat{f}(u)\|_s = \tau\|\Psi f(\Phi u)\|_s \leq \|\Phi f(\Phi u)\|_{s-1} \leq \Lambda_{s-1}(\|u\|_{s+1})\|u\|_{s+1}^2. \tag{28}$$

The first estimate of (28) follows from (11) and (27) with s instead of $s+1$ and the second one from (10) and (25). These properties yield the claimed bound on the numerical solution, first for u_{n+1} and with this also for $\dot{u}_{n+1}$. □

In the same way, but using the Lipschitz property (26) instead of (25), we can derive the following estimate for the difference of two numerical solutions (7).

Lemma 4.2 (Stability of a single time step)**.** *Let $s \geq 1$, and let the filter functions satisfy Assumptions 1 and 2. For numerical solutions $(u_n, \dot{u}_n) \in H^{s+1} \times H^s$ and $(v_n, \dot{v}_n) \in H^{s+1} \times H^s$ with*

$$|||(u_n, \dot{u}_n)|||_s \leq M \qquad \text{and} \qquad |||(v_n, \dot{v}_n)|||_s \leq M$$

we have

$$\big|\big|\big|(u_{n+1}, \dot{u}_{n+1}) - (v_{n+1}, \dot{v}_{n+1})\big|\big|\big|_s \leq C_M \big|\big|\big|(u_n, \dot{u}_n) - (v_n, \dot{v}_n)\big|\big|\big|_s. \qquad \square$$

Remark 4.3. The basic estimates (25) and (26) extend directly to a nonzero g in (1), and hence also the statements of Lemmas 4.1 and 4.2.

4.2 Outline of the proof of Theorem 3.2

Lemmas 4.1 and in particular Lemma 4.2 illustrate the difficulties that are encountered when trying to prove error bounds, say in $H^2 \times H^1$. Denoting by

$$(e_{n+1}, \dot{e}_{n+1}) = (u_{n+1}, \dot{u}_{n+1}) - (u(\cdot, t_{n+1}), \partial_t u(\cdot, t_{n+1})) \tag{29}$$

the global error of (7) after n time steps, and denoting, with the numerical flow φ_τ given by (9), by

$$(d_{n+1}, \dot{d}_{n+1}) = \varphi_\tau\big(u(\cdot, t_n), \partial_t u(\cdot, t_n)\big) - \big(u(\cdot, t_{n+1}), \partial_t u(\cdot, t_{n+1})\big) \tag{30}$$

the local error when starting at $(u(\cdot, t_n), \partial_t u(\cdot, t_n))$, one routinely decomposes the global error in $H^2 \times H^1$ as

$$|||(e_{n+1}, \dot{e}_{n+1})|||_1 \leq \big|\big|\big|\varphi_\tau\big(u_n, \dot{u}_n\big) - \varphi_\tau\big(u(\cdot, t_n), \partial_t u(\cdot, t_n)\big)\big|\big|\big|_1 + |||(d_{n+1}, \dot{d}_{n+1})|||_1.$$

By analyzing the error propagation of the method (stability of the method), one then aims for estimating the first term on the right-hand side by $\mathrm{e}^{C\tau}|||(e_n, \dot{e}_n)|||_1$. The stability estimate of Lemma 4.2, however, only yields a factor C_M instead of $\mathrm{e}^{C\tau}$, which makes this approach failing.

Our approach here is to replace the $H^2 \times H^1$-norm by a different but related measure for the error that allows us to prove a suitable stability estimate. This measure isn't a norm, and it depends on time. Its definition is inspired by energy techniques as used to analyze the exact solution: We introduce in Section 4.3 an energy-type functional $\mathcal{E}\colon H^2 \times H^1 \times H^2 \to \mathbb{R}$, and we will then use

$$\mathcal{E}_n(e_n, \dot{e}_n) = \mathcal{E}(e_n, \dot{e}_n, u_n)$$

instead of $|||(e_n, \dot{e}_n)|||_1$ as a measure for the global error $(e_n, \dot{e}_n)$.

The error accumulation in this quantity reads

$$\begin{aligned} \mathcal{E}_{n+1}(e_{n+1}, \dot{e}_{n+1}) = \mathcal{E}_{n+1}\Big(\varphi_\tau\big(u_n, \dot{u}_n\big) - \varphi_\tau\big(u(\cdot, t_n), \partial_t u(\cdot, t_n)\big)\Big) \\ + \Big(\mathcal{E}_{n+1}(e_{n+1}, \dot{e}_{n+1}) - \mathcal{E}_{n+1}\big(e_{n+1} - d_{n+1}, \dot{e}_{n+1} - \dot{d}_{n+1}\big)\Big). \end{aligned} \tag{31}$$

The difference in the second line of (31) accounts for the local error of the method. It is estimated in Section 4.5 below by adapting the proof for the semilinear case to the quasilinear case. The term in the first line is $\mathcal{E}_{n+1}$ evaluated at a difference of two numerical solutions, and hence describes the error propagation of the method. It turns out, in Section 4.3 below, that we can prove a suitable estimate for the error propagation in the quantity $\mathcal{E}$. The relation of $\mathcal{E}$ to the $H^2 \times H^1$-norm is then described in Section 4.4 below. Finally, the error accumulation is studied in Section 4.6 below.

4.3 Stability of the numerical method

A key step in the proof of Theorem 3.2 is to establish stability of the numerical method (7) in a suitable sense.

We introduce the energy-type quantity

$$\mathcal{E}(e, \dot{e}, u) = |||(e, \dot{e})|||_1^2 + \kappa \mathcal{U}\big(\Phi e, \Phi u\big), \tag{32}$$

where

$$\mathcal{U}(e, u) = \big\langle \cos(\tau\Omega)\partial_x^2 e, a(u)\partial_x^2 e\big\rangle_0 - \tfrac{1}{4}\tau^2\kappa \big\|\Psi_1\big(a(u)\partial_x^2 e\big)\big\|_1^2. \tag{33}$$

Up to the non-quadratic term $\mathcal{U}$, this energy $\mathcal{E}$ is essentially the $H^2 \times H^1$-norm of $(e, \dot{e})$. Under the Assumptions 1 and 2, the energy $\mathcal{E}$ is well-defined for $(e, \dot{e}) \in H^2 \times H^1$ and $u \in H^2$. This follows from the Cauchy–Schwarz inequality and (3) applied to the first term of $\mathcal{U}$ and from assumptions (10) and (11) applied as in (28) to the second term of $\mathcal{U}$.

The motivation to define the energy as in (32) is the calculation in the proof of the following lemma, where we compute the change in the energy along differences of numerical solutions.

Lemma 4.4 (Change in the energy)**.** *Let the filter functions satisfy Assumptions 1 and 2. For numerical solutions $(u_n, \dot{u}_n) \in H^2 \times H^1$ and $(v_n, \dot{v}_n) \in H^2 \times H^1$ we then have*

$$\begin{aligned}\mathcal{E}\big(u_{n+1} - v_{n+1}, \dot{u}_{n+1} - \dot{v}_{n+1}, u_{n+1}\big)& \\ = \mathcal{E}\big(u_n - v_n, \dot{u}_n - \dot{v}_n, u_n\big) + \kappa\, \mathcal{R}\big(\Phi u_{n+1}, \Phi u_n, \Phi v_{n+1}, \Phi v_n\big)&\end{aligned}$$

with the remainder

$$\mathcal{R}\big(u, u', v, v'\big) = \widetilde{\mathcal{R}}\big(u, u', v, v'\big) + \mathcal{R}^*\big(u, v\big) - \mathcal{R}^*\big(u', v'\big), \tag{34}$$

where

$$\widetilde{\mathcal{R}}\big(u, u', v, v'\big) = \big\langle u - v, f(u') - f(v')\big\rangle_1 - \big\langle u' - v', f(u) - f(v)\big\rangle_1 \tag{35}$$

and

$$\begin{aligned}\mathcal{R}^*(u, v) = \big\langle \cos(\tau\Omega)(u - v), a(u)\partial_x^2(u - v)\big\rangle_0 + \big\langle \cos(\tau\Omega)(u - v), \big(a(u) - a(v)\big)\partial_x^2 v\big\rangle_1& \\ + \tfrac{1}{2}\tau^2\kappa \big\langle \Psi_1\big(a(u)\partial_x^2(u - v)\big), \Psi_1\big((a(u) - a(v))\partial_x^2 v\big)\big\rangle_1& \\ + \tfrac{1}{4}\tau^2\kappa \big\|\Psi_1\big((a(u) - a(v))\partial_x^2 v\big)\big\|_1^2.&\end{aligned}$$

Proof. By the structure of the "matrix" in the second step of the method (7), we have

$$\|\Omega(u_{n+1} - v_{n+1})\|_1^2 + \|\dot{u}_{n+1}^- - \dot{v}_{n+1}^-\|_1^2 = \|\Omega(u_n - v_n)\|_1^2 + \|\dot{u}_n^+ - \dot{v}_n^+\|_1^2.$$

Hence, taking also the first and third step of the method into account, we get

$$\begin{aligned}\|\Omega(u_{n+1}-v_{n+1})\|_1^2 + \big\|\big(\dot u_{n+1}-\dot v_{n+1}\big) - \tfrac12\tau\kappa\big(\widehat f(u_{n+1})-\widehat f(v_{n+1})\big)\big\|_1^2 \\ = \|\Omega(u_n-v_n)\|_1^2 + \big\|\big(\dot u_n-\dot v_n\big) + \tfrac12\tau\kappa\big(\widehat f(u_n)-\widehat f(v_n)\big)\big\|_1^2.\end{aligned}$$

We then expand the second norm on the left and on the right. In the resulting mixed terms, we use the property

$$\langle v, \Psi_1 w\rangle_1 = \langle \Psi_1 v, w\rangle_1 = \langle \operatorname{sinc}(\tau\Omega)\Phi v, w\rangle_1, \tag{36}$$

which follows from Parseval's theorem and assumption (11), and we replace the resulting differences $\tau\operatorname{sinc}(\tau\Omega)(\dot u_{n+1}-\dot v_{n+1})$ and $\tau\operatorname{sinc}(\tau\Omega)(\dot u_n-\dot v_n)$ with the help of the relations

$$\begin{aligned}\tau\operatorname{sinc}(\tau\Omega)\dot u_{n+1} &= \cos(\tau\Omega)u_{n+1} + \tfrac12\tau^2\operatorname{sinc}(\tau\Omega)\kappa\widehat f(u_{n+1}) - u_n,\\ \tau\operatorname{sinc}(\tau\Omega)\dot u_n &= -\cos(\tau\Omega)u_n - \tfrac12\tau^2\operatorname{sinc}(\tau\Omega)\kappa\widehat f(u_n) + u_{n+1}\end{aligned}$$

and the same relations for v. The second of these relations is taken from (8), and the first one can be derived from the first one by the symmetry of the method (or from the numerical method in the form (7) by expressing u_n in terms of u_{n+1} and $\dot u_{n+1}$). Using again (36), the definition of the remainder $\mathcal R$ and $|||(e,\dot e)|||_1^2 = \|\Omega e\|_1^2 + \|\dot e\|_1^2$, this yields

$$\begin{aligned}\big|\!\big|\!\big|\big(u_{n+1}-v_{n+1}, \dot u_{n+1}-\dot v_{n+1}\big)\big|\!\big|\!\big|_1^2 + \kappa\widetilde{\mathcal U}\big(\Phi(u_{n+1}-v_{n+1}), \Phi u_{n+1}\big) \\ = \big|\!\big|\!\big|\big(u_n-v_n, \dot u_n-\dot v_n\big)\big|\!\big|\!\big|_1^2 + \kappa\widetilde{\mathcal U}\big(\Phi(u_n-v_n), \Phi u_n\big) \\ + \kappa\,\widetilde{\mathcal R}\big(\Phi u_{n+1}, \Phi u_n, \Phi v_{n+1}, \Phi v_n\big)\end{aligned}$$

with $\widetilde{\mathcal R}$ from (35) and

$$\widetilde{\mathcal U}(e,u) = -\big\langle \cos(\tau\Omega)e, f(u)-f(u-e)\big\rangle_1 - \tfrac14\tau^2\kappa\,\big\|\Psi_1\big(f(u)-f(u-e)\big)\big\|_1^2.$$

The statement of the lemma follows by setting

$$\mathcal R^*(u,v) = \mathcal U(u-v,u) - \widetilde{\mathcal U}(u-v,u).$$

To get the final form of $\mathcal R^*$, we use

$$f(u)-f(v) = a(u)\partial_x^2(u-v) + (a(u)-a(v))\partial_x^2 v$$

and $\langle\cdot,\cdot\rangle_1 = \langle\cdot,\cdot\rangle_0 - \langle\partial_x^2\cdot,\cdot\rangle_0$. □

We now estimate the remainder $\mathcal R$ of Lemma 4.4, which describes the change in the energy along numerical solutions. The crucial observation is that we gain a factor τ without requiring more regularity than $H^2\times H^1$ of the difference of the corresponding numerical solutions.

Lemma 4.5 (Bound of the change $\mathcal R$ in the energy)**.** *Let the filter functions satisfy Assumptions 1 and 2. For numerical solutions $(u_n,\dot u_n)\in H^2\times H^1$ and $(v_n,\dot v_n)\in H^3\times H^2$ with*

$$|||(u_n,\dot u_n)|||_1 \le M \qquad \text{and} \qquad |||(v_n,\dot v_n)|||_2 \le M$$

we have for the remainder $\mathcal R$ of Lemma 4.4 the bound

$$\big|\mathcal R\big(\Phi u_{n+1}, \Phi u_n, \Phi v_{n+1}, \Phi v_n\big)\big| \le C_M\tau\big|\!\big|\!\big|(u_n,\dot u_n)-(v_n,\dot v_n)\big|\!\big|\!\big|_1^2.$$

Proof. The main task is to get the factor τ in the estimate. This is done with the observation that we have

$$\|u_{n+1} - u_n\|_1 \leq C_M \tau, \tag{37}$$

which follows from (8) using the property (28) with $s = 1$ and $\|(\cos(\tau\Omega) - 1)u_n\|_1 = 2\|\sin(\frac{1}{2}\tau\Omega)^2 u_n\|_1 \leq \|\tau\Omega u_n\|_1 = \tau\|u_n\|_2$. Similarly, we have

$$\|v_{n+1} - v_n\|_2 \leq C_M \tau \tag{38}$$

by the higher regularity of $(v_n, \dot{v}_n)$ and also

$$\|u_{n+1} - u_n - (v_{n+1} - v_n)\|_1 \leq C_M \tau \big|\!\big|\!\big|(u_n, \dot{u}_n) - (v_n, \dot{v}_n)\big|\!\big|\!\big|_1. \tag{39}$$

Under the given regularity assumptions, the differences $u_{n+1} - u_n$ in H^1 and $v_{n+1} - v_n$ in H^2 thus allow us to gain a factor τ.

Our goal is therefore to recover in the remainder $\mathcal{R}$ defined in (34) such differences. In the following we set

$$\widehat{u}_n = \Phi u_n, \quad \widehat{v}_n = \Phi v_n. \tag{40}$$

In this notation and using that $\langle \cdot, \cdot \rangle_1 = \langle \cdot, \cdot \rangle_0 + \langle \partial_x \cdot, \partial_x \cdot \rangle_0$ we can express the remainder $\mathcal{R}$ in the following way: We have

$$\mathcal{R}\big(\widehat{u}_{n+1}, \widehat{u}_n, \widehat{v}_{n+1}, \widehat{v}_n\big) = \mathcal{R}_0 + \mathcal{R}_1 + \big(\mathcal{R}^*\big(\widehat{u}_{n+1}, \widehat{v}_{n+1}\big) - \mathcal{R}^*\big(\widehat{u}_n, \widehat{v}_n\big)\big),$$

where

$$\begin{aligned}\mathcal{R}_0 = \big\langle \widehat{u}_{n+1} - \widehat{v}_{n+1}, a(\widehat{u}_n)\partial_x^2 \widehat{u}_n - a(\widehat{v}_n)\partial_x^2 \widehat{v}_n \big\rangle_0 \\ - \big\langle \widehat{u}_n - \widehat{v}_n, a(\widehat{u}_{n+1})\partial_x^2 \widehat{u}_{n+1} - a(\widehat{v}_{n+1})\partial_x^2 \widehat{v}_{n+1} \big\rangle_0\end{aligned}$$

and

$$\begin{aligned}\mathcal{R}_1 = \big\langle \partial_x(\widehat{u}_{n+1} - \widehat{v}_{n+1}), \partial_x \big(a(\widehat{u}_n)\partial_x^2 \widehat{u}_n - a(\widehat{v}_n)\partial_x^2 \widehat{v}_n\big) \big\rangle_0 \\ - \big\langle \partial_x(\widehat{u}_n - \widehat{v}_n), \partial_x \big(a(\widehat{u}_{n+1})\partial_x^2 \widehat{u}_{n+1} - a(\widehat{v}_{n+1})\partial_x^2 \widehat{v}_{n+1}\big) \big\rangle_0.\end{aligned}$$

With the aid of integration by parts and adding zeroes we obtain that

$$\begin{aligned}\mathcal{R}_1 = -\big\langle \partial_x^2\big(\widehat{u}_{n+1} - \widehat{v}_{n+1}\big), a(\widehat{u}_n)\partial_x^2\big(\widehat{u}_n - \widehat{v}_n\big) + \big(a(\widehat{u}_n) - a(\widehat{v}_n)\big)\partial_x^2 \widehat{v}_n \big\rangle_0 \\ + \big\langle \partial_x^2\big(\widehat{u}_n - \widehat{v}_n\big), a(\widehat{u}_{n+1})\partial_x^2\big(\widehat{u}_{n+1} - \widehat{v}_{n+1}\big) + \big(a(\widehat{u}_{n+1}) - a(\widehat{v}_{n+1})\big)\partial_x^2 \widehat{v}_{n+1} \big\rangle_0.\end{aligned}$$

Note that by symmetry we have for the first term that

$$\big\langle \partial_x^2\big(\widehat{u}_{n+1} - \widehat{v}_{n+1}\big), a(\widehat{u}_n)\partial_x^2\big(\widehat{u}_n - \widehat{v}_n\big) \big\rangle_0 = \big\langle \partial_x^2\big(\widehat{u}_n - \widehat{v}_n\big), a(\widehat{u}_n)\partial_x^2\big(\widehat{u}_{n+1} - \widehat{v}_{n+1}\big) \big\rangle_0$$

which we can combine with the first term in the second row, i.e.,

$$\begin{aligned}\mathcal{R}_1 = \big\langle \partial_x^2\big(\widehat{u}_{n+1} - \widehat{v}_{n+1}\big), \big(a(\widehat{u}_{n+1}) - a(\widehat{u}_n)\big)\partial_x^2\big(\widehat{u}_n - \widehat{v}_n\big) \big\rangle_0 \\ - \big\langle \partial_x^2\big(\widehat{u}_{n+1} - \widehat{v}_{n+1}\big), \big(a(\widehat{u}_n) - a(\widehat{v}_n)\big)\partial_x^2 \widehat{v}_n \big\rangle_0 \\ + \big\langle \partial_x^2\big(\widehat{u}_n - \widehat{v}_n\big), \big(a(\widehat{u}_{n+1}) - a(\widehat{v}_{n+1})\big)\partial_x^2 \widehat{v}_{n+1} \big\rangle_0.\end{aligned}$$

Adding and subtracting the term $\langle \partial_x^2(\widehat{u}_n - \widehat{v}_n), (a(\widehat{u}_n) - a(\widehat{v}_n))\partial_x^2 \widehat{v}_n \rangle_0$ (and combining it with the term in the second row) furthermore yields that

$$\begin{aligned}\mathcal{R}_1 = {} & \langle \partial_x^2(\widehat{u}_{n+1} - \widehat{v}_{n+1}), (a(\widehat{u}_{n+1}) - a(\widehat{u}_n))\partial_x^2(\widehat{u}_n - \widehat{v}_n)\rangle_0 \\ & + \langle \partial_x^2(\widehat{u}_n - \widehat{v}_n - (\widehat{u}_{n+1} - \widehat{v}_{n+1})), (a(\widehat{u}_n) - a(\widehat{v}_n))\partial_x^2 \widehat{v}_n\rangle_0 \\ & - \langle \partial_x^2(\widehat{u}_n - \widehat{v}_n), (a(\widehat{u}_n) - a(\widehat{v}_n))\partial_x^2 \widehat{v}_n\rangle_0 \\ & + \langle \partial_x^2(\widehat{u}_n - \widehat{v}_n), (a(\widehat{u}_{n+1}) - a(\widehat{v}_{n+1}))\partial_x^2 \widehat{v}_{n+1}\rangle_0.\end{aligned}$$

Finally, adding and subtracting the term $\langle \partial_x^2(\widehat{u}_n - \widehat{v}_n), (a(\widehat{u}_{n+1}) - a(\widehat{v}_{n+1}))\partial_x^2 \widehat{v}_n \rangle_0$ (and combining it with the terms in the last two rows) and using integration by parts on the term in the second row, we obtain that

$$\begin{aligned}\mathcal{R}_1 = {} & \langle \partial_x^2(\widehat{u}_{n+1} - \widehat{v}_{n+1}), (a(\widehat{u}_{n+1}) - a(\widehat{u}_n))\partial_x^2(\widehat{u}_n - \widehat{v}_n)\rangle_0 \\ & - \langle \partial_x(\widehat{u}_n - \widehat{v}_n - (\widehat{u}_{n+1} - \widehat{v}_{n+1})), \partial_x\left((a(\widehat{u}_n) - a(\widehat{v}_n))\partial_x^2 \widehat{v}_n\right)\rangle_0 \\ & + \langle \partial_x^2(\widehat{u}_n - \widehat{v}_n), (a(\widehat{u}_{n+1}) - a(\widehat{v}_{n+1}) - (a(\widehat{u}_n) - a(\widehat{v}_n)))\partial_x^2 \widehat{v}_n\rangle_0 \\ & + \langle \partial_x^2(\widehat{u}_n - \widehat{v}_n), (a(\widehat{u}_{n+1}) - a(\widehat{v}_{n+1}))\partial_x^2(\widehat{v}_{n+1} - \widehat{v}_n)\rangle_0.\end{aligned}$$

With the aid of the bilinear estimates (3) we may thus bound the remainder $\mathcal{R}_1$ as follows: We have

$$\begin{aligned}|\mathcal{R}_1| \leq {} & C\|\widehat{u}_{n+1} - \widehat{v}_{n+1}\|_2 \|a(\widehat{u}_{n+1}) - a(\widehat{u}_n)\|_1 \|\widehat{u}_n - \widehat{v}_n\|_2 \\ & + C\|\widehat{u}_{n+1} - \widehat{v}_{n+1} - (\widehat{u}_n - \widehat{v}_n)\|_1 \|a(\widehat{u}_n) - a(\widehat{v}_n)\|_1 \|\widehat{v}_n\|_3 \\ & + C\|\widehat{u}_n - \widehat{v}_n\|_2 \|a(\widehat{u}_{n+1}) - a(\widehat{v}_{n+1}) - (a(\widehat{u}_n) - a(\widehat{v}_n))\|_1 \|\widehat{v}_n\|_2 \\ & + C\|\widehat{u}_n - \widehat{v}_n\|_2 \|a(\widehat{u}_{n+1}) - a(\widehat{v}_{n+1})\|_1 \|\widehat{v}_{n+1} - \widehat{v}_n\|_2.\end{aligned} \tag{41}$$

To estimate the quadruple term in a in the third line of (41), we consider the smooth function $H(u, e) = a(u + e) - a(u)$ and note that by (3) and since $H(u, 0) = 0$

$$\begin{aligned}\|H(u, e) - H(v, f)\|_1 &\leq \|H(u, e) - H(u, f)\|_1 + \|H(u, f) - H(v, f)\|_1 \\ &\leq \Lambda\big(\|u\|_1 + \|v\|_1 + \|e\|_1 + \|f\|_1\big)\big(\|e - f\|_1 + \|u - v\|_1 \|f\|_1\big)\end{aligned}$$

with a non-decreasing function $\Lambda(\cdot)$. With $u = \widehat{u}_n$, $e = \widehat{u}_{n+1} - \widehat{u}_n$, $v = \widehat{v}_n$ and $f = \widehat{v}_{n+1} - \widehat{v}_n$, this yields for the quadruple term in a in the third line of (41)

$$\begin{aligned}\big\|\big(a(\widehat{u}_{n+1}) - a(\widehat{u}_n)\big) - \big(a(\widehat{v}_{n+1}) - a(\widehat{v}_n)\big)\big\|_1 \leq {} & \Lambda\big(2\|\widehat{u}_n\|_1 + 2\|\widehat{v}_n\|_1 + \|\widehat{u}_{n+1}\|_1 + \|\widehat{v}_{n+1}\|_1\big) \\ & \cdot \Big(\big\|\widehat{u}_{n+1} - \widehat{u}_n - (\widehat{v}_{n+1} - \widehat{v}_n)\big\|_1 + \|\widehat{u}_n - \widehat{v}_n\|_1 \|\widehat{v}_{n+1} - \widehat{v}_n\|_1\Big).\end{aligned} \tag{42}$$

Thanks to the bound (10a) on the filter functions we obtain with the notation (40) that

$$\|\widehat{u}_n - \widehat{v}_n\|_s = \|\Phi(u_n - v_n)\|_s \leq \|u_n - v_n\|_s. \tag{43}$$

Plugging the bounds (42) and (43) together with the bounds on the difference $\|u_{n+1} - u_n\|_1$ given in (37), the difference $\|v_{n+1} - v_n\|_2$ given in (38) and the difference $\|u_{n+1} - u_n - (v_{n+1} - v_n)\|_1$ given in (39) as well as the bound on $\|u_{n+1} - v_{n+1}\|_2$ given in Lemmas 4.1 and 4.2 into (41) yields that

$$|\mathcal{R}_1| \leq C_M \tau \big|\!\big|\!\big|(u_n, \dot{u}_n) - (v_n, \dot{v}_n)\big|\!\big|\!\big|_1^2,$$

where we have used the given regularity assumptions (in particular that $v_n \in H^3$) and that a is a sufficiently smooth function.

Thus, the proof is completed upon computing the comparable bound on the more regular terms $\mathcal{R}_0$ and $\mathcal{R}^*(\widehat{u}_{n+1},\widehat{v}_{n+1})-\mathcal{R}^*(\widehat{u}_n,\widehat{v}_n)$ by a similar analysis. For example, the difference $\mathcal{R}^*(\widehat{u}_{n+1},\widehat{v}_{n+1})-\mathcal{R}^*(\widehat{u}_n,\widehat{v}_n)$ contains the difference

$$\begin{aligned}\mathcal{R}_1^* = \big\langle \cos(\tau\Omega)(\widehat{u}_{n+1}-\widehat{v}_{n+1}), a(\widehat{u}_{n+1})\partial_x^2(\widehat{u}_{n+1}-\widehat{v}_{n+1})\big\rangle_0 \\ - \big\langle \cos(\tau\Omega)(\widehat{u}_n-\widehat{v}_n), a(\widehat{u}_n)\partial_x^2(\widehat{u}_n-\widehat{v}_n)\big\rangle_0,\end{aligned}$$

which can be split as

$$\begin{aligned}\mathcal{R}_1^* = \big\langle \cos(\tau\Omega)\big(\widehat{u}_{n+1}-\widehat{u}_n-(\widehat{v}_{n+1}-\widehat{v}_n)\big), a(\widehat{u}_{n+1})\partial_x^2(\widehat{u}_{n+1}-\widehat{v}_{n+1})\big\rangle_0 \\ + \big\langle \cos(\tau\Omega)(\widehat{u}_n-\widehat{v}_n), \big(a(\widehat{u}_{n+1})-a(\widehat{u}_n)\big)\partial_x^2(\widehat{u}_{n+1}-\widehat{v}_{n+1})\big\rangle_0 \\ + \big\langle \cos(\tau\Omega)(\widehat{u}_n-\widehat{v}_n), a(\widehat{u}_n)\partial_x^2\big(\widehat{u}_{n+1}-\widehat{u}_n-(\widehat{v}_{n+1}-\widehat{v}_n)\big)\big\rangle_0.\end{aligned}$$

After partial integration in the last term, this can be estimated as above by

$$\begin{aligned}|\mathcal{R}_1^*| \le \|\widehat{u}_{n+1}-\widehat{u}_n-(\widehat{v}_{n+1}-\widehat{v}_n)\|_0\|a(\widehat{u}_{n+1})\|_1\|\widehat{u}_{n+1}-\widehat{v}_{n+1}\|_2 \\ + \|\widehat{u}_n-\widehat{v}_n\|_0\|a(\widehat{u}_{n+1})-a(\widehat{u}_n)\|_1\|\widehat{u}_{n+1}-\widehat{v}_{n+1}\|_2 \\ + \|\widehat{u}_n-\widehat{v}_n\|_1\|a(\widehat{u}_n)\|_1\|\widehat{u}_{n+1}-\widehat{u}_n-(\widehat{v}_{n+1}-\widehat{v}_n)\|_1,\end{aligned}$$

and hence

$$|\mathcal{R}_1^*| \le C_M\tau\big|\!\big|\!\big|(u_n,\dot{u}_n)-(v_n,\dot{v}_n)\big|\!\big|\!\big|_1^2.$$

Another exemplary term in the difference $\mathcal{R}^*(\widehat{u}_{n+1},\widehat{v}_{n+1})-\mathcal{R}^*(\widehat{u}_n,\widehat{v}_n)$ is

$$\mathcal{R}_2^* = \tfrac{1}{4}\tau^2\kappa\,\big\|\Psi_1\big((a(\widehat{u}_{n+1})-a(\widehat{v}_{n+1}))\partial_x^2\widehat{v}_{n+1}\big)\big\|_1^2,$$

for which we get with (3), (10) and Lemmas 4.1 and 4.2

$$|\mathcal{R}_2^*| \le C\tau^2\big\|a(\widehat{u}_{n+1})-a(\widehat{v}_{n+1})\big\|_1^2\|\widehat{v}_{n+1}\|_3^2 \le C_M\tau\big|\!\big|\!\big|(u_n,\dot{u}_n)-(v_n,\dot{v}_n)\big|\!\big|\!\big|_1^2. \qquad \square$$

In the situation outlined in Section 4.2, we get from Lemmas 4.4 and 4.5 the following estimate.

Proposition 4.6 (Stability)**.** *Let the filter functions satisfy Assumptions 1 and 2. If $(u,\partial_t u)$ is a solution to (4) in $H^3\times H^2$ with*

$$\big|\!\big|\!\big|\big(u(\cdot,t_n),\partial_t u(\cdot,t_n)\big)\big|\!\big|\!\big|_2 \le M,$$

and if $(u_n,\dot{u}_n)\in H^2\times H^1$ is a corresponding numerical solution with

$$|\!|\!|(u_n,\dot{u}_n)|\!|\!|_1 \le 2M,$$

then we have

$$\begin{aligned}\Big|\mathcal{E}\Big((u_{n+1},\dot{u}_{n+1})-\varphi_\tau\big(u(\cdot,t_n),\partial_t u(\cdot,t_n)\big),u_{n+1}\Big)\Big| \\ \le \big|\mathcal{E}(e_n,\dot{e}_n,u_n)\big| + C_M\tau|\kappa|\,|\!|\!|(e_n,\dot{e}_n)|\!|\!|_1^2\end{aligned}$$

with the global error $(e_n,\dot{e}_n)$ of (29).

Proof. Take $(v_n,\dot{v}_n)=(u(\cdot,t_n),\partial_t u(\cdot,t_n))$ and $(v_{n+1},\dot{v}_{n+1})=\varphi_\tau(u(\cdot,t_n),\partial_t u(\cdot,t_n))$ in Lemmas 4.4 and 4.5. $\square$

Remark 4.7. For a nonzero g in (1), the statement of Lemma 4.4 remains valid, but with a remainder $\mathcal{R}^*$ that contains additional terms with $g(u)$ instead of $a(u)\partial_x^2 u$. As $g(u)$ is more regular than $a(u)\partial_x^2 u$, the remainder estimate of Lemma 4.5 extends to these new terms, and hence Proposition 4.6 on the stability of the method also holds for nonzero g.

4.4 Controlling Sobolev norms with the energy

Our aim is to show that the energy (32) can be controlled by Sobolev norms and vice-versa. This is done by estimating the additional contribution $\kappa\mathcal{U}$ from above and below. For small $|\kappa|$, this is elementary, and the main result of this section, Proposition 4.10 below, can be derived directly from the properties (10) and (11) of the filters and (3) of Sobolev spaces using (27). For non-small κ, we have to work harder, and this is the main content of this section.

In the following, we set

$$\mathcal{L}(u) = \kappa\Phi a(u)\cos(\tau\Omega)\Phi - \tfrac{1}{4}\kappa^2\Phi a(u)\sin^2(\tau\Omega)\Phi^2 a(u)\Phi.$$

Using integration by parts and $\langle \sin(\tau\Omega)v, \sin(\tau\Omega)v\rangle_0 = \tau^2\|\mathrm{sinc}(\tau\Omega)v\|_1^2$, we obtain under the Assumption 2 that

$$\kappa\mathcal{U}(\Phi e, \Phi u) = \big\langle \mathcal{L}(\Phi u)\partial_x^2 e, \partial_x^2 e\big\rangle_0 \tag{44}$$

for the term $\mathcal{U}$ defined in (33). We have to prove an upper and a lower bound for this term, the latter being the crucial and difficult part. The key tool to prove the essential lower bound is given by the following lemma.

Lemma 4.8. *Fix $M > 0$, and let the filter functions satisfy Assumptions 1 and 3 with constants $0 < \delta < 1$ and $A_0 \geq 0$. Then, there exists $\tau_0 > 0$ such that for every $\tau \leq \tau_0$, for every $v \in L^2$ and for every $u \in H^2$ with*

$$\|u\|_2 \leq M, \qquad 1 + \kappa\, a(u(x)) \geq \tfrac{1}{2}\delta > 0, \qquad \kappa\, a(u(x)) \leq A_0 + \tfrac{1}{2}\delta,$$

we have that

$$\|v\|_0^2 + \big\langle \mathcal{L}(\Phi u)v, v\big\rangle_0 \geq \tfrac{1}{8}\delta\|v\|_0^2.$$

In order to prove the above Lemma we need the following estimate.

Lemma 4.9. *Let the filter functions satisfy Assumptions 1 and 3 with constants $0 < \delta < 1$ and $A_0 \geq 0$. We then have, for all $A \leq A_0 + \frac{1}{2}\delta$ with $1 + A \geq \frac{1}{2}\delta > 0$ and all $\xi \geq 0$,*

$$A\cos(\xi)\phi(\xi)^2 - \tfrac{1}{4}A^2\sin(\xi)^2\phi(\xi)^4 \geq -1 + \tfrac{1}{2}\delta. \tag{45}$$

Proof. We use $\cos(\xi) = \cos(\frac{1}{2}\xi)^2 - \sin(\frac{1}{2}\xi)^2$ and $\sin(\xi) = 2\cos(\frac{1}{2}\xi)\sin(\frac{1}{2}\xi)$ to rewrite (45) as

$$\big(1 + A\cos(\tfrac{1}{2}\xi)^2\phi(\xi)^2\big)\big(1 - A\sin(\tfrac{1}{2}\xi)^2\phi(\xi)^2\big) \geq \tfrac{1}{2}\delta. \tag{46}$$

As a function of A, the left-hand side of (46) is a parabola with a downwards opening or a linear function. To show that (46) holds for all $A \leq A_0 + \frac{1}{2}\delta$ with $1 + A \geq \frac{1}{2}\delta$, it thus suffices to prove it for the two boundary values $-1 + \frac{1}{2}\delta$ and $A_0 + \frac{1}{2}\delta$ of A.

For $A = A_0 + \frac{1}{2}\delta \geq 0$, we note that $1 + A\cos(\frac{1}{2}\xi)^2\phi(\xi)^2 \geq 1$ and

$$1 - A\sin(\tfrac{1}{2}\xi)^2\phi(\xi)^2 \geq 1 - A_0\sin(\tfrac{1}{2}\xi)^2\phi(\xi)^2 - \tfrac{1}{2}\delta \geq \tfrac{1}{2}\delta$$

by (10) and (12), and hence (46) holds for the right boundary value.

For $A = -1 + \frac{1}{2}\delta \leq 0$, we note that $1 - A\sin(\frac{1}{2}\xi)^2\phi(\xi)^2 \geq 1$ and

$$1 + A\cos(\tfrac{1}{2}\xi)^2\phi(\xi)^2 \geq 1 + \big(-1 + \tfrac{1}{2}\delta\big) = \tfrac{1}{2}\delta$$

by (10), and hence (46) also holds for the left boundary value. □

Proof of Lemma 4.8. We first observe that, by the Cauchy–Schwarz inequality, (3) and (10),

$$\left|\left\langle \mathcal{L}(\Phi u)v, v\right\rangle_0 - \left\langle \mathcal{L}(u)v, v\right\rangle_0\right| \leq C_M \|\Phi u - u\|_1 \|v\|_0^2 \leq C_M \tau \|v\|_0^2,$$

where we use in the second inequality that $|\phi(\xi) - 1| \leq \min(2, c_0\xi^2) \leq C\xi$ for $\xi \geq 0$ by (10a). This yields

$$\left\langle \mathcal{L}(\Phi u)v, v\right\rangle_0 \geq \left\langle \mathcal{L}(u)v, v\right\rangle_0 - C_M \tau \|v\|_0^2. \tag{47}$$

Next, we use semiclassical pseudodifferential calculus as presented in Appendix A. We first express the operator $\mathcal{L}(u)$ by quantization of certain symbols. We set, for $x \in \mathbb{T}$ and $\xi \in \mathbb{R}$ and with $\omega = \sqrt{\xi^2 + \tau^2}$,

$$b_1(x,\xi) = \phi(\omega), \quad b_2(x,\xi) = \kappa\, a(u(x)), \quad b_3(x,\xi) = \cos(\omega), \quad b_4(x,\xi) = \sin(\omega)^2\phi(\omega)^2.$$

As stated there is now somewhat unfortunately τ dependence in the ω symbol, however the dependence upon our semiclassical parameter τ arises as a very small, bounded perturbation and hence does not effect any of the semiclassical bounds used. Note in addition that this slight notational complication arises from the generality of treating Klein-Gordon type operators with our semi-classical formulation and the τ dependence in ω would not arise in the wave equation setting.

With the corresponding quantizations $\mathrm{Op}_{b_1}^\tau, \dots, \mathrm{Op}_{b_4}^\tau$ (see equation (67) in Appendix A), we then have

$$\mathcal{L}(u) = \mathrm{Op}_{b_1}^\tau \mathrm{Op}_{b_2}^\tau \mathrm{Op}_{b_3}^\tau \mathrm{Op}_{b_1}^\tau - \tfrac{1}{4} \mathrm{Op}_{b_1}^\tau \mathrm{Op}_{b_2}^\tau \mathrm{Op}_{b_4}^\tau \mathrm{Op}_{b_2}^\tau \mathrm{Op}_{b_1}^\tau .$$

Note that all the symbols $b_1, \dots, b_4$ are in $S_{\sigma,1} \cap S_{\sigma+1,0}$ for $\sigma = 1 > \frac{1}{2}$ since u is in H^2 and ϕ is has bounded derivative, and that we have

$$|b_j|_{\sigma,1} \leq C_M, \qquad |b_j|_{\sigma+1,0} \leq C_M \qquad \text{for} \qquad j = 1, \dots, 4$$

by (3b). By (3a), this also holds for finite products of these symbols. We refer to Appendix A for the definition of the symbol classes $S_{\sigma,0}$ and $S_{\sigma+1,1}$ and the corresponding seminorms $|\cdot|_{\sigma,1}$ and $|\cdot|_{\sigma+1,0}$.

By using Proposition A.3 repeatedly, we thus have that

$$\left\|\mathcal{L}(u)v - \mathrm{Op}_b^\tau(v)\right\|_0 \leq C_M \tau \|v\|_0$$

with the new symbol

$$b(x,\xi) = b_1(x,\xi)b_2(x,\xi)b_3(x,\xi)b_1(x,\xi) - \tfrac{1}{4} b_1(x,\xi)b_2(x,\xi)b_4(x,\xi)b_2(x,\xi)b_1(x,\xi).$$

This estimate and the Cauchy–Schwarz inequality imply

$$\left\langle \mathcal{L}(u)v, v\right\rangle_0 \geq \left\langle \mathrm{Op}_b^\tau\, v, v\right\rangle_0 - C_M \tau \|v\|_0^2. \tag{48}$$

Next we use Proposition A.4 to estimate the term with Op_b^τ in (48) further. Note that the symbol b is in $S_{\sigma+1,0} \cap S_{\sigma+1,1}$ for $\sigma = 1 > \frac{1}{2}$ since u is in H^2 and ϕ has bounded derivative, and that we have

$$|b|_{\sigma+1,0} \leq C_M, \qquad |b|_{\sigma+1,1} \leq C_M.$$

Note also that

$$1 + b(x,\xi) \geq \tfrac{1}{2}\delta \qquad \text{for all} \qquad x \in \mathbb{T},\ \xi \in \mathbb{R}$$

by Lemma 4.9 (with $A = \kappa\, a(u(x))$). By using Proposition A.4, we thus obtain that

$$\|v\|_0^2 + \left\langle \mathrm{Op}_b^\tau\, v, v\right\rangle_0 \geq \tfrac{1}{4}\delta\|v\|_0^2 - C_M\tau\|v\|_0^2.$$

Combining this estimate with (47) and (48) yields the statement of the lemma for sufficiently small τ. □

Proposition 4.10. *Fix $M > 0$ and $\delta > 0$, and let the filter functions satisfy Assumptions 1–3 with constants $0 < \delta < 1$ and $A_0 \geq 0$. Then, there exists $\tau_0 > 0$ such that for every $\tau \leq \tau_0$, for every $(e, \dot e) \in H^2 \times H^1$ and for every $u \in H^2$ with*

$$\|u\|_2 \leq M, \qquad 1 + \kappa\, a(u(x)) \geq \tfrac{1}{2}\delta > 0, \qquad \kappa\, a(u(x)) \leq A_0 + \tfrac{1}{2}\delta,$$

we have that the modified energy (32) controls the Sobolev norms, i.e.,

$$C_\delta |||(e, \dot e)|||_1^2 \leq \mathcal{E}(e, \dot e, u) \leq C_M |||(e, \dot e)|||_1^2 \tag{49}$$

for two positive constants C_δ, C_M.

Proof. First note that for τ sufficiently small and $0 < \delta < 1$ we have that

$$\left(\tfrac{1}{8}\delta - 1\right)\left\|\partial_x^2 e\right\|_0^2 \leq \kappa\mathcal{U}(\Phi e, \Phi u) \leq C_M\|e\|_2^2,$$

where the upper bound follows from (3) and the lower bound is a consequence of Lemma 4.8 (with $v = \partial_x^2 e$) using the representation of $\kappa\mathcal{U}(\Phi e, \Phi u)$ given in (44). The bound (49) on the modified energy then follows by its definition in (32). □

4.5 Local error bound

Similarly as in the semilinear case [11], but under higher regularity assumptions on the initial value, we can prove the following local error bound for the numerical method (7).

Lemma 4.11 (Local error bound in $H^2 \times H^1$)**.** *Let the filter functions satisfy Assumptions 1 and 2. If $(u, \partial_t u)$ is a solution to (4) in $H^5 \times H^4$ with*

$$\left|\left|\left|\big(u(\cdot, t), \partial_t u(\cdot, t)\big)\right|\right|\right|_4 \leq M \qquad \text{for} \qquad t_n \leq t \leq t_{n+1},$$

then we have

$$|||(d_{n+1}, \dot d_{n+1})|||_1 \leq C_M\tau^3|\kappa|$$

for the local error $(d_{n+1}, \dot d_{n+1})$ of (30).

Proof. Without loss of generality, we consider the case $n = 0$, that is, we consider the local error

$$(d_1, \dot d_1) = (u_1, \dot u_1) - \big(u(\cdot, \tau), \partial_t u(\cdot, \tau)\big).$$

As in the semilinear case [11], the proof relies on a comparison of the method in the form (8) with the variation-of-constants formula for the exact solution $(u(\cdot, \tau), \partial_t u(\cdot, \tau))$. For initial values (2) at time 0, the variation-of-constants formula reads

$$\begin{pmatrix} u(\cdot, \tau) \\ \partial_t u(\cdot, \tau) \end{pmatrix} = R(\tau)\begin{pmatrix} u_0 \\ \dot u_0 \end{pmatrix} + \kappa\int_0^\tau R(\tau - t)\begin{pmatrix} 0 \\ f(u(\cdot, t)) \end{pmatrix} \mathrm{d}t$$

with

$$R(t) = \begin{pmatrix} \cos(t\Omega) & t\,\mathrm{sinc}(t\Omega) \\ -\Omega\sin(t\Omega) & \cos(t\Omega) \end{pmatrix}.$$

Note that this formula makes sense in $H^2 \times H^1$ for solutions in $H^5 \times H^4$. Using this formula, the local error is seen to be of the form

$$\begin{pmatrix} u_1 - u(\cdot,\tau) \\ \dot{u}_1 - \partial_t u(\cdot,\tau) \end{pmatrix} = \tfrac{1}{2}\tau\kappa R(\tau) \begin{pmatrix} 0 \\ \widehat{f}(u_0) - f(u_0) \end{pmatrix} + \tfrac{1}{2}\tau\kappa \begin{pmatrix} 0 \\ \widehat{f}(u_1) - f(u_1) \end{pmatrix} \tag{50a}$$

$$+ \tfrac{1}{2}\tau\kappa \begin{pmatrix} 0 \\ f(u_1) - f(u(\cdot,\tau)) \end{pmatrix} \tag{50b}$$

$$+ \tfrac{1}{2}\tau\kappa \Bigg(R(\tau) \begin{pmatrix} 0 \\ f(u_0) \end{pmatrix} + R(0) \begin{pmatrix} 0 \\ f(u(\cdot,\tau)) \end{pmatrix} \Bigg) - \kappa \int_0^\tau R(\tau - t) \begin{pmatrix} 0 \\ f(u(\cdot,t)) \end{pmatrix} \mathrm{d}t. \tag{50c}$$

We estimate the three contributions (50a)–(50c) to the local error separately.

(a) The contributions to the local error in the first line (50a) are due to the introduction of filters in the nonlinearity $\widehat{f}(u) = \Psi_1 f(\Phi u)$. Using that $R(\tau)$ preserves the norm $|||\cdot|||_1$, we get

$$\left|\!\left|\!\left| R(\tau) \begin{pmatrix} 0 \\ \widehat{f}(u_0) - f(u_0) \end{pmatrix} \right|\!\right|\!\right|_1 = \|\widehat{f}(u_0) - f(u_0)\|_1.$$

We then split $\widehat{f}(u_0) - f(u_0) = \Psi_1(f(\Phi u_0) - f(u_0)) + (\Psi_1 f(u_0) - f(u_0))$ and use

$$\big\|\Psi_1\big(f(\Phi u) - f(u)\big)\big\|_1 \le C_M \|u\|_3 \|\Phi u - u\|_3 \le C_M \|u\|_3 \|\tau^2 \Omega^2 u\|_3 \le C_M \tau^2 \|u\|_5^2$$

by the Lipschitz property (26) and the assumptions (10) on the filter functions as well as

$$\|\Psi_1 f(u) - f(u)\|_1 \le C\tau^2 \|f(u)\|_3 \le C_M \tau^2 \|u\|_5^2$$

by (10) and (25). This shows that

$$\left|\!\left|\!\left| R(\tau) \begin{pmatrix} 0 \\ \widehat{f}(u_0) - f(u_0) \end{pmatrix} \right|\!\right|\!\right|_1 \le C_M \tau^2.$$

The term in (50a) with u_1 instead of u_0 can be dealt with in the same way using in addition that $\|u_1\|_5 \le C_M$ by Lemma 4.1 since $|||(u_0, \dot{u}_0)|||_4 \le M$. This finally yields

$$|||\text{term on right-hand side of (50a)}|||_1 \le C_M \tau^3 |\kappa|. \tag{51}$$

In the same way we also get

$$|||\text{term on right-hand side of (50a)}|||_2 \le C_M \tau^2 |\kappa|, \tag{52}$$

if we use $|1 - \psi_1(\xi)| \le C\xi$ and $|1 - \phi(\xi)| \le C\xi$ (which follow from (10)) instead of $|1 - \psi_1(\xi)| \le C\xi^2$ and $|1 - \phi(\xi)| \le C\xi^2$.

(b) The contribution to the local error in the third line (50c) is the quadrature error of the trapezoidal rule. With the corresponding second-order Peano kernel $K_2(\sigma) = \frac{1}{2}\sigma(\sigma - 1)$, it takes the form

$$\text{term (50c)} = -\tau^3 \kappa \int_0^1 K_2(\sigma) h''(\sigma\tau)\, \mathrm{d}\sigma \qquad \text{with} \qquad h(t) = R(\tau - t) \begin{pmatrix} 0 \\ f(u(\cdot,t)) \end{pmatrix}.$$

We thus have to estimate $|||h''(\sigma\tau)|||_1$. For that we use, for $\ell = 0, 1, 2$,

$$\left|\!\left|\!\left| \frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell} R(t) \begin{pmatrix} v \\ \dot{v} \end{pmatrix} \right|\!\right|\!\right|_1 = |||(v, \dot{v})|||_{1+\ell} \qquad \text{and} \qquad \left\| \frac{\mathrm{d}^{2-\ell}}{\mathrm{d}t^{2-\ell}} f\big(u(\cdot,t)\big) \right\|_{1+\ell} \le C_M$$

by (3) (and in the case $\ell = 0$ also (4) to replace $\partial_t^2 u$) since $(u, \partial_t u)$ is bounded in $H^5 \times H^4$. This yields

$$|||\text{term (50c)}|||_1 \le C_M \tau^3 |\kappa|. \tag{53}$$

In the same way we also get

$$|||\text{term (50c)}|||_2 \le C_M \tau^2 |\kappa|, \tag{54}$$

if we use the first-order Peano kernel and h' instead of the second-order Peano kernel and h''.

(c) The contribution to the local error in the second line (50b) concerns only the error in the velocities. Using that we thus already have $\|u_1 - u(\cdot,\tau)\|_3 \le C_M \tau^2 |\kappa|$ by the estimates (52) and (54) and that

$$\big\| f(u_1) - f\big(u(\cdot,\tau)\big) \big\|_1 \le C_M \|u_1 - u(\cdot,\tau)\|_3$$

by (26) and Lemma 4.1, we get

$$|||\text{term (50b)}|||_1 \le C_M \tau^3 |\kappa|.$$

Together with the estimates (51) and (53) in (a) and (b), this completes the proof of the stated local error bound. □

In view of (31), we are not so much interested in local errors $(d, \dot d)$ in the $H^2 \times H^1$-norm as estimated in the previous lemma, but instead in energy differences of the form $\mathcal{E}(e, \dot e, u) - \mathcal{E}(e - d, \dot e - \dot d, u)$, where $(d, \dot d)$ is a local error and $(e, \dot e)$ is a global error. This extension is done in the following lemma.

Lemma 4.12. *Let the filter functions satisfy Assumptions 1 and 2. If $(e, \dot e) \in H^2 \times H^1$, $(d, \dot d) \in H^2 \times H^1$ and $u \in H^2$ with*

$$\|e\|_2 \le M, \qquad \|d\|_2 \le M, \qquad \|u\|_2 \le M,$$

then we have

$$\big|\mathcal{E}(e, \dot e, u) - \mathcal{E}(e - d, \dot e - \dot d, u)\big| \le C_M \big(\tau^{-1}|\kappa|^{-1} |||(d, \dot d)|||_1^2 + \tau|\kappa|\, |||(e, \dot e) - (d, \dot d)|||_1^2\big).$$

Proof. The crucial property that we use in various forms is that, for any $\alpha > 0$,

$$2|\langle w, v\rangle| \le \alpha^{-1} \|w\|^2 + \alpha \|v\|^2 \tag{55}$$

for a scalar product $\langle \cdot, \cdot \rangle$ with associated norm $\|\cdot\|$, in particular

$$\big| \|v\|^2 - \|v - w\|^2 \big| = \big| \|w\|^2 + 2\langle w, v - w\rangle \big| \le (1 + \alpha^{-1}) \|w\|^2 + \alpha \|v - w\|^2.$$

This yields for the first term in the energy difference $\mathcal{E}(e, \dot e, u) - \mathcal{E}(e - d, \dot e - \dot d, u)$

$$\big| |||(e, \dot e)|||_1^2 - |||(e - d, \dot e - \dot d)|||_1^2 \big| \le (1 + \alpha^{-1}) |||(d, \dot d)|||_1^2 + \alpha |||(e - d, \dot e - \dot d)|||_1^2.$$

The second term of the energy is $\kappa \mathcal{U}(\widehat e, \widehat u)$ with $\widehat u = \Phi u$ and $\widehat e = \Phi e$. For the second term in $\mathcal{U}$, we get in the energy difference, with $\widehat u = \Phi u$, $\widehat e = \Phi e$ and $\widehat d = \Phi d$,

$$\begin{aligned} &\tfrac{1}{4}\tau^2|\kappa| \Big| \big\| \Psi_1\big(a(\widehat u)\partial_x^2 \widehat e\big) \big\|_1^2 - \big\| \Psi_1\big(a(\widehat u)\partial_x^2 (\widehat e - \widehat d)\big) \big\|_1^2 \Big| \\ &\qquad \le \tfrac{1}{4}\tau^2|\kappa| (1 + \alpha^{-1}) \big\| \Psi_1\big(a(\widehat u)\partial_x^2 \widehat d\big) \big\|_1^2 + \tfrac{1}{4}\tau^2|\kappa|\, \alpha \big\| \Psi_1\big(a(\widehat u)\partial_x^2 (\widehat e - \widehat d)\big) \big\|_1^2, \end{aligned}$$

and hence, by (10), (11) and (26) with $s = 0$ (similarly as in (28)), we get the bound $C_M(1+\alpha^{-1})\|\widehat{d}\|_2^2 + C_M\alpha\|\widehat{e}-\widehat{d}\|_2^2$ for this term. For the first term in $\mathcal{U}$, we have

$$\begin{aligned}
&\big\langle \cos(\tau\Omega)\widehat{e}, a(\widehat{u})\partial_x^2\widehat{e}\big\rangle_1 - \big\langle \cos(\tau\Omega)(\widehat{e}-\widehat{d}), a(\widehat{u})\partial_x^2(\widehat{e}-\widehat{d})\big\rangle_1 \\
&= \big\langle \cos(\tau\Omega)\widehat{d}, a(\widehat{u})\partial_x^2\widehat{e}\big\rangle_1 + \big\langle \cos(\tau\Omega)(\widehat{e}-\widehat{d}), a(\widehat{u})\partial_x^2\widehat{d}\big\rangle_1 \\
&= \big\langle \cos(\tau\Omega)\widehat{d}, a(\widehat{u})\partial_x^2\widehat{d}\big\rangle_1 + \big\langle \cos(\tau\Omega)\widehat{d}, a(\widehat{u})\partial_x^2(\widehat{e}-\widehat{d})\big\rangle_1 + \big\langle \cos(\tau\Omega)(\widehat{e}-\widehat{d}), a(\widehat{u})\partial_x^2\widehat{d}\big\rangle_1.
\end{aligned}$$

Using partial integration, (26) and (55), we get for these terms similarly as above the bound $C_M(1+\alpha^{-1})\|\widehat{d}\|_2^2 + C_M\alpha\|\widehat{e}-\widehat{d}\|_2^2$. By choosing $\alpha = \tau|\kappa|$, the statement of the lemma then follows from assumption (10). □

From Lemmas 4.11 and 4.12, we get the following bound for the local error in the form as it appears in (31).

Proposition 4.13 (Local error bound in the energy)**.** *Let the filter functions satisfy Assumptions 1 and 2. If $(u, \partial_t u)$ is a solution to (4) in $H^5 \times H^4$ with*

$$\big|\!\big|\!\big|\big(u(\cdot,t), \partial_t u(\cdot,t)\big)\big|\!\big|\!\big|_4 \le M \qquad \text{for} \qquad t_n \le t \le t_{n+1},$$

and if $(u_n, \dot{u}_n) \in H^2 \times H^1$ is a corresponding numerical solution with

$$|\!|\!|(u_n, \dot{u}_n)|\!|\!|_1 \le 2M,$$

then we have

$$\begin{aligned}
&\big|\mathcal{E}(e_{n+1}, \dot{e}_{n+1}, u_{n+1}) - \mathcal{E}(e_{n+1}-d_{n+1}, \dot{e}_{n+1}-\dot{d}_{n+1}, u_{n+1})\big| \\
&\qquad\qquad \le C_M\tau^5|\kappa| + C_M\tau|\kappa|\,|\!|\!|(e_n, \dot{e}_n)|\!|\!|_1^2
\end{aligned}$$

with the global errors $(e_n, \dot{e}_n)$ and $(e_{n+1}, \dot{e}_{n+1})$ of (29) and the local error $(d_{n+1}, \dot{d}_{n+1})$ of (30).

Proof. We apply Lemma 4.12 with $(e, \dot{e}) = (e_{n+1}, \dot{e}_{n+1})$, $(d, \dot{d}) = (d_{n+1}, \dot{d}_{n+1})$ and $u = u_{n+1}$ in combination with Lemma 4.11. Note that $\|d_{n+1}\|_2 \le C_M$ by Lemma 4.11, $\|u_{n+1}\|_2 \le C_M$ by Lemma 4.1 and $\|e_{n+1}\|_2 \le C_M$ by the assumption on the exact solution and by the just mentioned bound on u_{n+1}. Lemmas 4.11 and 4.12 then yield the stated estimate but with $(e_{n+1}, \dot{e}_{n+1}) - (d_{n+1}, \dot{d}_{n+1})$ instead of $(e_n, \dot{e}_n)$ on the right-hand side. To get the final statement, we use the definitions (29) of $(e_{n+1}, \dot{e}_{n+1})$ and (30) of $(d_{n+1}, \dot{d}_{n+1})$ and apply Lemma 4.2. □

Remark 4.14. The local error bound of Lemma 4.11 is only based on the estimates (25) and (26). As they extend to nonzero g in f and (1), also Lemma 4.11 and Proposition 4.13 extend to this case.

4.6 Error accumulation and proof of Theorem 3.2

We proceed as outlined in Section 4.2. Considering the numerical solution $(u_n, \dot{u}_n)$ given by (7) for $n = 0, 1, \ldots$, we set

$$\mathcal{E}_n(e, \dot{e}) = \mathcal{E}(e, \dot{e}, u_n).$$

We then decompose $\mathcal{E}_{n+1}(e_{n+1}, \dot{e}_{n+1})$ with the global error $(e_{n+1}, \dot{e}_{n+1})$ of (29) as in (31).

Adapting the usual inductive argument to prove global error bounds, we assume that the numerical solution $(u_j, \dot{u}_j)$ satisfies, for $j = 0, \dots, n$ and with the constants M, A_0 and δ of Assumption 4,

$$|||(u_j, \dot{u}_j)|||_1 \leq 2M \tag{56a}$$

and

$$1 + \kappa\, a(u_j) \geq \tfrac{1}{2}\delta \qquad \text{and} \qquad \kappa\, a(u_j) \leq A_0 + \tfrac{1}{2}\delta. \tag{56b}$$

This is clear for $j = 0$ by Assumption 4. Under these hypotheses, we will prove the error bound of Theorem 3.2 until time $t_n = n\tau$. We will then prove that (56) also holds for $j = n + 1$ to close the inductive argument.

Under the regularity assumption (20) on the exact solution and thanks to (56a), we get from Propositions 4.6 and 4.13

$$\big|\mathcal{E}_{j+1}(e_{j+1}, \dot{e}_{j+1})\big| \leq \big|\mathcal{E}_j(e_j, \dot{e}_j)\big| + C_M\tau|\kappa|\, |||(e_j, \dot{e}_j)|||_1^2 + C_M\tau^5|\kappa|$$

for $j = 0, \dots, n$. Thanks to (56b), we can then apply Proposition 4.10 (with $u = u_j$) to get

$$\big|\mathcal{E}_{j+1}(e_{j+1}, \dot{e}_{j+1})\big| \leq \big(1 + C_M\tau|\kappa|\big)\big|\mathcal{E}_j(e_j, \dot{e}_j)\big| + C_M\tau^5|\kappa|$$

for $j = 0, \dots, n$. Solving this recursion in the standard way yields the error bound

$$\big|\mathcal{E}_{j+1}(e_{j+1}, \dot{e}_{j+1})\big| \leq C_M\tau^4 \mathrm{e}^{C_M|\kappa|t_{j+1}} \tag{57}$$

for $j = 0, \dots, n$. By applying once again Proposition 4.10, we get the global error bound

$$|||(e_{j+1}, \dot{e}_{j+1})|||_1^2 \leq C_M\tau^4 \mathrm{e}^{C_M|\kappa|t_{j+1}}, \tag{58}$$

for $j = 0, \dots, n-1$.

In order to close the induction, we have to justify that (56) also holds for $j = n+1$. To do so, we note that the bound on a single time step given in Lemma 4.2, the local error bound of Lemma 4.11 and the bound (58) for $j = n - 1$, allow us to estimate[1]

$$|||(e_{n+1}, \dot{e}_{n+1})|||_1 \leq C_M |||(e_n, \dot{e}_n)|||_1 + |||(d_n, \dot{d}_n)|||_1 \leq C_{M,t_{n+1}}\tau^2.$$

This implies $\|u(\cdot, t_{n+1}) - u_{n+1}\|_2 \leq C_{M,t_{n+1}}\tau^2$, and hence (56a) also holds for $j = n+1$ by assumption (20) provided that τ is sufficiently small. It also implies $\|u(\cdot, t_{n+1}) - u_{n+1}\|_{L^\infty} \leq C_{M,t_{n+1}}\tau^2$, and hence, again for sufficiently small τ,

$$\|\kappa\, a(u(\cdot, t_{n+1})) - \kappa\, a(u_{n+1})\|_{L^\infty} \leq \tfrac{1}{2}\delta.$$

By assumptions (21) and (22), this shows that (56b) also holds for $j = n + 1$. This closes the induction and concludes the proof of Theorem 3.2.

5 Proof of the error bound for the full discretization

In this section, we study fully discrete methods (19). We first give the proof of Theorem 3.4 on their global error. The structure of this proof is the same as for the semi-discretization in time in Section 4. We study stability in Section 5.1 below,

[1] Note that this estimate is not a proof of (58) for $j = n$, because then the constants would explode. Instead, it is only used to justify (56) for $j = n + 1$.

control Sobolev norms with the energy in Section 5.2, estimate the local error Section 5.3 and put everything together in Section 5.4. All arguments are extensions to the fully discrete setting of the arguments of Section 4, which illustrates the importance of proving such semi-discrete error bounds first.

Throughout, we use, for $s \geq s' \geq 0$, the approximation property

$$\left\| v - \mathcal{P}^K(v) \right\|_{s'} \leq K^{-(s-s')} \|v\|_s \qquad \text{for} \qquad v \in H^s \tag{59}$$

of the L^2-orthogonal projection $\mathcal{P}^K$ of (16), and its stability

$$\left\| \mathcal{P}^K(v) \right\|_s \leq \|v\|_s \qquad \text{for} \qquad v \in H^s. \tag{60}$$

In addition, we use, for $s \geq s' \geq 0$ with $s - s' > \frac{1}{2}$, the approximation property

$$\left\| v - \mathcal{I}^K(v) \right\|_{s'} \leq C_{s,s'} K^{-(s-s')} \|v\|_s \qquad \text{for} \qquad v \in H^s \tag{61}$$

of the trigonometric interpolation $\mathcal{I}^K$, and its stability

$$\left\| \mathcal{I}^K(v) \right\|_s \leq C_s \|v\|_s \qquad \text{for} \qquad v \in H^s. \tag{62}$$

We emphasize that all estimates in the following are uniform in the spatial discretization parameter K.

5.1 Stability of the numerical method

Our aim is to show that the stability estimates of Section 4.3 carry over to the fully discrete situation.

Starting with the definition of the energy $\mathcal{E}$ of (32), we define its fully discrete version

$$\mathcal{E}^K(e, \dot{e}, u) = |||(e, \dot{e})|||_1^2 + \kappa \mathcal{U}^K(\Phi e, \Phi u) \tag{63}$$

with

$$\mathcal{U}^K(e, u) = \left\langle \cos(\tau\Omega)\partial_x^2 e, \mathcal{P}^K\left(a^K(u)\partial_x^2 e\right)\right\rangle_0 - \tfrac{1}{4}\tau^2\kappa \left\| \Psi_1 \mathcal{P}^K\left(a^K(u)\partial_x^2 e\right)\right\|_1^2. \tag{64}$$

The difference compared to the $\mathcal{E}$ of (32) are the additional projections $\mathcal{P}^K$ and the functions $a^K = \mathcal{I}^K \circ a$ instead of a.

The computation of Lemma 4.4 directly transfer to the new energy (63) of the fully discrete setting if we use that Ψ_1 and $\mathcal{P}^K$ commute and if we replace the function f by $\mathcal{P}^K \circ f^K$, where f^K is defined in (18).

In order to transfer the bound of the remainder term of Lemma 4.5 to the fully discrete setting, we use the bounds (60) and (62) on $\mathcal{P}^K$ and $\mathcal{I}^K$ (to estimate $a^K = \mathcal{I}^K \circ a$, which appears in f^K) and in addition the property

$$\left\langle v^K, \mathcal{P}^K(w)\right\rangle_s = \left\langle v^K, w\right\rangle_s \qquad \text{for} \qquad v^K \in \mathcal{V}^K,\ w \in H^s \tag{65}$$

with $s = 1$. This property is needed for the symmetry argument and the partial integrations in the proof of Lemma 4.5.[2]

From the fully discrete versions of Lemmas 4.4 and 4.5, we finally get the stability estimate of Proposition 4.6 also in the fully discrete setting for $\mathcal{E}^K$.

[2]It is not immediately clear, whether these steps can also be done if the (traditional) trigonometric interpolation is used instead of the projection $\mathcal{P}^K$ to define $\widehat{f}^K$ in the fully discrete method.

5.2 Controlling Sobolev norms with the energy

We show that the bounds on the energy $\mathcal{E}$ of Proposition 4.10 carry over to the fully discrete setting and the corresponding energy $\mathcal{E}^K$ of (63).

For the upper bound in (49), we proceed as in the proof of Proposition 4.10 and use in addition (60) and (62) to deal with the additional $\mathcal{P}^K$ and $\mathcal{I}^K$ (in $a^K = \mathcal{I}^K \circ a$).

For the lower bound in (49), we use that we have by (11), (60) and (65) with $s = 0$

$$\begin{aligned}\kappa \mathcal{U}^K(\Phi e, \Phi u) &\geq \kappa \big\langle \cos(\tau\Omega)\partial_x^2 \Phi e, a^K(\Phi u)\partial_x^2 \Phi e\big\rangle_0 - \tfrac{1}{4}\tau^2\kappa^2 \big\|\Psi_1\big(a^K(\Phi u)\partial_x^2\Phi e\big)\big\|_1^2 \\ &= \big\langle \mathcal{L}^K(\Phi u)\partial_x^2 e, \partial_x^2 e\big\rangle_0\end{aligned}$$

for a trigonometric polynomial e (note that, in comparison to (64), the projections $\mathcal{P}^K$ are absent on the right) with

$$\mathcal{L}^K(u) = \kappa \Phi a^K(u)\cos(\tau\Omega)\Phi - \tfrac{1}{4}\kappa^2 \Phi a^K(u)\sin^2(\tau\Omega)\Phi^2 a^K(u)\Phi.$$

The operator $\mathcal{L}^K$ is the same as the operator $\mathcal{L}$ of Section 5.2, except that a is replaced by $a^K = \mathcal{I}^K \circ a$. To obtain the lower bound in (49), we can thus proceed exactly as in the proof of Proposition 4.10 if we replace a by a^K in the statement of this proposition and restrict to trigonometric polynomials e.

5.3 Local error bound

The main difference compared to the semi-discrete setting arises in the local error bound of Section 4.5, which now has to take also the spatial error into account. We denote here and in the following by

$$u^K(\cdot,t) = \mathcal{P}^K\big(u(\cdot,t)\big), \qquad \partial_t u^K(\cdot,t) = \mathcal{P}^K\big(\partial_t u(\cdot,t)\big) \tag{66}$$

the L^2-orthogonal projection of the exact solution onto $\mathcal{V}^K$.

Lemma 5.1 (Local error bound in $H^2 \times H^1$)**.** *Let $s \geq 0$, and let the filter functions satisfy Assumptions 1 and 2. If $(u, \partial_t u)$ is a solution to (1) in $H^{5+s} \times H^{4+s}$ with*

$$\big|\!\big|\!\big|\big(u(\cdot,t), \partial_t u(\cdot,t)\big)\big|\!\big|\!\big|_{4+s} \leq M \qquad \text{for} \qquad t_n \leq t \leq t_{n+1},$$

then we have

$$\big|\!\big|\!\big|\big(d_{n+1}^K, \dot{d}_{n+1}^K\big)\big|\!\big|\!\big|_1 \leq C_M \tau^3 |\kappa| + C_M \tau K^{-2-s}|\kappa|$$

for the fully discrete local error

$$\big(d_{n+1}^K, \dot{d}_{n+1}^K\big) = \varphi_\tau^K\big(u^K(\cdot,t_n), \partial_t u^K(\cdot,t_n)\big) - \big(u^K(\cdot,t_{n+1}), \partial_t u^K(\cdot,t_{n+1})\big)$$

with the fully discrete numerical flow φ_τ^K.

Proof. The proof is similar to the one of Lemma 4.11. We restrict again to the case $n = 0$. Writing $\tilde{f}^K = \mathcal{P}^K \circ f$, we start from the fully discrete analog

$$\begin{aligned}\begin{pmatrix} u_1^K - u^K(\cdot,\tau) \\ \dot{u}_1^K - \partial_t u^K(\cdot,\tau)\end{pmatrix} &= \tfrac{1}{2}\tau\kappa R(\tau)\begin{pmatrix} 0 \\ \widehat{f}^K(u_0) - \tilde{f}^K(u_0)\end{pmatrix} + \tfrac{1}{2}\tau\kappa\begin{pmatrix} 0 \\ \widehat{f}^K(u_1^K) - \tilde{f}^K(u_1^K)\end{pmatrix} \\ &+ \tfrac{1}{2}\tau\kappa R(\tau)\begin{pmatrix} 0 \\ \widehat{f}^K(u_0^K) - \widehat{f}^K(u_0)\end{pmatrix} + \tfrac{1}{2}\tau\kappa\begin{pmatrix} 0 \\ \tilde{f}^K(u_1^K) - \tilde{f}^K(u(\cdot,\tau))\end{pmatrix} \\ &+ \tfrac{1}{2}\tau\kappa\Bigg(R(\tau)\begin{pmatrix} 0 \\ \tilde{f}^K(u_0)\end{pmatrix} + R(0)\begin{pmatrix} 0 \\ \tilde{f}^K(u(\cdot,\tau))\end{pmatrix}\Bigg) - \kappa\int_0^\tau R(\tau - t)\begin{pmatrix} 0 \\ \tilde{f}^K(u(\cdot,t))\end{pmatrix}\mathrm{d}t\end{aligned}$$

of the local error representation (50). In the derivation of this representation, we have used that $\mathcal{P}^K$ and the four components of R commute. The contributions to the local error can be estimated similarly as in the proof of Lemma 4.11 using in addition the properties (59)–(62) of $\mathcal{P}^K$ and $\mathcal{I}^K$ and the assumed regularity of the exact solution.

(a) In the terms of the first line, we decompose $\widehat{f}^K - \tilde{f}^K = (\widehat{f}^K - \mathcal{P}^K \circ f^K) + \mathcal{P}^K \circ (f^K - f)$. For the terms with $\widehat{f}^K - \mathcal{P}^K \circ f^K$ (which correspond to (50a)), we get an estimate $C_M \tau^3 |\kappa|$ in $H^2 \times H^1$ and $C_M \tau^2 |\kappa|$ in $H^3 \times H^2$ as in the proof of Lemma 4.11. For the terms with $\mathcal{P}^K \circ (f^K - f)$, we use in particular (61) in addition to the arguments of the proof of Lemma 4.11 to get an estimate $C_M \tau K^{-4-s} |\kappa|$ in $H^2 \times H^1$ and $C_M \tau K^{-3-s} |\kappa|$ in $H^3 \times H^2$.

(b) For the terms in the third line (which correspond to (50c)), we get as in the proof of Lemma 4.11 an estimate $C_M \tau^3 |\kappa|$ in $H^2 \times H^1$ and $C_M \tau^2 |\kappa|$ in $H^3 \times H^2$.

(c) For the new first term in the second line, we get an estimate $C_M \tau K^{-2-s} |\kappa|$ in $H^2 \times H^1$ and, using the properties (10) and (11) of the filters as in (28), an estimate $C_M K^{-2-s} |\kappa|$ in $H^3 \times H^2$. The second term in the second line corresponds to (50b) and can then be dealt with as in the proof of Lemma 4.11. □

Based on Lemma 5.1, we can then prove a corresponding local error bound in the energy as in Proposition 4.13.

5.4 Proof of Theorem 3.4

The error accumulation is done as in Section 4.6, but with the exact solution replaced by its projection (66). Note that we need (56b) for $a^K = \mathcal{I}^K \circ a$ instead of a; we use (59) and (61) to deal with that. This gives a the claimed global error bound of Theorem 3.4, but with u replaced by $u^K = \mathcal{P}^K \circ u$ in the error estimate. We then use once more (59) to get the precise error estimate of Theorem 3.4.

A Semiclassical pseudodifferential calculus

In this section we shall recall the basic results about pseudodifferential calculus that were needed in our proof. The presentation follows closely [20, Section 8]. We shall use the Fourier transform on the torus defined by

$$\hat{u}_j = (\mathcal{F}_x u)(j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x) \mathrm{e}^{-\mathrm{i}jx} \, \mathrm{d}x.$$

We consider symbols $a(x, \xi)$ (not to be confused with the function a in (1)) defined on $\mathbb{T} \times \mathbb{R}$ that are continuous in ξ, and we use the quantization

$$(\mathrm{Op}_a u)(x) = \sum_{j \in \mathbb{Z}} a(x, j) \hat{u}_j \mathrm{e}^{\mathrm{i}jx}. \tag{67}$$

We introduce for $\sigma \geq 0$ the following seminorms of symbols:

$$|a|_{\sigma,0} = \sup_{|\alpha| \leq \sigma} \big\| \mathcal{F}_x(\partial_x^\alpha a) \big\|_{L^2_j(\mathbb{Z}, L^\infty_\xi(\mathbb{R}))}, \qquad |a|_{\sigma,1} = \sup_{|\alpha| \leq \sigma} \big\| \mathcal{F}_x(\partial_x^\alpha \partial_\xi a) \big\|_{L^2_j(\mathbb{Z}, L^\infty_\xi(\mathbb{R}))}.$$

Note that

$$\big\| \mathcal{F}_x(\partial_x^\alpha a) \big\|^2_{L^2_j(\mathbb{Z}, L^\infty_\xi(\mathbb{R}))} = \sum_{j \in \mathbb{Z}} \big\| (\mathrm{i}j)^\alpha \hat{a}_j \big\|^2_{L^\infty(\mathbb{R})}$$

with the Fourier coefficients

$$\hat{a}_j(\xi) = (\mathcal{F}_x a)(j,\xi), \qquad j \in \mathbb{Z},\ \xi \in \mathbb{R}, \tag{68}$$

of a, and similarly

$$\big\|\mathcal{F}_x(\partial_x^\alpha \partial_\xi a)\big\|^2_{L^2_j(\mathbb{Z},L^\infty_\xi(\mathbb{R}))} = \sum_{j\in\mathbb{Z}} \big\|(\mathrm{i}j)^\alpha \hat{a}'_j\big\|^2_{L^\infty(\mathbb{R})}$$

with $\hat{a}'_j = \frac{\mathrm{d}}{\mathrm{d}\xi}\hat{a}_j = \widehat{(\partial_\xi a)}_j = (\mathcal{F}_x\partial_\xi a)(j,\xi)$. We shall say that $a \in S_{\sigma,0}$ if $|a|_{\sigma,0} < \infty$ and $a \in S_{\sigma,1}$ if $|a|_{\sigma,1} < \infty$. The use of these seminorms compared to some more classical ones allows us to avoid to lose too many derivatives while keeping very simple proofs. Note that we can easily relate $|a|_{\sigma,0}$ to more classical symbol seminorms up to losing more derivatives. For example, we have for every $\sigma \geq 0$

$$\sup_{|\alpha|\leq\sigma}\ \sup_{x\in\mathbb{T},\,\xi\in\mathbb{R}} |\partial_x^\alpha a(x,\xi)| \leq C|a|_{\sigma+s,0}$$

with $s > \frac{1}{2}$. The lower bound of $\frac{1}{2}$ is related to the Sobolev embedding $H^s \hookrightarrow L^\infty$ in $1d$ and should be generalized to $\frac{d}{2}$ for higher dimensional generalizations of our arguments here.

Writing a symbol $a(x,\xi)$ as a Fourier series with respect to its first variable x,

$$a(x,\xi) = \sum_{j\in\mathbb{Z}} \hat{a}_j(\xi)\mathrm{e}^{\mathrm{i}jx}$$

with Fourier coefficients (68), the quantization (67) takes (formally) the form

$$(\mathrm{Op}_a\, u)(x) = \sum_{l\in\mathbb{Z}} \Big(\sum_{k\in\mathbb{Z}} \hat{a}_{l-k}(k)\hat{u}_k\Big)\mathrm{e}^{\mathrm{i}lx}. \tag{69}$$

Its $L^2(\mathbb{T})$-norm (which we denote in this appendix by $\|\cdot\|_{L^2(\mathbb{T})}$ to avoid confusion with the other norms appearing here) is given by

$$\|\mathrm{Op}_a\, u\|^2_{L^2(\mathbb{T})} = \sum_{l\in\mathbb{Z}} \Big|\sum_{k\in\mathbb{Z}} \hat{a}_{l-k}(k)\hat{u}_k\Big|^2 \leq \sum_{l\in\mathbb{Z}} \Big|\sum_{k\in\mathbb{Z}} \|\hat{a}_{l-k}\|_{L^\infty(\mathbb{R})}|\hat{u}_k|\Big|^2.$$

Noting that the upper bound on the right is the squared $L^2(\mathbb{T})$-norm of the product of the functions with Fourier coefficients $\|\hat{a}_k\|_{L^\infty(\mathbb{R})}$ and $|\hat{u}_k|$, we get from (3a) the following L^2 continuity result.

Proposition A.1. *Assume that $\sigma > \frac{1}{2}$. Then, there exists $C > 0$ such that for every $a \in S_{\sigma,0}$ and for every $u \in L^2(\mathbb{T})$, we have $\mathrm{Op}_a\, u \in L^2(\mathbb{T})$ with*

$$\|\,\mathrm{Op}_a\, u\|_{L^2(\mathbb{T})} \leq C|a|_{\sigma,0}\|u\|_{L^2(\mathbb{T})}.$$

For a very similar result, we refer to [20, Proposition 8.1] which slightly refines in terms of the regularity of the symbols, the classical results of L^2 continuity for symbols in $S^0_{0,0}$ that are compactly supported in x, see for example [30].

We shall now state results of symbolic calculus, see also [20, Proposition 8.2].

Proposition A.2. *Assume that $\sigma > \frac{1}{2}$. Then, there exists $C > 0$ such that for every $a \in S_{\sigma+1,1}$, we have*

$$\left\|(\mathrm{Op}_a)^*(u) - \mathrm{Op}_{\overline{a}}(u)\right\|_{L^2(\mathbb{T})} \leq C|a|_{\sigma+1,1}|u|_{L^2(\mathbb{T})}, \tag{70}$$

where $(\mathrm{Op}_a)^*$ is the adjoint of the operator Op_a for the $L^2(\mathbb{T})$ scalar product. Moreover, we have for every $a \in S_{\sigma,1}$ and $b \in S_{\sigma+1,0}$ that $ab \in S_{\sigma,0}$ and

$$\left\|\mathrm{Op}_a \mathrm{Op}_b(u) - \mathrm{Op}_{ab}(u)\right\|_{L^2(\mathbb{T})} \leq C|a|_{\sigma,1}|b|_{\sigma+1,0}\,\|u\|_{L^2(\mathbb{T})}. \tag{71}$$

Proof. Let us first prove (70). We start by computing a symbol c with

$$\mathrm{Op}_a^* = \mathrm{Op}_c\,.$$

Using that, by (69),

$$\sum_{l\in\mathbb{Z}} \overline{\hat{a}_{l-j}(j)}\hat{u}_l = \left\langle \mathrm{Op}_a\, \mathrm{e}^{\mathrm{i}jx}, u\right\rangle_{L^2(\mathbb{T})} = \left\langle \mathrm{e}^{\mathrm{i}jx}, \mathrm{Op}_c\, u\right\rangle_{L^2(\mathbb{T})} = \sum_{l\in\mathbb{Z}} \hat{c}_{j-l}(l)\hat{u}_l,$$

we define such a symbol c by

$$\hat{c}_j(\xi) = \overline{\hat{a}_{-j}(\xi + j)}, \qquad j \in \mathbb{Z},\ \xi \in \mathbb{R}.$$

Assuming that $a \in S_{\sigma,0}$, we thus also have that $c \in S_{\sigma,0}$ with $|c|_{\sigma,0} = |a|_{\sigma,0}$. Next, by Taylor expansion, we can write

$$d(x,\xi) := c(x,\xi) - \overline{a(x,\xi)} = \sum_{j\in\mathbb{Z}} \int_0^1 j\, \overline{\hat{a}'_{-j}(\xi + sj)}\,\mathrm{d}s\, \mathrm{e}^{\mathrm{i}jx}.$$

We shall prove that $|d|_{\sigma,0} \leq |a|_{\sigma+1,1}$ and the result will follow from Proposition A.1. This estimate follows from

$$(\mathrm{i}j)^\alpha \hat{d}_j(\xi) = (\mathrm{i}j)^\alpha \int_0^1 j\, \overline{\hat{a}'_{-j}(\xi + sj)}\,\mathrm{d}s = -\mathrm{i}\int_0^1 \overline{(-\mathrm{i}j)^{\alpha+1}\hat{a}'_{-j}(\xi + sj)}\,\mathrm{d}s,$$

which implies $\|(\mathrm{i}j)^\alpha \hat{d}_j\|_{L^\infty(\mathbb{R})} \leq \|(-\mathrm{i}j)^{\alpha+1}\hat{a}'_{-j}\|_{L^\infty(\mathbb{R})}$, and ends the proof of (70).

Let us now prove (71). We first observe that for $\sigma > \frac{1}{2}$ and $a \in S_{\sigma,0}$, $b \in S_{\sigma,0}$, we have by (3a) (applied with functions u and v with Fourier coefficients $\hat{u}_j = \|\hat{a}_j\|_{L^\infty(\mathbb{R})}$ and $\hat{v}_j = \|\hat{b}_j\|_{L^\infty(\mathbb{R})}$) that

$$|ab|_{\sigma,0} \leq C|a|_{\sigma,0}|b|_{\sigma,0}, \tag{72}$$

and thus that $ab \in S_{\sigma,0}$.

Next, we compute a symbol e with

$$\mathrm{Op}_a \mathrm{Op}_b = \mathrm{Op}_e\,.$$

Such a symbol is obtained by writing Op_b as in (69) and Op_a as in (67), which yields

$$e(x,\xi) = \sum_{j\in\mathbb{Z}} a(x,\xi + j)\hat{b}_j(\xi)\mathrm{e}^{\mathrm{i}jx}.$$

We then get that

$$f(x,\xi) := e(x,\xi) - a(x,\xi)b(x,\xi) = \sum_{j\in\mathbb{Z}} \int_0^1 \partial_\xi a(x,\xi + sj)\,\mathrm{d}s\, j\, \hat{b}_j(\xi)\mathrm{e}^{\mathrm{i}jx}.$$

We next estimate a suitable seminorm of the symbol f to apply Proposition A.1. By taking the Fourier transform in x, we obtain that

$$(\mathrm{i}l)^\alpha \hat{f}_l(\xi) = (\mathrm{i}l)^\alpha \sum_{k\in\mathbb{Z}} \int_0^1 \hat{a}'_{l-k}(\xi + sk)\,\mathrm{d}s\, k\, \hat{b}_k(\xi).$$

We thus have for $\alpha \geq 0$

$$\big\|(\mathrm{i}l)^\alpha \hat{f}_l\big\|_{L^\infty(\mathbb{R})} \leq |l|^\alpha \sum_{k\in\mathbb{Z}} \big\|\hat{a}'_{l-k}\big\|_{L^\infty(\mathbb{R})} \big\|(\mathrm{i}k)\hat{b}_k\big\|_{L^\infty(\mathbb{R})},$$

from which we obtain for $\sigma > \frac{1}{2}$ by (3a) that

$$|f|_{\sigma,0} \leq C|a|_{\sigma,1}|b|_{\sigma+1,0}.$$

Since by definition of f, we have $\mathrm{Op}_a \mathrm{Op}_b - \mathrm{Op}_{ab} = \mathrm{Op}_f$, the result follows from Proposition A.1. □

We shall next define a semiclassical version of the above calculus which is the one of interest for us. For any symbol $a(x,\xi)$ as above, we set for $0 < \tau \leq 1$

$$a^\tau(x,\xi) = a(x,\tau\xi)$$

and we define

$$\mathrm{Op}^\tau_a u = \mathrm{Op}_{a^\tau} u.$$

For this calculus, we have the following result, see also [20, Proposition 8.3].

Proposition A.3. *Assume that $\sigma > \frac{1}{2}$. Then, there exists $C > 0$ such that for every $0 < \tau \leq 1$, we have*

- *for every $a \in S_{\sigma,0}$*
$$\big\|\mathrm{Op}^\tau_a u\big\|_{L^2(\mathbb{T})} \leq C|a|_{\sigma,0}\|u\|_{L^2(\mathbb{T})},$$
- *for every $a \in S_{\sigma,1}$ and for every $b \in S_{\sigma+1,0}$*
$$\big\|\mathrm{Op}^\tau_a \mathrm{Op}^\tau_b(u) - \mathrm{Op}^\tau_{ab}(u)\big\|_{L^2(\mathbb{T})} \leq C\tau|a|_{\sigma,1}|b|_{\sigma+1,0}\,\|u\|_{L^2(\mathbb{T})}$$
- *for every $a \in S_{\sigma+1,1}$*
$$\big\|(\mathrm{Op}^\tau_a)^*(u) - \mathrm{Op}^\tau_{\bar{a}}(u)\big\|_{L^2(\mathbb{T})} \leq C\tau|a|_{\sigma+1,1}\,\|u\|_{L^2(\mathbb{T})}.$$

Proof. The results are direct consequences of Propositions A.1 and A.2 since for any symbol a, we have by definition of a^τ that $|a^\tau|_{\sigma,0} = |a|_{\sigma,0}$ and $|a^\tau|_{\sigma,1} = \tau|a|_{\sigma,1}$. □

Let us finally state the semiclassical Gårding inequality.

Proposition A.4. *Assume that $\sigma > \frac{1}{2}$. For $a \in S_{\sigma+1,0} \cap S_{\sigma+1,1}$ assume further that there exists $\delta > 0$ such that*

$$a(x,\xi) \geq \delta \qquad \text{for all} \qquad x \in \mathbb{T},\ \xi \in \mathbb{R}.$$

Then, there exists $C > 0$ which depends only on $|a|_{\sigma+1,0}$, $|a|_{\sigma+1,1}$ and δ such that

$$\langle \mathrm{Op}^\tau_a u, u\rangle_{L^2(\mathbb{T})} \geq \tfrac{1}{2}\delta\|u\|^2_{L^2(\mathbb{T})} - C\tau\|u\|^2_{L^2(\mathbb{T})} \qquad \text{for all} \qquad 0 < \tau \leq 1.$$

Proof. We can write that

$$a(x,\xi) = \tfrac{1}{2}\delta + b(x,\xi)^2, \qquad b(x,\xi) = \big(a(x,\xi) - \tfrac{1}{2}\delta\big)^{1/2}.$$

We will show below that, since $a \geq \delta > 0$, we also have that $b \in S_{\sigma+1,0} \cap S_{\sigma+1,1}$ with

$$|b|_{\sigma+1,0} \leq C \qquad \text{and} \qquad |b|_{\sigma+1,1} \leq C, \tag{73}$$

where C depends only on $|a|_{\sigma+1,0}$, $|a|_{\sigma+1,1}$ and δ. By using Proposition A.3, we thus get that

$$\mathrm{Op}_a^\tau = \tfrac{1}{2}\delta + (\mathrm{Op}_b^\tau)^* \,\mathrm{Op}_b^\tau + R^\tau$$

with

$$\|R^\tau u\|_{L^2(\mathbb{T})} \leq C\tau \|u\|_{L^2(\mathbb{T})}.$$

The result follows easily.

It remains to show (73). We restrict here to $\sigma = 1$, which is the value of σ that is needed in Section 4. The proof for other values of σ is similar, but with longer formulas. In the following, we write

$$F(y) = \big(y - \tfrac{1}{2}\delta\big)^{1/2},$$

such that $b(x,\xi) = F(a(x,\xi))$ and we observe that F is a smooth function on $[\delta, +\infty[$.

For the first estimate of (73), we start from

$$|b|_{2,0} \leq |F(a)|_{0,0} + \big|\partial_x^2 F(a)\big|_{0,0} \leq |F(a)|_{0,0} + \big|F'(a)\partial_x^2 a\big|_{0,0} + \big|F''(a)(\partial_x a)^2\big|_{0,0}.$$

To estimate the products further, we use the estimate $|cd|_{0,0} \leq |c|_{0,0}|d|_{1,0}$, which follows from (3a) in the same way as (72). This yields

$$|b|_{2,0} \leq |F(a)|_{0,0} + |F'(a)|_{1,0}|\partial_x^2 a|_{0,0} + |F''(a)|_{1,0}|\partial_x a|_{1,0}|\partial_x a|_{0,0}. \tag{74}$$

To finish the proof, we just need to explain how to estimate $|G(a)|_{1,0}$ for some smooth G which is smooth on the image of a. We start from

$$\begin{aligned}|G(a)|_{1,0} &\leq \big\|\mathcal{F}_x\big((1+\partial_x)G(a)\big)\big\|_{L^2_j(\mathbb{Z},L^\infty_\xi(\mathbb{R}))} \leq C\big\|\mathcal{F}_x\big((1-\partial_x^2)G(a)\big)\big\|_{L^\infty_j(\mathbb{Z},L^\infty_\xi(\mathbb{R}))} \\ &\leq C\big\|\mathcal{F}_x\big((1-\partial_x^2)G(a)\big)\big\|_{L^\infty_\xi(\mathbb{R},L^2_j(\mathbb{Z}))} = C\|G(a)\|_{L^\infty_\xi(\mathbb{R},H^2_x(\mathbb{T}))},\end{aligned}$$

where the second estimate follows from

$$\begin{aligned}\sum_{j\in\mathbb{Z}} \big\|\big(\mathcal{F}_x(1+\partial_x)G(a)\big)(j)\big\|^2_{L^\infty_\xi(\mathbb{R})} &= \sum_{j\in\mathbb{Z}} \frac{1+j^2}{(1+j^2)^2}\big\|\big(\mathcal{F}_x(1-\partial_x^2)G(a)\big)(j)\big\|^2_{L^\infty_\xi(\mathbb{R})} \\ &\leq \Big(\sum_{j\in\mathbb{Z}} \frac{1}{1+j^2}\Big) \sup_{j\in\mathbb{Z}} \big\|\big(\mathcal{F}_x(1-\partial_x^2)G(a)\big)(j)\big\|^2_{L^\infty_\xi(\mathbb{R})},\end{aligned}$$

and the third estimate follows from interchanging the two L^∞-norms and estimating the $L^\infty_j(\mathbb{Z})$-norm by the $L^2_j(\mathbb{Z})$-norm.

Next, we can use (3b) to get that for every ξ,

$$\|G(a)\|_{H^2_x(\mathbb{T})} \leq \Lambda(\|a\|_{H^2_x(\mathbb{T})}))(1 + \|a\|_{H^2_x(\mathbb{T})})$$

where $\Lambda(\cdot)$ stands again for a continuous non-decreasing function that can change from line to line as a stand in for the dependence upon the algebra of the calculus established here. Therefore we finally obtain that

$$\|G(a)\|_{L^\infty_\xi(\mathbb{R},H^2_x(\mathbb{T}))} \leq \Lambda(\|a\|_{L^\infty_\xi(\mathbb{R},H^2_x(\mathbb{T}))})(1 + \|a\|_{L^\infty_\xi(\mathbb{R},H^2_x(\mathbb{T}))}) \leq \Lambda(|a|_{2,0}).$$

Using this estimate and the above estimate of $|G(a)|_{1,0}$ in (74) completes the proof of the first estimate of (73). The proof of the second estimate of (73) is very similar. □

Acknowledgement

We thank the referees for their valuable suggestions and comments. This work was supported by Deutsche Forschungsgemeinschaft (DFG) through project GA 2073/2-1 (LG), SFB 1114 (LG) and SFB 1173 (KS) and by National Science Foundation (NSF) under grants DMS-1454939 (JL), DMS-1312874 (JLM) and DMS-1352353 (JLM).

References

[1] W. Bao, X. Dong, Analysis and comparison of numerical methods for the Klein-Gordon equation in the nonrelativistic limit regime, *Numer. Math.* **120** (2012), 189–229.

[2] B. Cano, Conservation of invariants by symmetric multistep cosine methods for second-order partial differential equations, *BIT* **53** (2013), 29–56.

[3] B. Cano, M. J. Moreta, Multistep cosine methods for second-order partial differential systems, *IMA J. Numer. Anal.* **30** (2010), 431–461.

[4] M. Chirilus-Bruckner, W.-P. Düll, G. Schneider, NLS approximation of time oscillatory long waves for equations with quasilinear quadratic terms, *Math. Nachr.* **288** (2015), 158–166.

[5] C. Chong, G. Schneider, Numerical evidence for the validity of the NLS approximation in systems with a quasilinear quadratic nonlinearity, *ZAMM Z. Angew. Math. Mech.* **93** (2013), 688–696.

[6] D. Cohen, E. Hairer, C. Lubich, Conservation of energy, momentum and actions in numerical discretizations of non-linear wave equations, *Numer. Math.* **110** (2008), 113–143.

[7] X. Dong, Stability and convergence of trigonometric integrator pseudospectral discretization for N-coupled nonlinear Klein-Gordon equations, *Appl. Math. Comput.* **232** (2014), 752–765.

[8] W. Dörfler, H. Gerner, R. Schnaubelt, Local well-posedness of a quasilinear wave equation, *Appl. Anal.* **95** (2016), 2110–2123.

[9] W.-P. Düll, Justification of the nonlinear Schrödinger approximation for a quasilinear Klein–Gordon equation, *Comm. Math. Phys.* **355** (2017), 1189–1207.

[10] B. García-Archilla, J. M. Sanz-Serna, R. D. Skeel, Long-time-step methods for oscillatory differential equations, *SIAM J. Sci. Comput.* **20** (1999), 930–963.

[11] L. Gauckler, Error analysis of trigonometric integrators for semilinear wave equations, *SIAM J. Numer. Anal.* **53** (2015), 1082–1106.

[12] L. Gauckler, J. Lu, J. L. Marzuola, F. Rousset, K. Schratz, Trigonometric integrators for quasilinear wave equations, previous version of the present paper, arXiv:1702.02981v1, 2017.

[13] L. Gauckler, D. Weiss, Metastable energy strata in numerical discretizations of weakly nonlinear wave equations, *Discrete Contin. Dyn. Syst.* **37** (2017), 3721–3747.

[14] C. González, M. Thalhammer, Higher-order exponential integrators for quasi-linear parabolic problems. Part I: Stability, *SIAM J. Numer. Anal.* **53** (2015), 701–719.

[15] C. González, M. Thalhammer, Higher-order exponential integrators for quasi-linear parabolic problems. Part II: Convergence, *SIAM J. Numer. Anal.* **54** (2016), 2868–2888.

[16] V. Grimm, M. Hochbruck, Error analysis of exponential integrators for oscillatory second-order differential equations, *J. Phys. A*, **39** (2006), 5495–5507.

[17] M. D. Groves, G. Schneider, Modulating pulse solutions for quasilinear wave equations, *J. Differential Equations*, **219** (2005), 221–258.

[18] E. Hairer, C. Lubich, Long-time energy conservation of numerical methods for oscillatory differential equations, *SIAM J. Numer. Anal.* **38** (2000), 414–441.

[19] E. Hairer, C. Lubich, G. Wanner, *Geometric numerical integration. Structure-*

preserving algorithms for ordinary differential equations, second ed., vol. 31 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2006.

[20] D. Han-Kwan, F. Rousset, Quasineutral limit for Vlasov–Poisson with Penrose stable data, *Ann. Sci. Éc. Norm. Supér. (4)* **49** (2016), 1445–1495.

[21] M. Hochbruck, T. Pažur, Error analysis of implicit Euler methods for quasilinear hyperbolic evolution equations, *Numer. Math.* **135** (2017), 547–569.

[22] L. Hörmander, *Lectures on nonlinear hyperbolic differential equations*, vol. 26 of Mathématiques & Applications, Springer-Verlag, Berlin, 1997.

[23] T. J. R. Hughes, T. Kato, J. E. Marsden, Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity, *Arch. Rational Mech. Anal.* **63** (1977), 273–294.

[24] 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)*, 25–70. Lecture Notes in Math., vol. 448, Springer, Berlin, 1975.

[25] B. Kovács, C. Lubich, Stability and convergence of time discretizations of quasi-linear evolution equations of Kato type, *Numer. Math.* (2017), doi:10.1007/s00211-017-0909-3.

[26] P.-L. Lions, B. Perthame, and E. Tadmor, Kinetic formulation of the isentropic gas dynamics and p-system. *Comm. Math. Phys.* **163** (1994), 415–431.

[27] J. Lu, J. L. Marzuola, Strang splitting methods for a quasilinear Schrödinger equation: convergence, instability, and dynamics, *Commun. Math. Sci.* **13** (2015), 1051–1074.

[28] C. Lubich, A. Ostermann, Runge-Kutta approximation of quasi-linear parabolic equations, *Math. Comp.* **64** (1995), 601–627.

[29] C. D. Sogge, *Lectures on non-linear wave equations*, International Press Boston, 1995.

[30] M. E. Taylor, *Pseudodifferential operators*, vol. 34 of Princeton Mathematical Series, Princeton University Press, 1981.

[31] M. E. Taylor, *Pseudodifferential operators and nonlinear PDE*, vol. 100 of Progress in Mathematics, Birkhäuser Boston, Inc., Boston, MA, 1991.

[32] M. E. Taylor, *Partial differential equations III. Nonlinear equations.*, vol. 117 of Applied Mathematical Sciences, Springer, New York, 2011.