The Fourier spectral approach to the spatial discretization of quasilinear hyperbolic systems

Vincent Duchêne[1] and Johanna Ulvedal Marstrander[2]

[1]IRMAR, Univ. Rennes , F-35000 Rennes, France.
[2]Department of Mathematical Sciences, NTNU, 7491 Trondheim, Norway

July 2, 2025

Abstract

We discuss the rigorous justification of the spatial discretization by means of Fourier spectral methods of quasilinear first-order hyperbolic systems. We provide uniform stability estimates that grant spectral convergence of the (spatially) semi-discretized solutions towards the corresponding continuous solution provided that the underlying system satisfies some suitable structural assumptions. We consider a setting with sharp low-pass filters and a setting with smooth low-pass filters and argue that —at least theoretically— smooth low-pass filters are operable on a larger class of systems. While our theoretical results are supported with numerical evidence, we also pinpoint some behavior of the numerical method that currently has no theoretical explanation.

Keywords. fourier spectral methods, spectral convergence, hyperbolic systems
MSC codes. 65M12, 65M70, 76M22

Contents

1 Introduction

In this work we shall consider the spatial discretization by means of Fourier spectral methods of systems of the form

$$\partial_t \boldsymbol{U} + \sum_{j=1}^{d} A_j(\boldsymbol{U})\partial_{x_j}\boldsymbol{U} = \boldsymbol{0}, \quad \boldsymbol{U}|_{t=0} = \boldsymbol{U}^0. \tag{1.1}$$

where for all $j \in \{1, \dots, d\}$ and for all $\boldsymbol{U} \in \mathbb{R}^n$, $A_j(\boldsymbol{U})$ are matrices satisfying Assumption **A.1**.

Assumption A.1. *For all $j \in \{1, \dots, d\}$ and $\boldsymbol{U} \in \mathbb{R}^n$, $A_j(\boldsymbol{U})$ are n-by-n real-valued matrices. We assume that all entries of $A_j(\cdot)$ are polynomial.*

Remark 1.1. *Assuming that the system* (1.1) *has only polynomial nonlinearities may seem an over-restrictive assumption. This assumption is motivated by two considerations. Firstly, the Fourier spectral method may be efficiently implemented only within this framework; see Remark 1.2 below. Secondly, as far as we know, the literature lacks a theory analogous to para-differential calculus (see* e.g. *[19, Chapter 5]) on Sobolev spaces of periodic functions, which would provide the composition estimates in Proposition A.4 for general composition functions.*

We shall assume further additional structural assumptions on the system (1.1), depending on the needs. In particular we shall always consider Friedrichs-symmetrizable systems, which guarantees the hyperbolicity of the system, and local-in-time well-posedness of the initial-value problem in L^2-based Sobolev spaces of sufficiently high regularity index, $\boldsymbol{U}^0 \in H^s((2\pi\mathbb{T})^d)^n$ with $s > d/2+1$ (see Subsection 1.2). For simplicity, we consider in this work 2π-periodic functions in all spatial directions. The results extend straightforwardly to more general (periodic) frameworks, and analogous results in the n-dimensional Euclidean space could be obtained with some simple adaptations.

Let $\mathcal{T}_N$ be the space of trigonometric polynomials of degree N:

$$\mathcal{T}_N := \text{span}\{\exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}), \boldsymbol{k} \in \mathbb{Z}^d, |k_j| \leqslant N, j = 1, \dots, d\}, \text{ and } \mathcal{T}_N^n := \underbrace{\mathcal{T}_N \times \dots \times \mathcal{T}_N}_{n \text{ times}}.$$

Let $\mathsf{P}_N \colon L^2((2\pi\mathbb{T})^d)^n \to \mathcal{T}_N^n$ be the L^2-projection operator onto $\mathcal{T}_N^n$: $\mathsf{P}_N = \text{Diag}(P_N(D))$, where P_N is a Fourier multiplier with symbol $P_N = \mathbf{1}_{[\![-N,N]\!]^d}$. We refer to P_N as a **sharp low-pass filter**. For any $0 \leqslant r \leqslant s$ and $\boldsymbol{U} \in H^s((2\pi\mathbb{T})^d)^n$ Sobolev space of order s, P_N satisfies (see *e.g.* [5, (5.1.10),(5.8.4)])

$$|\boldsymbol{U} - \mathsf{P}_N \boldsymbol{U}|_{H^r} \leqslant C(d, s, r) \langle N \rangle^{r-s} |\boldsymbol{U}|_{H^s}. \tag{1.2}$$

This kind of estimates are referred to in the literature as *spectral convergence*, and we will follow this terminology.

The standard Fourier spectral method (see *e.g.* [16, 13, 18] or [1, Section 3]) for the spatial discretization of the problem (1.1) amounts to seeking solutions $\boldsymbol{U}_N \colon t \mapsto \mathcal{T}_N^n$ to the problem

$$\partial_t \boldsymbol{U}_N + \mathsf{P}_N \left(\sum_{j=1}^{d} A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N \right) = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N\boldsymbol{U}^0. \tag{1.3}$$

Remark 1.2. *Let us recall that one of the great assets of Fourier spectral methods is that spatial differentiation and multiplication can be very efficiently performed (up to machine precision) by means of Fast Fourier Transform (FFT/IFFT) and multiplication at spatial collocations points, provided suitable dealiasing operations are performed. In practice, if the entries of $A_j(\cdot)$ are all polynomial with maximal degree p, then one computes (following Orszag's rule [20]) $N\frac{p+2}{2}$ modes, and applying the projection P_N after multiplications at collocation points performs the necessary dealiasing. If entries of $A_j(\cdot)$ are not polynomials, then one typically uses* pseudo-*spectral schemes that follow the aforementioned strategy but cannot be formulated as in* (1.3).

We do not discuss in this work the full time-space discretization of (1.1), *that is well-suited numerical time integrators for* (1.3). *The implementation we used for numerical experiments is described in more detail in Subsection 4.2.*

In this work we shall discuss the long-time (*i.e.* uniform with respect to N) existence of solutions to (1.3) **and spectral convergence towards solutions to** (1.1) **as $N \to \infty$.**

Our results will depend on the structure of the system. As we shall see, the solution to (1.3) converges towards the corresponding solution to the problem (1.1) (assuming sufficient regularity) whenever the system is symmetric. If the system is only symmetrizable, the situation is more complicated. In order to deal with this situation we consider **smooth low-pass filters**, $\mathsf{S}_N \colon L^2((2\pi\mathbb{T})^d)^n \to \mathcal{T}_N^n$ where $\mathsf{S}_N = \mathrm{Diag}(S_N(D))$ is a Fourier multiplier with symbol $S_N(\cdot) = S(\cdot/N)$ where $S \in \mathcal{C}^1(\mathbb{R}^d)$ is even and satisfies

$$\begin{cases} S(\boldsymbol{k}) = 1 & \text{if } \max_{j=1,\ldots,d} |k_j| \leqslant 1/2, \\ S(\boldsymbol{k}) = 0 & \text{if } \min_{j=1,\ldots,d} |k_j| \geqslant 1, \\ S(\boldsymbol{k}) \in [0,1] & \text{otherwise.} \end{cases}$$

and $|\cdot| \nabla S^{1/2} \in L^\infty(\mathbb{R}^d)$. When $d = 1$, an example of such a function could be $S_1(\cdot) \coloneqq \max\left(0, \min\left(1, 2 - 2|\cdot|\right)\right)^2$. When $d \geqslant 2$, one can set $S_d(k_1, \ldots, k_d) \coloneqq S_1(k_1) \times \cdots \times S_1(k_d)$. The advantage of such smooth low-pass filters is that —contrarily to the sharp low-pass filter— they satisfy commutator estimates with gains of regularity uniformly with respect to N; see Proposition A.6. These are crucial to parts of the analysis. Notice that because $|1 - S(\boldsymbol{k})| \leqslant 1 = |1 - P_{1/2}(\boldsymbol{k})|$ when $\max_{j=1,\ldots,d} |k_j| \geqslant 1/2$ and $1 - S(\boldsymbol{k}) = 0 = 1 - P_{1/2}(\boldsymbol{k})$ otherwise, we infer from (1.2) the identical spectral convergence estimate

$$|\boldsymbol{U} - \mathsf{S}_N \boldsymbol{U}|_{H^r} \leqslant \left|\boldsymbol{U} - \mathsf{P}_{N/2}\boldsymbol{U}\right|_{H^r} \leqslant C(d, s, r)\langle N\rangle^{r-s} |\boldsymbol{U}|_{H^s}. \tag{1.4}$$

The spatial discretization of (1.1) using smooth low-pass filters could amount to finding a solution $\boldsymbol{U}_N \colon t \mapsto \mathcal{T}_N^n$ to

$$\partial_t \boldsymbol{U}_N + \mathsf{S}_N \left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N \right) = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N\boldsymbol{U}^0. \tag{1.5}$$

We also consider variants of this system such as

$$\partial_t \boldsymbol{U}_N + \sum_{j=1}^d (A_j^0 + \mathsf{S}_N(A_j^1(\boldsymbol{U}_N)[\circ]))\partial_{x_j}\boldsymbol{U}_N = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N\boldsymbol{U}^0. \tag{1.6}$$

where $A_j^0 = A_j(\boldsymbol{0})$ and $A_j^1(\boldsymbol{U}) = A_j(\boldsymbol{U}) - A_j^0$. Indeed, applying smooth low-pass filters to linear terms is unnecessary, especially when one uses exponential time integrators (see *e.g.* Program 27 in [25]). Notice the distinction between (1.5) and (1.6) is only necessary when using the smooth low-pass filter since when using the sharp low-pass filter P_N one has $\mathsf{P}_N\boldsymbol{U}_N = \boldsymbol{U}_N$ and P_N commutes with A_j^0.

Outline Let us now describe the structure and main results of this work. Symmetric systems are discussed in Section 2. We consider sharp and smooth low-pass filters in Subsection 2.1 and 2.2 respectively. We obtain convergence of the semi-discretized solutions in both cases, stated in Propositions 2.4 and 2.6. Symmetrizable systems are discussed in Section 3. Subsection 3.1 concerns smooth low-pass filters and we obtain analogous convergence results, stated in Proposition 3.4. The case with sharp low-pass filters is treated in Subsection 3.2 and in order to secure spectral convergence, more stringent structural assumptions on the system are required. This yields Proposition 3.9. Numerical experiments illustrating and investigating the sharpness of our theoretical results are provided in Section 4.

1.1 Motivation and related works

Our work was motivated by the study of Boussinesq and Whitham–Boussinesq systems that are nonlinear dispersive models for the propagation of surface gravity waves [17, 7]. The Fourier spectral method is especially indicated for the spatial discretization of these systems since the nonlinear contributions are quadratic (recall Remark 1.2) and the

dispersive contributions take the form of Fourier multipliers. The spectral convergence of discretized versions of Boussinesq models towards the corresponding continuous solutions was proved in [26] and [6]. These results have a shortcoming in that they lack uniformity in the non-dispersive limit. In order to clarify this point, let us consider the Benjamin–Bona–Mahony (BBM) and (a variant of) the Whitham equations, which read respectively

$$(\mathrm{Id} -\mu\partial_x^2)\partial_t u + \partial_x u + u\partial_x u = 0, \quad \text{and} \quad \partial_t u + \frac{\tanh(\sqrt{\mu}|D|)}{\sqrt{\mu}|D|}\big(\partial_x u + u\partial_x u\big) = 0.$$

The aforementioned Boussinesq (resp. Whitham–Boussinesq) systems can be loosely considered as systems generalizing the BBM (resp. Whitham) scalar equation, in the same way the Saint-Venant system discussed in Section 4 generalizes the inviscid Burgers equation. All these equations provide valid approximations of water waves provided (among other assumptions) the shallowness parameter $\mu > 0$ satisfies $\mu \ll 1$; see [17, 7].

Results proved in [26] and [6] and adapted to the simplest case of scalar equations are of the form

$$\big|u_N(t,\cdot)\big|_{H^s} \leqslant C_\mu(t)\big|u_N\big|_{t=0}\big|_{H^s} \quad \text{and} \quad \big|(u - u_N)(t,\cdot)\big|_{H^r} \leqslant C_\mu(t)N^{r-s}$$

where u is the solution to the scalar equation, and u_N the corresponding solution to the semi-discretized equation

$$(\mathrm{Id} -\mu\partial_x^2)\partial_t u_N + \partial_x u_N + \mathsf{P}_N(u_N\partial_x u_N) = 0, \quad \text{and} \quad \partial_t u_N + \frac{\tanh(\sqrt{\mu}|D|)}{\sqrt{\mu}|D|}\big(\partial_x u_N + \mathsf{P}_N(u_N\partial_x u_N)\big) = 0.$$

The aforementioned shortcoming is that $C_\mu(t)$ depends nonuniformly on μ in the non-dispersive limit $\mu \ll 1$, typically $C_\mu(t) \lesssim \exp(\mu^{-1}t)$. This is inconsistent with the standard well-posedness theory for initial data in Sobolev spaces $H^s(2\pi\mathbb{T})$, $s > 3/2$, which holds uniformly with respect to $\mu \in (0,1]$ (see *e.g.* [15, Proposition 6] and [22, 10, 21] for the corresponding results on the Boussinesq and Whitham–Boussinesq systems). The reason for this discrepancy is that for the spectral convergence results, stability estimates on the BBM and Whitham equations are obtained by viewing them as *semilinear systems* which can be formulated using the Duhamel formula

$$u(t,\cdot) = \int_0^t e^{-(t-\tau)L_\mu(D)} L_\mu(D)\big(\frac{u(\tau,\cdot)^2}{2}\big)\,\mathrm{d}\tau$$

where $L_\mu(D) = \frac{\partial_x}{\mathrm{Id} -\mu\partial_x^2}$ for the BBM equation, and $L_\mu(D) = \frac{\tanh(\sqrt{\mu}|D|)}{\sqrt{\mu}|D|}\partial_x$ for the Whitham equation. Notice that in both cases $L_\mu(D) \in \mathcal{B}(L^2(2\pi\mathbb{T}); L^2(2\pi\mathbb{T}))$, but that $\sup\{\big|L_\mu v\big|_{L^2} : \big|v\big|_{L^2} = 1\}$ is not uniformly bounded with respect to $\mu \in (0,1]$, which is the source of the issue when $\mu \ll 1$. On the other hand, one can view the BBM and Whitham equation as perturbations of the inviscid Burgers equation, with skew-symmetric dispersive terms that are inconspicuous for the energy method. This leads to stability estimates in $H^s(2\pi\mathbb{T})$, $s > 3/2$, that are uniform with respect to $\mu \ll 1$. Of course obtaining such results on dispersive *systems* requires a good understanding of the underlying (non-dispersive) quasilinear systems, which is the focus of the current work.

The rigorous analysis of semi-discretized or fully (space and time) discretized Fourier spectral schemes for semilinear equations is a very rich and active topic, which is impossible to summarize within a few lines; let us simply mention [4, 3] which are particularly relevant as they specifically consider the BBM equation and pay attention to the non-dispersive limit (although together with vanishing nonlinearity, that is the long wave limit). In contrast, to the best of our knowledge, there are only a handful of works dedicated to the rigorous analysis of the Fourier method for *quasilinear systems*, culminating with the work of Bardos and Tadmor [1] (following [23, 24, 11]; see also the review paper [12]). Here the authors consider the inviscid Burgers equation, as well as the one-dimensional isentropic Euler equation and the incompressible Euler equations. The first two systems belong to the class of equations we study. Specifically, the inviscid Burgers equation (as any scalar equation) has the symmetric structure we employ in Subsection 2.1, while the isentropic Euler equation in Lagrangian coordinates enjoys the Hamiltonian structure discussed in Subsection 3.2. The incompressible Euler equations

$$\partial_t \boldsymbol{u} + \mathbb{P}((\boldsymbol{u} \cdot \nabla)\boldsymbol{u}) = \boldsymbol{0}, \ \mathbb{P}(\boldsymbol{u}) = \boldsymbol{u}, \qquad \mathbb{P}(\boldsymbol{u}) \coloneqq \boldsymbol{u} - \nabla\Delta^{-1}\nabla \cdot \boldsymbol{u}$$

does not belong to the class of equations studied here due to the Leray projection operator $\mathbb{P}$ but it would not be difficult to extend our analysis (specifically concerning the symmetric situation) to this system. Let us mention that while our analysis is very similar, the estimates obtained in [1] are not as sharp as the ones obtained in our work due to different choices when performing stability estimates, as they take the form (see Theorem 3.1 therein)

$$\left|(\boldsymbol{U} - \boldsymbol{U}_N)(t, \cdot)\right|_{L^2} \leqslant C(t) N^{-s} \left|\boldsymbol{U}\right|_{t=0}\big|_{H^s} + N^{\frac{3}{4} - \frac{s}{2}} \max_{\tau \leqslant t} \left|\boldsymbol{U}(\tau, \cdot)\right|_{H^s},$$

while our results do not feature the second contribution. Quite interestingly, the authors also prove the *emergence of spurious oscillations* of the semi-discretized solution after the critical time of shock formation for the continuous solution of the inviscid Burgers equation. It would be interesting to study this problem when smooth low-pass filters are used and compare with the spectral viscosity method described therein.

1.2 Definitions and notations

In this section, we introduce a few notations used throughout the work.

Let $L^2((2\pi\mathbb{T})^d)$ be the Lebesgue space of real-valued, square-integrable functions on the 2π-periodic torus and

$$L^2((2\pi\mathbb{T})^d)^n = \underbrace{L^2((2\pi\mathbb{T})^d) \times \ldots \times L^2((2\pi\mathbb{T})^d)}_{n \text{ times}}.$$

We endow $L^2((2\pi\mathbb{T})^d)$ with the standard Lebesgue norm denoted $\left| \cdot \right|_{L^2}$, and the corresponding inner-product is denoted $\left(\cdot, \cdot\right)_{L^2}$. Similarly, we denote by $L^\infty((2\pi\mathbb{T})^d)$ the Lebesgue space of bounded functions and $W^{1,\infty}((2\pi\mathbb{T})^d)$ the space of Lipschitz continuous functions, endowed with their natural norms.

We use the notation $\langle\cdot\rangle = (1 + |\cdot|^2)^{1/2}$ and $\Lambda^s = (1 - \Delta)^{s/2}$, i.e. the Fourier multiplier with symbol $\langle\cdot\rangle^s$. For real $s \geqslant 0$, we denote the L^2-based periodic Sobolev spaces by $H^s((2\pi\mathbb{T})^d)^n$:

$$H^s((2\pi\mathbb{T})^d)^n = \{\boldsymbol{U} \in L^2((2\pi\mathbb{T})^d)^n, |\boldsymbol{U}|_{H^s} < \infty\}, \text{ where}$$

$$|\boldsymbol{U}|_{H^s}^2 = |\Lambda^s \boldsymbol{U}|_{L^2}^2 = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \langle\boldsymbol{k}\rangle^{2s} \left|\widehat{\boldsymbol{U}}_{\boldsymbol{k}}\right|^2$$

where $\widehat{\boldsymbol{U}}_{\boldsymbol{k}}$ is the $\boldsymbol{k}$-th fourier coefficient of $\boldsymbol{U}$. For X a Banach space and $I \subset \mathbb{R}$ an interval, the space of X-valued continuous functions on I is denoted $\mathcal{C}(I; X)$. Given $n \in \mathbb{N}$, the space of continuously n-th differentiable functions is denoted $\mathcal{C}^n(I; X)$.

As mentioned previously, we denote $\mathcal{T}_N$ the space of trigonometric polynomials of degree N:

$$\mathcal{T}_N := \text{span}\{\exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{x}), \boldsymbol{k} \in \mathbb{Z}^d, |k_j| \leqslant N, j = 1, \ldots, d\}, \text{ and } \mathcal{T}_N^n := \underbrace{\mathcal{T}_N \times \ldots \times \mathcal{T}_N}_{n \text{ times}},$$

and $\mathsf{P}_N := \text{Diag}(P_N(D))$ the Fourier multiplier with symbol $P_N := \mathbf{1}_{[\![-N,N]\!]^d}$ is the L^2-projection operator onto $\mathcal{T}_N^n$. Here, $[\![-N, N]\!] := \{-N, -N+1, \ldots, N-1, N\}$.

We set $\mathsf{S}_N = \text{Diag}(S_N(D))$ a Fourier multiplier with symbol $S_N(\cdot) = S(\cdot/N)$ where $S \in \mathcal{C}^1(\mathbb{R}^d)$ is even and satisfies

$$\begin{cases} S(\boldsymbol{k}) = 1 & \text{if } \max_{j=1,\ldots,d} |k_j| \leqslant 1/2, \\ S(\boldsymbol{k}) = 0 & \text{if } \min_{j=1,\ldots,d} |k_j| \geqslant 1, \\ S(\boldsymbol{k}) \in [0, 1] & \text{otherwise.} \end{cases}$$

and $|\cdot| \nabla S^{1/2} \in L^\infty(\mathbb{R}^d)$. Apart from these properties, the specific profile of the symbol S is inconsequential.

We denote by $C(\lambda_1, \lambda_2, \ldots)$ a positive "constant" depending on its parameters. Whenever such a parameter represents the norm of a function, C depends non-decreasingly on said norm. Whenever the parameter is a subset of the Euclidean space, C depends non-decreasingly on this parameter when set inclusion is used as (partial) ordering. Dependency on regularity indices $s \in \mathbb{R}$ or the dimension d are omitted when it is unessential or clear from the context. Note that such constants C will always be independent of the degree of the approximation N.

2 Symmetric quasilinear systems

In this section, we consider systems (1.1), where for all $j \in \{1, \ldots, d\}$, $A_j(\boldsymbol{U})$ satisfies the Assumption **A.1** and is additionally self-adjoint. We have the following standard result (see *e.g.* [2]).

Proposition 2.1 (Well-posedness)**.** *Let $s > 1 + d/2$, and $M > 0$. Suppose that for all $j \in \{1, \ldots, d\}$, $A_j(\boldsymbol{U}_N)$ satisfies the Assumption **A.1** and is self-adjoint. There exists $C > 0$ and $T > 0$ (depending only on s and M) such that for every $\boldsymbol{U}^0 \in H^s((2\pi\mathbb{T})^d)^n$ such that $\big|\boldsymbol{U}^0\big|_{H^s} \leqslant M$, there exists a unique $\boldsymbol{U} \in \mathcal{C}(I; H^s((2\pi\mathbb{T})^d)^n)$ maximal-in-time classical solution to (1.1), and moreover the open time interval $I \supset [0, T/\big|\boldsymbol{U}^0\big|_{H^s}]$ and for all $0 \leqslant t \leqslant T/\big|\boldsymbol{U}^0\big|_{H^s}$,*

$$\big|\boldsymbol{U}\big|_{H^s} \leqslant \big|\boldsymbol{U}^0\big|_{H^s} \exp(C\big|\boldsymbol{U}^0\big|_{H^s} t) \leqslant 2\big|\boldsymbol{U}^0\big|_{H^s}.$$

2.1 Discretization with sharp low-pass filters

Recall that the spatial discretization of the problem amounts to finding a solution $\boldsymbol{U}_N : t \mapsto \mathcal{T}_N^n$ of the problem

$$\partial_t \boldsymbol{U}_N + \mathsf{P}_N \left(\sum_{j=1}^d A_j(\boldsymbol{U}_N) \partial_{x_j} \boldsymbol{U}_N \right) = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N \boldsymbol{U}^0, \tag{2.1}$$

where $\mathsf{P}_N = \mathrm{Diag}(P_N(D))$, with $P_N(\cdot) = \mathbf{1}_{[\![-N,N]\!]^d}(\cdot)$.

We want to show that the semi-discretized solution $\boldsymbol{U}_N$ to (2.1) converges as $N \to \infty$ towards $\boldsymbol{U}$ the corresponding solution to the underlying system (1.1). To do so, we will first show that the semi-discretized solution exists and is bounded on a time interval independent of N. This is Proposition 2.2. Then we use this bound to compare the semi-discretized and continuous solution on the interval of existence in Proposition 2.4. Finally, we refine this result by showing that if N is large enough the existence of the semi-discretized solution and the estimate on the difference hold on any compact subset of the interval of existence of the solution to (1.1).

Proposition 2.2 (Uniform estimates)**.** *Let $s > d/2 + 1$ and $M > 0$. Suppose that for all $j \in \{1, \ldots, d\}$, $A_j(\boldsymbol{U}_N)$ satisfies the Assumption **A.1** and is self-adjoint. There exists $C > 0$ and $T > 0$ depending only on s and M such that for every $N \in \mathbb{N}$ and for every $\boldsymbol{U}^0 \in H^s((2\pi\mathbb{T})^d)^n$ such that $\big|\boldsymbol{U}^0\big|_{H^s} \leqslant M$, there exists a unique $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$ maximal-in-time classical solution to (2.1) and $\boldsymbol{U}_N|_{t=0} = \mathsf{P}_N \boldsymbol{U}^0$. The open time interval $I_N \supset [0, T/\big|\boldsymbol{U}^0\big|_{H^s}]$ and for all $0 \leqslant t \leqslant T/\big|\boldsymbol{U}^0\big|_{H^s}$,*

$$\big|\boldsymbol{U}_N\big|_{H^s} \leqslant \big|\boldsymbol{U}^0\big|_{H^s} \exp(C\big|\boldsymbol{U}^0\big|_{H^s} t) \leqslant 2\big|\boldsymbol{U}^0\big|_{H^s}.$$

Moreover, for any $s' \geqslant s$, one has $\boldsymbol{U}_N \in \mathcal{C}^1(I_N; H^{s'}((2\pi\mathbb{T})^d)^n)$ and for any $0 < T^* \in I_N$ and $M^* > 0$ such that $\sup_{t \in [0,T^*]} |\boldsymbol{U}_N(t, \cdot)|_{H^s} \leqslant M^*$ there exists $C^* > 0$ depending only on s, s' and M^* such that for all $0 \leqslant t \leqslant T^*$,

$$\big|\boldsymbol{U}_N\big|_{H^{s'}} \leqslant \big|\boldsymbol{U}^0\big|_{H^{s'}} \exp(C^* M^* t).$$

The key ingredient to show Proposition 2.2 is the following apriori estimate.

Lemma 2.3. *Let $s' \geqslant s > d/2 + 1$. Let $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$ be solution to (2.1) on $I_N \subset \mathbb{R}$ open time interval. Then $\boldsymbol{U}_N \in \mathcal{C}^1(I_N; H^{s'}((2\pi\mathbb{T})^d)^n)$ and for any $t \in I_N$, one has*

$$\frac{\mathrm{d}}{\mathrm{d}t} \big|\boldsymbol{U}_N\big|_{H^{s'}} \leqslant C\left(\big|\boldsymbol{U}_N\big|_{H^s}\right) \big|\boldsymbol{U}_N\big|_{H^s} \big|\boldsymbol{U}_N\big|_{H^{s'}}. \tag{2.2}$$

Proof. Notice $\boldsymbol{U}_N = \mathsf{P}_N \boldsymbol{U}_N$ and hence we have smoothness in space, $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^\sigma((2\pi\mathbb{T})^d)^n)$ for any $\sigma \in \mathbb{R}$. We infer smoothness in time, $\boldsymbol{U}_N \in \mathcal{C}^1(I_N; H^\sigma((2\pi\mathbb{T})^d)^n)$, using eq. (2.1), product and composition estimates

in $H^s((2\pi\mathbb{T})^d)^n$ —Propositions A.3 and A.4— and that $\mathsf{P}_N : H^s((2\pi\mathbb{T})^d)^n \to H^\sigma((2\pi\mathbb{T})^d)^n$ is bounded. Denote $\dot{\boldsymbol{U}}_N := \Lambda^{s'}\boldsymbol{U}_N$. Using that $\boldsymbol{U}_N$ satisfies the system (2.1), that P_N is symmetric for the L^2 inner-product and $\mathsf{P}_N\dot{\boldsymbol{U}}_N = \dot{\boldsymbol{U}}_N$, and finally that A_j are self-adoint for all $j \in \{1, \dots, d\}$ and integration by parts, we have

$$
\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left|\dot{\boldsymbol{U}}_N\right|_{L^2}^2 &= \left(\Lambda^{s'}\partial_t\boldsymbol{U}_N, \dot{\boldsymbol{U}}_N\right)_{L^2} = -\left(\mathsf{P}_N(\sum_{j=1}^d[\Lambda^{s'}, A_j(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{U}_N), \dot{\boldsymbol{U}}_N\right)_{L^2} - \left(\mathsf{P}_N(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\dot{\boldsymbol{U}}_N), \dot{\boldsymbol{U}}_N\right)_{L^2} \\
&= -\left(\sum_{j=1}^d[\Lambda^{s'}, A_j(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{U}_N, \dot{\boldsymbol{U}}_N\right)_{L^2} - \left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N\right)_{L^2} \\
&= -\left(\sum_{j=1}^d[\Lambda^{s'}, A_j(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{U}_N, \dot{\boldsymbol{U}}_N\right)_{L^2} + \frac{1}{2}\left(\sum_{j=1}^d[\partial_{x_j}, A_j(\boldsymbol{U}_N)]\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N\right)_{L^2}.
\end{aligned}
$$

Using the commutator and composition estimates of Propositions A.5, A.4 with $s_0 = s - 1$ for the first contribution, the continuous Sobolev embedding $H^s((2\pi\mathbb{T})^d)^n \subset W^{1,\infty}((2\pi\mathbb{T})^d)^n$ (Proposition A.1) for the second as well the Cauchy-Schwarz inequality gives the bound

$$
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left|\dot{\boldsymbol{U}}_N\right|_{L^2}^2 \leqslant C(|\boldsymbol{U}_N|_{H^s})\,|\boldsymbol{U}_N|_{H^s}\left|\dot{\boldsymbol{U}}_N\right|_{L^2}^2,
$$

which yields the desired inequality. $\qquad\square$

Now we proceed to prove Proposition 2.2.

Proof of Proposition 2.2. For each $N > 0$, existence and uniqueness of a local-in-time solution $\boldsymbol{U}_N$ to (2.1) follows from its formulation as a system of ODEs in the Banach space $H^s((2\pi\mathbb{T})^d)^n$ (using eq. (2.1), product and composition estimates in $H^{s-1}((2\pi\mathbb{T})^d)^n$ —Propositions A.3 and A.4— and that $\mathsf{P}_N : H^{s-1}((2\pi\mathbb{T})^d)^n \to H^s((2\pi\mathbb{T})^d)^n$ is bounded) and the Cauchy–Lipschitz theorem. We denote I_N the maximal interval of existence, and notice as in the proof of Lemma 2.3 that $\boldsymbol{U}_N \in \mathcal{C}^1(I_N; H^{s'}((2\pi\mathbb{T})^d)^n)$ for any $s' \geqslant s$.

The second part of the proposition is an immediate consequence of Lemma 2.3 and Grönwall's inequality, using that $|\boldsymbol{U}_N|_{t=0}|_{H^{s'}} = \left|\mathsf{P}_N\boldsymbol{U}^0\right|_{H^{s'}} \leqslant \left|\boldsymbol{U}^0\right|_{H^{s'}}$. To show the first part of the proposition, we use a standard continuity argument. Let

$$
\varphi_N : t \mapsto \sup_{t' \in [0,t]}\left|\boldsymbol{U}_N(t', \cdot)\right|_{H^s} \quad \text{and} \quad J_N := \left\{t \in I_N \cap \mathbb{R}^+ : \varphi_N(t) \leqslant 2\left|\boldsymbol{U}^0\right|_{H^s}\right\}.
$$

Since $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$, $\varphi_N \in \mathcal{C}(I_N \cap \mathbb{R}^+; \mathbb{R})$ is non-decreasing and $J_N = \varphi_N^{-1}([0, 2\left|\boldsymbol{U}^0\right|_{H^s}])$ is a closed interval. Let us prove that one can set $T > 0$ independently of N such that $J_N \cap [0, T/\left|\boldsymbol{U}^0\right|_{H^s}]$ is an open subset of $[0, T/\left|\boldsymbol{U}^0\right|_{H^s}]$. Let $t \in J_N \cap [0, T/\left|\boldsymbol{U}^0\right|_{H^s}]$. By the second part of the proposition, we can set $C = 2C^*$ depending only on s and $\left|\boldsymbol{U}^0\right|_{H^s}$ such that

$$
|\boldsymbol{U}_N(t, \cdot)|_{H^s} \leqslant \left|\boldsymbol{U}^0\right|_{H^s}\exp\left(C\left|\boldsymbol{U}^0\right|_{H^s} t\right) \leqslant \left|\boldsymbol{U}^0\right|_{H^s}\exp(CT). \tag{2.3}
$$

Choosing $T = \ln(3/2)/C$, we find that $\varphi(t) \leqslant \frac{3}{2}\left|\boldsymbol{U}^0\right|_{H^s}$, and hence —by the continuity of φ_N— there exists $\delta > 0$ such that $[t - \delta, t + \delta] \in J_N$. This concludes the proof that $J_N \cap [0, T/\left|\boldsymbol{U}^0\right|_{H^s}]$ is a non-empty, connected, closed and open subset of $[0, T/\left|\boldsymbol{U}^0\right|_{H^s}]$. Hence $I_N \supset J_N \supset [0, T/\left|\boldsymbol{U}^0\right|_{H^s}]$ and estimate (2.3) concludes the proof. $\qquad\square$

Having established a bound on the semi-discretized solution $\boldsymbol{U}_N$, we proceed to estimate the difference between the solution $\boldsymbol{U}_N$ to (2.1) and $\boldsymbol{U}$, solution to (1.1).

Proposition 2.4 (Convergence)**.** *Let $s > d/2 + 1$, $M > 0$ and suppose that $\boldsymbol{U}^0 \in H^s((2\pi\mathbb{T})^d)^n$ with $\left|\boldsymbol{U}^0\right|_{H^s} \leqslant M$. Suppose that for all $j \in \{1, \dots, d\}$, $A_j(\boldsymbol{U}_N)$ satisfies the Assumption A.1 and is self-adjoint. Denote $\boldsymbol{U} \in$*

$\mathcal{C}(I; H^s((2\pi\mathbb{T})^d)^n)$ the maximal-in-time classical solution to (1.1), and $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$ the maximal-in-time classical solution to (2.1). Let $T > 0$ be the minimum value of Propositions 2.1 and 2.2. For all $0 \leqslant t \leqslant T/|\boldsymbol{U}^0|_{H^s}$, there is a $C > 0$, depending only on s and M such that for any $0 \leqslant r \leqslant s$,

$$\left|(\boldsymbol{U} - \boldsymbol{U}_N)(t, \cdot)\right|_{H^r} \leqslant C \left|\boldsymbol{U}^0\right|_{H^s} N^{r-s}.$$

Moreover, for every compact subset $I^* \subset I$, there is an $N_0 \in \mathbb{N}$ and $C^* > 0$, depending only on s, $|I^*|$ and $M^* := \sup_{t \in I^*} |\boldsymbol{U}(t, \cdot)|_{H^s}$ such that for all $N \geqslant N_0$, $I_N \supset I^*$ and for any $0 \leqslant r \leqslant s$,

$$\sup_{t \in I^*}\left|(\boldsymbol{U} - \boldsymbol{U}_N)(t, \cdot)\right|_{H^r} \leqslant C^* M^* N^{r-s}.$$

Proof. We shall assume that $\boldsymbol{U}^0 \in H^{s+1}((2\pi\mathbb{T})^d)^n$, so that $\boldsymbol{U} \in \mathcal{C}(I; H^{s+1}((2\pi\mathbb{T})^d)^n) \cap \mathcal{C}^1(I; H^s((2\pi\mathbb{T})^d)^n)$. The general case $\boldsymbol{U}^0 \in H^s((2\pi\mathbb{T})^d)^n$ follows by regularization of the initial data, continuity of the flow map and passing to the limit.

We shall first prove the result for $t \in [0, T/|\boldsymbol{U}^0|_{H^s}]$. Recall that $T > 0$ is the minimum value of Propositions 2.1 and 2.2, so that we have $|\boldsymbol{U}(t, \cdot)|_{H^s} \leqslant 2|\boldsymbol{U}^0|_{H^s}$ and $|\boldsymbol{U}_N(t, \cdot)|_{H^s} \leqslant 2|\boldsymbol{U}^0|_{H^s}$ for all $N \in \mathbb{N}$. Denote $\boldsymbol{D}_N := \boldsymbol{U} - \boldsymbol{U}_N$ and notice

$$\partial_t \boldsymbol{D}_N + \sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{D}_N + \sum_{j=1}^d (A_j(\boldsymbol{U}) - A_j(\boldsymbol{U}_N))\partial_{x_j}\boldsymbol{U} = -(\mathrm{Id} - \mathsf{S}_N)\left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N\right).$$

Now we apply the smooth low-pass filter S_N and use that $\mathsf{S}_N(\mathrm{Id} - \mathsf{P}_N) = 0$:

$$\partial_t \mathsf{S}_N \boldsymbol{D}_N + \mathsf{S}_N^{1/2}\left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\mathsf{S}_N^{1/2}\boldsymbol{D}_N\right) + \mathsf{S}_N^{1/2}\left(\sum_{j=1}^d [\mathsf{S}_N^{1/2}, A_j(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{D}_N\right)$$
$$+ \mathsf{S}_N\left(\sum_{j=1}^d (A_j(\boldsymbol{U}) - A_j(\boldsymbol{U}_N))\partial_{x_j}\boldsymbol{U}\right) = \boldsymbol{0}.$$

Testing against $\boldsymbol{D}_N$, using that S_N is symmetric for the $L^2((2\pi\mathbb{T})^d)^n$ inner-product and commutes with ∂_t and after some rearranging we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left|\mathsf{S}_N^{1/2}\boldsymbol{D}_N\right|_{L^2}^2 = -\left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\mathsf{S}_N^{1/2}\boldsymbol{D}_N, \mathsf{S}_N^{1/2}\boldsymbol{D}_N\right)_{L^2} - \left(\sum_{j=1}^d [\mathsf{S}_N^{1/2}, A_j(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{D}_N, \mathsf{S}_N^{1/2}\boldsymbol{D}_N\right)_{L^2}$$
$$- \left(\mathsf{S}_N^{1/2}\left(\sum_{j=1}^d (A_j(\boldsymbol{U}) - A_j(\boldsymbol{U}_N))\partial_{x_j}\boldsymbol{U}\right), \mathsf{S}_N^{1/2}\boldsymbol{D}_N\right)_{L^2}.$$

Using integration by parts and the properties of $A_j(\boldsymbol{U}_N)$ —Assumption A.1 together with the composition estimate of Proposition A.4 and self-adjointness— and the continuous Sobolev embedding $H^s((2\pi\mathbb{T})^d)^n \subset W^{1,\infty}((2\pi\mathbb{T})^d)^n$ (Proposition A.1), the first term on the right-hand side is bounded by $C(|\boldsymbol{U}_N|_{H^s})|\boldsymbol{U}_N|_{H^s}\left|\mathsf{S}_N^{1/2}\boldsymbol{D}_N\right|_{L^2}^2$. Using the commutator estimate of Proposition A.6 and the composition estimate of Proposition A.4, the second term on the right is bounded by $C(|\boldsymbol{U}_N|_{H^s})|\boldsymbol{U}_N|_{H^s}|\boldsymbol{D}_N|_{L^2}\left|\mathsf{S}_N^{1/2}\boldsymbol{D}_N\right|_{L^2}$. By the composition estimate of Proposition A.4, the third term is bounded by $C(|\boldsymbol{U}|_{H^s}, |\boldsymbol{U}_N|_{H^s})|\boldsymbol{U}|_{H^s}|\boldsymbol{D}_N|_{L^2}\left|\mathsf{S}_N^{1/2}\boldsymbol{D}_N\right|_{L^2}$, where we again use Assumption A.1 as well as the boundedness of $\mathsf{S}_N^{1/2}$. In all three estimates, we also used the Cauchy-Schwarz inequality. Altogether, we get

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(|\mathsf{S}_N^{1/2}\boldsymbol{D}_N|_{L^2}^2\right) \leqslant C(|\boldsymbol{U}|_{H^s}, |\boldsymbol{U}_N|_{H^s})\left(|\boldsymbol{U}_N|_{H^s} + |\boldsymbol{U}|_{H^s}\right)\left(|\mathsf{S}_N^{1/2}\boldsymbol{D}_N|_{L^2} + |\boldsymbol{D}_N|_{L^2}\right)|\mathsf{S}_N^{1/2}\boldsymbol{D}_N|_{L^2}.$$

Now we remark that since the symbol S_N satisfies $S_N(\cdot) \in [0,1]$ and $S_N(\boldsymbol{k}) = 1$ if $\max_{j=1,\dots,d} |k_j| \leqslant N/2$ we have $\left| (1 - S_N^{1/2}(\cdot)) \langle \cdot \rangle^{-s} \right|_{L^\infty} \leqslant \langle N/2 \rangle^{-s}$ and hence

$$\left| \boldsymbol{D}_N \right|_{L^2} \leqslant \left| \mathsf{S}_N^{1/2} \boldsymbol{D}_N \right|_{L^2} + \left| (\mathrm{Id} - \mathsf{S}_N^{1/2}) \boldsymbol{D}_N \right|_{L^2} \leqslant \left| \mathsf{S}_N^{1/2} \boldsymbol{D}_N \right|_{L^2} + \langle N/2 \rangle^{-s} \left| \boldsymbol{D}_N \right|_{H^s}.$$

Hence since $\left| \boldsymbol{D}_N \right|_{H^s} \leqslant \left| \boldsymbol{U} \right|_{H^s} + \left| \boldsymbol{U}_N \right|_{H^s} \leqslant 4 \left| \boldsymbol{U}^0 \right|_{H^s}$ by the triangle inequality we have

$$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \left(\left| \mathsf{S}_N^{1/2} \boldsymbol{D}_N \right|_{L^2}^2 \right) \leqslant C(\left| \boldsymbol{U}^0 \right|_{H^s}) \left| \boldsymbol{U}^0 \right|_{H^s} \left(\left| \mathsf{S}_N^{1/2} \boldsymbol{D}_N \right|_{L^2} + \left| \boldsymbol{U}^0 \right|_{H^s} N^{-s} \right) \left| \mathsf{S}_N^{1/2} \boldsymbol{D}_N \right|_{L^2}$$

and we infer by Grönwall's Lemma that

$$\left| \mathsf{S}_N^{1/2} \boldsymbol{D}_N(t,\cdot) \right|_{L^2} \leqslant C(\left| \boldsymbol{U}^0 \right|_{H^s}) \left| \boldsymbol{U}^0 \right|_{H^s} \left(\left| \mathsf{S}_N^{1/2} \boldsymbol{D}_N \right|_{t=0} \right|_{L^2} + t N^{-s} \left| \boldsymbol{U}^0 \right|_{H^s} \right) \exp \left(C(\left| \boldsymbol{U}^0 \right|_{H^s}) \left| \boldsymbol{U}^0 \right|_{H^s} t \right).$$

The desired estimate for $r = 0$ follows by using that $\mathsf{S}_N^{1/2} \boldsymbol{D}_N \big|_{t=0} = \mathsf{S}_N^{1/2} (\mathrm{Id} - \mathsf{P}_N) \boldsymbol{U}^0 = 0$, $t \in [0, T/\left| \boldsymbol{U}^0 \right|_{H^s}]$ and again

$$\left| \boldsymbol{D}_N \right|_{L^2} \leqslant \left| \mathsf{S}_N^{1/2} \boldsymbol{D}_N \right|_{L^2} + \langle N/2 \rangle^{-s} \left| \boldsymbol{U}^0 \right|_{H^s}.$$

The general case $0 \leqslant r \leqslant s$ follows by the interpolation inequality, Proposition A.2, and using once again that $\left| \boldsymbol{U} - \boldsymbol{U}_N \right|_{H^s} \leqslant 4 \left| \boldsymbol{U}^0 \right|_{H^s}$ by the triangle inequality.

Let us now prove the proposition for general I^* compact subset of I. Without loss of generality, we can assume $0 \in I^*$ and we will focus on positive times, $t \in I^* \cap \mathbb{R}^+$, negative times being obtained by time-symmetry. Let $d/2 + 1 < s < s'$ and assume $\boldsymbol{U}^0 \in H^{s'}$ (this is for notational convenience, as s' will eventually play the role of s in the statement of the proposition). Denote $M^* = 2 \sup_{t \in I^*} \left| \boldsymbol{U} \right|_{H^{s'}}$ and C^* the constant depending on s, s', M^* as in Proposition 2.2, and finally $M' = 2 \left| \boldsymbol{U}^0 \right|_{H^{s'}} \exp(C^* M^* T^*)$, where $T^* := \sup(I^*)$. We set

$$J_N = \left\{ t \in I_N \cap \mathbb{R}^+ \; : \; \sup_{t' \in [0,t]} \left| \boldsymbol{U}_N(t',\cdot) \right|_{H^s} \leqslant M^* \text{ and } \sup_{t' \in [0,t]} \left| \boldsymbol{U}_N(t',\cdot) \right|_{H^{s'}} \leqslant M' \right\},$$

and our aim is to prove that there exists $N_0 \in \mathbb{N}$ such that for any $N \geqslant N_0$, $J_N \supset I^* \cap \mathbb{R}^+$. We use the continuity argument. By continuity of $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^{s'}((2\pi\mathbb{T})^d)^n)$, we have that J_N is a non-empty closed interval. Let us now prove that $J_N \cap I^* \cap \mathbb{R}^+$ is an open subset of $I^* \cap \mathbb{R}^+$. Let $t \in J_N \cap I^* \cap \mathbb{R}^+$. We can follow the proof of the first part of the proposition replacing s with s' and the bound $\left| \boldsymbol{U} \right|_{H^s} + \left| \boldsymbol{U}_N \right|_{H^s} \leqslant 4 \left| \boldsymbol{U}^0 \right|_{H^s}$ with $\left| \boldsymbol{U} \right|_{H^{s'}} + \left| \boldsymbol{U}_N \right|_{H^{s'}} \leqslant M^* + M'$ to infer that there exists C, depending only on s', M^* and T^* such that for all $0 \leqslant r \leqslant s'$,

$$\sup_{t' \in [0,t]} \left| \boldsymbol{U}(t',\cdot) - \boldsymbol{U}_N(t',\cdot) \right|_{H^r} \leqslant C(M^* + M') N^{r - s'}.$$

Applying this estimate with $r = s$, it follows that there exists $N_0 \in \mathbb{N}$ depending only on s, s', M^* and T^* such that for any $N \geqslant N_0$,

$$\sup_{t' \in [0,t]} \left| \boldsymbol{U}_N(t',\cdot) \right|_{H^s} \leqslant \sup_{t' \in [0,t]} \left| \boldsymbol{U}(t',\cdot) \right|_{H^s} + \sup_{t' \in [0,t]} \left| \boldsymbol{U}(t',\cdot) - \boldsymbol{U}_N(t,\cdot) \right|_{H^s} \leqslant \frac{2}{3} M^*.$$

Moreover, using the second part of Proposition 2.2, we have immediately

$$\sup_{t' \in [0,t]} \left| \boldsymbol{U}_N(t',\cdot) \right|_{H^{s'}} \leqslant \frac{1}{2} M'.$$

This shows, using again the continuity of $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^{s'}((2\pi\mathbb{T})^d)^n)$, that there exists $\delta > 0$ such that $[t - \delta, t + \delta] \subset J_N$, and hence that $J_N \cap I^* \cap \mathbb{R}^+$ is a non-empty, connected, closed and open subset of $I^* \cap \mathbb{R}^+$. Hence $J_N \supset I^* \cap \mathbb{R}^+$. Moreover, the desired estimate has been proven along the argument. This concludes the proof. $\qquad \square$

2.2 Discretization with smooth low-pass filters

Consider now solutions to the systems semi-discretized with a smooth low-pass filter, given by (1.5) and (1.6), and which we recall here for the sake of clarity.

$$\partial_t \boldsymbol{U}_N + \mathsf{S}_N \left(\sum_{j=1}^d A_j(\boldsymbol{U}_N) \partial_{x_j} \boldsymbol{U}_N \right) = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N \boldsymbol{U}^0, \tag{2.4}$$

$$\partial_t \boldsymbol{U}_N + \sum_{j=1}^d (A_j^0 + \mathsf{S}_N(A_j^1(\boldsymbol{U}_N)[\circ])) \partial_{x_j} \boldsymbol{U}_N = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N \boldsymbol{U}^0, \tag{2.5}$$

where S_N is the smooth low-pass filter described in Section 1, and $A_j(\cdot) = A_j^0 + A_j^1(\cdot)$, $A_j^0 = A_j(\boldsymbol{0})$.

For symmetric systems, there is no great difference between the semi-discretization with sharp versus smooth low-pass filters. The results are the same, although the proofs must be adapted slightly. We outline the results and proofs, but refer to the previous section for technical details.

Our result regarding existence and boundedness of solutions $\boldsymbol{U}_N$ to the semi-discrete problems (uniformly with respect to N) extends to the case of smooth low-pass filters.

Proposition 2.5 (Uniform estimates). *The statement of Proposition 2.2 holds replacing (2.1) with (2.4) or (2.5).*

Proof. We consider first the system (2.5). As for the case with the sharp low-pass filter, the key ingredient is an apriori estimate

$$\frac{\mathrm{d}}{\mathrm{d}t} |\boldsymbol{U}_N|_{H^{s'}} \leqslant C(|\boldsymbol{U}_N|_{H^s}) |\boldsymbol{U}_N|_{H^s} |\boldsymbol{U}_N|_{H^{s'}} \tag{2.6}$$

for $s' \geqslant s > d/2 + 1$ and $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$ solution to (2.5) on $I_N \subset \mathbb{R}$ open time interval. To show this we first notice that, applying $(\mathrm{Id} - \mathsf{P}_N)$ to (2.5) and using that $(\mathrm{Id} - \mathsf{P}_N)\mathsf{S}_N = 0$, we have $\partial_t(\mathrm{Id} - \mathsf{P}_N)\boldsymbol{U}_N + \sum_{j=1}^d A_j^0 \partial_{x_j}(\mathrm{Id} - \mathsf{P}_N)\boldsymbol{U}_N = 0$ and $(\mathrm{Id} - \mathsf{P}_N)\boldsymbol{U}_N|_{t=0} = 0$. By uniqueness of the solution to this initial-value problem, we infer $\boldsymbol{U}_N = \mathsf{P}_N \boldsymbol{U}_N$ and hence $\boldsymbol{U}_N \in \mathcal{C}^1(I; H^\sigma((2\pi\mathbb{T})^d)^n)$ for all $\sigma \in \mathbb{R}$. Then we apply $\Lambda^{s'}$ to the system (2.5), denote $\dot{\boldsymbol{U}}_N \coloneqq \Lambda^{s'} \boldsymbol{U}_N$ and infer

$$\partial_t \dot{\boldsymbol{U}}_N + \sum_{j=1}^d (A_j^0 + \mathsf{S}_N^{1/2} A_j^1(\boldsymbol{U}_N)[\mathsf{S}_N^{1/2}\circ]) \partial_{x_j} \dot{\boldsymbol{U}}_N$$

$$= -\mathsf{S}_N \left(\sum_{j=1}^d [\Lambda^{s'}, A_j^1(\boldsymbol{U}_N)] \partial_{x_j} \boldsymbol{U}_N \right) - \mathsf{S}_N^{1/2} \left(\sum_{j=1}^d [\mathsf{S}_N^{1/2}, A_j^1(\boldsymbol{U}_N)] \partial_{x_j} \dot{\boldsymbol{U}}_N \right). \tag{2.7}$$

Using that $\mathsf{S}_N : L^2 \to L^2$ is bounded, as well as the product, composition and commutator estimates — Propositions A.3, A.4, A.5 and A.6 — the terms on the right-hand side can be estimated in $L^2((2\pi\mathbb{T})^d)^n$ as

$$|\mathrm{RHS}|_{L^2} \leqslant C(|\boldsymbol{U}_N|_{H^s}) |\boldsymbol{U}_N|_{H^s} \left| \dot{\boldsymbol{U}}_N \right|_{L^2}.$$

By assumption, $A_j(\boldsymbol{U}) = A_j^0 + A_j(\boldsymbol{U})$ is a symmetric matrix for all $\boldsymbol{U} \in \mathbb{R}^n$ and $j \in \{1, \dots, d\}$. In particular, this implies that $A_j^0 = A_j(\boldsymbol{0})$ and hence $A_j^1(\boldsymbol{U})$ are both symmetric matrices, which in turn implies that the operator $A_j^0 + S_N^{1/2} A_j^1(\boldsymbol{U}_N) S_N^{1/2}$ is symmetric for the $L^2((2\pi\mathbb{T})^d)^n$ inner-product for all $j \in \{1, \dots, d\}$. Arguing in the usual manner using integration by parts, we have

$$\left(\sum_{j=1}^d (A_j^0 + \mathsf{S}_N^{1/2} A_j^1(\boldsymbol{U}_N)[\mathsf{S}_N^{1/2}\circ]) \partial_{x_j} \dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N \right)_{L^2} = -\frac{1}{2} \sum_{j=1}^d \left([\partial_{x_j}, \mathsf{S}_N^{1/2} A_j^1(\boldsymbol{U}_N) \mathsf{S}_N^{1/2}] \dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N \right)_{L^2}$$

$$\leqslant C(|\boldsymbol{U}_N|_{H^s}) |\boldsymbol{U}_N|_{H^s} \left| \dot{\boldsymbol{U}}_N \right|_{L^2}^2,$$

where we used the continuous Sobolev embedding $H^s((2\pi\mathbb{T})^d)^n \subset W^{1,\infty}((2\pi\mathbb{T})^d)^n$ (Proposition A.1) and that $\mathsf{S}_N^{1/2} : L^2 \to L^2$ is bounded. Testing the identity (2.7) against $\dot{\boldsymbol{U}}_N$, using the Cauchy-Schwarz inequality on the right-hand side and inserting the two estimates above yields the desired inequality (2.6). With this estimate in hand, the rest of the proof follows exactly the proof of Proposition 2.2.

Considering now the system (2.4), we notice that the inequality (2.6) is obtained as above, using additionally that $\mathsf{S}_N A_j^0 = \mathsf{S}_N^{1/2} A_j^0 \mathsf{S}_N^{1/2}$ is self-adjoint, and the result follows. $\qquad\square$

The spectral convergence of solutions $\boldsymbol{U}_N$ to the semi-discrete problems towards the corresponding solution to the continuous problem (1.1) as $N \to \infty$ is the identical in the case of sharp or smooth low-pass filters.

Proposition 2.6 (Convergence). *The statement of Proposition 2.4 holds replacing (2.1) with (2.4) or (2.5).*

Proof. The proof follows the proof of Proposition 2.4, with one modification. Consider $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$ solution to (2.5) (considering instead $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$ solution to (2.4) amounts to replacing $A_j^1(\boldsymbol{U}_N)$ with $A_j(\boldsymbol{U}_N)$ in the right-hand side of the following identity, with no consequence). Then the difference $\boldsymbol{D}_N \coloneqq \boldsymbol{U} - \boldsymbol{U}_N$ satisfies

$$\partial_t \boldsymbol{D}_N + \sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{D}_N + \sum_{j=1}^d (A_j(\boldsymbol{U}) - A_j(\boldsymbol{U}_N))\partial_{x_j}\boldsymbol{U} = (\mathrm{Id} - \mathsf{P}_N)\left(\sum_{j=1}^d A_j^1(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N\right).$$

Instead of applying S_N as in the proof of Proposition 2.4, we apply $\mathsf{S}_{N/2}$, noting that $\mathsf{S}_{N/2}(\mathrm{Id} - \mathsf{S}_N) = 0$. Then we have

$$\partial_t \mathsf{S}_{N/2}\boldsymbol{D}_N + \mathsf{S}_{N/2}^{1/2}\left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N\right) + \mathsf{S}_{N/2}^{1/2}\left(\sum_{j=1}^d [\mathsf{S}_{N/2}^{1/2}, A_j(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{D}_N\right)$$
$$+ \mathsf{S}_{N/2}\left(\sum_{j=1}^d (A_j(\boldsymbol{U}) - A_j(\boldsymbol{U}_N))\partial_{x_j}\boldsymbol{U}\right) = \boldsymbol{0}.$$

Because $\mathsf{S}_{N/2}$ satisfies the same commutator estimates as S_N and $\|\mathrm{Id} - \mathsf{S}_{N/2}^{1/2}\|_{H^s \to L^2} \leqslant \langle N/4\rangle^{-s}$, we may then proceed exactly as in the proof of Proposition 2.4. $\qquad\square$

3 Symmetrizable quasilinear systems

In this section, we consider systems that are symmetrizable in the sense of Friedrichs. That is, we assume that there is an open set $\mathcal{U} \subset \mathbb{R}^n$ containing the origin and an operator $S(\boldsymbol{U})$ which is a symmetrizer for the system (1.1):

Assumption A.2. *There exists $S(\boldsymbol{U})$ a Friedrichs-symmetrizer for the system (1.1), that is, there exists an open set $\mathcal{U} \subset \mathbb{R}^n$ with $\boldsymbol{0} \in \mathcal{U}$ such that for all $\boldsymbol{U} \in \mathcal{U}$, $S(\boldsymbol{U})$ is real-valued, symmetric positive definite, and for all $j \in \{1, \dots, d\}$, $S(\boldsymbol{U})A_j(\boldsymbol{U})$ is symmetric. We assume that all entries of $S(\cdot)$ are polynomial.*

We have the following standard result [2].

Proposition 3.1 (Well-posedness). *Suppose that the system (1.1) satisfies Assumptions A.1 and A.2. Let $s > 1 + d/2$, $M > 0$ and $\mathcal{K} \subset \mathcal{U}$ compact. There exists $C > 0$ and $T > 0$ (depending only on s, M and $\mathcal{K}$) such that for every $\boldsymbol{U}^0 \in H^s((2\pi\mathbb{T})^d)^n$ such that $|\boldsymbol{U}^0|_{H^s} \leqslant M$ and taking values in $\mathcal{K}$, there exists a unique $\boldsymbol{U} \in \mathcal{C}(I; H^s((2\pi\mathbb{T})^d)^n)$ maximal-in-time classical solution to (1.1) and $\boldsymbol{U}\big|_{t=0} = \boldsymbol{U}^0$, and moreover the open time interval $I \supset [0, T/|\boldsymbol{U}^0|_{H^s}]$ and for all $t \in [0, T/|\boldsymbol{U}^0|_{H^s}]$,*

$$|\boldsymbol{U}|_{H^s} \leqslant C|\boldsymbol{U}^0|_{H^s} \exp(C|\boldsymbol{U}^0|_{H^s}t).$$

3.1 Discretization with smooth low-pass filters

We consider the spatial discretization with smooth low-pass filters first, as it is more similar to the previous sections. Recall that the spatial discretization with the smooth low-pass filter S_N amounts to solving

$$\partial_t \boldsymbol{U}_N + \mathsf{S}_N \left(\sum_{j=1}^d A_j(\boldsymbol{U}_N) \partial_{x_j} \boldsymbol{U}_N \right) = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N \boldsymbol{U}^0. \tag{3.1}$$

As discussed in the introduction, one would typically prefer in practice the variant

$$\partial_t \boldsymbol{U}_N + \sum_{j=1}^d (A_j^0 + \mathsf{S}_N(A_j^1(\boldsymbol{U}_N)[\circ])) \partial_{x_j} \boldsymbol{U}_N = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N \boldsymbol{U}^0, \tag{3.2}$$

where $A_j^0 := A_j(\boldsymbol{0})$ and $A_j^1(\boldsymbol{U}) := A_j(\boldsymbol{U}) - A_j^0$. However, we face a difficulty that while any symmetrizer for the system (1.1), $S(\boldsymbol{U})$, readily provides a suitable symmetrizer of the semi-discretized system (3.1), such is not the case for system (3.2), and additional assumptions are needed.

Assumption A.3 (Compatibility of the symmetrizer)**.** *Supposing the Assumptions A.1 and A.2 hold, and decomposing $A_j(\boldsymbol{U}) = A_j^0 + A_j^1(\boldsymbol{U})$ where $A_j^0 := A_j(\boldsymbol{0})$ and $S(\boldsymbol{U}) = S^0 + S^1(\boldsymbol{U})$ where $S^0 := S(\boldsymbol{0})$, we have*

$$\forall j \in \{1, \dots, d\}, \ \forall \boldsymbol{U} \in \mathcal{U}, \quad S^0 A_j^0, \ S^0 A_j^1(\boldsymbol{U}) + S^1(\boldsymbol{U}) A_j^0 \text{ and } S^1(\boldsymbol{U}) A_j^1(\boldsymbol{U}) \text{ are symmetric.}$$

Remark 3.2. *As can be seen by Taylor-expanding $S(\boldsymbol{U}) A_j(\boldsymbol{U})$ about the origin, Assumption A.3 holds in particular when $A_j(\boldsymbol{U})$ and $S(\boldsymbol{U})$ are linear, that is $A_j^1(\boldsymbol{U})$ and $S^1(\boldsymbol{U})$ are homogeneous polynomials of order 1 in the entries of $\boldsymbol{U}$; see also Remark 3.5.*

We have the following bound on solutions $\boldsymbol{U}_N$ to the semi-discretized problems (3.1) or (3.2).

Proposition 3.3 (Uniform estimates)**.** *Suppose that the system (1.1) satisfies Assumptions A.1 and A.2. Let $s > d/2 + 1$, $M > 0$ and $\mathcal{K} \subset \mathcal{U}$ compact. There exists $C > 0$ and $T > 0$ (depending only on s, M, and $\mathcal{K}$) such that for every $N \in \mathbb{N}$ and for every $\boldsymbol{U}^0 \in H^s((2\pi\mathbb{T})^d)^n$ such that $|\boldsymbol{U}^0|_{H^s} \leq M$ and taking values in $\mathcal{K}$, there exists a unique $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$ maximal-in-time classical solution to (3.1) and $\boldsymbol{U}|_{t=0} = \mathsf{P}_N \boldsymbol{U}^0$. The open time interval $I_N \supset [0, T/|\boldsymbol{U}^0|_{H^s}]$ and for all $0 \leq t \leq T/|\boldsymbol{U}^0|_{H^s}$,*

$$|\boldsymbol{U}_N|_{H^s} \leq C|\boldsymbol{U}^0|_{H^s} \exp(C|\boldsymbol{U}^0|_{H^s} t) \leq 2C|\boldsymbol{U}^0|_{H^s}.$$

Moreover, for any $s' \geq s$, one has $\boldsymbol{U}_N \in \mathcal{C}^1(I_N; H^{s'}((2\pi\mathbb{T})^d)^n)$ and for any $0 < T^* \in I_N$, $M^* > 0$ and $\mathcal{K}^* \subset \mathcal{U}$ compact such that $\sup_{t \in [0, T^*]} |\boldsymbol{U}_N(t, \cdot)|_{H^s} \leq M^*$ and $\boldsymbol{U}_N([0, T^*] \times \mathbb{R}^d) \subset \mathcal{K}^*$ there exists $C^* > 0$ depending only on s, s', M^* and $\mathcal{K}^*$ such that for all $0 \leq t \leq T^*$,

$$|\boldsymbol{U}_N|_{H^{s'}} \leq C^* |\boldsymbol{U}^0|_{H^{s'}} \exp(C^* M^* t).$$

The same results holds replacing (3.1) with (3.2) if additionally Assumption A.3 holds.

Proof. As in Propositions 2.2 and 2.5, the key ingredient is an apriori estimate, and we focus on the derivation of such estimates for solutions $\boldsymbol{U}_N \in \mathcal{C}^1(I_N; H^\sigma((2\pi\mathbb{T})^d)^n)$ for all $\sigma \in \mathbb{R}$ (recall $\boldsymbol{U}_N = \mathsf{P}_N \boldsymbol{U}_N$ since $\mathsf{S}_N = \mathsf{S}_N \mathsf{P}_N$). We consider system (3.2) which is the most involved, the case of system (3.1) being obtained in the same way. Indeed, while the symmetrizer of the system (1.1), $S(\boldsymbol{U})$, readily offers a suitable symmetrizer of the semi-discretized system (3.1), for (3.2) we need to consider a modification, namely

$$\tilde{S}(\boldsymbol{U})[\circ] = S^0 + \mathsf{S}_N^{1/2}(S^1(\boldsymbol{U}) \mathsf{S}_N^{1/2}[\circ]). \tag{3.3}$$

Recall $\mathsf{S}_N = \mathrm{Diag}(S_N(D))$ and the symbol S_N is nonnegative, hence $\mathsf{S}_N^{1/2} = \mathrm{Diag}(S_N(D)^{1/2})$ is well-defined. Let us first prove that for any $\mathcal{K} \subset \mathcal{U}$ compact, there exists $0 < \alpha \leqslant \beta < \infty$ such that for all $\boldsymbol{U} \subset \mathcal{K}$ and $\boldsymbol{V} \in \mathbb{R}^n$ one has

$$\alpha \left| \boldsymbol{V} \right|_{L^2}^2 \leqslant \left(\tilde{S}(\boldsymbol{U})\boldsymbol{V}, \boldsymbol{V} \right)_{L^2} \leqslant \beta \left| \boldsymbol{V} \right|_{L^2}^2. \tag{3.4}$$

Note first that the result holds for $S(\boldsymbol{U})$ by Assumption **A.2**, and hence for $S^0 = S(\mathbf{0})$ since $\mathbf{0} \in \mathcal{U}$. The result then follows from the identities

$$\begin{aligned}
\left(\tilde{S}(\boldsymbol{U})\boldsymbol{V}, \boldsymbol{V} \right)_{L^2} &= (S^0 \boldsymbol{V}, \boldsymbol{V})_{L^2} + \left(S^1(\boldsymbol{U})\mathsf{S}_N^{1/2}\boldsymbol{V}, \mathsf{S}_N^{1/2}\boldsymbol{V} \right)_{L^2} \\
&= \left(S^0 (\mathrm{Id} - \mathsf{S}_N)^{1/2}\boldsymbol{V}, (\mathrm{Id} - \mathsf{S}_N)^{1/2}\boldsymbol{V} \right)_{L^2} + \left(S(\boldsymbol{U})\mathsf{S}_N^{1/2}\boldsymbol{V}, \mathsf{S}_N^{1/2}\boldsymbol{V} \right)_{L^2}
\end{aligned}$$

and

$$\left| \boldsymbol{V} \right|_{L^2}^2 = \left| (\mathrm{Id} - \mathsf{S}_N)^{1/2}\boldsymbol{V} \right|_{L^2}^2 + \left| \mathsf{S}_N^{1/2}\boldsymbol{V} \right|_{L^2}^2.$$

Now we want to show that for any $s' \geqslant s > 1 + d/2$ and $\boldsymbol{U}_N$ solution to (3.2) taking values in $\mathcal{K} \subset \mathcal{U}$ compact one has

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\tilde{S}(\boldsymbol{U}_N)\Lambda^{s'}\boldsymbol{U}_N, \Lambda^{s'}\boldsymbol{U}_N \right)_{L^2} \leqslant C(\mathcal{K}, |\boldsymbol{U}_N|_{H^s}) |\boldsymbol{U}_N|_{H^s} |\boldsymbol{U}_N|_{H^{s'}}^2. \tag{3.5}$$

Indeed, if (3.5) holds, applying (3.4) in (3.5) and using Grönwall's Lemma yields the following estimate

$$\begin{aligned}
\alpha^{1/2}(\mathcal{K}^*) |\boldsymbol{U}_N(t,\cdot)|_{H^{s'}} \leqslant \mathcal{F}_{s'}(t) &\leqslant \mathcal{F}_{s'}(0) \exp\left(\alpha^{-1}(\mathcal{K}^*) \int_0^t C(\mathcal{K}^*, |\boldsymbol{U}_N(\tau,\cdot)|_{H^s}) |\boldsymbol{U}_N(\tau,\cdot)|_{H^s} \, \mathrm{d}\tau \right) \\
&\leqslant \beta^{1/2}(\mathcal{K}^*) |\boldsymbol{U}_N(0,\cdot)|_{H^{s'}} \exp\left(t \times \alpha^{-1}(\mathcal{K}^*)C(\mathcal{K}^*; \mathcal{M}_s(t))\mathcal{M}_s(t) \right)
\end{aligned}$$

where $\mathcal{K}^*$ is compact with $\mathcal{K} \subset \mathcal{U}^* \subset \mathcal{K}^* \subset \mathcal{U}$ and $\mathcal{U}^*$ open, and where we denote

$$\mathcal{F}_{s'}(\tau) \coloneqq \left. \left(\tilde{S}(\boldsymbol{U}_N)\Lambda^{s'}\boldsymbol{U}_N, \Lambda^{s'}\dot{\boldsymbol{U}}_N \right)_{L^2}^{1/2} \right|_{t=\tau} \quad \text{and} \quad \mathcal{M}_s(t) \coloneqq \sup_{\tau \in [0,t]} |\boldsymbol{U}_N(\tau,\cdot)|_{H^s}.$$

The estimate is valid as long as $\boldsymbol{U}_N(\tau,\cdot)$ takes values in $\mathcal{K}^*$ for all $\tau \in [0,t]$. This is ensured by

$$\left| \boldsymbol{U}_N(\tau,\cdot) - \boldsymbol{U}_N^0 \right|_{L^\infty} \leqslant \int_0^t |\partial_t \boldsymbol{U}_N(\tau,\cdot)|_{L^\infty} \, \mathrm{d}\tau \leqslant t \times C(\mathcal{K}^*; \mathcal{M}_s(t))\mathcal{M}_s(t)$$

which follows from the system (3.2) and continuous Sobolev embedding $H^s((2\pi\mathbb{T})^d)^n \subset W^{1,\infty}((2\pi\mathbb{T})^d)^n$ (Proposition A.1). By means of these estimates, we can employ the continuity argument as in the proof of Proposition 2.2 to conclude the proof.

Let us now prove the estimate (3.5). Apply $\Lambda^{s'}$ to the system (3.2) and denote $\dot{\boldsymbol{U}}_N \coloneqq \Lambda^{s'}\boldsymbol{U}_N$ to infer

$$\partial_t \dot{\boldsymbol{U}}_N + \sum_{j=1}^d (A_j^0 + \mathsf{S}_N(A_j^1(\boldsymbol{U}_N)[\circ]))\partial_{x_j}\dot{\boldsymbol{U}}_N = -\mathsf{S}_N\left(\sum_{j=1}^d [\Lambda^{s'}, A_j^1(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{U}_N \right).$$

Applying the operator $\tilde{S}(\boldsymbol{U}_N)$ to the system and using the self-adjointness of $\tilde{S}(\boldsymbol{U}_N)$ we get

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\tilde{S}(\boldsymbol{U}_N)\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N \right)_{L^2} = &\frac{1}{2}\left([\partial_t, \tilde{S}(\boldsymbol{U}_N)]\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N \right)_{L^2} - \sum_{j=1}^d \left(\tilde{S}(\boldsymbol{U}_N)(A_j^0 + \mathsf{S}_N(A_j^1(\boldsymbol{U}_N)[\circ]))\partial_{x_j}\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N \right)_{L^2} \\
&- \sum_{j=1}^d \left(\tilde{S}(\boldsymbol{U}_N)\mathsf{S}_N\left([\Lambda^{s'}, A_j^1(\boldsymbol{U}_N)]\partial_{x_j}\dot{\boldsymbol{U}}_N \right), \dot{\boldsymbol{U}}_N \right)_{L^2}.
\end{aligned}$$

By using that ∂_t commutes with S^0 and $\mathsf{S}_N^{1/2}$ we get

$$\left|\left([\partial_t, \tilde{S}(\boldsymbol{U}_N)]\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N\right)_{L^2}\right| = \left|\left([\partial_t, S^1(\boldsymbol{U}_N)]\mathsf{S}_N^{1/2}\dot{\boldsymbol{U}}_N, \mathsf{S}_N^{1/2}\dot{\boldsymbol{U}}_N\right)_{L^2}\right| \leqslant C\left(\mathcal{K}, |\boldsymbol{U}_N|_{L^\infty}\right)|\partial_t \boldsymbol{U}_N|_{L^\infty}\left|\mathsf{S}_N^{1/2}\dot{\boldsymbol{U}}_N\right|_{L^2}^2.$$

It follows from the system (3.2), continuous Sobolev embedding $H^s((2\pi\mathbb{T})^d)^n \subset W^{1,\infty}((2\pi\mathbb{T})^d)^n$ (Proposition A.1) and product and composition estimates (Propositions A.3 and A.4) that

$$\left|\left([\partial_t, \tilde{S}(\boldsymbol{U}_N)]\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N\right)_{L^2}\right| \leqslant C\left(\mathcal{K}, |\boldsymbol{U}_N|_{H^s}\right)|\boldsymbol{U}_N|_{H^s}|\boldsymbol{U}_N|_{H^{s'}}^2.$$

For the third term, we use the boundedness of the operators $\tilde{S}(\boldsymbol{U}_N)$ and S_N and the commutator estimate in Proposition A.5 and find

$$\left|\left(\tilde{S}(\boldsymbol{U}_N)\mathsf{S}_N\left([\Lambda^{s'}, A_j^1(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{U}_N\right), \dot{\boldsymbol{U}}_N\right)_{L^2}\right| \leqslant C\left(\mathcal{K}, |\boldsymbol{U}_N|_{H^s}\right)|\boldsymbol{U}_N|_{H^s}|\boldsymbol{U}_N|_{H^{s'}}^2.$$

Estimating the remaining term requires more care, since $\tilde{S}(\boldsymbol{U}_N)$ is not a perfect symmetrizer for the system. Observe

$$\begin{aligned}
\tilde{S}(\boldsymbol{U}_N)(A_j^0 + \mathsf{S}_N(A_j^1(\boldsymbol{U}_N))) =& S^0 A_j^0 + \mathsf{S}_N^{1/2}(S^1(\boldsymbol{U}_N)A_j^0 + S^0 A_j^1(\boldsymbol{U}_N))\mathsf{S}_N^{1/2} + \mathsf{S}_N S^1(\boldsymbol{U}_N)A_j^1(\boldsymbol{U}_N)\mathsf{S}_N \\
&+ \mathsf{S}_N^{1/2} S^0[\mathsf{S}_N^{1/2}, A_j^1(\boldsymbol{U}_N)] - \mathsf{S}_N^{1/2}[\mathsf{S}_N^{1/2}, S^1(\boldsymbol{U}_N)]\mathsf{S}_N A_j^1(\boldsymbol{U}_N) + \mathsf{S}_N S^1(\boldsymbol{U}_N)[\mathsf{S}_N, A_j^1(\boldsymbol{U}_N)] \\
=& \mathrm{Sym}(\boldsymbol{U}_N) + \mathrm{Com}(\boldsymbol{U}_N).
\end{aligned}$$

Notice that all terms in the first line —the sum being denoted $\mathrm{Sym}(\boldsymbol{U}_N)$— are symmetric operators by Assumption A.3, while the remaining terms —the sum being denoted $\mathrm{Com}(\boldsymbol{U}_N)$— involve commutators between $\mathsf{S}_N^{1/2}$ and either $S^1(\boldsymbol{U}_N)$ or $A_j^1(\boldsymbol{U}_N)$. Integration by parts and symmetry considerations implies that

$$\left|(\mathrm{Sym}(\boldsymbol{U}_N)\partial_{x_j}\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N)_{L^2}\right| \leqslant C(\mathcal{K}, |\boldsymbol{U}_N|_{H^s})|\boldsymbol{U}_N|_{H^s}|\boldsymbol{U}_N|_{H^{s'}}^2.$$

For the other terms we use that $\mathsf{S}_N^{1/2}$ satisfies the commutator estimate in Proposition A.6 to infer

$$\left|(\mathrm{Com}(\boldsymbol{U}_N)\partial_{x_j}\dot{\boldsymbol{U}}_N, \dot{\boldsymbol{U}}_N)_{L^2}\right| \leqslant C(\mathcal{K}, |\boldsymbol{U}_N|_{H^s})|\boldsymbol{U}_N|_{H^s}|\boldsymbol{U}_N|_{H^{s'}}^2.$$

Combining all the estimates shows (3.5), which concludes the proof. $\qquad\square$

Having established uniform bounds for solutions $\boldsymbol{U}_N$ to the semi-discretized system (3.1) and (3.2), we now show the convergence towards corresponding solutions of the underlying continuous problem (1.1) as $N \to \infty$.

Proposition 3.4 (Convergence). *Suppose that the system* (1.1) *satisfies Assumptions A.1 and A.2. Let $s > d/2 + 1$ and $\boldsymbol{U}^0 \in H^s((2\pi\mathbb{T})^d)^n$ such that $\boldsymbol{U}^0$ takes values into the hyperbolic domain $\mathcal{U}$. Denote $\boldsymbol{U} \in \mathcal{C}(I; H^s((2\pi\mathbb{T})^d)^n)$ the maximal-in-time classical solution to* (1.1), *and $\boldsymbol{U}_N \in \mathcal{C}(I_N; H^s((2\pi\mathbb{T})^d)^n)$ the maximal-in-time classical solution to* (3.1). *For every compact subset $I^* \subset I$, there is an $N_0 \in \mathbb{N}$ and $C^* > 0$, depending only on s, $|I^*|$, $\sup_{t \in I^*}|\boldsymbol{U}(t, \cdot)|_{H^s}$ and $\mathcal{K}^* \subset \mathcal{U}$ compact such that $\boldsymbol{U}(I^* \times \mathbb{R}^d) \subset \mathcal{K}^*$ such that for all $N \geqslant N_0$, one has $I_N \supset I^*$ and for any $0 \leqslant r \leqslant s$,*

$$\sup_{t \in I^*}\left|(\boldsymbol{U} - \boldsymbol{U}_N)(t, \cdot)\right|_{H^r} \leqslant C^* M^* N^{r-s}.$$

The same results holds replacing (3.1) *with* (3.2) *if additionally Assumption A.3 holds.*

Proof. The proof is similar to that of Propositions 2.4 and 2.6, and we only sketch the main arguments. We consider the system (3.1); system (3.2) can be treated in a similar way after introducing the symmetrizer $\tilde{S}(\boldsymbol{U})$ defined in (3.3)

as in the proof of Proposition 3.3. Denote $\boldsymbol{D}_N := \boldsymbol{U} - \boldsymbol{U}_N$ and use that $(\mathrm{Id} - \mathsf{S}_N)\mathsf{S}_{N/2}^{1/2} = 0$ to infer

$$
\partial_t \mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N + \sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N + \sum_{j=1}^d [\mathsf{S}_{N/2}^{1/2}, A_j(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{D}_N
$$
$$
+ \mathsf{S}_{N/2}^{1/2}\left(\sum_{j=1}^d (A_j(\boldsymbol{U}) - A_j(\boldsymbol{U}_N))\partial_{x_j}\boldsymbol{U}\right) = \boldsymbol{0}.
$$

Applying the symmetrizer $S(\boldsymbol{U}_N)$ and testing against $\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N$ yields,

$$
\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(S(\boldsymbol{U}_N)\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N, \mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N\right)_{L^2} = {}& \frac{1}{2}\left([\partial_t, S(\boldsymbol{U}_N)]\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N, \mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N\right)_{L^2} \\
&- \sum_{j=1}^d \left(S(\boldsymbol{U}_N)A_j(\boldsymbol{U}_N)\partial_{x_j}\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N, \mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N\right)_{L^2} \\
&- \sum_{j=1}^d \left(S(\boldsymbol{U}_N)[\mathsf{S}_{N/2}^{1/2}, A_j(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{D}_N, \mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N\right)_{L^2} \\
&- \sum_{j=1}^d \left(S(\boldsymbol{U}_N)\mathsf{S}_{N/2}^{1/2}\left((A_j(\boldsymbol{U}) - A_j(\boldsymbol{U}_N))\partial_{x_j}\boldsymbol{U}\right), \mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N\right)_{L^2}.
\end{aligned}
$$

The first two terms can be estimated in the standard manner using that $S(\boldsymbol{U})$ is a symmetrizer, *i.e.* Assumption A.2. The third term is estimated using the regularizing properties of the commutator with $\mathsf{S}_{N/2}^{1/2}$, Proposition A.6. The fourth term is estimated thanks to Assumption A.1. Altogether we obtain

$$
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(S(\boldsymbol{U}_N)\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N, \mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N\right)_{L^2} \leqslant C(|\boldsymbol{U}|_{H^s}, |\boldsymbol{U}_N|_{H^s})\left(|\boldsymbol{U}|_{H^s} + |\boldsymbol{U}_N|_{H^s}\right)\left(|\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N|_{L^2} + |\boldsymbol{D}_N|_{L^2}\right)|\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N|_{L^2}.
$$

With this estimate in hand, we can follow the proof of Proposition 2.4. Using that $\|\mathrm{Id} - \mathsf{S}_{N/2}^{1/2}\|_{H^s \to L^2} \leqslant \langle N/4 \rangle^{-s}$, the coercivity of $S(\boldsymbol{U}_N)$ —see (3.4)— and the uniform estimates for $\boldsymbol{U}_N$ stated in Proposition 3.3 we infer (by Grönwall's lemma), denoting $M^\star := 2\sup_{t \in I^\star}|\boldsymbol{U}|_{H^{s'}}$ and assuming $\sup_{t \in I^\star}|\boldsymbol{U}_N(t, \cdot)|_{H^s} \leqslant M^\star$ and $\boldsymbol{U}_N(I^\star \times \mathbb{R}^n) \subset \mathcal{K}^\star \subset \mathcal{U}$,

$$
|\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N(t, \cdot)|_{L^2} \leqslant C(\mathcal{K}^\star, M^\star)M^\star\left(|\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N|_{t=0}|_{L^2} + tN^{-s}|\boldsymbol{U}^0|_{H^s}\right)\exp\left(C(\mathcal{K}^\star, M^\star)M^\star t\right),
$$

and we have moreover $\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N|_{t=0} = 0$ and

$$
|\boldsymbol{D}_N|_{L^2} \leqslant |\mathsf{S}_{N/2}^{1/2}\boldsymbol{D}_N|_{L^2} + \langle N/4 \rangle^{-s}|\boldsymbol{U}^0|_{H^s}.
$$

This yields the desired estimate for $r = 0$, and the general case follows by interpolation. A continuity argument concludes the proof as in Proposition 2.4. □

Remark 3.5. *Let us comment on the restrictive Assumption A.3 arising in the study of system* (3.2). *This assumption allows to ensure that we can construct a symmetrizer operator $\tilde{S}(\boldsymbol{U})$ that is*

1. *bounded, coercive and self-adjoint for the $L^2(\mathbb{R}^d)^n$ inner-product, and*

2. *such that $\tilde{S}(\boldsymbol{U})(A_j^0 + \mathsf{S}_N A_j^1(\boldsymbol{U}))$ is self-adjoint up to regularizing operators of order -1.*

Let us notice that if Assumption A.3 does not hold, it is possible to modify the semi-discretized scheme (3.2) in a way that allows for a symmetrizer operator satisfying at least partially the above requirements. Specifically, consider the system

$$\partial_t \boldsymbol{U}_N + \sum_{j=1}^d \sum_{\ell=0}^{m_j} \mathsf{S}_N^\ell(A_j^\ell(\boldsymbol{U}_N)[\circ])\partial_{x_j}\boldsymbol{U}_N = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N \boldsymbol{U}^0, \tag{3.6}$$

and associated symmetrizer

$$\tilde{S}(\boldsymbol{U})[\circ] \coloneqq \sum_{\ell=0}^m \mathsf{S}_N^{\ell/2}(S^\ell(\boldsymbol{U})\mathsf{S}_N^{\ell/2}[\circ]).$$

where we used the convention $\mathsf{S}_N^0 = \mathrm{Id}$ and $\mathsf{S}_N^\ell = \underbrace{\mathsf{S}_N \circ \cdots \circ \mathsf{S}_N}_{\ell \ times}$, and decompositions

$$S(\boldsymbol{U}) = \sum_{\ell=0}^m S^\ell(\boldsymbol{U}), \qquad A_j(\boldsymbol{U}) = \sum_{\ell=0}^{m_j} A_j^\ell(\boldsymbol{U}) \qquad (j \in \{1, \dots, d\})$$

where entries of $S^\ell(\cdot)$ and $A_j^\ell(\cdot)$ are homogeneous polynomials of order ℓ.

By Taylor expansion about $\boldsymbol{U} = \boldsymbol{0}$ of the symmetric matrices $S(\boldsymbol{U})$ and $S(\boldsymbol{U})A_j(\boldsymbol{U})$ we find that $S^\ell(\boldsymbol{U})$ and $\sum_{\ell_1+\ell_2=\ell} S^{\ell_1}(\boldsymbol{U})A_j^{\ell_2}(\boldsymbol{U})$ are symmetric. This shows that $\tilde{S}(\boldsymbol{U})$ is self-adjoint for the $L^2(\mathbb{R}^d)^n$ inner-product and, denoting $\tilde{A}_j(\boldsymbol{U})[\circ] \coloneqq \sum_{\ell=0}^{m_j} \mathsf{S}_N^\ell(A_j^\ell(\boldsymbol{U}_N)[\circ])$, $\tilde{S}(\boldsymbol{U})\tilde{A}_j(\boldsymbol{U})$ is self-adjoint up to regularizing operators of order -1. Notice however that some additional restrictions on smallness of $\boldsymbol{U} \in \mathcal{U}$ may be necessary to enforce the coercivity of the operator $\tilde{S}(\boldsymbol{U})$.

3.2 Discretization with sharp low-pass filters

In this section we consider the case of spatial discretization through the sharp low-pass filter $\mathsf{P}_N = \mathrm{Diag}(P_N(D))$ where $P_N(\cdot) = \mathbf{1}_{[\![-N,N]\!]^d}(\cdot)$:

$$\partial_t \boldsymbol{U}_N + \mathsf{P}_N\left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N\right) = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N\boldsymbol{U}^0. \tag{3.7}$$

The analysis of the previous section fails, due to the lack of good commutator properties of the operator P_N. Specifically, when applying a symmetrizer $S(\boldsymbol{U})$ to the underlying system (1.1) we may write

$$S(\boldsymbol{U}_N)\partial_t\boldsymbol{U}_N + \sum_{j=1}^d S(\boldsymbol{U}_N)A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N = -\sum_{j=1}^d S(\boldsymbol{U}_N)(\mathrm{Id} - \mathsf{P}_N)\big(A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N\big).$$

In order to control the energy functional $\mathcal{F}_s(\boldsymbol{U}_N) \coloneqq (S(\boldsymbol{U}_N)\Lambda^s\boldsymbol{U}_N, \Lambda^s\boldsymbol{U}_N)^{1/2} \approx |\boldsymbol{U}_N|_{H^s}$ as in Proposition 3.3, we wish to control (uniformly with respect to N) the quantity $J_N(\boldsymbol{U}_N, \Lambda^s\boldsymbol{U}_N)$ where

$$J_N(\boldsymbol{U}, \boldsymbol{V}) \coloneqq \Big(\mathsf{P}_N\Big(S(\boldsymbol{U})(\mathrm{Id} - \mathsf{P}_N)\big(A_j(\boldsymbol{U})(\partial_{x_j}\mathsf{P}_N\boldsymbol{V})\big)\Big), \boldsymbol{V}\Big)_{L^2}.$$

As we shall see in an example in Section 4, it turns out we cannot improve in general the bound $J_N = \mathcal{O}(N|\boldsymbol{V}|_{L^2}^2)$. Notice that in the symmetric cases discussed in Section 2.1, namely when $S = \mathrm{Id}$, we have $J_N = 0$.

In this section we consider symmetrizable systems satisfying the following assumption.

Assumption A.5. *There exists $S(\cdot)$ and an open set $\mathcal{U} \subset \mathbb{R}^n$ with $\boldsymbol{0} \in \mathcal{U}$ such that for all $\boldsymbol{U} \in \mathcal{U}$, $S(\boldsymbol{U})$ is real-valued, symmetric positive definite, and for all $j \in \{1, \dots, d\}$ there exists S_j^0 real-valued symmetric matrix with constant coefficients such that for all $\boldsymbol{U} \in \mathcal{U}$,*

$$A_j(\boldsymbol{U}) = S_j^0 S(\boldsymbol{U}).$$

We assume that all entries of $S(\cdot)$ are polynomial.

Remark 3.6. *Assumption A.5 is a special case of symmetrizable systems, as it implies Assumptions A.1 and A.2.*

Remark 3.7 (Hamiltonian systems). *Assumption A.5 is motivated by the Hamiltonian structure of the underlying system. Indeed, denote*

$$\mathsf{J} := \sum_{j=1}^{d} S_j^0 \partial_{x_j}$$

the constant-coefficient skew-symmetric (for the $L^2(\mathbb{R}^d)^n$ inner-product) operator and $\mathscr{H} : \boldsymbol{U} \in \mathcal{U} \to \mathbb{R}$ coercive functional such that for all $\boldsymbol{U} \in \mathcal{U}$,

$$\mathrm{Hess}(\mathscr{H}(\boldsymbol{U})) = S(\boldsymbol{U}).$$

Then we remark that (1.1) *takes the Hamiltonian form*

$$\partial_t \boldsymbol{U} + \mathsf{J}\big(\nabla_{\boldsymbol{U}}\mathscr{H}(\boldsymbol{U})\big) = \boldsymbol{0},$$

where $\nabla_{\boldsymbol{U}}\mathscr{H} : \mathcal{U} \to \mathbb{R}^n$ is the Jacobian of $\mathscr{H}$.

Moreover, noticing that (3.7) *also enjoys a Hamiltonian structure,*

$$\partial_t \boldsymbol{U}_N + \mathsf{J}\big(\nabla_{\boldsymbol{U}}\mathscr{H}_N(\boldsymbol{U}_N)\big) = \boldsymbol{0}$$

where $\mathscr{H}_N(\boldsymbol{U}) = \mathscr{H}(\mathsf{P}_N\boldsymbol{U})$, we find that $\mathrm{Hess}(\mathscr{H}_N(\boldsymbol{U}_N)) = \mathsf{P}_N S(\boldsymbol{U}_N)\mathsf{P}_N$ is a symmetrizer of the system (3.7).

Under this assumption, we have the following bound on solutions $\boldsymbol{U}_N$ to the semi-discretized problems (3.7).

Proposition 3.8 (Uniform estimates). *Under the Assumption A.5, the statement of Proposition 3.3 holds replacing* (3.1) *with* (3.7).

Proof. We follow very closely the proof of Proposition 3.3, and only sketch how the necessary estimates can be obtained. Apply $\mathsf{P}_N S(\boldsymbol{U}_N)\mathsf{P}_N \Lambda^{s'}$ to the system (3.7) and use the identity $A_j(\boldsymbol{U}_N) = S_j^0 S(\boldsymbol{U}_N)$ to infer

$$\mathsf{P}_N\big(S(\boldsymbol{U}_N)\mathsf{P}_N(\partial_t \Lambda^{s'}\boldsymbol{U}_N)\big) + \sum_{j=1}^{d} \mathsf{P}_N\Big(S(\boldsymbol{U}_N)\big(\mathsf{P}_N S_j^0 \mathsf{P}_N(S(\boldsymbol{U}_N)\partial_{x_j}\Lambda^{s'}\boldsymbol{U}_N)\big)\Big)$$

$$= -\mathsf{P}_N\left(\sum_{j=1}^{d} S(\boldsymbol{U}_N)\mathsf{P}_N\big([\Lambda^{s'}, S_j^0 S(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{U}_N\big)\right)$$

where we used that $\mathsf{P}_N^2 = \mathsf{P}_N$ commutes with S_j^0 and $\Lambda^{s'}$. We can now test the identity against $\Lambda^{s'}\boldsymbol{U}_N$ and use the self-adjointness of P_N, S_j^0 and $S(\boldsymbol{U}_N)$ as well as the identity $\boldsymbol{U}_N = \mathsf{P}_N\boldsymbol{U}_N$ to infer

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big(S(\boldsymbol{U}_N)\Lambda^{s'}\boldsymbol{U}_N, \Lambda^{s'}\boldsymbol{U}_N\Big)_{L^2} = \frac{1}{2}\Big([\partial_t, S(\boldsymbol{U}_N)]\Lambda^{s'}\boldsymbol{U}_N, \Lambda^{s'}\boldsymbol{U}_N\Big)_{L^2}$$

$$- \sum_{j=1}^{d}\Big(S_j^0 \mathsf{P}_N\big([\partial_{x_j}, S(\boldsymbol{U}_N)]\Lambda^{s'}\boldsymbol{U}_N\big), \mathsf{P}_N\big(S(\boldsymbol{U}_N)\Lambda^{s'}\boldsymbol{U}_N\big)\Big)_{L^2}$$

$$- \sum_{j=1}^{d}\Big(S(\boldsymbol{U}_N)\mathsf{P}_N\big([\Lambda^{s'}, S_j^0 S(\boldsymbol{U}_N)]\partial_{x_j}\boldsymbol{U}_N\big), \Lambda^{s'}\boldsymbol{U}_N\Big)_{L^2}.$$

We then proceed as in Proposition 3.3 and obtain the energy estimate valid as long as $\boldsymbol{U}_N$ takes values into $\mathcal{K} \subset \mathcal{U}$ compact:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big(S(\boldsymbol{U}_N)\Lambda^{s'}\boldsymbol{U}_N, \Lambda^{s'}\boldsymbol{U}_N\Big)_{L^2} \leqslant C\,|\boldsymbol{U}_N|_{H^s}\,|\boldsymbol{U}_N|_{H^{s'}}^2,$$

where the constant C depends only on $s, s', \mathcal{K}$ and non-decreasingly on $|\boldsymbol{U}_N|_{H^s}$. We also have immediately the coercivity of $S(\boldsymbol{U})$: for any $\boldsymbol{U} \subset \mathcal{K}$ and $\boldsymbol{V} \in \mathbb{R}^n$ one has

$$\alpha\,|\boldsymbol{V}|_{L^2}^2 \leqslant (S(\boldsymbol{U})\boldsymbol{V}, \boldsymbol{V}) \leqslant \beta\,|\boldsymbol{V}|_{L^2}^2,$$

where $0 < \alpha \leqslant \beta < \infty$ depend uniquely on $\mathcal{K}$. These two ingredients yield the desired result. $\qquad\square$

Having established uniform bounds for solutions $\boldsymbol{U}_N$ to the semi-discretized system (3.7), we infer the convergence towards corresponding solutions of the underlying continuous problem (1.1) as $N \to \infty$.

Proposition 3.9 (Convergence). *Under Assumption A.5, the statement of Proposition 3.4 holds replacing (3.1) with (3.7).*

Proof. The proof is identical to that of Proposition 3.4. □

4 Numerical experiments for the Saint-Venant system

We shall illustrate our findings and investigate numerically the standard Saint-Venant (or shallow water) system

$$\begin{cases} \partial_t \eta + \nabla \cdot \big((1+\eta)\boldsymbol{u}\big) = 0, \\ \partial_t \boldsymbol{u} + \nabla \eta + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} = \boldsymbol{0}, \end{cases} \tag{4.1}$$

which describes the propagation of shallow water waves in the flat-bottom situation; see [17]. Specifically, the scalar variable η describes the elevation of the surface of a layer of homogeneous, incompressible and inviscid fluid and the variable $\boldsymbol{u}$ represents the layer-averaged horizontal velocity of fluid particles (both depending on time and horizontal space). The gravitational constant has been set to $g = 1$.

It will be interesting to consider the following variant (when $d = 2$)

$$\begin{cases} \partial_t \eta + \nabla \cdot \big((1+\eta)\boldsymbol{u}\big) = 0, \\ \partial_t \boldsymbol{u} + \nabla \eta + \frac{1}{2}\nabla(|\boldsymbol{u}|^2) = \boldsymbol{0}. \end{cases} \tag{4.2}$$

While the two systems are identical when $d = 1$, only the second has a Hamiltonian structure when $d = 2$; see below. As exhibited in the following section, the hyperbolicity domain of the Hamiltonian system (4.2) is a strict subset of the hyperbolicity domain of the standard system (4.2).

Thanks to these features, numerical experimenyts on the Saint-Venant systems allow to showcase our numerical findings, that we summarize here for the sake of readability.

- Numerical experiments validate our results concerning the spectral convergence of the (semi-)discretized solutions as $N \to \infty$, when Assumptions A.1, A.2 and A.3 hold and smooth low-pass filters are used.

- In the case of sharp low-pass filters, we have *not* been able to observe numerical instabilities when Assumptions A.1, A.2 and A.3 hold but Assumption A.5 fails.

- Contrarily to sharp low-pass filters, smooth low-pass filters are able to instate a form of stability even outside the domain of hyperbolicity, that is when Assumption A.2 fails.

4.1 Analysis of the Saint-Venant system

We can apply the analysis of the previous sections to systems (4.1) and (4.2) due to the following result.

Lemma 4.1. *System (4.1) is a symmetrizable hyperbolic system in the sense of Assumptions A.1 and A.2 with hyperbolic domain $\mathcal{U} \coloneqq \{(\eta, \boldsymbol{u}) \in \mathbb{R}^{1+d} : 1 + \eta > 0\}$ and*

$$S((\eta, \boldsymbol{u})) = \begin{pmatrix} 1 & \boldsymbol{0}^\top \\ \boldsymbol{0} & (1+\eta)\,\mathrm{Id} \end{pmatrix} \tag{4.3}$$

where Id is the identity matrix in $\mathbb{R}^d$. Moreover, the additional Assumption A.3 holds (see Remark 3.2).

System (4.2) satisfies Assumption A.5 (and hence Assumptions A.1 and A.2; see Remark 3.6) with hyperbolic domain $\mathcal{U}_{\mathscr{H}} \coloneqq \{(\eta, \boldsymbol{u}) \in \mathbb{R}^{1+d} : 1 + \eta - |\boldsymbol{u}|^2 > 0\}$ and

$$S_{\mathscr{H}}((\eta, \boldsymbol{u})) = \begin{pmatrix} 1 & \boldsymbol{u}^\top \\ \boldsymbol{u} & (1+\eta)\,\mathrm{Id} \end{pmatrix} \text{ and } S_j^0 = \begin{pmatrix} 0 & \boldsymbol{e}_j^\top \\ \boldsymbol{e}_j & \boldsymbol{0} \end{pmatrix} \text{ where } \boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{4.4}$$

(when $d = 2$, set $\boldsymbol{e}_j = 1$ for the analogous definitions for $d = 1$) and Hamiltonian energy

$$\mathscr{H}((\eta, \boldsymbol{u})) = \frac{1}{2} \int_{(2\pi\mathbb{T})^d} \eta^2 + (1 + \eta)|\boldsymbol{u}|^2 \, d\boldsymbol{x}.$$

Remark 4.2. *As aforementioned, systems* (4.1) *and* (4.2) *are identical when* $d = 1$, *and hence enjoy both properties. Notice that, when* $d = 2$, *the domain of hyperbolicity of system* (4.2), $\mathcal{U}_{\mathscr{H}}$, *is strictly embedded in the domain of hyperbolicity of system* (4.1), $\mathcal{U}$, *while only the former satisfies Assumption* **A.5**, *associated with its Hamiltonian formulation.*

Proof. The systems (4.1) and (4.2) can be reformulated as

$$\partial_t \boldsymbol{U} + \sum_{j=1}^d A_j(\boldsymbol{U})\partial_{x_j}\boldsymbol{U} = \boldsymbol{0}, \tag{4.5}$$

with $\boldsymbol{U} = (\eta, \boldsymbol{u})$ and

$$A_j((\eta, \boldsymbol{u})) = \begin{pmatrix} u_j & (1 + \eta)\boldsymbol{e}_j^\top \\ \boldsymbol{e}_j & u_j \, \mathrm{Id} \end{pmatrix}$$

for system (4.1) and

$$A_j((\eta, \boldsymbol{u})) = \begin{pmatrix} u_j & (1 + \eta)\boldsymbol{e}_j^\top \\ \boldsymbol{e}_j & \boldsymbol{e}_j\boldsymbol{u}^\top \end{pmatrix}$$

for system (4.2). It is then straightforward to check the assumptions. $\qquad\square$

Recall the spatial discretization of the system (4.5) with the smooth low-pass filter S_N,

$$\partial_t \boldsymbol{U}_N + \mathsf{S}_N\left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N\right) = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N\boldsymbol{U}^0, \tag{4.6}$$

or

$$\partial_t \boldsymbol{U}_N + \sum_{j=1}^d (A_j^0 + \mathsf{S}_N(A_j^1(\boldsymbol{U}_N)[\circ]))\partial_{x_j}\boldsymbol{U}_N = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N\boldsymbol{U}^0, \tag{4.7}$$

and the spatial discretization with the sharp low-pass filter P_N,

$$\partial_t \boldsymbol{U}_N + \mathsf{P}_N\left(\sum_{j=1}^d A_j(\boldsymbol{U}_N)\partial_{x_j}\boldsymbol{U}_N\right) = \boldsymbol{0}, \quad \boldsymbol{U}_N|_{t=0} = \mathsf{P}_N\boldsymbol{U}^0. \tag{4.8}$$

It follows immediately from Lemma 4.1 that we have convergence of the numerical scheme in all three cases.

Proposition 4.3. *(Convergence) The statement of Proposition 3.4, concerning spectral convergence of solutions to the semi-discrete systems* (4.6) *and* (4.7), *holds for the Saint-Venant system* (4.1) *whenever* $\boldsymbol{U}^0$ *takes values in* $\mathcal{U} := \{(\eta, \boldsymbol{u}) \in \mathbb{R}^{1+d} \, : \, 1 + \eta > 0\}$. *If additionally* $\boldsymbol{U}^0$ *takes values in* $\mathcal{U}_{\mathscr{H}} := \{(\eta, \boldsymbol{u}) \in \mathbb{R}^{1+d} \, : \, 1 + \eta - |\boldsymbol{u}|^2 > 0\}$, *then Proposition 3.4 also holds for the Hamiltonian Saint-Venant system* (4.2).

Whenever $\boldsymbol{U}^0$ *takes values in* $\mathcal{U}_{\mathscr{H}}$, *we furthermore have that the statement of Proposition 3.9, concerning spectral convergence of solutions to the semi-discrete system* (4.8), *holds for the system* (4.2).

Proof. Lemma 4.1 ensures that the systems (4.1) and (4.2) satisfy the assumptions of Proposition 3.4 with symmetrizers $S(\boldsymbol{U})$ and $S_{\mathscr{H}}(\boldsymbol{U})$ respectively, and that (4.2) additionally satisfies the assumptions of Proposition 3.9. The domains $\mathcal{U}$ and $\mathcal{U}_{\mathscr{H}}$ correspond respectively to the domains for which the symmetrizers $S(\boldsymbol{U})$ and $S_{\mathscr{H}}(\boldsymbol{U})$ in (4.3), (4.4) are positive definite, as required in Assumptions **A.2** and **A.5**. $\qquad\square$

Remark 4.4. *As discussed in the previous section, Section 3.2, we require more stringent structural assumptions to show convergence for symmetrizable systems when discretizing with the sharp low-pass filter* P_N. *The Saint-Venant system when* $d = 1$ *illustrates that we may also have to require a stricter condition on the inital data* ($\boldsymbol{U}^0$ *taking values in* $\mathcal{U}_{\mathscr{H}}$), *even for systems that satify the structural assumptions.*

This is because any *symmetrizer for the underlying system in the sense of Assumption A.2 satisfying the additional compatibility Assumption A.3 can be used to construct a symmetrizer to the semi-discrete systems* (4.6) *and* (4.7). *For the the semi-discretization with* P_N (4.8) *on the other hand, we use the symmetrizer directly related to the structure of the system through Assumption A.5 to construct a symmetrizer for* (4.8).

Let us illustrate the discussion in Remark 4.4. As discussed in the beginning of Section 3.2, considering a semi-discrete system with sharp low-pass filter emanating from a symmetrizable continuous system with symmetrizer $S(\boldsymbol{U})$, one wishes to control the energy functional $\mathcal{F}_s(\boldsymbol{U}_N) = (S(\boldsymbol{U}_N)\Lambda^s\boldsymbol{U}_N, \Lambda^s\boldsymbol{U}_N)^{1/2} \approx \left|\boldsymbol{U}_N\right|_{H^s}$, which in turn requires to control (uniformly with respect to N) the quantity $J_N(\boldsymbol{U}_N, \Lambda^s\boldsymbol{U}_N)$ where

$$J_N(\boldsymbol{U}, \boldsymbol{V}) := \left(\mathsf{P}_N\Big(S(\boldsymbol{U})(\mathrm{Id} - \mathsf{P}_N)\big(A_j(\boldsymbol{U})(\partial_{x_j}\mathsf{P}_N\boldsymbol{V})\big)\Big), \boldsymbol{V}\right)_{L^2}.$$

In the specific case of the Saint-Venant system (4.1) when $d = 1$, one has

$$\mathcal{F}_s(\boldsymbol{U}) = \int_{(2\pi\mathbb{T})} (\Lambda^s\eta)^2 + (1 + \eta)(\Lambda^s u)^2 \, \mathrm{d}x$$

and

$$S(\boldsymbol{U}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 + \eta \end{pmatrix}, \quad A(\boldsymbol{U}) = \begin{pmatrix} u & 1 + \eta \\ 1 & u \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \eta \\ u \end{pmatrix}.$$

Let $\boldsymbol{U} := (\eta_p, u_p)$ where $\eta_p := -\frac{1}{2}\cos(px)$ and $u_p := \sin(px)$, and $\boldsymbol{V}_N := (0, \sin((N - q)x))$ with $0 \leqslant q < p \ll N$. A direct calculation yields

$$(\mathrm{Id} - \mathsf{P}_N)\big(A(\boldsymbol{U})(\partial_x\mathsf{P}_N\boldsymbol{V}_N)\big) = \frac{N - q}{4} \begin{pmatrix} -\cos((N - q + p)x) \\ 2\sin((N - q + p)x) \end{pmatrix} + \mathcal{O}(1),$$

so that

$$\mathsf{P}_N\Big(S(\boldsymbol{U})(\mathrm{Id} - \mathsf{P}_N)\big(A(\boldsymbol{U})(\partial_{x_j}\mathsf{P}_N\boldsymbol{V}_N)\big)\Big) = \frac{N - q}{8} \begin{pmatrix} 0 \\ -\sin((N - q)x) \end{pmatrix} + \mathcal{O}(1),$$

and hence

$$J_N(\boldsymbol{U}, \boldsymbol{V}_N) = -\frac{\pi}{8}(N - q) + \mathcal{O}(1).$$

This shows that one cannot propagate for positive time (at least in a direct manner) a uniform-in-N control of the energy functional $\mathcal{F}_s(\boldsymbol{U}_N)$ for $\boldsymbol{U}_N$ the solution emerging from initial data $\boldsymbol{U}_N^0 := \boldsymbol{U} + \boldsymbol{V}/(N - q)^s$, despite the fact that $\boldsymbol{U}_N^0 \in \mathcal{U}$ since $1 + \eta_N \geqslant 1/2 > 0$, and $\left|\boldsymbol{U}_N^0\right|_{H^s} \approx 1$. Notice also that $\boldsymbol{U}_N^0 \notin \mathcal{U}_{\mathscr{H}}$ since $1 + \eta(\frac{\pi}{2p}) - |u(\frac{\pi}{2p})|^2 = 0$, but one could enforce $\boldsymbol{U}_N^0 \in \mathcal{U}_{\mathscr{H}}$ (while keeping valid all previous statements) by considering for instance $u_p = \frac{1}{2}\sin(px)$. In that case, the Hamiltonian structure allows to propagate the functional $\mathcal{F}_{\mathscr{H},s}(\boldsymbol{U}) \approx \left|\boldsymbol{U}_N\right|_{H^s}$ with

$$\mathcal{F}_{\mathscr{H},s}(\boldsymbol{U}) := (S_{\mathscr{H}}(\boldsymbol{U})\Lambda^s\boldsymbol{U}, \Lambda^s\boldsymbol{U})^{1/2} = \int_{(2\pi\mathbb{T})} (\Lambda^s\eta)^2 + (1 + \eta)(\Lambda^s u)^2 + 2u(\Lambda^s\eta)(\Lambda^s\eta) \, \mathrm{d}x.$$

4.2 Numerical experiments in dimension one

We seek numerical approximations to (4.1) with $d = 1$ (or, equivalently, (4.2)), η_M, u_M, in terms of finite Fourier sums of the form

$$f(x) = \sum_{k=-M+1}^{M} a_k \exp(\mathrm{i}kx),$$

20

and similarly for u_M. The vectors $\boldsymbol{\eta} = (\eta_M(x_1), \ldots, \eta_M(x_{2M}))$, $\boldsymbol{u} = (u_M(x_1), \ldots, u_M(x_{2M}))$ contain the values of η_M, u_M at regularly spaced collocation points $x_n = -\pi + \pi n/M, n = 1, \ldots, 2M$. We use the discrete Fourier Transform, computed efficiently with a Fast Fourier transform (FFT), to find

$$\eta_M(x) = \sum_{k=-M+1}^{M} \hat{\boldsymbol{\eta}}_k \exp(\mathrm{i}kx), \quad u_M(x) = \sum_{k=-M+1}^{M} \hat{\boldsymbol{u}}_k \exp(\mathrm{i}kx),$$

where $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}_{-M+1}, \ldots, \hat{\boldsymbol{\eta}}_M), \hat{\boldsymbol{u}} = (\hat{\boldsymbol{u}}_{-M+1}, \ldots, \hat{\boldsymbol{u}}_M)$ are the coefficients of the Fast Fourier transform of $\boldsymbol{\eta}, \boldsymbol{u}$, and $\hat{k} = k/\pi, k = -M + 1, \ldots, M$ are the discrete Fourier modes. Abusing notation, we will incorrectly refer to $\hat{\boldsymbol{\eta}}_k, \hat{\boldsymbol{u}}_k$ as Fourier coefficients (they are related to the coefficients c_j of infinite Fourier series $f(x) = \sum_{j \in \mathbb{Z}} c_j \exp(\mathrm{i}jx)$ through $\hat{\boldsymbol{f}}_k = \sum_{j \in \mathbb{Z}} c_{k+2jM}$). For functions $f \in H^s(2\pi\mathbb{T})$, the error due to this aliasing effect is of order M^{-s}.

Spatial differentiation is now obtained by multiplying the Fourier coefficients with $\mathrm{i}k$. Nonlinear operations are computed pointwise on collocation points x_n, via inverse Fast Fourier transform. This procedure leads, in general, to aliasing errors. For polynomial nonlinearities (such as for the Saint-Venant system), one can use so-called dealiasing techniques to remove these errors. For quadratic nonlinearities, one may for example use Orszag's 3/2-rule [20], which consists of adding a sufficient number of Fourier modes with coefficients set to zero. For more information on spectral methods and dealiasing techniques, we refer to [5] and [25].

In our numerical codes, to remove aliasing errors from the nonlinear terms while still working with vectors $\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{u}}$ of fixed length, we shall set the highest $1/3$ of the Fourier modes to zero. As we numerically compute approximate solutions of the semi-discretized equations (4.7) and (4.8), this procedure is naturally performed when applying sharp or smooth low-pass filters, $\mathsf{P}_N, \mathsf{S}_N$, with $N < 2M/3$. For the smooth low-pass filter, we use the example from the introduction, that is $\mathsf{S}_N = \mathrm{Diag}(S_N(D))$ with $S_N(\cdot) = S(\cdot/N)$ and $S(\cdot) = \max(0, \min(1, 2 - 2|\cdot|))^2$. For the discretization with the smooth low-pass filter, we shall consider only the version (4.7) where the low-pass filter is only applied to nonlinear terms.

This procedure of semi-discretization in space yields a system of differential equations in time for the Fourier coefficients $\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{u}}$. We approximately solve this initial-value problem using an explicit Runge-Kutta 4 method. All numerical simulations are made using the Julia package WaterWaves1D [9] and can be reproduced using the scripts available at `WaterWaves1D.jl/examples/StudySaintVenant.jl`.

From now on, we denote the number of collocation points by $2M$, and let $N = \lfloor 2M/3 \rfloor$, that is, the greatest integer smaller than $2M/3$. Our numerical scheme maintains the highest $1/3$ of the Fourier modes to zero at each time-step, see the discussion above. Abusing notation, we will refer to the fully-discretized numerical solution as $\boldsymbol{U}_N = (\eta_N, u_N)$, since only $2N$ Fourier modes are nonzero. This convention means that N plays the same role in this section as in the previous, analytical sections. We will compute the numerical solution with $2M = 2^j, j = 6, \ldots, 15$ collocation points and use time step $dt = 1e - 5$. The time step is an order of magnitude smaller than needed to avoid stability issues, and small enough to ensure the error due to the spatial discretization dominates. We will use the solution computed with $2M = 2^{15}$ and sharp low-pass filter as a reference solution $\boldsymbol{U}_{\mathrm{ref}} = (\eta_{\mathrm{ref}}, u_{\mathrm{ref}})$, and compute the relative error of the numerical solutions $\boldsymbol{U}_N = (\eta_N, u_N)$ by comparing with the reference solution:

$$E_s(\boldsymbol{U}_N) = \frac{|\boldsymbol{U}_N - \boldsymbol{U}_{\mathrm{ref}}|_{H^s}}{|\boldsymbol{U}_{\mathrm{ref}}|_{H^s}}.$$

The norms will be computed approximately using the Fourier coefficients of the numerical solutions.

We solve numerically the Saint-Venant system (4.1) in one spatial dimension. For the tested initial data in $\mathcal{U}_{\mathscr{H}}$, numerical results are in agreement with the analysis. To study the experimental convergence, we consider the following initial data for $\alpha > 0$,

$$\eta_0(x) = \frac{1}{2} \exp(-|x|^\alpha) \exp(-4x^2), \qquad u_0(x) = 0. \tag{4.9}$$

Notice $\boldsymbol{U}^0 = (\eta_0, u_0)$ satisfies both $1 + \eta_0 > 0$ and the stricter condition $1 + \eta_0 - (u_0)^2 > 0$. The initial surface is a heap of water situated at the origin. Both η_0, u_0 decay to machine precision at $-\pi, \pi$ and can therefore be seen as periodic. Moreover, $\boldsymbol{U}^0 \in H^{\alpha+1/2}(2\pi\mathbb{T})^2$. We let $\alpha = 1.5$ and simulate the time-evolution up to a final time $T = 0.5$

with either sharp or smooth lowpass filters applied to the nonlinear terms. A plot of the initial data as well as the decay of its Fourier coefficients (through $E^s(\boldsymbol{U}^0) = \frac{|\boldsymbol{U}_{\text{ref}} - \mathsf{P}_M \boldsymbol{U}_{\text{ref}}|_{H^s}}{|\boldsymbol{U}_{\text{ref}}|_{H^s}}$) is shown in Figure 1.
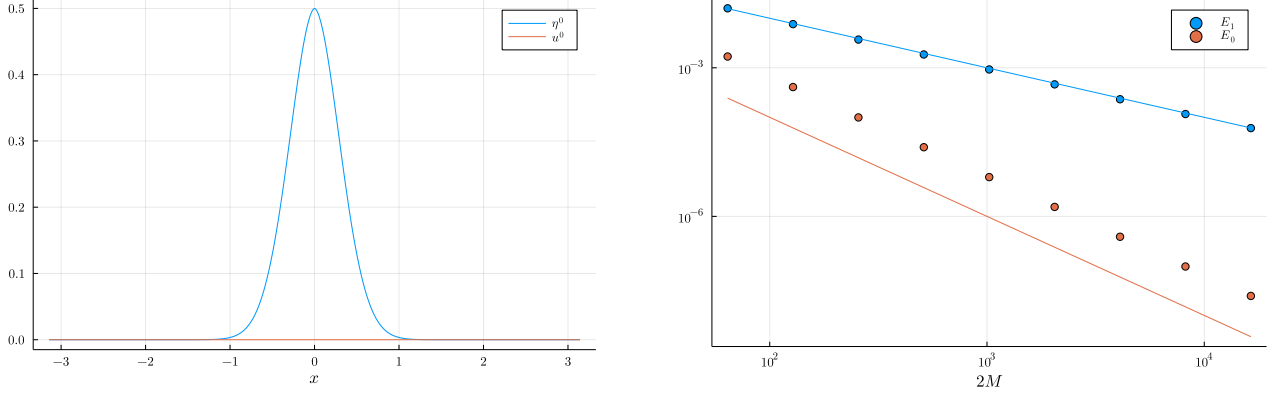
(a) Plot of the initial data (4.9), where η_0 is the initial surface profile, and u_0 is the initial velocity.

(b) Decay of the Fourier coefficients for the initial data (4.9). The blue and orange points show $E_s(\boldsymbol{U}^0)$ for $s = 0, 1$ respectively. To illustrate, the blue and orange lines have slopes -2 and -1 respectively.

Figure 1: Experiments with initial data (4.9).

Figure 2a show log-log plots of the error E_s at time $T = 0.5$ for the numerical solution for $2M = 2^j, j = 6, \ldots, 15$ computed using sharp and smooth low-pass filters. Figure 2a shows the relative error measured in the L^2-norm, $E_0(\boldsymbol{U}_N)$ and in the H^1-norm, $E_1(\boldsymbol{U}_N)$. The experimental order of convergence, given by

$$EOC_s = \frac{\log(E_s(\boldsymbol{U}_N)/E_s(\boldsymbol{U}_{N+1}))}{\log(2)},$$

is given in Figure 2b. With the initial data in $H^2(2\pi\mathbb{T})^d$, we expect from the analysis that the L^2-error should decay as $\mathcal{O}(N^{-2})$ and the H^1-error should decay as $\mathcal{O}(N^{-1})$ when using both the sharp and smooth low-pass filter. This aligns with our numerical results. The absolute error is slightly larger when using the smooth low-pass filter, which is to be expected since applying S_N removes more information in each time step than does P_N. Taking as initial data (4.9) with other values of α (we have tested $\alpha = 1, 2.5, 3$) also yields the expected results.

Let us now consider initial data $(\eta_0, u_0) \in \mathcal{U}\backslash\mathcal{U}_{\mathcal{H}}$, that is satisfying $1 + \eta_0 > 0$, but *not* the stricter condition $1 + \eta_0 - (u_0)^2 > 0$. In particular, we will consider the initial data from the example at the end of Section 4.1, with $q = 0, p = 1, s = 2$:

$$\boldsymbol{U}_N^0 = (\eta_0, u_{0,N}), \quad \eta_0^N(x) = -\frac{1}{2}\cos(x), \quad u_{0,N}(x) = \sin(x) + \frac{\sin(Nx)}{N^2} \tag{4.10}$$

Notice $f(x) = 1 - \frac{1}{2}\cos(x) - \sin^2(x) < 0$ for $x \in (-\frac{\pi}{2}, -\frac{\pi}{3}) \cup (\frac{\pi}{3}, \frac{\pi}{2})$. On the other hand, $1 + \eta_0(x) > 0$ for all $x \in 2\pi\mathbb{T}$. We therefore expect the numerical scheme to converge when using the smooth low-pass filter, but that there may be instabilities when discretizing with the sharp low-pass filter. Convergence plots for the numerical solutions in both cases are shown in Figure 3b. In line with the analysis, we have convergence when using the smooth low-pass filter. However, this is also true for the sharp low-pass filter. *Despite many attempts, we have not been able to observe numerical instabilities for initial data we have tested satisfying $1 + \eta_0 > 0$ but violating $1 + \eta_0 - (u_0)^2 > 0$.*

Interestingly, we do observe a difference between sharp and smooth low-pass filters when (barely) violating the non-cavitation assumption $1 + \eta_0 > 0$. Let

$$\eta_{0,N}(x) = -\cos(x), \quad u_{0,N}(x) = \sin(x) + \frac{\sin(Nx)}{N^2}. \tag{4.11}$$

	Sharp low-pass filter		Smooth low-pass filter	
$2M$	EOC_0	EOC_1	EOC_0	EOC_1
2^6	1.69	0.7	2.75	1.7
2^7	2.02	1.01	1.93	1.00
2^8	2.02	1.01	1.97	0.99
2^9	2.04	1.03	1.95	0.97
2^{10}	2.03	1.02	1.94	0.96
2^{11}	2.03	1.02	1.98	0.98
2^{12}	2.01	0.99	1.98	0.99
2^{13}	1.95	0.92	1.98	0.98
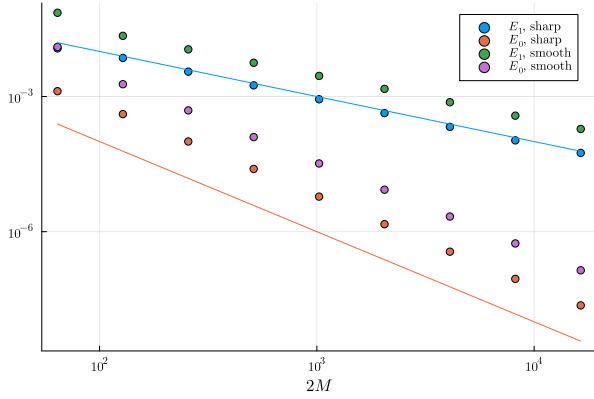
(a) Plot illustrating the convergence of the numerical schemes (4.7) and (4.8) as the number of collocation points $2M$ increases. The plot shows the relative error of the numerical solution for initial data (4.9) in $H^2(2\pi\mathbb{T})^2$ measured in the L^2-norm, E_0 and in the H^1-norm, E_1 for $2M = 2^j, j = 6, \ldots, 14$ when using either sharp or smooth low-pass filters. To illustrate, the blue and orange lines have slopes -2 and -1 respectively. The numerical scheme exhibits spectral convergence with both sharp and smooth low-pass filters.

(b) Experimental order of convergence for the numerical solution with initial data (4.9) for both sharp and smooth low-pass filters. The Experimental order of convergence is measured in the L^2-norm, EOC_0, and in the H^1-norm, EOC_1.

Figure 2: Experiments with initial data (4.9).
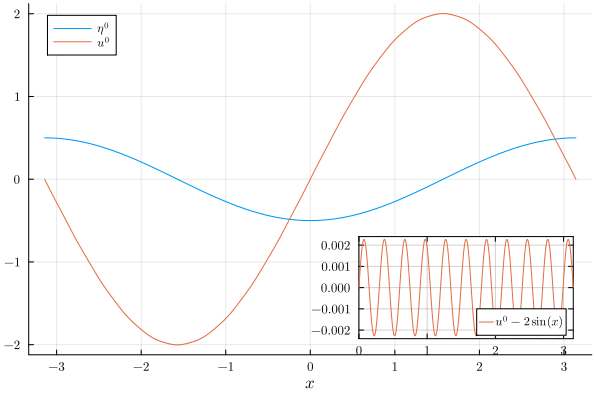
(a) Plot of the initial data (4.10), where η_0 is the initial surface profile, and u_0 is the initial velocity.
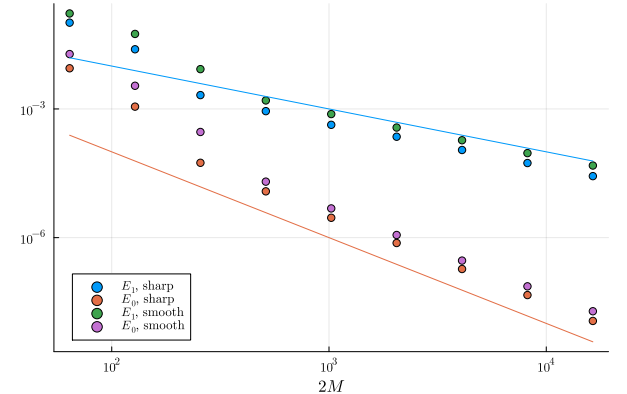
(b) Plot illustrating the convergence of the numerical schemes (4.7) and (4.8) as the number of collocation points $2M$ increases. The plot shows the relative error of the numerical solution measured in the L^2-norm, E_0 and in the H^1-norm, E_1 for $2M = 2^j, j = 6, \ldots, 14$ when using either sharp or smooth low-pass filters. To illustrate, the blue and orange lines have slopes -2 and -1 respectively. The numerical scheme exhibits spectral convergence with both sharp and smooth low-pass filters.

Figure 3: Experiments with initial data (4.10).

23
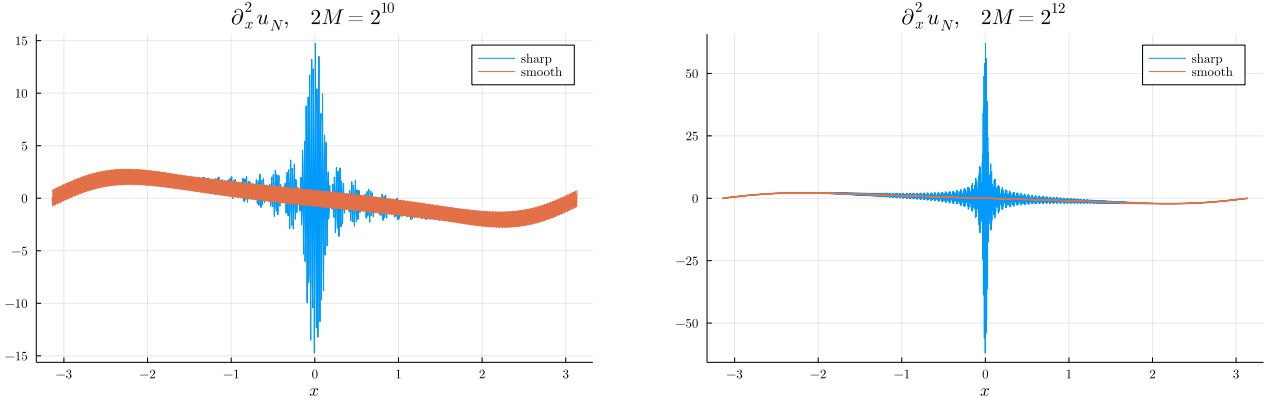
(a) Plot of the second derivative of the velocity u_N at time $T = 0.1$ with initial data (4.11) with $2M = 2^{10}$. The orange and blue lines show the solutions found by using the the numerical schemes (4.7) and (4.8) with smooth and sharp low-pass filters respectively.

(b) Plot of the second derivative of the velocity u_N at time $T = 0.1$ with initial data (4.11) with $2M = 2^{12}$. The orange and blue lines show the solutions found by using the numerical schemes (4.7) and (4.8) with smooth and sharp low-pass filters respectively.

Figure 4: Experiments with initial data (4.11).

Figure 4 shows the second derivative of the velocity at time $T = 0.1$ for smooth and sharp low-pass filters for $2M = 2^{10}$ and $2M = 2^{12}$. The second derivative blows up at the point $x = 0$ where the non-cavitation assumption is violated when using the sharp low-pass filter, but not when using the smooth low-pass filter.

4.3 Numerical experiments in dimension two

The numerical simulations of the Saint-Venant system with $d = 2$ are analogous to the case when $d = 1$ and are made in the same Julia package WaterWaves1D [9]. They are also reproducible using the scripts available at `WaterWaves1D.jl/examples/StudySaintVenant.jl`. We let $2M$ denote the number of collocation points in each of the two spatial dimensions, x and y, which form a grid with $4M^2$ collocation points. As in the previous section, we set $N = \lfloor 2M/3 \rfloor$ and let $\mathsf{P}_N, \mathsf{S}_N$ be as described in the introduction. In particular, notice S_N is now a composition of one dimensional low-pass filters: $\mathsf{S}_N := \mathrm{Diag}(S_N(D))$ with $S_N(k_1, k_2) := S_1(k_1/N)S_1(k_2/N)$, where we set $S_1(\cdot) := \max\left(0, \min\left(1, 2 - 2|\cdot|\right)\right)^2$. We denote numerical approximations by η_N, u_N, v_N. The values at collocation points $\boldsymbol{\eta}, \boldsymbol{u}, \boldsymbol{v}$ and associated discrete Fourier modes $\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}$ are $2M$-by-$2M$ matrices.

Our main interest in studying numerically the systems in two dimensions is to examine whether we observe any difference between the Hamiltonian and non-Hamiltonian version of the Saint-Venant system, respectively (4.2) and (4.1). While there is indeed a difference with respect to stability of the numerical schemes —we observe instabilities with the sharp low-pass filter when violating $1 + \eta_0 - (u_0^2 + v_0^2) > 0$ for (4.2), but not for (4.1)— this can be explained by the difference between the hyperbolicity domains $\mathcal{U}_{\mathscr{H}}$ and $\mathcal{U}$ rather than on the presence or absence of a Hamiltonian structure. Because the setting of dimension $d = 2$ is computationally costlier, we use timestep $dt = 5e - 4$ in our numerical experiments and, when calculating relative errors, we take as a reference solution the numerical solution computed with $2M = 2^{10}$.

We test the numerical method on initial data in $H^s((2\pi\mathbb{T})^2)^3$ of the form

$$\eta_0(x, y) = (h_0 - 1)\cos(x)\cos(y),$$
$$u_0(x, y) = u_\mathrm{l}\sin(x)\cos(y) + u_\mathrm{h}\frac{\sin(Nx)\cos(Ny)}{N^s},$$
$$v_0(x, y) = v_\mathrm{l}\cos(x)\sin(y) + v_\mathrm{h}\frac{\cos(Nx)\sin(Ny)}{N^s},$$

(4.12)

where $h_0 > 0, u_\mathrm{l}, v_\mathrm{l}, s \geqslant 0$ are real numbers. When $u_\mathrm{l} = v_\mathrm{l}$ and $u_\mathrm{h} = v_\mathrm{h}$, the initial data is irrotational, and systems (4.1) and (4.2) are equivalent.
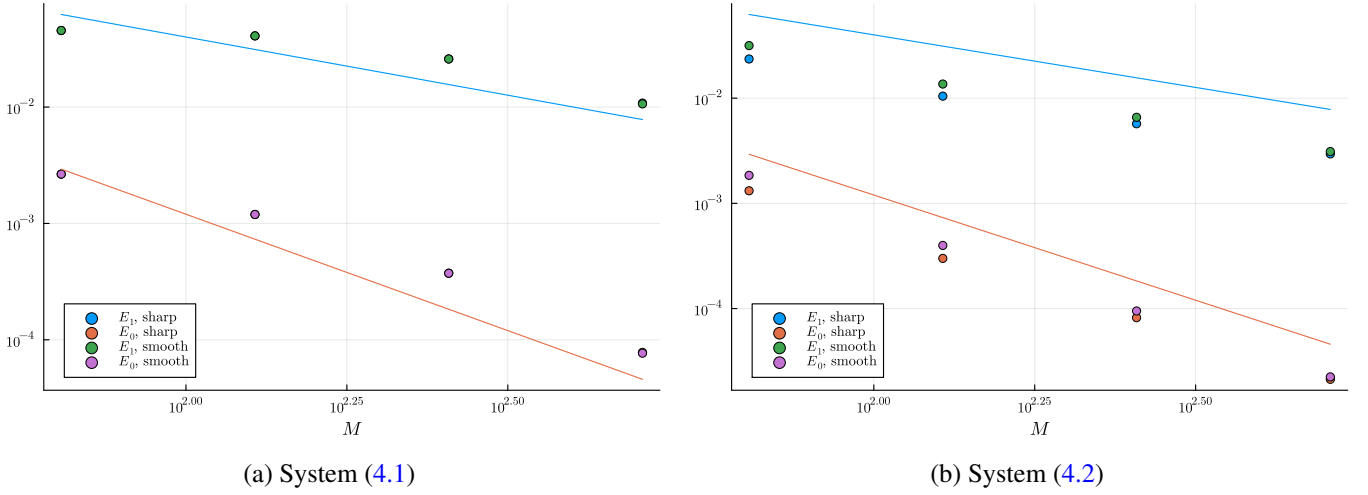
(a) System (4.1)

(b) System (4.2)

Figure 5: Plot illustrating the convergence of the numerical schemes (4.8) and (4.7) for the systems (4.1) (in the left) and (4.2) (in the right) in two spatial dimensions as the number of collocation points $2M$ increases. The plot shows the relative error of the numerical solution for initial data (4.12) with $h_0 = 0.5, u_l = -v_l = 0.5, u_h = -v_h = 1$ and $s = 2$ at time $T = 0.1$. The initial data is in in $H^2((2\pi\mathbb{T})^2)^3$ and the relative error is measured in the L^2-norm, E_0 and in the H^1-norm, E_1 for $2M = 2^j, j = 6, \ldots, 9$ when using either sharp or smooth low-pass filters. To illustrate, the blue and orange lines have slopes -2 and -1 respectively. The numerical scheme exhibits spectral convergence with both sharp and smooth low-pass filters, for both systems.

For the standard, non-Hamiltonian Saint-Venant system (4.1) our numerical results when $d = 2$ align with the numerical results when $d = 1$. That is, the numerical approximation converges with order s for tested initial data in $H^s(2\pi\mathbb{T})$ as long as $1 + \eta_0 > 0$ when using both the smooth and sharp low-pass filters. Figure 5a shows a log-log plot of the relative error measured in the L^2-norm, $E_0(\boldsymbol{U}_N)$ and in the H^1-norm, $E_1(\boldsymbol{U}_N)$, at time $T = 0.1$ for $2M = 2^j, j = 5, \ldots, 9$ and $N = \lfloor 2M/3 \rfloor$ with initial data (4.12) with $h_0 = 0.5, u_l = -v_l = 0.5, u_h = -v_h = 1$ and $s = 2$. Analogous results also hold for other values of s (we have tested $s = 2.5, 3$).

We have not observed any instabilities in the numerical approximation of system (4.1) due to the use of sharp low-pass filter for any of the initial data we tested satisfying $1 + \eta_0 > 0$. Just as in the case of dimension one, we observe instabilities when violating the non-cavitation assumption $1 + \eta_0 > 0$ for the sharp low-pass filter, but not the smooth low-pass filter, see Figure 6.

For the Hamiltonian Saint-Venant system (4.2), the numerical results are in line with the analysis. Whenever the hyperbolicity condition $1 + \eta_0 - (u_0^2 + v_0^2) > 0$ is satisfied, the numerical scheme converges with the expected rate for both the smooth and the sharp low-pass filter. This is illustrated in Figure 5b for initial data (4.12) with $h_0, u_l = 0.5, v_l = -0.5, u_h = -v_h = 1$ and $s = 2$.

We observe instabilities in the numerical approximation of system (4.1) when using the sharp low-pass filter for initial data violating $1 + \eta_0 - (u_0^2 + v_0^2) > 0$ but not $1 + \eta_0 > 0$. This is shown in Figure 7. There, we take as initial data (4.12) with $h_0 = 0.5, u_l = 2, v_l = -2, s = 2, u_h = -v_h = 1$. We relate these instabilities to the lack of well-posedness of the underlying system (4.2) when the hyperbolicity condition $1 + \eta_0 - (u_0^2 + v_0^2) > 0$ fails; see Lemma 4.1. The numerical scheme with the smooth low-pass filter does not exhibit instabilities for the tested values of M, but limited computational power prevents us from testing very large values of M in dimension two.
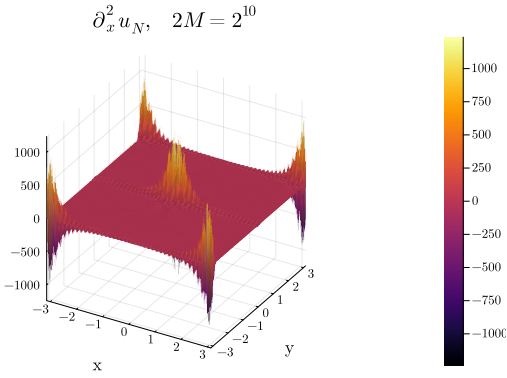
A   Technical tools

The following results are standard, and proofs in the Euclidean space (*e.g.* [14, Theorem 8.3.1] for product estimates) straightforwardly adapt to the periodic setting.
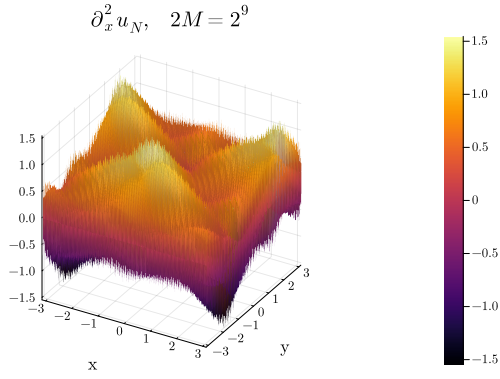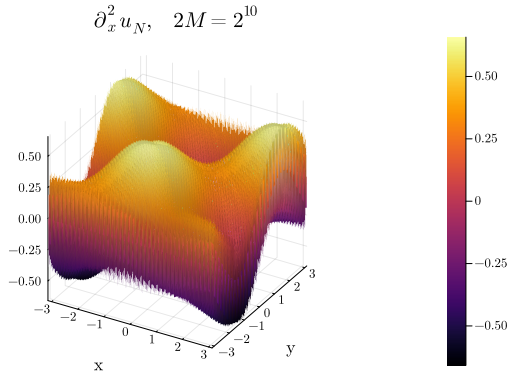
(a) Solution computed with the sharp low-pass filter and $2M = 2^9$.
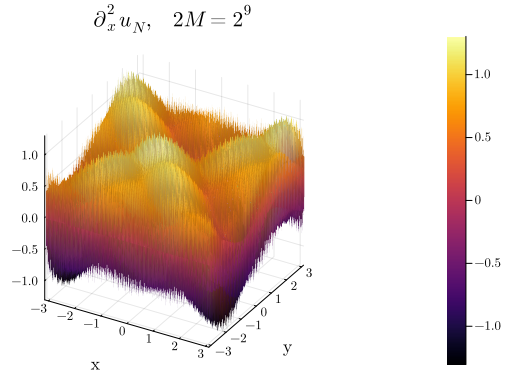
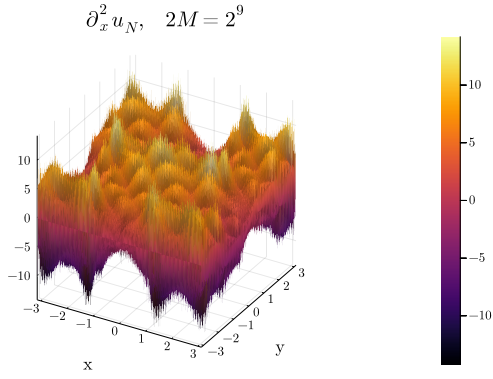(b) Solution computed with the smooth low-pass filter and $2M = 2^9$.

(c) Solution computed with the sharp low-pass filter and $2M = 2^{10}$.
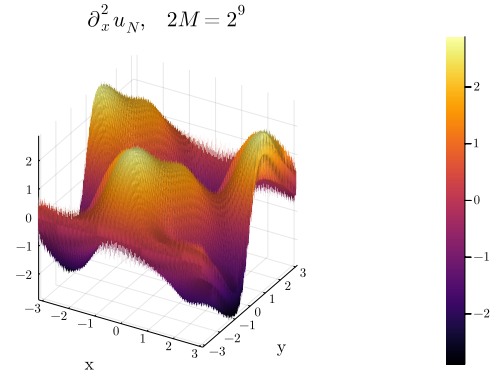
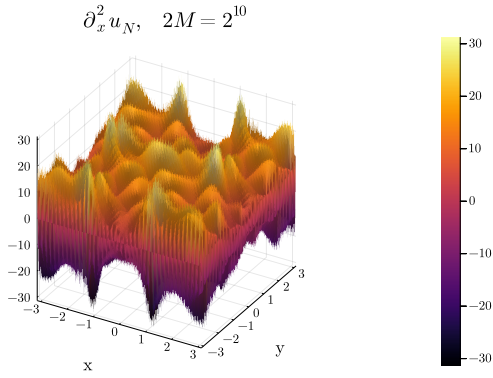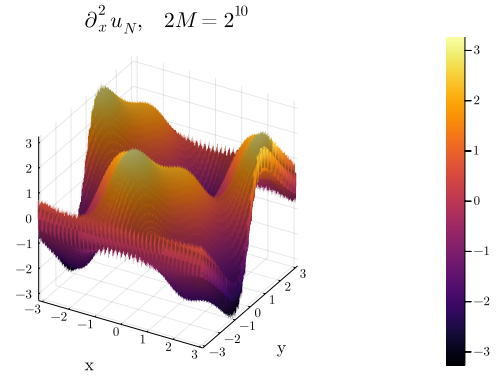(d) Solution computed with the smooth low-pass filter and $2M = 2^{10}$.

Figure 6: Plots of the second derivative $\partial_x^2 u_N$ of the numerical solution to (4.1) for $d = 2$ at time $T = 0.1$, computed with either the sharp or smooth low-pass filter. The initial data is (4.12) with $s = 2$, $u_l = 0.5$, $v_l = -0.5$, $u_h = -v_h = 1$, negative minimial depth $h_0 = -0.1$ and $N = \lfloor 2M/3 \rfloor$ for $2M = 2^9$ or $2M = 2^{10}$.

(a) Solution computed with the sharp low-pass filter and $2M = 2^9$.

(b) Solution computed with the smooth low-pass filter and $2M = 2^9$.

(c) Solution computed with the sharp low-pass filter and $2M = 2^{10}$.

(d) Solution computed with the smooth low-pass filter and $2M = 2^{10}$.

Figure 7: Plots of the second derivative $\partial_x^2 u_N$ of the numerical solution to (4.2) for $d = 2$ at time $T = 0.1$, computed with either the sharp or smooth low-pass filter. The initial data is (4.12) with $s = 2, u_l = 2, v_l = -2, u_h = -v_h = 1$, positive minimal depth $h_0 = 0.5$ and $N = \lfloor 2M/3 \rfloor$ for $2M = 2^9$ or $2M = 2^{10}$.

Proposition A.1 (Continuous embedding)**.** *Let $s \in \mathbb{R}$, $s > d/2$ and $f \in H^s((2\pi\mathbb{T})^d)$. Then $f \in L^\infty((2\pi\mathbb{T})^d)$ and*

$$\big|f\big|_{L^\infty} \leqslant C(s)\big|f\big|_{H^s}.$$

Proposition A.2 (Interpolation inequality)**.** *Let $s_1, s_2 \in \mathbb{R}$ and $f \in H^{s_1}((2\pi\mathbb{T})^d) \cap H^{s_2}((2\pi\mathbb{T})^d)$. Then for any $0 \leqslant \theta \leqslant 1$, $f \in H^{\theta s_1 + (1-\theta)s_2}((2\pi\mathbb{T})^d)$ and*

$$\big|f\big|_{H^{\theta s_1 + (1-\theta)s_2}} \leqslant \big|f\big|_{H^{s_1}}^{\theta}\big|f\big|_{H^{s_2}}^{1-\theta}.$$

Proposition A.3 (Product estimates)**.** *Let $s_0 > d/2, s \geqslant -s_0$ and $f \in H^s((2\pi\mathbb{T})^d) \bigcap H^{s_0}((2\pi\mathbb{T})^d), g \in H^s((2\pi\mathbb{T})^d)$. Then $fg \in H^s((2\pi\mathbb{T})^d)$ and*

$$\big|fg\big|_{H^s} \leqslant C(s_0, s)\Big(\big|f\big|_{H^{s_0}}\big|g\big|_{H^s} + \big|f\big|_{H^s}\big|g\big|_{H^{s_0}}\Big).$$

If moreover $s \leqslant s_0$ then

$$\big|fg\big|_{H^s} \leqslant C(s_0, s)\big|f\big|_{H^{s_0}}\big|g\big|_{H^s}.$$

Assuming polynomial nonlinearities, the following proposition is a straightforward consequence of product estimates. Extending this result to general (smooth) functions P requires an analysis that is outside of the scope of the present paper.

Proposition A.4 (Composition estimates)**.** *Let $s_0 > d/2, s \geqslant -s_0$, $f, g \in H^s((2\pi\mathbb{T})^d) \bigcap H^{s_0}((2\pi\mathbb{T})^d)$ and $P \in \mathbb{R}[X]$ a polynomial. Then $P(f), P(g) \in H^s((2\pi\mathbb{T})^d)$ and*

$$\big|P(f) - P(g)\big|_{H^s} \leqslant C(P, s_0, s, \big|f\big|_{H^{\max(s_0,s)}}, \big|g\big|_{H^{\max(s_0,s)}})\big|f - g\big|_{H^s},$$
$$\big|P(f) - P(0)\big|_{H^s} \leqslant C(P, s_0, s, \big|f\big|_{H^{s_0}})\big|f\big|_{H^s}.$$

Proposition A.5 (Commutator estimates with symbols of order s)**.** *Let $s_0 > d/2, s \geqslant 0$ and $\Lambda^s = (\mathrm{Id} - \Delta)^{s/2}$. Let $f \in H^s((2\pi\mathbb{T})^d) \bigcap H^{s_0+1}((2\pi\mathbb{T})^d), g \in H^{s-1}((2\pi\mathbb{T})^d) \bigcap H^{s_0}((2\pi\mathbb{T})^d)$. Then*

$$\big|[\Lambda^s, f]g\big|_{L^2} \leqslant C(s_0, s)(\big|f\big|_{H^s}\big|g\big|_{H^{s_0}} + \big|f\big|_{H^{s_0+1}}\big|g\big|_{H^{s-1}}).$$

The following result is shown on $\mathbb{R}^d$ in [8, Lemma 4.5]. It is straightforward to adapt the proof to the periodic setting.

Proposition A.6 (Commutator estimates with operators of order zero)**.** *Let $s_0 > d/2, s \geqslant 0$ and $G(D)$ be a Fourier multiplier with symbol G satisfying $\big|G\big|_{L^\infty}, \big|\,|\cdot|\nabla G\big|_{L^\infty} \leqslant C_G$. Let $f \in H^{s_0+1}((2\pi\mathbb{T})^d) \bigcap H^s((2\pi\mathbb{T})^d)$, $g \in H^{s-1}((2\pi\mathbb{T})^d)$. Then*

$$\big|[G(D), f]g\big|_{H^s} \leqslant C(s_0, s)\, C_G\, \big|f\big|_{H^{\max(s_0+1,s)}}\big|g\big|_{H^{s-1}}.$$

Remark A.7. *Notice that if G is such that $\big|G\big|_{L^\infty}, \big|\,|\cdot|\nabla G\big|_{L^\infty} \leqslant C_G$, then $G_N := G(\cdot/N)$ satisfies*

$$\big|G_N\big|_{L^\infty}, \big|\,|\cdot|\nabla G_N\big|_{L^\infty} \leqslant C_G$$

uniformly with respect to N.

Acknowledgements

References

[1] C. Bardos and E. Tadmor. Stability and spectral convergence of Fourier method for nonlinear problems: on the shortcomings of the 2/3 de-aliasing method. *Numer. Math.*, 129(4):749–782, 2015.

[2] S. Benzoni-Gavage and D. Serre. *Multidimensional hyperbolic partial differential equations. First-order systems and applications*. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, Oxford, 2007.

[3] M. Cabrera Calvo, F. Rousset, and K. Schratz. Time integrators for dispersive equations in the long wave regime. *Math. Comp.*, 91(337):2197–2214, 2022.

[4] M. Cabrera Calvo and K. Schratz. Uniformly accurate splitting schemes for the Benjamin-Bona-Mahony equation with dispersive parameter. *BIT*, 62(4):1625–1647, 2022.

[5] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods. Fundamentals in single domains.* Sci. Comput. Berlin: Springer, 2006.

[6] V. A. Dougalis, A. Duran, and L. Saridaki. On the numerical approximation of Boussinesq/Boussinesq systems for internal waves. *Numer. Methods Partial Differential Equations*, 39(5):3677–3704, 2023.

[7] V. Duchêne. Many Models for Water Waves. Open Math Notes, OMN:202109.111309, 2021.

[8] V. Duchêne and B. Melinand. Rectification of a deep water model for surface gravity waves. *Pure Appl. Anal.*, 6(1):73–128, 2024.

[9] V. Duchêne and P. Navaro. Waterwaves1d.jl (version v0.1.0). Zenodo. https://doi.org/10.5281/zenodo.7142921.

[10] L. Emerald. Local well-posedness result for a class of non-local quasi-linear systems and its application to the justification of Whitham-Boussinesq systems. arXiv preprint:2206.09213.

[11] J. Goodman, T. Hou, and E. Tadmor. On the stability of the unsmoothed Fourier method for hyperbolic equations. *Numer. Math.*, 67(1):93–129, 1994.

[12] D. Gottlieb and J. S. Hesthaven. Spectral methods for hyperbolic problems. volume 128, pages 83–131. 2001. Numerical analysis 2000, Vol. VII, Partial differential equations.

[13] D. Gottlieb and S. A. Orszag. *Numerical analysis of spectral methods: theory and applications*. CBMS-NSF Regional Conference Series in Applied Mathematics, No. 26. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1977.

[14] L. Hörmander. *Lectures on nonlinear hyperbolic differential equations*, volume 26 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 1997.

[15] C. Klein, F. Linares, D. Pilod, and J.-C. Saut. On Whitham and related equations. *Stud. Appl. Math.*, 140(2):133–177, 2018.

[16] H.-O. Kreiss and J. Oliger. Comparison of accurate methods for the integration of hyperbolic equations. *Tellus*, 24:199–215, 1972.

[17] D. Lannes. *The water waves problem*, volume 188 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2013. Mathematical analysis and asymptotics.

[18] A. Majda, J. McDonough, and S. Osher. The Fourier method for nonsmooth initial data. *Math. Comp.*, 32(144):1041–1081, 1978.

[19] G. Métivier. *Para-differential calculus and applications to the Cauchy problem for nonlinear systems*, volume 5 of *Centro di Ricerca Matematica Ennio De Giorgi (CRM) Series*. Edizioni della Normale, Pisa, 2008.

[20] S. A. Orszag. On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *J. Atmos. Sci.*, 28:1074, 1971.

[21] M. O. Paulsen. Long time well-posedness of Whitham-Boussinesq systems. *Nonlinearity*, 35(12):6284–6348, 2022.

[22] J.-C. Saut and L. Xu. The Cauchy problem on large time for surface waves Boussinesq systems. *J. Math. Pures Appl. (9)*, 97(6):635–662, 2012.

[23] E. Tadmor. Stability analysis of finite difference, pseudospectral and Fourier-Galerkin approximations for time-dependent problems. *SIAM Rev.*, 29(4):525–555, 1987.

[24] E. Tadmor. Convergence of spectral methods for nonlinear conservation laws. *SIAM J. Numer. Anal.*, 26(1):30–44, 1989.

[25] L. N. Trefethen. *Spectral methods in MATLAB*, volume 10 of *Software, Environments, and Tools*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[26] J. C. Xavier, M. A. Rincon, D. G. Alfaro Vigo, and D. E. Amundsen. Stability analysis for a fully discrete spectral scheme for Boussinesq systems. *Appl. Anal.*, 97(4):610–632, 2018.