School of Mathematical, Physical and Computational Sciences

Finite element methods as geometric structure preserving algorithms

by

James Iain Jackaman

Thesis submitted for the degree of Doctor of Philosophy

Department of Mathematics and Statistics

September 2018

Declaration of original authorship

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged. Chapter 2 consists of a literature review. The literature review continues through §3–3.1.0 inclusive, the remainder of Chapter 3 is based on the joint work with Pryer [109] at the University of Reading. Chapters 5 and 7 are based on the papers [108] and [110] respectively, and are also joint work with Pryer. Chapter 8 is based on the paper [107], and is joint work with Papamikos and Pryer. The work with Papamikos was conducted during his tenure at the University of Reading. For all foregoing papers I am the principal contributor.

James Iain Jackaman

Abstract

Here we investigate finite element techniques aimed at preserving the underlying geometric structures for various problems, and, in doing so, develop new geometric structure preserving methods. We initially focus on systems of Hamiltonian ODEs, examining the place of existing methods as geometric numerical integrators. We then develop a new geometrical finite element method for Hamiltonian ODEs with a view to generalise it to be the temporal discretisation of a space-time adaptive finite element method.

We go on to investigate how well finite element methods can preserve the structure of Hamiltonian PDEs, which are a large class of physically relevant PDEs possessing a conserved physical invariant, the Hamiltonian functional, which often physically represents the energy of the problem. Examples of this kind of problem include, but are not limited to, oceanographical models of wave propagation such as KdV type equations and the nonlinear Schrödinger equations, and the semi-geostrophic equations for atmospheric modelling. We construct a general methodology for the design of finite element schemes for such problems and go on to develop multiple schemes in this framework for not only Hamiltonian PDEs but also systems of Hamiltonian PDEs. Within the study of finite element methods for Hamiltonian PDEs we prove both a priori and a posteriori error bounds, in addition to examining the role of spatial adaptivity for our schemes.

Acknowledgements

I would like to thank my supervisor, Tristan Pryer, for all of his support throughout my PhD, and my collaborators Georgios Papamikos (University of Reading) and Thomas Melvin (UK Met Office). I would also like to thank all of my family and friends who have continually supported me, especially through the final year of my PhD. Additionally, I would like to thank EPSRC and the University of Reading for financial support under the grant "EPSRC Centre for Doctoral Training in the Mathematics of Planet Earth at Imperial College London and the University of Reading" EP/L016613/1. Last, but not least, I would like to thank both examiners, Steve Langdon and Andreas Dedner, for their careful reading of this thesis and constructive feedback.

Table of Contents

Chapter 1

Introduction

A vast number of problems arising from physics possess various algebraic and geometric structures. For example, interstellar motion modelled by the Hénon Heines model [90] through systems of ordinary differential equations *(ODEs)*, which possesses a conserved energy and closed orbits. Or alternatively, atmospheric fluid dynamics which is typically modelled by Navier-Stokes or Euler partial differential equations *(PDEs)*, see [177], which possess physical structures such as conservation of mass, energy, among a multitude of others.

When discretising any physical problem, it is of paramount importance to take these underlying physical properties into account. Not only do they yield more physically relevant solutions, but the conservation of such properties often leads to numerical stability. It is crucial to note that the converse is not true, numerical stability does not imply that any physical properties are preserved, as can be seen through the study of the backward Euler method in [102].

In this thesis we focus on finite element methods and their ability to preserve the underlying structure of a problem. In the ODE setting the underlying structures we consider fall under the umbrella of *geometric integration* (as discussed in Chapter 2). We apply the term "structure" in the PDE setting to refer to the natural generalisation of the notion of geometric integration. The term "finite element method" was coined in [44], with original applications in aeronautical engineering, however, the development of finite element methods goes back to [96, 50] and their work on developing "finite element" basis functions. Due to the analytical nature of their framework, see [15], and their flexibility, finite element methods have flourished, see [171, 6, 65, 176, 33]. Significant progress has been made on the study of the conservative properties of finite element methods for conservation

laws for PDEs, see for example [47], here we extend the understanding of the conservative properties of finite element methods for a wide variety of underlying problems.
We divide this work into two parts based on the type of problem, and hence structure, we consider. Initially, we consider the application of finite element methods to ODEs, then their application as the spatial discretisation of Hamiltonian PDEs.

1.1 Finite element methods for ODEs

While typically the finite element method is utilised for the solving of PDEs, it is also well developed for systems of first order ODEs, and has been extensively analysed when constructing both a continuous and discontinuous finite element solution, see [113, 73, 67, 66, 68]. A large class of systems of first order ODEs, known as Hamiltonian systems, see [130], can be studied as prototypical examples of structure preserving ODEs.
Such problems are well understood for classical time stepping methods for ODEs and are known as *geometric numerical integrators*, see [85, 127, 162]. For this class of problem two key structures are considered numerically. The conservation of the energy, and the preservation of a "symplectic" structure on the flow map of the solution. While the former is easier to visualise, and has well developed numerical methods, see [152, 81, 137, 42, c.f.], the latter is typically aimed to be conserved. This is primarily due to the ability to design classes of symplectic methods of arbitrarily high order, see [160]. In addition, it is known that a method conserving the symplectic structure is "close" to conserving the energy, see [28].
Due to the maturity of geometric numerical integration, it may initially appear that establishing the finite element method in geometric numerical integration is an exercise in futility, however this is not the case. The development of finite element geometric integrators affords insight into the temporal discretisation of PDEs. Spatial finite element methods provide a powerful tool for the discretisation of the spatial component of a PDE, however the temporal discretisations typically fall within a different framework, see [91, 100, 60, 173, c.f.]. Through discretising temporally within the finite element framework it is possible to unify the analysis of space and time, see [99, 174, 140, 120]. For conservative problems, to yield physically relevant simulations over long time, it is important to choose compatible finite element temporal discretisations which conserve invariants of the underlying problem.
A powerful tool for the accurate simulation of space-time finite element schemes is space-

time adaptivity, see [71, 140, 164]. However, for the development of space-time adaptive algorithms, in general, discontinuity is required to avoid hanging nodes, see [22]. Note that through the careful coupling of refinement and coarsening with the polynomial degree of a finite element approximation it is possible to construct continuous space-time adaptive algorithms, see [168], but we shall not press this point here. Regardless, the temporally discontinuous nature of any space-time adaptive algorithm precludes the ability to conserve geometric structures over time as for a well used discontinuous temporal algorithm to be conservative it needs to be solved globally, which is not practical for evolution problems. For spatial problems a new type of finite element method has been designed, known as the recovered finite element method *(RFEM)*, see [74, 58]. This method allows for a discontinuous underlying solution which possesses a continuous reconstruction. The design of a temporal RFEM method affords the ability to design adaptive space-time algorithms, avoid the issue of hanging nodes, and maintain a structure preserving continuous reconstruction. The development of the aforementioned method will be a focal point in Chapter 3.

Some work into the understanding of finite element methods as geometric numerical integrators has already been conducted. See, for example, [87], where it is shown that the continuous finite element method exactly conserves the energy of Hamiltonian ODEs. Additionally, the same finite element method is "close" to a standard symplectic time stepping method, and is therefore close to preserving a discrete symplectic mapping, see [105]. Note that, in general, it is not possible to conserve both a discrete symplectic mapping and the energy, see [185].

The notion of a symplectic mapping is not solely restricted to the ODE setting. In [135] the notion of a *multisymplectic* integrator was introduced, and formally says that a PDE is symplectic if it is *on average* symplectic in time and space. This has lead to the development of discrete multisymplectic integrators, which typically comprise of Euler/Preissman box schemes, see [34, 49, 48]. While it is possible to compare appropriate finite element methods to the box schemes this is not the natural framework for a finite element method due to their continuous nature. Indeed, a more appropriate notion of multisymplectic within the finite element framework is described in [138]. The authors consider hybridised discontinuous Galerkin methods, which can be implemented locally subject to prescribed boundary fluxes and notice that, as within a finite element the numerical solution is continuous, the multisymplectic structure of the method arises from the appropriate handling of the boundary fluxes. They place multiple existing methods into the hybridised discontinuous Galerkin framework and examine their multisymplectic properties.

1.2 Finite element methods for Hamiltonian PDEs

Throughout our study of finite element methods for PDEs we primarily focus on Hamiltonian PDEs. Hamiltonian PDEs are a specific class of PDE endowed with physically relevant algebraic and geometric structures [147]. They arise from a variety of areas, not least meteorological [166], such as the semi-geostrophic equations [155], and oceanographical, such as the Korteweg-de Vries (KdV) and nonlinear Schrödinger equations [144]. The KdV and nonlinear Schrödinger equations are particularly special examples, in that they are bi-Hamiltonian [131]. This means they have two different Hamiltonian formulations which, in turn, is one way to understand the notion of integrability of these problems. Regardless, the applications and the need to quantify the dynamics of the general Cauchy problem motivate the development of accurate long time simulations for reliable prediction of dynamics in both meteorology and oceanography.

A difficulty in the design of schemes for this class of problems is that the long term dynamics of solutions can be destroyed by the addition of *artificial numerical diffusion*. The reason for inclusion of this in a given scheme is the desirable stability properties this endows on the approximation, however, this typically destroys all information in the long term dynamics of the system through smearing of solutions.

Here we shall primarily focus on KdV type equations. In previous numerical studies of the scalar KdV and modified KdV equations [182, 183, c.f.], it has been observed that classical finite volume and discontinuous Galerkin *(dG)* schemes with "standard" numerical fluxes introduce numerical artifacts. These typically appear through numerical regularisation effects included for stability purposes which are not adapted to the variational structure of the problem. The result of these artifacts is an inconsistency in the discrete energy.

Hamiltonian problems are inherently conservative, that is, the underlying Hamiltonian functional is conserved over time. Other equations, including those of integrable type, may have additional structures that manifest themselves through additional conserved quantities. In particular, mass and momentum are such quantities. In [31, 114] the authors propose and analyse a dG scheme for generalised KdV equations. The scheme itself is very carefully designed to be conservative, in that the invariant corresponding to the *momentum* is inherited by the discretisation. This yields L_2 stability quite naturally in the numerical scheme and extremely good long time dynamics. It is also possible to design schemes that conserve the energy itself, see [180] and §4.2 onward, however, it does not seem possible to design schemes to conserve more than two of these invariants for nonlinear problems.

A primary goal of this work is the derivation of Galerkin discretisations aimed at preserving the underlying algebraic properties satisfied by the PDE system whilst avoiding the introduction of stabilising diffusion terms. Our schemes are therefore consistent with the Hamiltonian formulation of the original problems. It is important to note that our approach is not an adaptation of entropy conserving schemes developed for systems of conservation laws, see [172, 88], rather we study the algebraic properties of the PDE and formulate the discretisation to inherit this specific structure. The methods are of arbitrarily high order of accuracy in space and provide relevant approximations free from numerical artifacts. Similar techniques have proven useful in the study of dispersive phase flow problems [77, 75] and we anticipate they will be extremely useful in dynamic model adaptivity [79].

The schemes proposed here form classes of nonconforming finite element method which have proven successful in the design of schemes for elliptic, parabolic and hyperbolic type PDEs that are constructed *without* enforcing global continuity on the discrete solution. One of the strengths of the method stems from the flexibility offered in the flux choice over the element endpoints. This has proven a powerful tool in the design of stable schemes for convection-dominated problems [45]. See also [12] for an accessible overview and history of these methods for elliptic problems. For high order spatial operators, for example, the dispersion operator in KdV, dG methods are a useful alternative to C^1 elements whose derivation and implementation can become very complicated, see [19, 151].

The KdV equation [119] has been extensively studied numerically, see [159, 1, 7, 163, 180, 115, c.f.]. In addition, the local discontinuous Galerkin method has proven quite successful for the linearised problem, see [183, 182, 128, 98], where superconvergent approximation schemes can be designed which are also conservative. Note also the recent work [43] where a hybrid discontinuous Galerkin scheme has been presented for the stationary linearised problem. These methods superconverge at the nodes which, via a post processing procedure, lead to a uniformly superconvergent approximation by reconstructing the approximation using high order interpolation about the nodes, see [46, 141].

A notion of integrability appropriate for the KdV equation is that the equation possesses *infinitely* many conservation laws, see [124]. Two of these correspond to the two Hamiltonians that admit KdV into the bi-Hamiltonian framework. The two Hamiltonian functions of the KdV equation physically represent the *momentum* and *energy* of the PDE, which constitute two of the three lowest order invariants. The linearised KdV problem allows for the design of discretisations that conserve the mass and both the underlying Hamiltonians, *all three* of the fundamental, or base, conservation laws, see §4.2. This is due

to both conserved quantities being quadratic allowing for the creation of *compatible* spatial and temporal discretisations. In the nonlinear case one of the Hamiltonians remains quadratic, while the other is cubic. It is not known how to design schemes that are able to conserve all three base conservation laws in this case, even in the semi-discrete setting, neither spatial nor temporal. This forces upon the user a choice of which invariant to conserve and motivates Chapter 5, to examine the properties of the discretisations arising from choosing to conserve one conserved quantity over the other.

To highlight the good behaviour of the proposed schemes, we develop a priori error bounds for a subset of our schemes. Fundamentally, these a priori bounds depend upon the energy arguments, see [175, 129, 157], and in particular the energy of the underlying problem *inducing a norm*.

Further, in the sequel, we give an a posteriori error analysis making use of a hybrid framework consisting of elliptic reconstruction techniques [133, 121, 122] together with those developed for hyperbolic conservation laws [76] to allow derivation of optimal a posteriori error bounds in the energy norm. Note that the arguments we use are quite different to that of [114] where the authors construct a dispersive reconstruction to allow for a posteriori control in L_2.

Equipped with a posteriori bounds, we examine the delicate interplay between adaptivity and conservation. We show that standard adaptive procedures applied to our scheme not only fail to conserve the invariants but can also become unstable. The instability is caused by an incompatibility in the Hamiltonian structure of the problem and the *mesh change operator*. This is the mechanism with which we transfer information between meshes of different refinement levels. We rectify this incompatibility by proposing a modified mesh change operator that correctly preserves the invariants under mesh refinement and is dissipative under mesh coarsening, ensuring numerical stability. We note a similar observation was given in [59] where the author studies the linear Schödinger equation, although stability is not guaranteed. However, constructing dissipative operators for conservation problems is not the only way to guarantee numerical stability for adaptive algorithms, and in fact employing conservative mesh change operators through Lagrange multipliers has proven successful in the literature, see [63, 143].

We note that issues have been observed even when examining dissipative PDEs adaptively. A classical example shown in [62, §4] shows modifying the underlying mesh for a difference discretisation of a one dimensional heat equation can result in a completely inconsistent numerical scheme. Further stability issues have been observed for multi-dimensional discretisations of the heat equation, for example in [24] the authors observe instabilities

occurring in a Crank-Nicolson finite element discretisation upon refinement that was resolved through the introduction of an appropriate mesh change term. Further examples are given in [29].
When examining systems of Hamiltonian equations much less work has been carried out, for example [17] give a near conservative method for a system of Schrödinger–KdV type and [30] study a system of KdV equations. In the sequel we consider a system of Hamiltonian equations known as the vectorial modified KdV equation *(vmKdV)* [9]. To the author's knowledge there has not been any numerical work on this system, nor such Hamiltonian systems in general. Similarly to the scalar case, our proposed scheme is consistent with (one of) the Hamiltonian formulation of the original problem. The methods are of arbitrarily high order of accuracy in space and provide relevant approximations free from numerical artifacts. Multiple technical complications arise extending the methodology from the scalar to vectorial case, not least of which is the accurate transfer of energy between vector components of the solution which can be resolved through the introduction of a Lagrange multiplier, see [16].

1.3 Thesis structure

This thesis can be divided into two main parts. The first revolves around the study of ODEs §2–3. In Chapter §2 we provide an overview of the literature on geometric integration for Hamiltonian ODEs and lay the necessary foundations to discuss finite element methods as geometric integrators. In Chapter §3 we recall the geometric integration properties of existing finite element methods discussed in [105]. Additionally, we construct a priori error bounds for linear problems. In §3.2 we introduce a new finite element method, the temporal RFEM method, which facilitates the design of a discontinuous underlying solution which possesses a continuous reconstruction. We discuss the implementation of this method as a temporally adaptive algorithm with space-time adaptivity in mind. A priori bounds are computed for this new method, and numerical experiments are presented.
In the second part, §4–8, we investigate the application of finite element methods to Hamiltonian PDEs. We begin in §4.1 by proposing a general methodology for the construction of discretisations of Hamiltonian PDEs such that the discretisation inherently conserves the corresponding Hamiltonian functional. We apply this methodology to the linearised KdV equation in §4.2, which is bi-Hamiltonian. We find that for the linear problem we can design a spatial discretisation which conserves *both* Hamiltonian functionals of the

problem (namely momentum and energy). The conservation of both functionals allows us to develop an a priori bound for the spatially discrete scheme.

As it does not appear possible to design discretisations which conserve multiple Hamiltonian functionals for nonlinear KdV type problems, we investigate which invariant yields more accurate numerical results in Chapter 5. We compare a scheme designed using the methodology in §4.1 which conserves the Hamiltonian corresponding to energy against the scheme proposed in [31] which conserves the Hamiltonian corresponding to momentum. We remark here that both conservative discretisations are not immediately compatible with adaptivity, leading us to Chapter 6 where we introduce an adaptive algorithm for Hamiltonian problems where we use a spatial finite element discretisation in space and a finite difference, or method of lines, discretisation in time. For clarity, we restrict our study here to the method for linearised KdV proposed in §4.2. We show that the conservative properties of the adaptive scheme depend entirely on the *mesh change operator*, i.e., the operator which maps finite element functions between different spatial meshes. In particular, the mesh change operator must be designed such that it inherently preserves the invariants of the nonadaptive method. Here we introduce a mass conserving and momentum dissipating mesh change operator.

In Chapter 7 we investigate dispersive KdV type equations. Here, the energy of the scheme *induces a norm*, and as such the conservative methods we design exactly preserve an appropriate energy norm over time. Note that, even in the linear case, this scheme does not conserve momentum. Conservation of the energy norm over time allows us to prove optimal a priori and a posteriori error bounds in the energy norm for the linear case in §7.2 and §7.3 respectively. Additionally, we have laid out the framework for these bounds to be generalised to the nonlinear case. We also investigate adaptivity for this scheme in §7.4, proposing an energy dissipating mesh change operator and examine the numerical error for a variety of different mesh change operators in §7.5.2.

We go on in Chapter 8 to investigate a *system* of Hamiltonian PDEs known as the vectorial modified KdV equation, proposing a fully discrete energy conserving scheme comprising of the amalgamation of a spatial finite element method and energy conservative time stepping. For all chapters in this part we present appropriate numerical experiments at the end of the respective chapters.

1.4 Novel contributions

Before proceeding we shall summarise the novel contributions of the work within this thesis. In Chapter 3 we propose a new type of temporal finite element method which is inspired by the spatial finite element methods developed in [74], but is fundamentally different in both construction and implementation. This new temporal finite element facilitates a methodology for the construction of structure preserving space-time adaptive finite element methods.

We outline a new general methodology for the construction of conservative finite element schemes for Hamiltonian PDEs in Chapter 4, noting that similar structures have been built into existing finite difference schemes in the literature. Using this methodology we construct multiple schemes for various Hamiltonian PDEs, with a primary focus on KdV-type equations. We prove new a priori, and a posteriori error bounds for such schemes in Chapter 4 and Chapter 7, as well as a numerical comparison of a new energy conserving scheme compared against a well studied momentum conserving example in Chapter 5.

In addition, we develop new stable adaptive algorithms in Chapter 6 and Chapter 7. The former relies on known techniques, however the latter involves the development of a new stable operator for mapping the numerical solution between meshes.

Finally, we develop the first scheme for the vectorial modified KdV equation, a system of Hamiltonian PDEs. For this systems of Hamiltonian PDEs multiple additional complications arise. Our discretisation inherits the underlying Hamiltonian structure of the continuous problem in the sense that it conserves the energy of the system over time.

Chapter 2

Geometric numerical integration for ODEs

Here we summarise several key results from the research area of geometric numerical integration. While no new results are given in this chapter, it serves not only as an introduction to structure preserving numerical methods but is the genesis point for this thesis. All results presented in this chapter can be found in [85, 161, 127]. Several ideas here have also been investigated in the masters thesis [104]. As such, we shall not typically present the proofs of results here, instead providing references to where the proofs were originally presented. Geometric numerical integration of ODEs provides powerful tools for the design of temporal discretisations for PDEs. We shall investigate structure preserving temporal discretisations of PDEs from Chapter 4 onwards.

2.1 The continuous problem

Let $\boldsymbol{u}(t) = \boldsymbol{u} \in (C^1([0,T]))^D$ for $t \in [0,T]$ with T some predetermined constant, and where D is a positive integer. Then, we define a general ODE such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = \boldsymbol{F}(\boldsymbol{u}, t), \tag{2.1}$$

subject to the initial data $\boldsymbol{u}(0) = \boldsymbol{u}_0$ for $\boldsymbol{u}_0 \in \mathbb{R}^D$, and where $\frac{\mathrm{d}}{\mathrm{d}t}$ represents a temporal first derivative. Note that here $\boldsymbol{F}(\boldsymbol{u}, t)$ can include general linear operators acting on $\boldsymbol{u}$ which physically represent either spatial operators, or discretisations of spatial operators. Typically, for an ODE to preserve geometric properties of the continuous level we require

it to be *autonomous*, that is to say that it does not explicitly depend on time, i.e., we seek $\boldsymbol{u}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{u}). \tag{2.2}$$

To mark this difference we have changed notation for the right hand side of (2.2) adding additional clarity to the type of operator we are considering. Within this chapter we restrict ourselves to the study of autonomous problems of the form (2.2).
Of course, there is no general structure we expect to preserve for a general autonomous ODE of the form (2.2). For example, when $D = 1$, we could choose $\boldsymbol{f}(u) = u^2$ and have u blow up in finite time, or $\boldsymbol{f}(u) = -u$ and have u decay exponentially. As such, we introduce the notion of a Hamiltonian system.

Definition 2.1.1 (A Hamiltonian system)**.** *Assume that D is* even*, then let $\boldsymbol{u} \in (C^1([0,T]))^D$, $\mathcal{H} : \mathbb{R}^D \to \mathbb{R}$ with $\mathcal{H}(\boldsymbol{u}) \in C^1([0,T])$, and $J \in \mathbb{R}^{D\times D}$ be a* constant *skew-symmetric matrix. A Hamiltonian system is given by seeking $\boldsymbol{u}$ such that*

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = J\nabla\mathcal{H}(\boldsymbol{u}), \tag{2.3}$$

subject to the initial condition $\boldsymbol{u}(0) = \boldsymbol{u_0}$. Throughout we assume that J is invertible, which is expected for physical Hamiltonian systems.

In the sequel whenever discussing structure preservation of ODEs we are implicitly restricting our study to Hamiltonian systems. For particular examples of Hamiltonian systems see §3.3.0.

Remark 2.1.2 (Skew-symmetric inner product)**.** *By definition, a skew-symmetric matrix $J \in \mathbb{R}^{D\times D}$ satisfies*

$$J^T = -J.$$

This definition implies that for a given vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ we have that

$$\boldsymbol{v}\cdot J\boldsymbol{w} = -\boldsymbol{w}\cdot J\boldsymbol{v},$$

as $\boldsymbol{v}\cdot J\boldsymbol{w} = J^T\boldsymbol{v}\cdot\boldsymbol{w}$. This tells us that the skew-symmetric matrix J induces a skew-symmetric inner product. In fact, this inner product defines a Poisson bracket, see [85], similarly to that we discuss in §4.1 and §8.1.1.

Here we shall focus on two key geometric properties possessed by Hamiltonian systems.

The first of which is the conservation of the Hamiltonian functional $\mathcal{H}(\boldsymbol{u})$, which physically represents energy, over time.

Theorem 2.1.3 (Conservation of the Hamiltonian over time)**.** *The Hamiltonian function (or total energy) $\mathcal{H}(\boldsymbol{u})$, corresponding to the system* (2.3)*, is conserved, i.e.,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(\boldsymbol{u}) = 0.$$

Proof. In view of the chain rule we see that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(\boldsymbol{u}) &= \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\cdot\nabla\mathcal{H}(\boldsymbol{u}) \\ &= J\nabla\mathcal{H}(\boldsymbol{u})\cdot\nabla\mathcal{H}(\boldsymbol{u}),\end{aligned}$$

after the application of (2.3). Through the skew-symmetric inner product induced by J as discussed in Remark 2.1.2 we observe that the Hamiltonian function is preserved over time.

□

Hamiltonian systems also conserve a symplectic structure, which is characterised by the behaviour of the solution vector $\boldsymbol{u}$ over time. When the vector dimension $D = 2$ the symplectic structure corresponds to area conservation in the solution space (u_1, u_2). For a detailed introduction to the notion of symplectic mappings for Hamiltonian systems where $D = 2$ see [161, §2]. In arbitrary dimension the notion of symplecticity encompasses volume conservation in the solution space, but it is a stronger condition, see [127, §3.5].
There are two main (equivalent) ways of viewing the notion of symplecticity. The first of which is through the flow map of the solution, see for example [94]. The second, which we shall present here, is through differential calculus.

Definition 2.1.4 (Exterior calculus notation and fundamental properties, [127])**.** *Let $\mathbf{d}\boldsymbol{u}$ denote the Cartan exterior derivative of $\boldsymbol{u}$, and additionally let $\wedge$ denote the skew-symmetric wedge product, which is also known as the exterior product. We shall not provide concise definitions of these operators here as they require the introduction of multiple additional concepts, and instead refer the reader to [170, 72] for a gentle introduction. Throughout we shall define the exterior derivative and wedge product through the following identities:*
Let $\mathbf{d}\boldsymbol{a}$, $\mathbf{d}\boldsymbol{b}$ and $\mathbf{d}\boldsymbol{c}$ be Cartan exterior derivatives over a subdomain of $\mathbb{R}^d$ and $\alpha, \beta \in \mathbb{R}$.

Then the wedge product, $\wedge$, respects a skew-symmetry

$$\mathbf{d}\boldsymbol{a} \wedge \mathbf{d}\boldsymbol{b} = -\,\mathbf{d}\boldsymbol{b} \wedge \mathbf{d}\boldsymbol{a}, \tag{2.4}$$

bilinearity

$$\mathbf{d}\boldsymbol{a} \wedge (\alpha\,\mathbf{d}\boldsymbol{b} + \beta\,\mathbf{d}\boldsymbol{c}) = \alpha\,\mathbf{d}\boldsymbol{a} \wedge \mathbf{d}\boldsymbol{b} + \beta\,\mathbf{d}\boldsymbol{a} \wedge \mathbf{d}\boldsymbol{c}, \tag{2.5}$$

and a rule of matrix multiplication

$$\mathbf{d}\boldsymbol{a} \wedge (A\,\mathbf{d}\boldsymbol{b}) = \left(A^T\,\mathbf{d}\boldsymbol{a}\right) \wedge \mathbf{d}\boldsymbol{b}, \tag{2.6}$$

for $A \in \mathbb{R}^{d\times d}$.

Theorem 2.1.5 (Symplectic structure preservation over time)**.** *Let $\boldsymbol{u}$ denote the solution of the Hamiltonian system* (2.3)*, then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}\right) = 0, \tag{2.7}$$

i.e, the symplectic structure $\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}$ is preserved over time.

Proof of Theorem 2.1.5. Through application of the product rule, and bilinearity of the wedge product (2.5), we find

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}\right) &= \mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \wedge J^{-1}\,\mathbf{d}\boldsymbol{u} + \mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \\
&= \mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \wedge J^{-1}\,\mathbf{d}\boldsymbol{u} - J^{-1}\,\mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \wedge \mathbf{d}\boldsymbol{u} \\
&= 2\,\mathbf{d}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}\right) \wedge J^{-1}\,\mathbf{d}\boldsymbol{u},
\end{aligned}$$

through (2.4), (2.6) and the skew-symmetry of J^{-1}. Applying the ODE (2.3) we find that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}\right) &= 2J^{-1}\nabla\nabla^{\mathrm{T}}\mathcal{H}\left(\boldsymbol{u}\right)\,\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u} \\
&= -2J^{-1}J^{-1}\nabla\nabla^{\mathrm{T}}\mathcal{H}\left(\boldsymbol{u}\right)\,\mathbf{d}\boldsymbol{u} \wedge \mathbf{d}\boldsymbol{u},
\end{aligned}$$

where $\nabla\nabla^{\mathrm{T}}\mathcal{H}\left(\boldsymbol{u}\right)$ is the *Hessian* of $\mathcal{H}\left(\boldsymbol{u}\right)$, through (2.6) and the skew-symmetry of J^{-1}. Note that the square of a skew-symmetric matrix is symmetric. Additionally the Hessian $\nabla\nabla^{\mathrm{T}}\mathcal{H}\left(\boldsymbol{u}\right)$ is symmetric as the partial derivatives of $\mathcal{H}\left(\boldsymbol{u}\right)$ with respect to u_i are smooth

for $i = 1, ..., D$, through Schwarz's theorem. So we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{d}\boldsymbol{u} \wedge J^{-1}\,\mathbf{d}\boldsymbol{u}\right) = -2A\,\mathbf{d}\boldsymbol{u} \wedge \mathbf{d}\boldsymbol{u},$$

where $A = J^{-1}J^{-1}\nabla\nabla^{\mathrm{T}}\mathcal{H}(\boldsymbol{u})$ is symmetric. Through application of (2.4) and (2.6) we observe that

$$A\,\mathbf{d}\boldsymbol{u} \wedge \mathbf{d}\boldsymbol{u} = -A\,\mathbf{d}\boldsymbol{u} \wedge \mathbf{d}\boldsymbol{u},$$

allowing us to conclude the proof.

□

2.2 Discrete structure preservation

Throughout this section we shall recall numerical algorithms from the literature and investigate how well they preserve discrete equivalents of the structure we investigated on the continuous level.

We partition our temporal interval $[0, T]$ such that $0 := t_0 < t_1 < \cdots < t_N =: T$ where we define the length of each subinterval as $\tau_n := t_{n+1} - t_n$ for $n = 0, ..., N-1$. Here all numerical approximations we consider consist solely of point evaluations at the nodes t_n. We shall denote the point value approximating $\boldsymbol{u}(t_n)$ as $\boldsymbol{u}^n$ throughout, and shall always fix the initial value of our numerical approximation as $\boldsymbol{u}^0 = \boldsymbol{u}(0)$.

As our discrete approximations consist only of a collection of point values we are required to introduce discrete notions of conservation of the Hamiltonian and symplectic structure over time.

Recall in the continuous setting that conservation of the Hamiltonian is written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{H}(\boldsymbol{u})\right) = 0. \tag{2.8}$$

While we no longer have the notation of a time derivative we can instead define conservation of the Hamiltonian through a difference quotient, i.e.,

$$\frac{\mathcal{H}(\boldsymbol{u}^{n+1}) - \mathcal{H}(\boldsymbol{u}^n)}{\tau_n} = 0. \tag{2.9}$$

Note that through integrating the left hand side of (2.8) we observe that

$$\int_{t_n}^{t_{n+1}} \frac{\mathrm{d}}{\mathrm{d}t} \left(\mathcal{H} \left(\boldsymbol{u} \right) \right) \mathrm{d}t = \frac{\mathcal{H} \left(\boldsymbol{u} \left(t_{n+1} \right) \right) - \mathcal{H} \left(\boldsymbol{u} \left(t_n \right) \right)}{\tau_n},$$

through the fundamental theorem of calculus, so (2.9) holds for continuous problems but is weaker.

We define the notion of preservation of the symplectic mapping in the same way as energy conservation, i.e., through difference quotients. We formally define the notion of preservation of the symplectic structure locally as

$$\frac{\mathbf{d}\boldsymbol{u}^{n+1} \wedge J^{-1} \mathbf{d}\boldsymbol{u}^{n+1} - \mathbf{d}\boldsymbol{u}^{n} \wedge J^{-1} \mathbf{d}\boldsymbol{u}^{n}}{\tau_n} = 0, \tag{2.10}$$

for more details on (2.10) see [162] or [127]. Similarly to the discrete notion of conservation of the Hamiltonian, the discrete notion of symplecticity holds for continuous problems but is a weaker notion.

Theorem 2.2.1 (The incompatibility between conserving the Hamiltonian and preserving the symplectic structure in a numerical method, [185])**.** *Let* $\boldsymbol{u}^n$ *be a discrete approximation to* (2.3) *for* $n = 0, ..., N$*, and assume that the Hamiltonian function* $\mathcal{H}$ *is higher than quadratic in order. If* $\boldsymbol{u}^n$ *conserves the Hamiltonian in the sense of* (2.9) *and preserves the symplectic structure* (2.10) *then* $\boldsymbol{u}^n \equiv \boldsymbol{u} \left(t_n \right)$ *for* $n = 0, ..., N$*. That is to say to preserve both the Hamiltonian and the symplectic mapping our numerical approximation* must *be exact.*

Remark 2.2.2 (Choosing which structure to conserve)**.** *As we found in Theorem 2.2.1, in general it is not possible to conserve both the Hamiltonian and the symplectic structure numerically. There has been much debate as to which invariant is more favourable to conserve, see for example [167]. While no conclusions can be drawn on this point which satisfy all in the field of geometric numerical integration, the preservation of the symplectic mapping appears to be favoured in the literature. This may be, in part, due to the following theorem.*

Theorem 2.2.3 (Deviation in the Hamiltonian for symplectic methods over long time, [28])**.** *Let* $\boldsymbol{u}^n$ *for* $n = 0, ..., N$ *be a symplectic numerical method, i.e., it satisfies* (2.10)*, then,*

$$\mathcal{H} \left(\boldsymbol{u}^N \right) = \mathcal{H} \left(\boldsymbol{u}^0 \right) + \mathcal{O} \left(\tau_{max}^2 \right)$$

over an exponentially long time T, where $\tau_{max} := \max_n \tau_n$.

Remark 2.2.4 (Symmetry in Hamiltonian ODEs)**.** *In addition to conservation of the Hamiltonian and the symplectic structure many Hamiltonian ODEs are also symmetric, or time reversible. The symmetry of Hamiltonian ODEs is observed through the* flow map *of the problem, where the flow map describes how the solution changes over time. The notion of symmetry arises from the flow map being* time reversible*, that is to say that assuming the vector $\boldsymbol{u}$ flows forwards in time, if time is then reversed the solution will flow backwards in time along the same trajectory with the same velocity. We shall aim to design numerical methods which are symmetric in time.*
Let

$$\boldsymbol{u}^{n+1} = \Psi_{\tau_n}(\boldsymbol{u}^n) \tag{2.11}$$

define a one step numerical scheme, where Ψ_{τ_n} represents the discrete flow of the numerical approximation from $\boldsymbol{u}^n \to \boldsymbol{u}^{n+1}$, then the method is symmetric if

$$\Psi_{\tau_n}(\boldsymbol{u}^n) = \Psi_{-\tau_n}(\boldsymbol{u}^n)^{-1}. \tag{2.12}$$

We note that while the solution is vectorial the action of the flow map does not change the dimension of the solution. In the discrete setting for a one step method the inverse of the flow map can formally be found by interchanging $\boldsymbol{u}^{n+1}$ and $\boldsymbol{u}^n$ in (2.11)*. A method being symmetric is equivalent to saying that it is self-adjoined. Physically, a method being symmetric will give rise to symmetries of the solution in phase space when visualised over all time, which we expect to observe physically.*

We shall now discuss particular classes of numerical methods and investigate their structure preserving properties.

2.2.1 Runge-Kutta methods

Runge-Kutta methods are an incredibly popular class of temporal discretisation, for both ODEs and as temporal discretisations of PDEs, see [111, 47, 160]. Their popularity is not only due to the ease of their implementation, or the concise formulation of the methods developed by Butcher, see [37]. They have been analysed extensively, and it is possible to construct stable methods of "arbitrary" order, see [39]. For a detailed history of Runge-Kutta methods see [38].

Definition 2.2.5 (Runge-Kutta method)**.** *Let a_{ij} and b_i for $i, j = 1, ..., s$ be real numbers. An s-stage Runge-Kutta method is given by*

$$\begin{aligned} \boldsymbol{u}^{n+1} =& \boldsymbol{u}^n + \tau_n \sum_{i=1}^{s} b_i \boldsymbol{p}^i \\ \boldsymbol{p}^i =& \boldsymbol{f}\left(\boldsymbol{u}^n + \tau_n \sum_{j=1}^{s} a_{ij} \boldsymbol{p}^j\right) \qquad \text{for } i = 1, ..., s. \end{aligned} \tag{2.13}$$

We shall now investigate the geometric properties it is possible to preserve by Runge-Kutta methods.

Theorem 2.2.6 (Condition for symplecticity in Runge-Kutta methods, [160])**.** *Let $\boldsymbol{u}$ describe a Hamiltonian ODE of the form* (2.3)*, i.e., $\boldsymbol{f}(\boldsymbol{u}) = J\nabla\mathcal{H}(\boldsymbol{u})$. Further let a_{ij} and b_i be the coefficients of the Runge-Kutta method* (2.13) *for $i, j = 1, ..., s$, then if*

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0,$$

then the method is symplectic, i.e.,

$$\mathbf{d}\boldsymbol{u}^{n+1} \wedge J^{-1}\, \mathbf{d}\boldsymbol{u}^{n+1} = \mathbf{d}\boldsymbol{u}^n \wedge J^{-1}\, \mathbf{d}\boldsymbol{u}^n.$$

Proof. Taking the differentials of (2.13) and substituting the differential of the numerical scheme in to the left hand side of (2.7) and utilising (2.4), (2.5) and (2.6) we obtain the desired result. The details of this proof can be found in [160].

□

Remark 2.2.7 (Hamiltonian conservation by Runge-Kutta methods)**.** *It was found in [40] that Runge-Kutta methods do not, in general, conserve Hamiltonians of a higher degree than quadratic. However, it is possible to construct a Runge-Kutta method which preserves a particular polynomial Hamiltonian function, see [42]. We shall not press this point here.*

Theorem 2.2.8 (Symmetry condition for Runge-Kutta methods, [179])**.** *Let a_{ij} and b_i for $i, j = 1, ..., s$ determine the Runge-Kutta method* (2.13)*, then the method is symmetric if*

$$a_{s+1-i,s+1-j} + a_{ij} = b_j \qquad \forall i, j = 1, ..., s.$$

Example 2.2.9 (Gauss-Legendre methods: A symplectic family of Runge-Kutta methods)**.** *The Gauss-Legendre family of Runge-Kutta methods are obtained by seeking the*

constants c_i for $i = 1, ..., s$ such that c_i are given by the zeroes of

$$\frac{\mathrm{d}^s}{\mathrm{d}x^s}\left(x^s\left(x-1\right)^s\right).$$

Then the coefficients a_{ij} and b_i for $i, j = 1, ...s$ are determined by

$$a_{ij} = \int_0^{c_i} \mathcal{L}_j(t)\mathrm{d}t, \qquad b_i = \int_0^1 \mathcal{L}_i(t)\mathrm{d}t,$$

where $\mathcal{L}_i(t)$ is the Lagrange polynomial $\mathcal{L}_i(t) = \prod_{j\neq i} \frac{t-c_j}{c_i-c_j}$
Through fixing s we can observe through direct calculation that the condition for a method to be symplectic given in Theorem 2.2.6 is satisfied. For a proof of this for general s see [160].

2.2.2 Collocation methods

Collocation methods are a popular class of methods which can trace their origin back to [86], see [85, §II.1.2]. However, since the popularisation of the work of [84] they are not typically studied independently as they fall within the framework of the Runge-Kutta method. While this may appear to make the independent introduction of this class of methods superfluous we present them here as we utilise them in Chapter 3.

Definition 2.2.10 (Collocation method)**.** *Let c_i for $i = 1, ..., s$ be distinct real numbers, with $c_0 = 0$ and $c_s = 1$. The collocation polynomial $\boldsymbol{w}(t)$ is a degree s polynomial satisfying*

$$\boldsymbol{w}(t_n) = \boldsymbol{u}^n$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{w}\left(t_n + c_i\tau_n\right) = \boldsymbol{f}\left(\boldsymbol{w}\left(t_n + c_i\tau_n\right)\right).$$

The corresponding collocation method is given by

$$\boldsymbol{u}^{n+1} = \boldsymbol{w}\left(t_n + \tau_n\right). \tag{2.14}$$

Note that these methods do not solely consist of point values. In fact, in view of $\boldsymbol{w}$ we obtain a continuous piecewise polynomial solution, similarly to a finite element function.

Theorem 2.2.11 (Collocation methods as Runge-Kutta methods, [181])**.** *The collocation method* (2.14) *defined by the real constants $c_1, ..., c_s$ is equivalent to the s stage Runge-Kutta*

method (2.13) *with coefficients*

$$a_{ij} = \int_0^{c_i} \mathcal{L}_j(t)\mathrm{d}t, \qquad b_i = \int_0^1 \mathcal{L}_i(t)\mathrm{d}t,$$

where $\mathcal{L}_i(t) = \prod_{j\neq i} \frac{t-c_j}{c_i-c_j}$ *is the Lagrange polynomial.*

Remark 2.2.12 (Hamiltonian conservation of collocation methods)**.** *As collocation methods can be viewed as a subset of Runge-Kutta methods through Theorem 2.2.11 they* cannot *conserve an arbitrary Hamiltonian function over time, see Remark 2.2.7.*

Theorem 2.2.13 (Criteria for symmetric collocation methods)**.** *Let* c_i *for* $i = 1, ..., s$ *describe the collocation method* (2.14)*, then if* $c_i = 1 - c_{s+1-i}$ *then the method is* symmetric*.*

Proof. As no original reference could be found for this proof we shall present it here. Recall from Remark 2.2.4 that a numerical method is symmetric if it is invariant under reversal of time and inversion of the discrete flow map, see (2.12). Inverting the discrete flow map we can rewrite the collocation method in Definition 2.2.10 as

$$\begin{aligned} \boldsymbol{w}(t_{n+1}) &= \boldsymbol{u}^{n+1} \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{w}\left(t_{n+1} + c_i\tau_n\right) &= \boldsymbol{f}\left(\boldsymbol{w}(t_{n+1} + c_i\tau_n)\right) \\ \boldsymbol{u}^n &= \boldsymbol{w}\left(t_{n+1} + \tau_n\right). \end{aligned}$$

Through reversing the flow of time, i.e., mapping $\tau_n \to -\tau_n$, we observe that the collocation method is preserved if and only if $c_i = 1 - c_{s+1-i}$.

□

2.2.3 Discrete gradient methods

Discrete gradient methods, as their name suggests, are numerical methods designed such that they preserve a discrete Hamiltonian, see [137, 139]. Such methods cannot be presented in one class as concisely as Runge-Kutta and collocation methods. We can define a discrete gradient method for a Hamiltonian ODE as follows.

Definition 2.2.14 (Discrete gradient method)**.** *Let* $\boldsymbol{u}$ *be the solution to the Hamiltonian problem* (2.3)*, then a discrete gradient method is of the form*

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \tau_n J\widetilde{\nabla}\boldsymbol{K}, \tag{2.15}$$

where $\widetilde{\nabla}\boldsymbol{K}$ *denotes a discrete gradient, approximating the continuous function* $\nabla\mathcal{H}(\boldsymbol{u})$*, which satisfies the identity*

$$\widetilde{\nabla}\boldsymbol{K}\cdot\left(\boldsymbol{u}^{n+1}-\boldsymbol{u}^{n}\right)=\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)-\mathcal{H}\left(\boldsymbol{u}^{n}\right). \tag{2.16}$$

Theorem 2.2.15 (Hamiltonian conservation of discrete gradient methods)**.** *Let* $\boldsymbol{u}^n$ *for* $n=0,...,N$ *be described by the discrete gradient method* (2.15) *for a Hamiltonian problem* (2.3)*, then the Hamiltonian function is conserved nodally, i.e.,*

$$\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)=\mathcal{H}\left(\boldsymbol{u}^{n}\right).$$

Proof. Through application of (2.16) and (2.15) we see that

$$\begin{aligned}\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)-\mathcal{H}\left(\boldsymbol{u}^{n}\right)&=\widetilde{\nabla}\boldsymbol{K}\cdot\left(\boldsymbol{u}^{n+1}-\boldsymbol{u}^{n}\right)\\&=\tau_n\widetilde{\nabla}\boldsymbol{K}\cdot J\widetilde{\nabla}\boldsymbol{K}\\&=0,\end{aligned}$$

as J is skew-symmetric.

□

Remark 2.2.16 (Preservation of a symplectic structure)**.** *As discrete gradient methods are energy conserving in nature they cannot preserve a discrete symplectic structure by Theorem 2.2.1. For numerical methods which preserve a discrete symplectic map we know that numerical solution remains close to conserving energy, see Theorem 2.2.3. Unfortunately, there are no analytic results stating that the converse is true in the literature.*

There are three main families of discrete gradient methods which are introduced in [81], [89] and [103]. Here we introduce a member of the family discussed in [81] as an illustrative example.

Example 2.2.17 (An example of a discrete gradient method)**.** *Let* $\boldsymbol{u}=(u_1,u_2,...,u_D)$ *describe a Hamiltonian problem of the form* (2.3)*, then a discrete gradient method is given by*

$$u_i^{n+1}=u_i^n+\tau_n J\frac{\nabla\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)_i-\nabla\mathcal{H}\left(\boldsymbol{u}^{n}\right)_i}{u_i^{n+1}-u_i^n}, \tag{2.17}$$

for $i=1,...,D$*. Clearly, this discrete gradient satisfies* (2.16)*. Additionally,* (2.17) *is symmetric, as interchanging* $\boldsymbol{u}^{n+1}$ *and* $\boldsymbol{u}^n$ *then mapping* $\tau_n\to-\tau_n$ *returns the original method.*

2.2.4 Composition methods

Composition methods are *not* a class of method in their own right. A composition method refers to any numerical method in which the flow of the numerical solution over a single step is composed of multiple numerical schemes. Such methods have very useful properties in geometric integration, as well as improved stability, see [101]. It is possible to compose methods which do not preserve geometric properties as a methodology of constructing geometric numerical integrators. Here we focus on composition methods which already possess geometric properties in view of the following remark.

Remark 2.2.18 (Composition of geometric integrators)**.** *Let Ψ_{τ_n} describe the numerical flow of $\boldsymbol{u}^n$ and $\gamma_1, \gamma_2 > 0$. If the numerical flow either conserves a symplectic mapping or the Hamiltonian, then the composition of this flow $\Psi_{\gamma_1\tau_n} \circ \Psi_{\gamma_2\tau_n}$ must also conserve the same symplectic mapping or Hamiltonian due to the local nature of the discrete geometric structure.*

It is possible, through composition, to *increase* the order of a numerical method as has been developed in [184, 136].

Theorem 2.2.19 (Minimum order of a composition method, [85, §II.4])**.** *Consider the composition method*

$$\boldsymbol{u}^{n+1} = \Psi_{\gamma_s\tau_n} \circ \cdots \circ \Psi_{\gamma_1\tau_n} \circ \boldsymbol{u}^n, \tag{2.18}$$

where $\Psi_{\tau_n}(\boldsymbol{u}^n)$ describes the numerical flow of an order p method and $2 < s \in \mathbb{Z}$. If

$$\begin{aligned} \gamma_1 + \cdots + \gamma_s &= 1 \\ \gamma_1^{p+1} + \ldots \gamma_s^{p+1} &= 0, \end{aligned} \tag{2.19}$$

then the composition method (2.18) *is at least order p+1. Further, if $\Psi_{\tau_n}(\boldsymbol{u}^n)$ is symmetric, and*

$$\gamma_{s+1-i} = \gamma_i, \tag{2.20}$$

for $i = 1, ..., s$ then the composition (2.18) *is at least order $p + 2$.*

Remark 2.2.20 (Construction of arbitrarily high degree geometric integrators)**.** *In view of Theorem 2.2.19 and Remark 2.2.18 we can construct arbitrarily high order geometric numerical integrators. As well as preserving either a symplectic mapping or the Hamiltonian function it is also important to preserve symmetry if the method we compose with is*

itself symmetric, to preserve the symmetry we need to satisfy (2.20)*. The minimum number of times we need to compose a method with itself to preserve symmetry and increase the order of accuracy is* 3*, as we have three constraints in* (2.19) *and* (2.20)*, so we can choose* $s = 3$*. The parameters* γ_i *as discussed in Theorem 2.2.19 are then given uniquely by*

$$\gamma_1 = \gamma_3 = \frac{1}{3}2^{\frac{1}{3}} + \frac{1}{6}2^{\frac{2}{3}} + \frac{2}{3}, \qquad \gamma_2 = -\frac{2}{3}2^{\frac{1}{3}} - \frac{1}{3}2^{\frac{2}{3}} - \frac{1}{3}.$$

While we shall not investigate composition methods in the sequel they provide us with a powerful tool for generalising the temporal methods we develop from Chapter 5 onwards. We shall investigate multiple nonlinear PDEs and design fully discrete numerical schemes which conserve physical invariants of the problems. In space our numerical methods can be of arbitrarily high order, however the coupled temporal schemes shall always be $\mathcal{O}(\tau_{max}^2)$. The methodology outlined here allows us to extend the temporal discretisation to be arbitrarily high in order, although this comes at significant computational cost.

2.3 Conclusion

We gave a brief overview of some of the literature in the area of geometric numerical integration. We introduced the concepts of preservation of symplectic mapping and conservation of energy for Hamiltonian ODEs, and outlined selected numerical schemes which preserve these properties. We also outlined a methodology for the construction of higher order geometric numerical integrators from lower order integrators.

Chapter 3

Finite element methods for ODEs

Here we discuss finite element discretisations for ODEs, both recalling methods from the literature and developing a new finite element method. In the spirit of the previous chapter we shall investigate geometric properties of these methods, similarly to the investigation conducted in [105].

We also investigate the stability and convergence properties here, primarily to gain insight into the properties of our newly developed method.

In this chapter we focus on two main types of ODE, both encompassed by (2.1). The first are *explicit nonlinear functions* of $\boldsymbol{u}$, i.e., $\boldsymbol{F}(\boldsymbol{u},t) = \boldsymbol{f}(\boldsymbol{u})$. This is a very large class of problems which include Hamiltonian systems, i.e., $\boldsymbol{f}(\boldsymbol{u}) = J\nabla\mathcal{H}(\boldsymbol{u})$ where J is a constant skew-symmetric matrix, and $\mathcal{H}(\boldsymbol{u})$ is a scalar function which physically corresponds to the energy of the system, see Definition 2.1.1. These problems have interesting physical properties, such as conservation of the Hamiltonian function, and preservation of the symplectic structure of the flow map, but due to their skew-symmetric structure they are difficult to analyse. The second class of problems we consider are *linear ODEs with forcing*, i.e., $\boldsymbol{F}(\boldsymbol{u},t) = \mathfrak{f}(t) - \mathfrak{A}\boldsymbol{u}$, where $\mathfrak{f}(t)$ represents the forcing term and $\mathfrak{A}$ is the sum of a symmetric linear operator A and a skew-symmetric linear operator B, i.e., $\mathfrak{A} = c_1A + c_2B$ with c_1 and c_2 known constants. This class of problems is mainly utilised to describe the temporal aspect of evolution PDEs with a combination of symmetric and skew-symmetric operators in space. Here $\mathfrak{A}$ can either represent a continuous spatial operator or an appropriate discretisation of a spatial operator. Note that when required we will divide this case into two sub-cases, when the problem is purely *symmetric*, i.e., when $c_2 = 0$, and when the problem is purely *skew-symmetric*, i.e., when $c_1 = 0$. The symmetric case treats equations such as the heat equation, and the skew-symmetric case

can treat equations such as the wave equation, linear KdV, or even a linear Hamiltonian ODE such as the harmonic oscillator.

3.1 Known methods and their geometric properties

The ideas in this section have been presented in the masters thesis [105], we recall them here for completeness in addition to them informing our study of temporal finite element methods later in this chapter. All results from [105] are referenced appropriately.
Before introducing temporal finite element methods we must first introduce an abundance of notation. Recall from §2.2 that we partitioned our temporal interval $[0,T]$ such that $0 := t_0 < t_1 < \cdots < t_N =: T$. We define a *temporal finite element* as $I_n := (t_n, t_{n+1})$ which possesses an element length $\tau_n := t_{n+1} - t_n$. We shall often write $\max_n \tau_n = \tau_{max}$. Throughout this section we shall write the discrete counterpart to a continuous function through capitalisation, i.e., for a temporal continuous function $\boldsymbol{u} = \boldsymbol{u}(t)$ we write the finite element function approximating it as $\boldsymbol{U} = \boldsymbol{U}(t)$. Note that when there is no ambiguity we shall not explicitly write the dependency of functions. In the sequel, when considering spatial finite element methods, we shall also use capitalisation to refer to *spatial* finite element functions.

Definition 3.1.1 (Temporal finite element spaces)**.** *Let $\mathbb{P}_q(I_n)$ denote the space of polynomials of degree q on an interval $I_n \subset \mathbb{R}$, then the discontinuous finite element space is*

$$\mathbb{V}_q([0,T]) = \{\boldsymbol{W} : \boldsymbol{W}|_{I_n} \in (\mathbb{P}_q(I_n))^D, n = 0, ..., N-1\},$$

further to this the continuous finite element space is defined analogously with global continuity enforced, i.e.,

$$\mathbb{V}_q^C([0,T]) = \mathbb{V}_q([0,T]) \cap \left(\mathcal{C}^0([0,T])\right)^D.$$

When there is no ambiguity we shall not explicitly write the domain of our finite element spaces. Additionally, in the sequel $\mathbb{V}_q^C(I_n)$ represents the localisation of the continuous finite element space to a single element. Here, the initial conditions of functions in this space are fixed by the endpoint of functions in $\mathbb{V}_q^C(I_{n-1})$, or an appropriate initial condition.

With the definition of the temporal finite element space in mind, we define the mesh function $\tau \in \mathbb{V}_0$ as the piecewise constant finite element function representing the length

of an element, i.e.,

$$\tau|_{I_n} = \tau_n.$$

Remark 3.1.2 (A 1D finite element de Rham complex between spaces,[13])**.** *A de Rham complex, in the general setting, refers to a complex of differential forms acting on a smooth manifold which map all functions in a given space to all functions in a subsequent space through an appropriate exterior derivative. In fact, as outlined in [13], there are discrete counterparts to these continuous mappings in one, two and three dimensional spaces. In the one dimensional case the de Rham mapping is comparatively straightforward. The de Rham complex maps functions in the continuous finite element space* $\mathbb{V}_{q+1}^C$ *to functions in the discontinuous finite element space of one degree lower* $\mathbb{V}_q$*. The exterior derivative which employs this mapping is the standard derivative operator* $\frac{\mathrm{d}}{\mathrm{d}t}$*.*

The first temporal finite element we shall investigate is the continuous, or conforming, Galerkin method. This method has been extensively analysed, see [67, 73].

Definition 3.1.3 (Continuous Galerkin method)**.** *Let* $\mathbb{V}_{q+1}^C$ *and* $\mathbb{V}_q$ *be the vectorial finite dimensional spaces given in Definition 3.1.1. The continuous Galerkin* (cG) *finite element approximation is given by seeking* $\boldsymbol{U} \in \mathbb{V}_{q+1}^C$ *such that*

$$\begin{aligned} \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{V}\mathrm{d}t = \int_0^T \boldsymbol{F}\left(\boldsymbol{U}, t\right) \cdot \boldsymbol{V}\mathrm{d}t \qquad \forall \boldsymbol{V} \in \mathbb{V}_q, \\ \boldsymbol{U}(0) = \boldsymbol{u}_0. \end{aligned} \tag{3.1}$$

Remark 3.1.4 (Difference between test and trial spaces for the cG method)**.** *For readers with a background in spatial finite element methods continuous Galerkin may initially seem like a misnomer as it utilises a discontinuous test space of one degree lower, so is a Petrov-Galerkin type method. In fact this is the natural continuous finite element formulation due to the lack of symmetry in derivatives, and the degrees of freedom on the left hand side of* (3.1) *match as* $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}, \boldsymbol{V} \in \mathbb{V}_q$*. This follows from the discrete de Rham complex discussed in Remark 3.1.2.*

Remark 3.1.5 (cG time stepping)**.** *The discontinuous nature of the test functions allow us to rewrite the cG method over a single element, with the initial value of the trial function* $\boldsymbol{U}$ *on the element being the final value of the trial function on the previous element due*

to continuity. Explicitly by choosing the test function

$$\boldsymbol{V}(t) = \begin{cases} \widetilde{\boldsymbol{V}}(t) & \text{for } t \in I_n \\ 0 & \text{else,} \end{cases}$$

where $\widetilde{\boldsymbol{V}}$ *is an arbitrary function in* $\left(\mathbb{P}_q\left(I_n\right)\right)^D$*, over an arbitrary interval, we can write the cG method locally as follows: Let* $\boldsymbol{U}(t_n)$ *be given, then seek* $\boldsymbol{U} \in \mathbb{V}_{q+1}^C\left(I_n\right)$ *such that*

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \widetilde{\boldsymbol{V}} \mathrm{d}t = \int_{I_n} \boldsymbol{F}\left(\boldsymbol{U}, t\right) \cdot \widetilde{\boldsymbol{V}} \mathrm{d}t \qquad \forall \widetilde{\boldsymbol{V}} \in \left(\mathbb{P}_q\left(I_n\right)\right)^D. \tag{3.2}$$

Theorem 3.1.6 (Preservation of the Hamiltonian function for the cG method, [105, 87])**.** *Let* $\boldsymbol{F}\left(\boldsymbol{U}, t\right) = J\nabla\mathcal{H}\left(\boldsymbol{U}\right)$ *where* J *is constant, i.e., we restrict ourselves to the case of a Hamiltonian system, as described in Definition 2.1.1, then the cG method described in Definition 3.1.3 preserves the Hamiltonian function at the nodes. That is to say that for* $n = 0, ..., N-1$

$$\mathcal{H}\left(\boldsymbol{U}(t_{n+1})\right) = \mathcal{H}\left(\boldsymbol{U}(t_n)\right).$$

Proof. Through the fundamental theorem of calculus we observe that

$$\mathcal{H}\left(\boldsymbol{U}(t_{n+1})\right) - \mathcal{H}\left(\boldsymbol{U}(t_n)\right) = \int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{H}\left(\boldsymbol{U}\right)\right) \mathrm{d}t = \int_{I_n} \nabla\mathcal{H}\left(\boldsymbol{U}\right) \cdot \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \mathrm{d}t.$$

Additionally, through choosing $\widetilde{\boldsymbol{V}} = \Pi_{\mathbb{V}_q}\left(\nabla\mathcal{H}\left(\boldsymbol{U}\right)\right)$, where $\Pi_{\mathbb{V}_q}$ is the L_2 projection into $\mathbb{V}_q$, in the local cG method (3.2) we find

$$\begin{aligned} \int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \nabla\mathcal{H}\left(\boldsymbol{U}\right) \mathrm{d}t &= \int_{I_n} J\nabla\mathcal{H}\left(\boldsymbol{U}\right) \cdot \Pi_{\mathbb{V}_q}\left(\nabla\mathcal{H}\left(\boldsymbol{U}\right)\right) \mathrm{d}t \\ &= \int_{I_n} J\Pi_{\mathbb{V}_q}\left(\nabla\mathcal{H}\left(\boldsymbol{U}\right)\right) \cdot \Pi_{\mathbb{V}_q}\left(\nabla\mathcal{H}\left(\boldsymbol{U}\right)\right) \mathrm{d}t \end{aligned}$$

through the definition of the L_2 projection, and as J is constant. In view of the skew symmetry of J we can write

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \nabla\mathcal{H}\left(\boldsymbol{U}\right) \mathrm{d}t = 0,$$

allowing us to conclude.

□

Remark 3.1.7 ("Exact" conservation of the Hamiltonian)**.** *While the cG method exactly*

preserves the Hamiltonian function this is not necessarily true for an implementation of the method. To implement the method we are required to make a quadrature approximation, and if this quadrature approximation is not *exact, then the Hamiltonian will not be exactly preserved. In practice, we will only be able to preserve the Hamiltonian exactly if our Hamiltonian function is polynomial and the quadrature method is chosen to be of high enough order. When dealing with non-polynomial Hamiltonian functions it is also possible to take an "overkill" approach, i.e., we can take a very high order quadrature approximation so the quadrature error is of the same order of magnitude as machine precision.*

Remark 3.1.8 (A comparison between the cG method and discrete gradient methods)**.** *As we discussed in §2.2.3, a discrete gradient method* (2.15) *satisfies the identity*

$$\widetilde{\nabla}\boldsymbol{K}\cdot\left(\boldsymbol{u}^{n+1}-\boldsymbol{u}^{n}\right)=\mathcal{H}\left(\boldsymbol{u}^{n+1}\right)-\mathcal{H}\left(\boldsymbol{u}^{n}\right),$$

for some discrete gradient $\widetilde{\nabla}\boldsymbol{K}$*. The cG method* (3.2)*, while not strictly falling within this framework, satisfies the similar identity in the continuous framework*

$$\int_{I_n}\nabla\mathcal{H}\left(\boldsymbol{U}\right)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\mathrm{d}t=\mathcal{H}\left(\boldsymbol{U}(t_{n+1})\right)-\mathcal{H}\left(\boldsymbol{U}(t_n)\right),$$

where $\boldsymbol{U}$ *is the cG solution, which is crucial in the proof of Theorem 3.1.6.*

Theorem 3.1.9 (cG as a Runge-Kutta method, [105])**.** *Let* $\boldsymbol{F}\left(\boldsymbol{U},t\right)=\boldsymbol{f}\left(\boldsymbol{U}\right)$ *be an explicit function of* $\boldsymbol{U}$*, then under a* $q+1$ *point quadrature choice the cG scheme,* (3.1)*, agrees with a Runge-Kutta method, see Definition 2.2.5, at the nodes. Furthermore this Runge-Kutta method is given by*

$$\begin{aligned}b_i&=\int_0^1\mathcal{L}_i(t)\mathrm{d}t\\a_{ij}&=\int_0^{c_i}\mathcal{L}_j(t)\mathrm{d}t\end{aligned}\tag{3.3}$$

for $i,j=1,...,q+1$*, where* $c_i=\sum_{j=1}^{q+1}a_{ij}$ *are our quadrature points and* $\mathcal{L}_i=\prod_{j\neq i}\frac{t-c_j}{c_i-c_j}$ *is a Lagrange polynomial.*

Proof. Under a $q+1$ point quadrature we can write the cG method, (3.1), as

$$\begin{aligned}\sum_{n=0}^{N-1}\sum_{iq=1}^{q+1}c_{iq}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}(t_{n,iq})\cdot\boldsymbol{V}(t_{n,iq})&=\sum_{n=0}^{N-1}\sum_{iq=1}^{q+1}c_{iq}\boldsymbol{f}\left(\boldsymbol{U}(t_{n,iq})\right)\cdot\boldsymbol{V}(t_{n,iq})\\\boldsymbol{U}(0)&=\boldsymbol{u}_0.\end{aligned}\tag{3.4}$$

As $\boldsymbol{V} \in \mathbb{V}_q$ we can choose

$$\boldsymbol{V} = \begin{cases} 1 & t = t_{n,iq} \\ 0 & \text{all other quadrature points,} \end{cases}$$

for an arbitrary quadrature point iq in an arbitrary interval I_n. Under this choice of test function we can write (3.4) as

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}(t_{n,iq}) = \boldsymbol{f}\left(\boldsymbol{U}(t_{n,iq})\right). \tag{3.5}$$

As (3.5) is for an arbitrary quadrature point on an arbitrary interval we can conclude that $\boldsymbol{U}$ is a collocation polynomial, see Definition 2.2.10. This implies that at the nodes our cG formulation agrees with a collocation method defined by the quadrature points. This collocation method is equivalent to the Runge-Kutta method (3.3) by Theorem 2.2.11.

□

Remark 3.1.10 (Symplectic implementation of cG, [105])**.** *Applying particular quadrature choices (which introduce leading order errors) we obtain a scheme equivalent on the nodes to families of symplectic methods. For example choosing the* $q+1$ *point Gauss quadrature, see [85, P34], the cG method* (3.1) *agrees with the Gauss Runge-Kutta family on the nodes. These methods are known to be symplectic, as they satisfy Theorem 2.2.6.*

We now shift our focus to a discontinuous Galerkin method often employed for temporal ODEs, which has also been extensively studied in the literature, see [113, 66, 68]. Before we can introduce this method we first must define additional notation.

Definition 3.1.11 (Discontinuous finite element notation)**.** *Due to the discontinuous nature of the finite element space finite element functions are permitted to be multi-valued at the nodes of the elements. With this in mind we write*

$$\boldsymbol{U}_n^+ := \boldsymbol{U}(t_n^+) := \lim_{t \searrow t_n} \boldsymbol{U}(t), \qquad \boldsymbol{U}_n^- := \boldsymbol{U}(t_n^-) := \lim_{t \nearrow t_n} \boldsymbol{U}(t),$$

to describe the values of the function on the right and left of a discontinuity respectively. Additionally we define the jump of a function at the node t_n *to be*

$$[\![\boldsymbol{U}_n]\!] = \boldsymbol{U}_n^- - \boldsymbol{U}_n^+ \tag{3.6}$$

and the average as

$$\{\boldsymbol{U}_n\} = \frac{1}{2}\left(\boldsymbol{U}_n^- + \boldsymbol{U}_n^+\right). \tag{3.7}$$

Definition 3.1.12 (Upwind discontinuous Galerkin method)**.** *Let $\mathbb{V}_q$ be the space of piecewise polynomial functions of degree q as described in Definition 3.1.1. The upwind discontinuous Galerkin* (dG) *approximation is given by seeking $\boldsymbol{U} \in \mathbb{V}_q$ such that*

$$\begin{aligned} \sum_{n=0}^{N-1} \int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{V} \mathrm{d}t = \sum_{n=0}^{N-1} \int_{I_n} \boldsymbol{f}\left(\boldsymbol{U}, \boldsymbol{V}\right) \mathrm{d}t + [\![\boldsymbol{U}_n]\!] \cdot \boldsymbol{V}_n^+ \qquad \forall \boldsymbol{V} \in \mathbb{V}_q \\ \boldsymbol{U}^-(0) = \boldsymbol{u}_0. \end{aligned} \tag{3.8}$$

Remark 3.1.13 (dG time stepping)**.** *Similarly to the cG method, the upwind dG method can be implemented in a time-stepping fashion utilising the discontinuous nature of the test function. To be concise, we can reduce the upwind scheme to a single element by choosing our test function such that*

$$\boldsymbol{V}(t) = \begin{cases} \widetilde{\boldsymbol{V}}(t) & \text{for } t \in I_n \\ 0 & \text{else,} \end{cases}$$

where $\widetilde{V}$ is an arbitrary function in $\left(\mathbb{P}_q\left(I_n\right)\right)^D$ allowing us to write

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \widetilde{\boldsymbol{V}} \mathrm{d}t = \int_{I_n} \boldsymbol{f}\left(\boldsymbol{U}, \widetilde{\boldsymbol{V}}\right) \mathrm{d}t + [\![\boldsymbol{U}_n]\!] \cdot \widetilde{\boldsymbol{V}}_n^+ \qquad \forall \widetilde{\boldsymbol{V}} \in \left(\mathbb{P}_q\left(I_n\right)\right)^D. \tag{3.9}$$

Note that we cannot typically write dG methods in a time stepping fashion. The choice of flux needs to be chosen carefully such that on an arbitrary element it only depends on the value of the solution on the previous element. Many standard flux choices for spatial finite element methods require information from both neighbouring elements and the methods need to be solved globally. These global methods are not practical temporal discretisations.

Remark 3.1.14 (Dissipation of the upwind dG method)**.** *In [68] the upwind dG method* (3.9) *is viewed as a* discrete dynamical system *for problems with a linear right hand side. It is found that the associated dynamical system is dissipative in the sense that the numerical solution over time can not increase. In addition, we will discover in the sequel, if the right hand side is chosen to describe a purely skew-symmetric problem then the upwind dG method is dissipative in the same sense that the solution over time does not increase, see Theorem 3.1.22.*

Due to the dissipative nature of (3.9) *it does* not *fall within the framework of geometric numerical integration when applied to Hamiltonian systems.*

Error bounds for both the cG and upwind dG methods have been fully developed in the literature for linear problems, these bounds can be described as follows.

Theorem 3.1.15 (A priori error bound for the cG method (3.1), [67])**.** *Let* $\boldsymbol{U}$ *be the solution of the cG method* (3.1)*, where* $\boldsymbol{u} \in C^{q+2}([0,T])$ *denotes the corresponding exact solution given in* (2.1)*. Under the assumption that* $\boldsymbol{F}(\boldsymbol{U},t)$ *is linear we have for* $q = 0, 1$ *that*

$$\sup_{[0,T]} |\boldsymbol{U} - \boldsymbol{u}| \leq C(T)\tau_{max}^{q+2} \sup_{[0,T]} \left| \frac{\mathrm{d}^{q+2}}{\mathrm{d}^{q+2}t} \boldsymbol{u} \right|,$$

where the constant $C(T)$ *depends exponentially on* T*.*

Theorem 3.1.16 (A priori error bound for the upwind dG method (3.8), [113])**.** *Let* $\boldsymbol{U}$ *be the solution of the upwind dG method* (3.8)*, where* $\boldsymbol{u} \in C^{q+1}([0,T])$ *represents the corresponding exact solution given in* (2.1)*. Under the assumption that* $\boldsymbol{F}(\boldsymbol{U},t)$ *is linear we have for* $q = 0, 1$ *that*

$$\sup_{[0,T]} |\boldsymbol{U} - \boldsymbol{u}| \leq C(T)\tau_{max}^{q+1} \sup_{[0,T]} \left| \frac{\mathrm{d}^{q+1}}{\mathrm{d}^{q+1}t} \boldsymbol{u} \right|,$$

where the constant $C(T)$ *depends exponentially on* T*.*

Note that while for the cG method (3.1) we seek a polynomial solution of degree $q+1$, for the upwind dG method (3.8) we seek a polynomial solution of degree q, ergo both schemes converge at the same rate.
While these error bounds are well established, we selectively develop equivalent bounds in the sequel as they inform the analysis of the new finite element method we propose in §3.2.

3.1.1 Stability

Here we investigate the stability properties of the temporal finite element methods presented in this section. Note that while similar results are often cited in the literature they are rarely presented in full. We consider the stability of symmetric linear ODEs, skew-symmetric linear ODEs, and Hamiltonian ODEs independently.

3.1.1.1 Symmetric linear ODEs

In this subsection we restrict ourselves to *symmetric* linear ODEs, that is we choose the right hand side of the ODE (2.1) such that $\boldsymbol{F}(\boldsymbol{u}, t) = \mathfrak{f} - A\boldsymbol{u}$, where $\mathfrak{f} = \mathfrak{f}(t)$ and A is a symmetric linear operator. The operator A represents a symmetric spatial operator, which can be either continuous in space or discrete. An example of an appropriate continuous spatial operator could be the second spatial derivative, and a discrete operator could be a symmetric finite element discretisation. Physically $\mathfrak{f}$ represents the forcing of the system. Our symmetric linear ODE can be written in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} + A\boldsymbol{u} = \mathfrak{f}.$$

We further assume that this operator is symmetric and consequently induces a norm, in addition to this at various points in the sequel we will assume that the operator A induces a bilinear form which is coercive, i.e., for $\boldsymbol{w} \in \mathbb{R}^D$

$$A\boldsymbol{w} \cdot \boldsymbol{w} \geq C_A \left|\boldsymbol{w}\right|_A^2 \tag{3.10}$$

and/or continuous, i.e., for $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^D$

$$\left|A\boldsymbol{w}_1 \cdot \boldsymbol{w}_2\right| \leq c_A \left|\boldsymbol{w}_1\right|_A \left|\boldsymbol{w}_2\right|_A, \tag{3.11}$$

where $\left|\cdot\right|_A$ represents some appropriate semi-norm depending on A. For example, A could represent, or approximate, the second order spatial Laplacian

$$A = -\triangle.$$

In this case the semi-norm $\left|\cdot\right|_A$ is in fact the spatial H^1 norm, see for example [33]. If A is a matrix then the coercivity and continuity conditions correspond to A being nondegenerate and bounded.

Theorem 3.1.17 (Stability of the cG method for symmetric problems)**.** *Consider the cG method* (3.1) *applied to a symmetric linear ODE, i.e.,* $\boldsymbol{F}(\boldsymbol{U}, t) = \mathfrak{f} - A\boldsymbol{U}$ *where* $\mathfrak{f}$ *is bounded and* A *is a symmetric operator. The solution to the cG method satisfies*

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|_{L_2([0,T])}^2 + A\boldsymbol{U}(T) \cdot \boldsymbol{U}(T) \leq A\boldsymbol{U}(0) \cdot \boldsymbol{U}(0) + \left\|\mathfrak{f}\right\|_{L_2([0,T])}^2.$$

Remark 3.1.18 (Stability in multiple norms)**.** *Under the assumption that A induces a coercive and continuous form, see* (3.10) *and* (3.11)*, we can obtain the following stability bound*

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|^2_{L_2([0,T])} \leq c_A \left|\boldsymbol{U}(0)\right|^2_A + \|\mathfrak{f}\|^2_{L_2([0,T])}.$$

Additionally we have stability in the norm induced by A, i.e.,

$$\left|\boldsymbol{U}(T)\right|^2_A \leq \left|\boldsymbol{U}(0)\right|^2_A + \|\mathfrak{f}\|^2_{L_2([0,T])}.$$

Proof of Theorem 3.1.17. For a symmetric linear problem we may write the cG method as follows: seek $\boldsymbol{U} \in \mathbb{V}^C_{q+1}$ such that

$$\begin{aligned}\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\cdot\boldsymbol{V}\mathrm{d}t + \int_0^T A\boldsymbol{U}\cdot\boldsymbol{V}\mathrm{d}t = \int_0^T \mathfrak{f}\cdot\boldsymbol{V}\mathrm{d}t \qquad \forall \boldsymbol{V}\in\mathbb{V}_q\\ \boldsymbol{U}(0) = \boldsymbol{u}_0.\end{aligned}$$

Choosing $\boldsymbol{V} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}$ we have

$$\int_0^T \left|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right|^2 + A\boldsymbol{U}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\mathrm{d}t = \int_0^T \mathfrak{f}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\mathrm{d}t,$$

which, after using the symmetry of A, can be written as

$$\begin{aligned}\int_0^T \left|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right|^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(A\boldsymbol{U}\cdot\boldsymbol{U}\right)\mathrm{d}t &= \int_0^T \mathfrak{f}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\mathrm{d}t\\ &\leq \|\mathfrak{f}\|_{L_2([0,T])}\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|_{L_2([0,T])}\\ &\leq \frac{1}{4\epsilon}\|\mathfrak{f}\|^2_{L_2([0,T])} + \epsilon\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|^2_{L_2([0,T])},\end{aligned}$$

through Hölder's inequality and Cauchy's inequality with ϵ. Choosing $\epsilon = \frac{1}{2}$ and applying the fundamental theorem of calculus

$$\left\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}\right\|^2_{L_2([0,T])} + A\boldsymbol{U}(T)\cdot\boldsymbol{U}(T) \leq A\boldsymbol{U}(0)\cdot\boldsymbol{U}(0) + \|\mathfrak{f}\|^2_{L_2([0,T])},$$

as required.

□

Theorem 3.1.19 (Stability of the upwind dG method for symmetric problems)**.** *Let $\boldsymbol{U}$ be the solution of the upwind dG method* (3.8)*. Further assume that $f(\boldsymbol{U}) = \mathfrak{f} - A\boldsymbol{U}$, where $\mathfrak{f} := \mathfrak{f}(t)$ is bounded and A is symmetric. Additionally assume that A is a coercive operator, i.e., for all $\boldsymbol{W} \in \mathbb{V}_q$*

$$\|\boldsymbol{W}\|^2_{L_2(I_n)} \leq C_A \int_{I_n} A\boldsymbol{W} \cdot \boldsymbol{W} \mathrm{d}t,$$

where we have assumed that $|\boldsymbol{W}| \leq C\,|\boldsymbol{W}|_A$, then

$$\left|\boldsymbol{U}^-_{n+1}\right|^2 + \int_0^T A\boldsymbol{U} \cdot \boldsymbol{U} \mathrm{d}t \leq |\boldsymbol{u}_0|^2 + \frac{C_A}{2} \|\mathfrak{f}\|^2_{L_2((0,T))},$$

where $\boldsymbol{u}_0$ denotes the initial data.

Before presenting the proof of Theorem 3.1.19 we first require the following result.

Lemma 3.1.20 (A useful dG inequality)**.** *Let $\boldsymbol{W} \in \mathbb{V}_q$, then*

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{W} \cdot \boldsymbol{W} \mathrm{d}t + [\![\boldsymbol{W}_n]\!] \cdot \boldsymbol{W}^+_n \geq \frac{1}{2}\left|\boldsymbol{W}^-_{n+1}\right|^2 - \frac{1}{2}\left|\boldsymbol{W}^-_n\right|^2.$$

Proof. Applying the definition of the jump and the fundamental theorem of calculus we find

$$\begin{aligned}\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{W} \cdot \boldsymbol{W} \mathrm{d}t + [\![\boldsymbol{W}_n]\!] \cdot \boldsymbol{W}^+_n &= \int_{I_n} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{W}|^2 \mathrm{d}t + \left(\boldsymbol{W}^+_n - \boldsymbol{W}^-_n\right) \cdot \boldsymbol{W}^+_n \\ &= \frac{1}{2}\left(\left|\boldsymbol{W}^-_{n+1}\right|^2 - \left|\boldsymbol{W}^+_n\right|^2\right) + \left|\boldsymbol{W}^+_n\right|^2 - \boldsymbol{W}^-_n \cdot \boldsymbol{W}^+_n.\end{aligned}$$

Through Cauchy's inequality we have that

$$-\left|\boldsymbol{W}^-_n \cdot \boldsymbol{W}^+_n\right| \geq -\frac{1}{2}\left|\boldsymbol{W}^-_n\right|^2 - \frac{1}{2}\left|\boldsymbol{W}^+_n\right|^2,$$

which tells us

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{W} \cdot \boldsymbol{W} \mathrm{d}t + [\![\boldsymbol{W}_n]\!] \cdot \boldsymbol{W}^+_n \geq \frac{1}{2}\left|\boldsymbol{W}^-_{n+1}\right|^2 - \frac{1}{2}\left|\boldsymbol{W}^-_n\right|^2,$$

as required.

□

Proof of Theorem 3.1.19. Choosing $\boldsymbol{V} = \boldsymbol{U}$ in the localised dG method (3.9) we have

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{U} + A\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t + [\![\boldsymbol{U}_n]\!] \cdot \boldsymbol{U}_n^+ = \int_{I_n} \mathfrak{f} \cdot \boldsymbol{U}\mathrm{d}t.$$

Applying Lemma 3.1.20 we observe that

$$\left|\boldsymbol{U}_{n+1}^-\right|^2 - \left|\boldsymbol{U}_n^-\right|^2 + \int_{I_n} A\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t \leq \int_{I_n} \mathfrak{f} \cdot \boldsymbol{U}\mathrm{d}t.$$

Through Hölder's and Cauchy's inequality with ϵ we see

$$\begin{aligned}\left|\boldsymbol{U}_{n+1}^-\right|^2 - \left|\boldsymbol{U}_n^-\right|^2 + \int_{I_n} A\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t \leq& \frac{1}{4\epsilon} \|\mathfrak{f}\|_{L_2(I_n)}^2 + \epsilon \|\boldsymbol{U}\|_{L_2(I_n)}^2 \\ \leq& \frac{1}{4\epsilon} \|\mathfrak{f}\|_{L_2(I_n)}^2 + \epsilon C_A \int_{I_n} A\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t,\end{aligned}$$

after applying coercivity of A. Choosing $\epsilon = \frac{1}{2C_A}$ we find

$$\left|\boldsymbol{U}_{n+1}^-\right|^2 - \left|\boldsymbol{U}_n^-\right|^2 + \frac{1}{2}\int_{I_n} A\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t \leq \frac{C_A}{2} \|\mathfrak{f}\|_{L_2(I_n)}^2.$$

As this bound holds for arbitrary n we can iterate back in time to $n = 0$ finding

$$\left|\boldsymbol{U}_{n+1}^-\right|^2 + \frac{1}{2}\int_0^T A\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t \leq \left|\boldsymbol{U}_0^-\right|^2 + \frac{C_A}{2} \|\mathfrak{f}\|_{L_2([0,T])}^2,$$

noting that $\boldsymbol{U}_0^-$ is, by definition, the initial condition we can conclude.

□

3.1.1.2 Skew-symmetric linear ODEs

Here we restrict the right hand side function of the general ODE (2.1) to $\boldsymbol{F}(\boldsymbol{U}, t) = -B\boldsymbol{U}$, where B is a linear skew-symmetric operator. Here the operator B can represent either: a continuous spatial operator, a discretised spatial operator, or a constant operator. Possible PDEs which can be represented when B is a spatial operator include the Airy equation and linear KdV. We have assumed that there is no forcing term, as forcing destroys a lot of the structure of these problems.

Theorem 3.1.21 (Stability of the cG method for skew-symmetric problems)**.** *Consider the cG method* (3.1) *applied to a skew-symmetric linear ODE, i.e.,* $\boldsymbol{F}(\boldsymbol{U}, t) = -B\boldsymbol{U}$ *where*

B is a skew-symmetric operator. The solution to the cG method satisfies

$$|\boldsymbol{U}(T)| = |\boldsymbol{U}(0)|.$$

Proof. The following argument follows directly from the "conservative" nature of the ODE, and the fact that the discretisation mimics this conservative structure. Choosing $\boldsymbol{V} = \Pi_{\mathbb{V}_q}(\boldsymbol{U})$, where $\Pi_{\mathbb{V}_q}$ represents the L_2 projection into $\mathbb{V}_q$, in (3.1) for skew-symmetric linear ODEs we find

$$\begin{aligned} 0 &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \Pi_{\mathbb{V}_q}(\boldsymbol{U}) + B\boldsymbol{U} \cdot \Pi_{\mathbb{V}_q}(\boldsymbol{U})\mathrm{d}t \\ &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{U} + B\Pi_{\mathbb{V}_q}(\boldsymbol{U}) \cdot \Pi_{\mathbb{V}_q}(\boldsymbol{U})\mathrm{d}t, \end{aligned}$$

through the definition of the L_2 projection into $\mathbb{V}_q$ as $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \in \mathbb{V}_q$ by Remark 3.1.2. Through skew-symmetry of B, i.e., as $B\Pi_{\mathbb{V}_q}(\boldsymbol{U}) \cdot \Pi_{\mathbb{V}_q}(\boldsymbol{U}) = 0$ we find that

$$\begin{aligned} 0 &= \int_0^T \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{U}|^2 \,\mathrm{d}t \\ &= |\boldsymbol{U}(T)|^2 - |\boldsymbol{U}(0)|^2, \end{aligned}$$

through the fundamental theorem of calculus as required.

□

Theorem 3.1.22 (Stability of the upwind dG method for skew-symmetric problems)**.** *Consider the upwind dG method* (3.8) *applied to the skew-symmetric linear ODE, i.e., $\boldsymbol{F}(\boldsymbol{U}, t) = -B\boldsymbol{U}$ where B is a skew-symmetric operator. The solution to the upwind dG method satisfies*

$$\left|\boldsymbol{U}_{n+1}^-\right|^2 \leq \left|\boldsymbol{U}_n^-\right|^2,$$

i.e., the numerical solution is stable over time. Note that for the true solution we have that

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{u}|^2 \,\mathrm{d}t = 0,$$

so this bound allows for artificial diffusion, but does not ensure it as we have not explicitly excluded the possibility that $\left|\boldsymbol{U}_{n+1}^-\right|^2 = \left|\boldsymbol{U}_n^-\right|^2$.

Proof. Choosing $\boldsymbol{V} = \boldsymbol{U}$ in the localised upwind dG method (3.9) we have that

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t + [\![\boldsymbol{U}_n]\!] \cdot \boldsymbol{U}_n^+ + \int_{I_n} B\boldsymbol{U} \cdot \boldsymbol{U}\mathrm{d}t = 0.$$

Applying Lemma 3.1.20 and the skew-symmetry of B we find

$$\frac{1}{2}\left|\boldsymbol{U}_{n+1}^{-}\right| \leq \frac{1}{2}\left|\boldsymbol{U}_{n}^{-}\right|,$$

as required.

□

3.1.1.3 Hamiltonian systems

In the continuous setting the natural notion of non-linear stability arises from conservation of the Hamiltonian, i.e., $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(\boldsymbol{u}) = 0$. For the cG method (3.1) applied to Hamiltonian systems a discrete notion of non-linear stability follows directly from Theorem 3.1.6 which tells us that

$$\mathcal{H}(\boldsymbol{U}_{n+1}) = \mathcal{H}(\boldsymbol{U}_{n}),$$

where $\boldsymbol{U}$ is the solution of the cG method (3.1) applied to a Hamiltonian system.
For the upwind dG method (3.8) stability does not follow as immediately in general, as the Hamiltonian is not conserved at the nodes. However, the diffuse nature of the method, see Remark 3.1.14, leads to stability. Note that the diffuse nature of the method has only been proven in the case that $J\nabla\mathcal{H}(\boldsymbol{u})$ is linear. To prove the diffuse nature of nonlinear problems we can follow the methodology outlined in §3.1.1.2.

3.1.2 Convergence

Here we develop a new a priori error bound for the cG method (3.1) for symmetric linear problems. Similar order bounds can be developed for skew-symmetric linear problems, however, for brevity we shall not discuss this case here. As presented in Theorem 3.1.15, a priori error bounds already exist for this method which utilise duality arguments. Here we employ a simpler approach. While we are proving this a priori bound for the cG method we are primarily introducing the techniques so they can be applied to the new temporal finite element method we introduce in §3.2. Throughout we shall assume that the exact solution $\boldsymbol{u}$ is sufficiently smooth, to be concise we will require that $\boldsymbol{u} \in C^{q+2}([0,T])$.

Definition 3.1.23 (Interpolation operator for the cG method)**.** *Let* $\boldsymbol{w} \in \left(\mathcal{C}^0([0,T]) \cap L_2([0,T])\right)^D$*, then we define the interpolation operator* $\mathcal{I}(\boldsymbol{w}) \in \mathbb{V}_{q+1}^{C}(I_n)$

locally such that

$$\int_{I_n} \left(\mathcal{I}\left(\boldsymbol{w}\right) - \boldsymbol{w}\right) \cdot \boldsymbol{\phi} \mathrm{d}t = 0 \qquad \forall \boldsymbol{\phi} \in \left(\mathbb{P}_{q-1}\left(I_n\right)\right)^D, \tag{3.12}$$

for $n = 0, ..., N-1$*, and* $\mathcal{I}\left(\boldsymbol{w}(t_n)\right) = \boldsymbol{w}(t_n)$ *for* $n = 0, .., N$*. Note that this interpolation is uniquely defined as* (3.12) *locally fixes* q *degrees of freedom and nodal conditions fix the remaining degree of freedom on each element. When* $q = 0$ *our interpolation operator is uniquely defined through being exact at the nodes.*

Lemma 3.1.24 (Convergence of the cG interpolation operator)**.** *Let* $\boldsymbol{w} \in \left(\mathcal{C}^{q+1}([0,T]) \cap L_2([0,T])\right)^D$*, further let the interpolation operator* $\mathcal{I}$ *be as discussed in Definition 3.1.23, then*

$$\left\|\boldsymbol{w} - \mathcal{I}\left(\boldsymbol{w}\right)\right\|_{L_2(I_n)} \leq C\tau_n^{q+2} \left\| \frac{\mathrm{d}^{q+2}}{\mathrm{d}^{q+2}t} \boldsymbol{w} \right\|_{L_2(I_n)}.$$

Proof. Lemma 3.1.24 follows from the interpolation operator given in Definition 3.1.23 being exact for functions in $\mathbb{V}_{q+1}^C\left([0,T]\right)$.

□

Theorem 3.1.25 (A discrete error for the cG method for symmetric linear ODEs)**.** *Let* $\boldsymbol{U}$ *be the solution to the cG method* (3.1) *with right hand side* $\boldsymbol{F}\left(\boldsymbol{U}, t\right) = \mathfrak{f}\left(t\right) - A\boldsymbol{U}$*, additionally let* $\mathcal{I}\left(\boldsymbol{u}\right)$ *be the interpolation of the exact solution of the ODE* $\boldsymbol{u}$ *as described in Definition 3.1.23. The cG solution satisfies the bound*

$$\frac{1}{2}\left|\boldsymbol{U}_n - \mathcal{I}\left(\boldsymbol{u}\right)_n\right|^2 + \left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{U} - \mathcal{I}\left(\boldsymbol{u}\right)\right)\right|_A\right\|_{L_2([0,t_n])}^2 \leq C\left\|\left|\mathcal{I}\left(\boldsymbol{u}\right) - \boldsymbol{u}\right|_A\right\|_{L_2([0,t_n])}^2,$$

with $\Pi_{\mathbb{V}_q}$ *denoting the* L_2 *projection into the finite element space* $\mathbb{V}_q$*. Note that here the constant* C *depends on the coercivity and continuity constants given in* (3.10) *and* (3.11) *respectively.*

Proof of Theorem 3.1.25. We begin by splitting the error of the cG solution $\boldsymbol{U}$ with respect to the exact solution $\boldsymbol{u}$, by adding and subtracting the cG interpolant described in Definition 3.1.23 applied to the exact solution $\mathcal{I}\left(\boldsymbol{u}\right)$, i.e.,

$$\boldsymbol{U} - \boldsymbol{u} = \left(\boldsymbol{U} - \mathcal{I}\left(\boldsymbol{u}\right)\right) + \left(\mathcal{I}\left(\boldsymbol{u}\right) - \boldsymbol{u}\right) =: \boldsymbol{\theta} + \boldsymbol{\rho}.$$

Adding and subtracting the aforementioned interpolant appropriately to the local cG method (3.2) we observe that

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}\cdot\boldsymbol{V} + A\boldsymbol{\theta}\cdot\boldsymbol{V} + \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}\left(\boldsymbol{u}\right)\cdot\boldsymbol{V} + A\mathcal{I}\left(\boldsymbol{u}\right)\cdot\boldsymbol{V}\,\mathrm{d}t = \int_{I_n} \mathfrak{f}\cdot\boldsymbol{V}\,\mathrm{d}t \qquad \forall \boldsymbol{V}\in\mathbb{V}_q. \quad (3.13)$$

Additionally, we can write the general linear ODE (2.1) with $\boldsymbol{F}\left(\boldsymbol{U},t\right) = \mathfrak{f}\left(t\right) - A\boldsymbol{U}$ variationally as

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} + A\boldsymbol{u}\cdot\boldsymbol{V}\,\mathrm{d}t = \int_{I_n} \mathfrak{f}\cdot\boldsymbol{V}\,\mathrm{d}t \qquad \forall \boldsymbol{V}\in\mathbb{V}_q. \quad (3.14)$$

Eliminating $\mathfrak{f}$ from (3.13) and (3.14) we can write

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}\cdot\boldsymbol{V} + A\boldsymbol{\theta}\cdot\boldsymbol{V}\,\mathrm{d}t = -\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\rho}\cdot\boldsymbol{V} + A\boldsymbol{\rho}\cdot\boldsymbol{V}\,\mathrm{d}t. \quad (3.15)$$

Through integration by parts we have that

$$\begin{aligned}\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\rho}\cdot\boldsymbol{V}\,\mathrm{d}t =& -\int_{I_n} \boldsymbol{\rho}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{V}\,\mathrm{d}t + \boldsymbol{\rho}_{n+1}\cdot\boldsymbol{V}_{n+1} - \boldsymbol{\rho}_n\cdot\boldsymbol{V}_n \\ =&0,\end{aligned} \quad (3.16)$$

after applying the definition of the interpolant, as $\boldsymbol{\rho} := \mathcal{I}\left(\boldsymbol{u}\right) - \boldsymbol{u}$. Note that in the case $q = 0$ we have that $\int_{I_n}\boldsymbol{\rho}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{V}\mathrm{d}t$ is trivially zero. In view of (3.16), we can choose $\boldsymbol{V} = \Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)$ in (3.15) allowing us to write

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}\cdot\boldsymbol{\theta} + A\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\cdot\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\mathrm{d}t = -\int_{I_n}\boldsymbol{\rho}\cdot\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\mathrm{d}t,$$

through the definition of the L_2 projection, as $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}\in\mathbb{V}_q$. Through Hölder's inequality, and then continuity of A we have

$$\int_{I_n}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\left|\boldsymbol{\theta}\right|^2\right) + A\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\cdot\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\mathrm{d}t \leq C_A\left\|\left|\boldsymbol{\rho}\right|_A\right\|_{L_2(I_n)}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|_{L_2(I_n)}$$

where C_A is given by (3.11). Further, through coercivity (3.10) and Cauchy's inequality with ϵ

$$\frac{1}{2}\left|\boldsymbol{\theta}_{n+1}\right|^2 + c_A\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|^2_{L_2(I_n)} \leq \frac{1}{2}\left|\boldsymbol{\theta}_n\right|^2 + \frac{C_A^2}{4\epsilon}\left\|\left|\boldsymbol{\rho}\right|_A\right\|^2_{L_2(I_n)} + \epsilon\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|^2_{L_2(I_n)}.$$

Choosing $\epsilon = \frac{c_A}{2}$ and iterating back to the initial point in time

$$\frac{1}{2}\left|\boldsymbol{\theta}_{n+1}\right|^2 + \frac{c_A}{2}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|^2_{L_2([0,t_{n+1}])} \leq \frac{1}{2}\left|\boldsymbol{\theta}_0\right|^2 + \frac{C_A}{2c_A}\left\|\left|\boldsymbol{\rho}\right|_A\right\|^2_{L_2([0,t_{n+1}])}.$$

Observing that $\boldsymbol{\theta}_0 = \boldsymbol{\rho}_0 = 0$ allows us to conclude.

□

Corollary 3.1.26 (Convergence of the cG method for symmetric linear ODEs)**.** *Let $\boldsymbol{U}$ be the solution to the cG method* (3.1) *with the symmetric right hand side $\boldsymbol{F}\left(\boldsymbol{U},t\right) = \mathfrak{f}\left(t\right) - A\boldsymbol{U}$ where A is a symmetric linear operator. Additionally, let $\boldsymbol{u} \in C^{q+2}([0,T])$ be the exact solution of the ODE* (2.1)*, then*

$$\left|\boldsymbol{U}_n - \boldsymbol{u}_n\right|^2 + \left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{U}-\boldsymbol{u}\right)\right|_A\right\|^2_{L_2([0,t_n])} \leq C\left\|\tau^{q+2}\left|\frac{\mathrm{d}^{q+2}}{\mathrm{d}^{q+2}t}\boldsymbol{u}\right|_A\right\|^2_{L_2([0,t_n])}.$$

Proof. Recall that through the introduction of the interpolant $\mathcal{I}\left(\boldsymbol{u}\right)$ described in Definition 3.1.23 we write

$$\boldsymbol{U} - \boldsymbol{u} = \left(\boldsymbol{U} - \mathcal{I}\left(\boldsymbol{u}\right)\right) + \left(\mathcal{I}\left(\boldsymbol{u}\right) - \boldsymbol{u}\right) =: \boldsymbol{\theta} + \boldsymbol{\rho}.$$

Through the triangle inequality, and Cauchy's inequality, we can write

$$\begin{aligned}\left|\boldsymbol{U}_n - \boldsymbol{u}_n\right|^2 \leq& \frac{3}{2}\left|\boldsymbol{\theta}_n\right|^2 + \frac{3}{2}\left|\boldsymbol{\rho}_n\right|^2 \\ \leq& \frac{3}{2}\left|\boldsymbol{\theta}_n\right|^2,\end{aligned} \tag{3.17}$$

through Definition 3.1.23. Following a similar argument we see that

$$\begin{aligned}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{U}-\boldsymbol{u}\right)\right|_A\right\|^2_{L_2([0,t_n])} \leq& \frac{3}{2}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|^2_{L_2([0,t_n])} + \frac{3}{2}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\rho}\right)\right|_A\right\|^2_{L_2([0,t_n])} \\ \leq& \frac{3}{2}\left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{\theta}\right)\right|_A\right\|^2_{L_2([0,t_n])} + \frac{3}{2}\left\|\left|\boldsymbol{\rho}\right|_A\right\|^2_{L_2([0,t_n])},\end{aligned} \tag{3.18}$$

through the stability of the L_2 projector. Combining (3.17) and (3.18), and applying Theorem 3.1.25 we find that

$$\left|\boldsymbol{U}_n - \boldsymbol{u}_n\right|^2 + \left\|\left|\Pi_{\mathbb{V}_q}\left(\boldsymbol{U}-\boldsymbol{u}\right)\right|_A\right\|^2_{L_2([0,t_n])} \leq C\left\|\left|\boldsymbol{\rho}\right|_A\right\|_{L_2([0,t_n])}.$$

We may conclude through the application of Lemma 3.1.24.

□

3.2 The recovered finite element method

The notion of recovered finite element methods *(RFEM)* have been recently introduced in [74] for the two dimensional Poisson problem, with a view to being extended for general spatial finite element methods. The authors define a reconstruction $\mathcal{E}$ which maps a discrete function from a nonconforming to a conforming finite element space. Their operator is similar to reconstructions applied when post-processing nonconforming approximations, see [74, Page 3]. The key difference is that the reconstruction is hard-coded into the numerical scheme, i.e., a nonconforming solution $\boldsymbol{U}$ is sought such that some variational formulation holds where the variational formulation is described in terms of the conforming reconstruction $\mathcal{E}(\boldsymbol{U})$. A recovered finite element method can also be developed for temporal finite element methods, although it is fundamentally different due to the evolutionary nature of the problem and lack of symmetry in the derivatives. We will see here for first order temporal problems that the resultant RFEM method for ODEs of the form (2.1) can be viewed as a generalisation of the standard cG method (3.1).

The concise definition of RFEM relies heavily on the function space which the reconstruction operator maps from, i.e., the function space of the underlying discontinuous solution. Throughout we will denote this space as $\mathbb{V}_p$ where p corresponds to the polynomial degree of the space. Our reconstruction operator $\mathcal{E}$ will act on each vector component of the input independently, returning a vector of the same structure. That is to say that when we write $\mathcal{E}(\boldsymbol{U})$ we are just succinctly writing $(\mathcal{E}(U_1), \mathcal{E}(U_2), ..., \mathcal{E}(U_D))^T$.

The continuous reconstruction $\mathcal{E}(\boldsymbol{U})$ is not necessarily sought on the same mesh as the underlying solution $\boldsymbol{U}$. In fact, a key motivator behind the development of this method is the ability to choose a different mesh for the discontinuous solution and continuous reconstruction. This allows us to employ a discontinuous solution $\boldsymbol{U}$ over an *adaptive* mesh, which possesses an associated continuous *structure preserving* reconstruction over a *fixed* mesh.

Before defining RFEM we first must formally define the mesh of a method. The mesh $\mathcal{M}$ is uniquely determined through a collection of nodal points, i.e.,

$$\mathcal{M} := \{t_0, t_1, ..., t_N\}, \tag{3.19}$$

where t_n are as described at the beginning of §3.1. Additionally, we can modify our notation slightly to highlight the mesh dependency of a finite element space as follows.

Definition 3.2.1 (Temporal finite element spaces with variable meshes)**.** *Let $\mathcal{M}$ be as*

given in (3.19)*, then we define the elements* $J_n = (t_n, t_{n+1})$*. Further let* $\mathbb{P}_q(J_n)$ *denote the space of polynomials of degree* q *on the element* J_n*, then we define the discontinuous finite element space as*

$$\mathbb{V}_q(\mathcal{M}) = \{\boldsymbol{W} : \boldsymbol{W}|_{J_n} \in (\mathbb{P}_q(J_n))^D, n = 0, ..., N-1\},$$

further to this the continuous finite element space is defined analogously with global continuity enforced, i.e.,

$$\mathbb{V}_q^C(\mathcal{M}) = \mathbb{V}_q(\mathcal{M}) \cap \mathcal{C}^0([0,T]).$$

This is a slight abuse of notation as we sometimes also write the finite element space as a function of the domain, however, the mesh implicitly defines the domain, and as such we shall not enforce a clear distinction here.

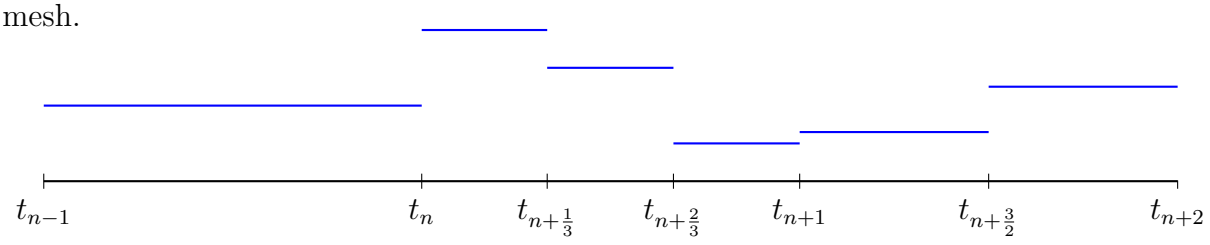
With Definition 3.2.1 in mind we can define RFEM as follows.

Definition 3.2.2 (Recovered finite element method)**.** *Let the mesh* $\widetilde{\mathcal{M}}$ *be a refinement of the mesh* $\mathcal{M}$*, i.e.,* $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$*, then we define recovered finite element methods for ODEs as follows: Let the reconstruction operator be* $\mathcal{E} : \mathbb{V}_p(\widetilde{\mathcal{M}}) \to \mathbb{V}_{q+1}^C(\mathcal{M})$ *for some* $p \leq q$*, then seek* $\boldsymbol{U} \in \mathbb{V}_p\left(\widetilde{\mathcal{M}}\right)$ *such that*

$$\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\boldsymbol{U}) \cdot \boldsymbol{V}\mathrm{d}t = \int_0^T \boldsymbol{F}(\mathcal{E}(\boldsymbol{U}), t) \cdot \boldsymbol{V}\mathrm{d}t \qquad \forall \boldsymbol{V} \in \mathbb{V}_q\left(\widetilde{\mathcal{M}}\right). \tag{3.20}$$

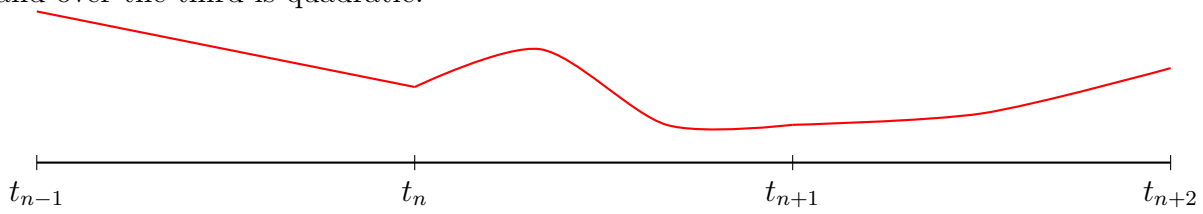
Consider the case where $p = 0$, and $q = 3$, possible functions of $\mathbb{V}_p(\widetilde{\mathcal{M}})$ and $\mathbb{V}_q^C(\mathcal{M})$ can be described by Figure 3.1 and Figure 3.2.

Figure 3.1: An example of a function in $\mathbb{V}_0(\widetilde{\mathcal{M}})$, defined over a subset of the associated mesh.



Remark 3.2.3 (A motivator behind the development of RFEM for ODEs)**.** *While the recovered finite element method, as discussed in this section, is a purely temporal method*

Figure 3.2: An example of a function in $\mathbb{V}_3^C(\mathcal{M})$, defined over a subset of the associated mesh. Notice that the solution over the first element is linear, over the second is cubic and over the third is quadratic.



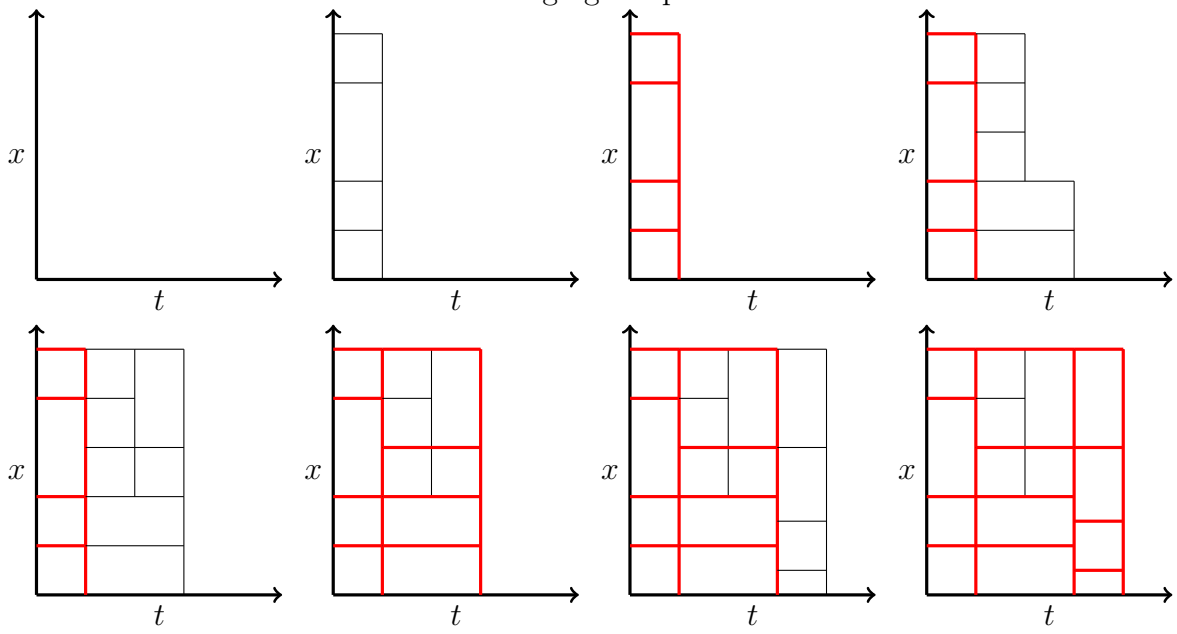
the development of this method is motivated by its extension to space-time numerical methods. Through coupling the RFEM discretisation in time with an appropriate discontinuous spatial discretisation we can obtain a discontinuous *numerical approximation on a fine mesh with a* continuous *reconstruction which preserves the structure of the problem over a coarser temporal mesh. The discontinuity of the underlying solution allows us to implement space-time adaptivity. Note that the numerical scheme will be adaptive* only on the fine mesh, *and the coarse mesh will be fixed a priori. We will find in the sequel that the increase in temporal resolution on the finer mesh can be heuristically compared to increasing the polynomial degree of the continuous reconstruction of the numerical solution. An example of a potential refined and coarse space-time mesh are given in Figure 3.3. Note that in the literature the adaptivity with spatial RFEM is fundamentally different, in that the conforming reconstruction is is defined over a coarser mesh than the underlying solution, see [58].*

A fundamental, and indeed defining, property of RFEM is the choice of reconstruction operator. Throughout we shall focus of study on the following reconstruction operator.

Definition 3.2.4 (The RFEM reconstruction operator)**.** *Assume that $W \in \mathbb{V}_p(\widetilde{\mathcal{M}})$, and let $\mathcal{N}$ denote the Lagrange nodes of $\mathbb{V}_{q+1}^C(\mathcal{M})$. We split the set of nodes into three groups.*

1. *Let $\mathcal{N}_0$ denote a set containing the first node at time $t = 0$.*
2. *Let $\mathcal{N}_e$ denote the set of nodal values at the nodes of the elements.*
3. *Let $\mathcal{N}_i$ denote the set of nodal values on the interiors of the elements.*

Figure 3.3: The time-stepping implementation of the fine mesh (thin black lines), on which the underlying discontinuous solution lives superimposed with the coarse mesh (thick red lines) on where the conforming reconstruction lives. Notice in that the second time step (in the third sub-figure) is *not* uniform in space, so a discontinuous approximation is required on the fine mesh to avoid issues with hanging temporal nodes.



We define the reconstruction $\mathcal{E}(W)$ *such that for* $\gamma \in \mathcal{N}$ *it satisfies*

$$\mathcal{E}(W)(\gamma) = \begin{cases} W(\gamma^-) & \text{for } \gamma \in \mathcal{N}_e \backslash \mathcal{N}_0 \\ \{W(\gamma)\} & \text{for } \gamma \in \mathcal{N}_i \\ w_0 & \text{for } \gamma \in \mathcal{N}_0, \end{cases}$$

where w_0 *denotes the initial condition of the corresponding continuous problem.*

Remark 3.2.5 (A comparison of Definition 3.2.4 and existing RFEM reconstruction operators)**.** *For spatial RFEM the conforming reconstruction operator is chosen such that for arbitrary Lagrange nodes of the solution space* $\gamma \in \mathcal{N}$

$$\widetilde{\mathcal{E}(W)}(\gamma) = \{W(\gamma)\},$$

see [74, 58]. This choice is desirable as it is symmetry preserving and is well studied, see [116]. In the temporal case we cannot choose such a reconstruction operator as it requires us to solve globally*. To remedy this in Definition 3.2.4 we have chosen a reconstruction operator which can be evaluated* locally *in the sense that the solution over the element* I_n *depends solely on information in* I_n *and on the previous element* I_{n-1}*. We do not need to know any information from "future" elements.*

Remark 3.2.6 (A time stepping implementation)**.** *In general, it is crucial that temporal discretisations can be implemented in a time stepping fashion as opposed to solving globally. As our choice of reconstruction operator is localisable it suffices for us to describe the method over a single element of* $\mathcal{M}$*. Let* $\mathcal{E} : \mathbb{V}_p\left(\widetilde{\mathcal{M}}\big|_{(t_n,t_{n+1})}\right) \to \mathbb{V}^C_{q+1}([t_n, t_{n+1}])$ *for some* $p \leq q$*, then the local method is given by seeking* $\boldsymbol{U} \in \mathbb{V}_p\left(\widetilde{\mathcal{M}}\big|_{I_n}\right)$ *such that*

$$\int_{I_n} \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\boldsymbol{U}) \cdot \boldsymbol{V}\,\mathrm{d}t = \int_{I_n} \boldsymbol{F}(\mathcal{E}(\boldsymbol{U}), t) \cdot \boldsymbol{V}\,\mathrm{d}t \qquad \forall \boldsymbol{V} \in \mathbb{V}_q\left(\widetilde{\mathcal{M}}\big|_{I_n}\right). \tag{3.21}$$

This formulation is equivalent to its global counterpart (3.20)*, as can be seen similarly to Remark 3.1.5. It is paramount to note that communication with the solution on the previous element is conducted through the* continuity *of the reconstruction of the solution* $\mathcal{E}(\boldsymbol{U})$*.*

While the underlying approximation $\boldsymbol{U}$ for RFEM is discontinuous, the continuous reconstruction of the solution possesses the following desirable geometric property.

Lemma 3.2.7 (Geometric properties of RFEM)**.** *The* continuous reconstruction *of the local RFEM method* (3.21) *satisfies the same geometric properties as the cG method. That is to say if* $\boldsymbol{F}(\boldsymbol{U},t) = J\nabla\mathcal{H}(\boldsymbol{U})$*, then the reconstruction* $\mathcal{E}(\boldsymbol{U})$ *preserves the Hamiltonian across the nodes, i.e.,*

$$\mathcal{H}(\mathcal{E}(\boldsymbol{U}(t_{n+1}))) = \mathcal{H}(\mathcal{E}(\boldsymbol{U}(t_n)))\,.$$

Similarly under a $q+1$ *point Gauss quadrature, the continuous reconstruction is symplectic. This result follows analogously to the proof of Theorem 3.1.9 in view of Remark 3.1.10.*

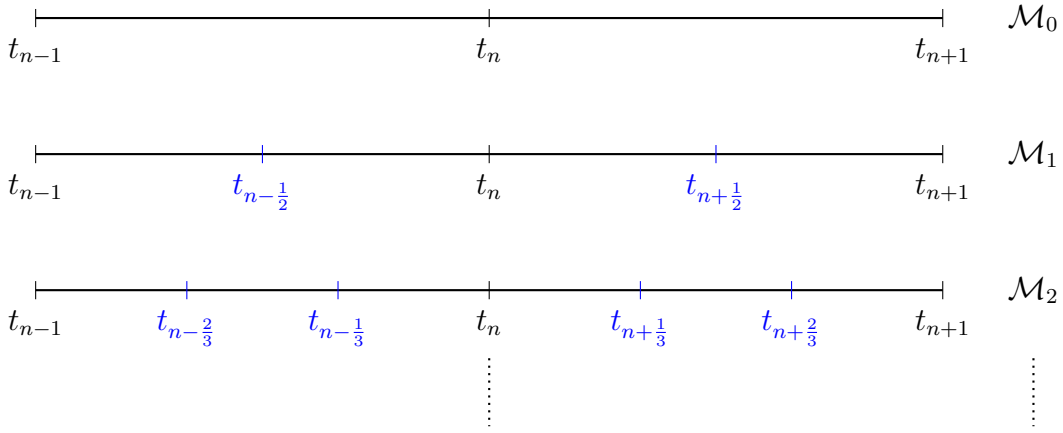
Proof. The proof of Lemma 3.2.7 is identical to the proof of Theorem 3.1.6. Note that we can choose the test function $\boldsymbol{V} = J\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\boldsymbol{U})$ due to the discrete de Rham complex discussed in Remark 3.1.2 and as $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$.

□

A key property of RFEM is its ability to simultaneously possess a nonadaptive *conservative* continuous reconstruction and an adaptive discontinuous solution. To examine this property we need first define various potential candidates for $\widetilde{\mathcal{M}}$.

Definition 3.2.8 (Potential mesh refinements)**.** *We define the mesh* $\mathcal{M}_0$ *through the collection of points* $\{t_0, t_1, ..., t_N\}$ *given at the beginning of §3.2 to define our base mesh. We further define* $\mathcal{M}_1$ *through the collection of points* $\{t_0, t_{\frac{1}{2}}, t_1, ..., t_{N-\frac{1}{2}}, t_N\}$ *where* $t_{n+\frac{1}{2}} = \frac{1}{2}(t_n + t_{n+1})$*. This mesh can be viewed as a refinement of* $\mathcal{M}_0$ *where we add one equidistant point between every preexisting point. Through adding* i *new equidistant points between every point of* $\mathcal{M}_0$ *we yield* $\mathcal{M}_i$ *for* $i \in \mathbb{N}$*. For a pictographic representation see Figure 3.4.*

Figure 3.4: The structure of $\mathcal{M}_i$ over two elements as discussed in Definition 3.2.8.

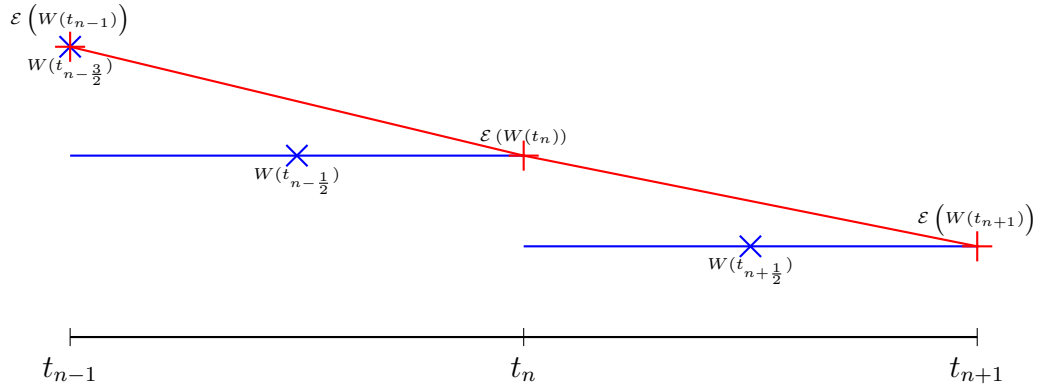


Remark 3.2.9 (A preliminary remark on the practical implementation of RFEM)**.** *In practice, to compute the RFEM approximation in a time stepping fashion, as given in* (3.21)*, we must first assemble the forms as if solving for the conforming reconstruction* $\mathcal{E}(\boldsymbol{U})$*. In order to solve for* $\boldsymbol{U}$ *we multiply the terms in the form which depend on* $\mathcal{E}(\boldsymbol{U})$ *with a* local transition matrix*, which represents the action of* $\mathcal{E}$*. The behaviour of the method is highly dependent on this transition matrix, as such we make it the focal point of the immediate sequel.*

3.2.1 Implementation of RFEM when $\widetilde{\mathcal{M}} \equiv \mathcal{M}$

Assume that the continuous reconstruction $\mathcal{E}(\boldsymbol{U})$ and $\boldsymbol{U}$ are defined over the same mesh, i.e., that $\widetilde{\mathcal{M}} \equiv \mathcal{M}$. Throughout we shall restrict ourselves to an arbitrary component of the vector $\boldsymbol{U}$ which we denote W, for clarity of exposition. In this situation RFEM (3.20) is well posed when $p = q$, as the kernel of $\mathcal{E}$ as given in Definition 3.2.4 is zero. While it is possible to chose $p < q$ we shall not focus on this case here, as it increases the computational complexity of the method.
We present the cases where $p = q = 0$ and $p = q = 1$ independently. Firstly, if $p = q = 0$ we can view graphically in Figure 3.5. Notice that the reconstruction depends on the value

Figure 3.5: The underlying solution W and the reconstruction $\mathcal{E}(W)$ over two arbitrary elements.



of the underlying solution on the previous interval. When considering the first interval this value of the "underlying solution" on the previous interval is given by the initial condition. We implement this operator through the manipulation of its values at the degrees of freedom. We define the local transition matrix $\overrightarrow{T} : \mathbb{V}_0([t_n, t_{n+1})) \to \mathbb{V}_1^C(I_n)$ which maps the values of the degrees of freedom of W to $\mathcal{E}(W)$ on a given element. More

concisely, through defining the local degrees of freedom of W (in addition to the ultimate degree of freedom on the previous element) as

$$\overrightarrow{W} = \begin{pmatrix} W\left(t_{n-\frac{1}{2}}\right) \\ W\left(t_{n+\frac{1}{2}}\right) \end{pmatrix}$$

and the local degrees of freedom of the reconstruction $\mathcal{E}(W)$ as

$$\overrightarrow{\mathcal{E}} = \begin{pmatrix} \mathcal{E}(W(t_n)) \\ \mathcal{E}(W(t_{n+1})) \end{pmatrix},$$

the local transition matrix satisfies

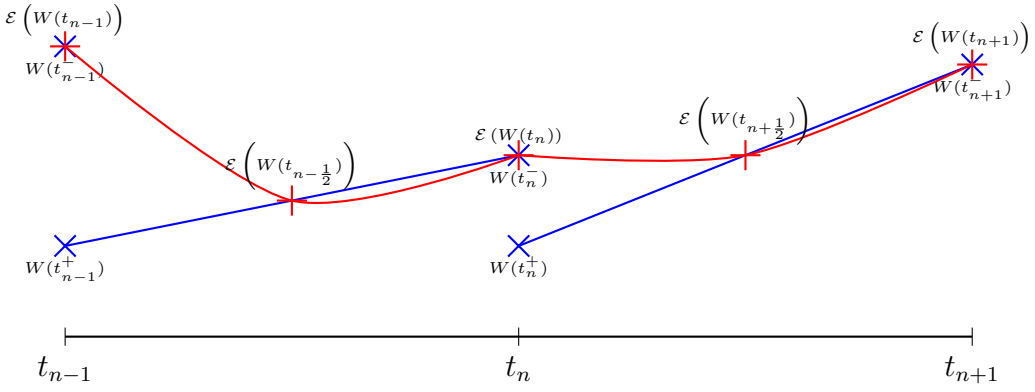
$$\overrightarrow{\mathcal{E}} = \overrightarrow{T}\overrightarrow{W}. \tag{3.22}$$

In view of Definition 3.2.4 we observe that

$$\overrightarrow{T} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

i.e., the transition matrix is the identity.
If instead we assume that $p = q = 1$ then the reconstruction operator $\mathcal{E} : \mathbb{V}_q \to \mathbb{V}_{q+1}^C$ can be viewed graphically in Figure 3.6. The values of W and $\mathcal{E}(W)$ at the degrees of freedom

Figure 3.6: The underlying solution W and the reconstruction $\mathcal{E}(W)$ over two arbitrary elements.



locally are now given by

$$\overrightarrow{W} = \begin{pmatrix} W\left(t_n^-\right) \\ W\left(t_n^+\right) \\ W\left(t_{n+1}^-\right) \end{pmatrix}$$

and

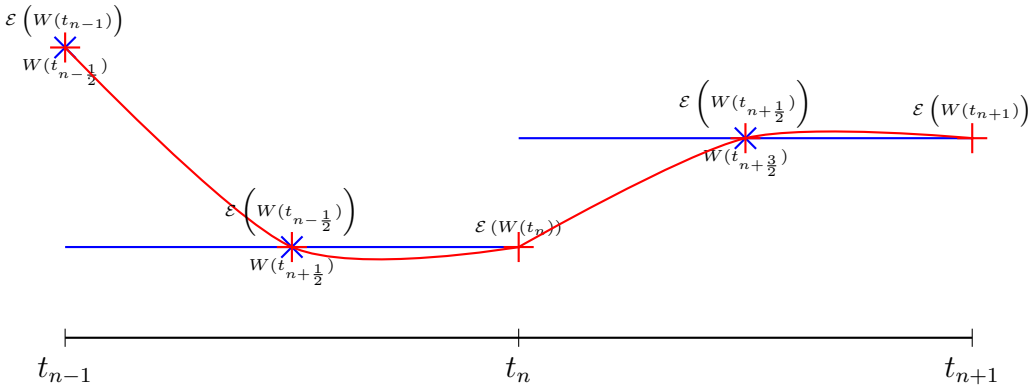
$$\overrightarrow{\mathcal{E}} = \begin{pmatrix} \mathcal{E}\left(W(t_n)\right) \\ \mathcal{E}\left(W(t_{n+\frac{1}{2}})\right) \\ \mathcal{E}\left(W(t_{n+1})\right) \end{pmatrix}$$

respectively. With this in mind, through Definition 3.2.4, the local transition matrix is given by

$$\overrightarrow{T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

While we shall not explore this case numerically for $\widetilde{\mathcal{M}} \equiv \mathcal{M}$, we can again choose $p < q$. For example, if $p = 0$ and $q = 1$ then the reconstruction operator behaves as suggested in Figure 3.7. Through the definition of $\mathcal{E}$ we observe that we can write the local transition

Figure 3.7: The underlying solution $W \in \mathbb{V}_0$ and the reconstruction $\mathcal{E}(W) \in \mathbb{V}_2^C$ over two arbitrary elements.



matrix $\overrightarrow{T}$, which satisfies (3.22), as

$$\overrightarrow{T} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

with

$$\overrightarrow{W} = \begin{pmatrix} W\left(t_{n-\frac{1}{2}}\right) \\ W\left(t_{n+\frac{1}{2}}\right) \end{pmatrix}, \qquad \overrightarrow{\mathcal{E}} = \begin{pmatrix} \mathcal{E}\left(W\left(t_n\right)\right) \\ \mathcal{E}\left(W\left(t_{n+\frac{1}{2}}\right)\right) \\ \mathcal{E}\left(W\left(t_{n+1}\right)\right) \end{pmatrix}.$$

Note that this transition matrix is injective when acting on $\overrightarrow{W}$, but not surjective.

Lemma 3.2.10 (Relation of RFEM to known methods)**.** *Let $\boldsymbol{U}$ be the solution to* (3.20) *where $\mathcal{E}(\boldsymbol{U})$ is given by Definition 3.2.4. Additionally assume that $\widetilde{\mathcal{M}} \equiv \mathcal{M}$ and that $p = q$, then the reconstruction of the RFEM solution is* equivalent *to the solution to the cG method* (3.1)*.*

Remark 3.2.11 (The importance of the relationship between p and q when $\widetilde{\mathcal{M}} \equiv \mathcal{M}$)**.** *The assumption that $p = q$ in Lemma 3.2.10 is crucial. If we consider, for example, $p < q$ then the reconstruction operator $\mathcal{E}$ is not bijective. As such, we expect $\mathcal{E}(\boldsymbol{U})$ to describe the numerical solution less accurately than the cG approximation given in Definition 3.1.3 of degree q over the same mesh.*
If instead, we assumed that $p > q$ then the reconstruction operator will have multiple elements in its kernel, and the associated RFEM approximation would not be well posed without the addition of stabilising terms.

Proof of Lemma 3.2.10. Lemma 3.2.10 follows from the fact that for all $\boldsymbol{U} \in \mathbb{V}_q(\mathcal{M})$ there exists $\mathcal{E}(\boldsymbol{U}) \in \mathbb{V}_{q+1}^C(\mathcal{M})$, and conversely for all $\mathcal{E}(\boldsymbol{U}) \in \mathbb{V}_{q+1}^C(\mathcal{M})$ there exists $\boldsymbol{U} \in \mathbb{V}_q(\mathcal{M})$. This statement is equivalent to saying that $\mathcal{E}$ is an invertible mapping, which arises from the fact that an additional degree of freedom is fixed in the reconstruction by enforcing continuity and some initial data.

□

To illustrate Lemma 3.2.10 consider the case where $q = 0$. Additionally as the reconstruction acts on each component of a vector independently it is sufficient to consider a function $W \in \mathbb{V}_0(\mathcal{M})$. Over an arbitrary interval $t \in I_n$ we can express W as the polynomial

$$W(t) = W_{n+\frac{1}{2}},$$

where $W_{n+\frac{1}{2}}$ is a constant. We can describe an arbitrary linear polynomial over the same interval as

$$P(t) = a_0 + a_1 t,$$

In the case where $q = 0$ the values of the reconstruction on this interval are determined uniquely by

$$\begin{aligned}\mathcal{E}(W)(t_n) =& W_n^- \\ \mathcal{E}(W)(t_n) =& W_{n+\frac{1}{2}}.\end{aligned}$$

Note that W_n^- is enforced by either continuity from the previous element or as an initial condition. Our linear reconstruction can then be written as

$$\begin{aligned}\mathcal{E}(W) =& \frac{t - t_{n+1}}{t_n - t_{n+1}} W_n^- + \frac{t - t_n}{t_{n+1} - t_n} W_{n+\frac{1}{2}} \\ =& \frac{t_{n+1} W_n^- - t_n W_{n+\frac{1}{2}}}{t_{n+1} - t_n} + t \frac{W_{n+\frac{1}{2}} - W_n^-}{t_{n+1} - t_n},\end{aligned}$$

which is exactly the general linear polynomial $P(t)$ with $a_0 = \frac{t_{n+1} W_n^- - t_n W_{n+\frac{1}{2}}}{t_{n+1} - t_n}$ and $a_1 = \frac{W_{n+\frac{1}{2}} - W_n^-}{t_{n+1} - t_n}$. As $W_{n+\frac{1}{2}}$ is arbitrary, and W_n^- is defined such that continuity, or initial, is enforced we have that $P(t)$ is an arbitrary function of $\mathbb{V}_1^C(\mathcal{M})$ after the enforcement of some appropriate initial data.

3.2.2 Implementation of RFEM when $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$

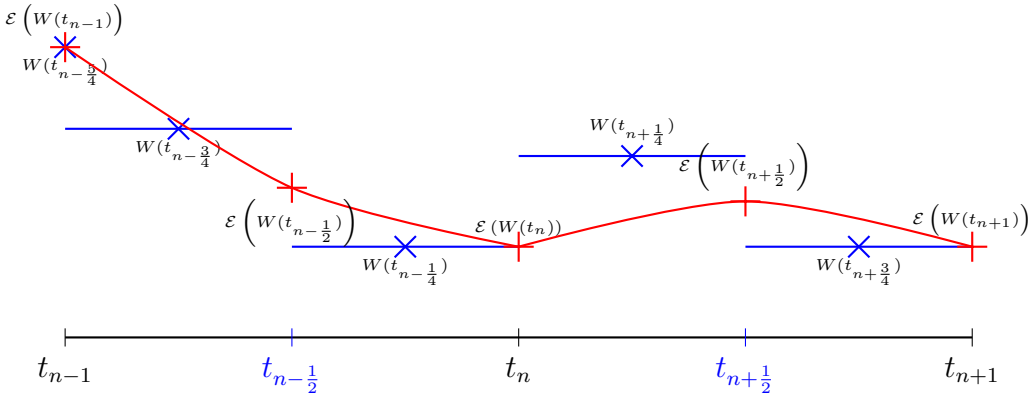
Here we relax the assumption that our underlying method $\boldsymbol{U}$ and continuous reconstruction $\mathcal{E}(\boldsymbol{U})$ are defined over the same mesh. To yield a well posed method we must observe a delicate balance between the mesh refinement and q. We shall see that if $\widetilde{\mathcal{M}} = \mathcal{M}_i$, as described in Definition 3.2.8, this amounts to requiring that $q = p + i$.

For simplicity, we consider the case that $p = 0$, and restrict $\widetilde{\mathcal{M}}$ to be the hierarchy of meshes $\mathcal{M}_i$ described in Definition 3.2.8.

Consider $\widetilde{\mathcal{M}} \equiv \mathcal{M}_1$ and choose $q = 1$, then through the definition of $\mathcal{E}$ we can view $W \in \mathbb{V}_0(\mathcal{M}_1)$ and $\mathcal{E}(W) \in \mathbb{V}_2^C(\mathcal{M}_0)$ graphically in Figure 3.8. Similarly to §3.2.1 we can write W and $\mathcal{E}(W)$ through their values at the degrees of freedom as

$$\overrightarrow{W} = \begin{pmatrix} W\left(t_{n-\frac{1}{4}}\right) \\ W\left(t_{n+\frac{1}{4}}\right) \\ W\left(t_{n+\frac{3}{4}}\right) \end{pmatrix}, \qquad \overrightarrow{\mathcal{E}} = \begin{pmatrix} \mathcal{E}(W(t_n)) \\ \mathcal{E}\left(W\left(t_{n+\frac{1}{2}}\right)\right) \\ \mathcal{E}(W(t_{n+1})) \end{pmatrix},$$

Figure 3.8: The finite element mesh of $\mathcal{E}(W)$ (black lines) superimposed on the finite element mesh of W (black or blue lines), with the underlying solution $W \in \mathbb{V}_0(\mathcal{M}_1)$ over the fine mesh and the reconstruction $\mathcal{E}(W) \in \mathbb{V}_2^C(\mathcal{M}_0)$ over the coarse mesh.



allowing us to define the local transition matrix $\overrightarrow{T}$ satisfying (3.22) as

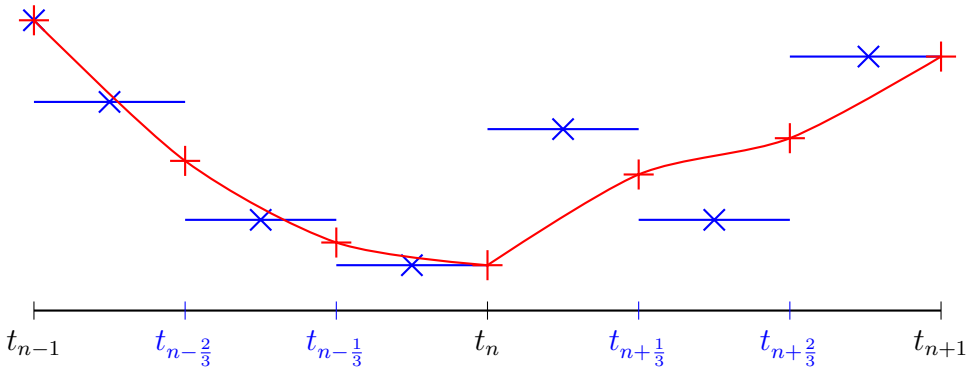
$$\overrightarrow{T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

Observe that that the local transition matrix defines $\mathcal{E}$ as a bijective operator. As such, the resulting numerical method (3.20) is well posed here. Similarly to in §3.2.1 we can increase q and still yield a well posed method, however we cannot expect to achieve best approximability with respect to q as while we have increased q the underlying solution does not possess any additional information.

If we instead we choose $\widetilde{\mathcal{M}} \equiv \mathcal{M}_2$, then we must also choose $q = 2$. As such, we obtain an underlying solution W and continuous reconstruction $\mathcal{E}(W)$ like that shown in Figure 3.9. The local transition matrix in this case is given by

$$\overrightarrow{T} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.23}$$

Figure 3.9: The finite element mesh of $\mathcal{E}(W)$ (black lines) superimposed on the finite element mesh of W (black or blue lines), with the underlying solution $W \in \mathbb{V}_0(\mathcal{M}_2)$ over the fine mesh and the reconstruction $\mathcal{E}(W) \in \mathbb{V}_3^C(\mathcal{M}_0)$ over the coarse mesh.



where

$$\overrightarrow{W} = \begin{pmatrix} W\left(t_{n-\frac{1}{6}}\right) \\ W\left(t_{n+\frac{1}{6}}\right) \\ W\left(t_{n+\frac{1}{2}}\right) \\ W\left(t_{n+\frac{5}{6}}\right) \end{pmatrix}, \qquad \overrightarrow{\mathcal{E}} = \begin{pmatrix} \mathcal{E}(W(t_n)) \\ \mathcal{E}\left(W\left(t_{n+\frac{1}{3}}\right)\right) \\ \mathcal{E}\left(W\left(t_{n+\frac{2}{3}}\right)\right) \\ \mathcal{E}(W(t_{n+1})) \end{pmatrix}.$$

Again we see that the reconstruction operator defined through the local transition matrix (3.23) is a bijective operator. Similarly if we increase q we obtain an injective operator, as we have more degrees of freedom for the reconstruction than the underlying solution. Repeating this procedure we obtain similar results where $\widetilde{\mathcal{M}} \equiv \mathcal{M}_i$ for $p = 0$ and $q = i$, leading us to the following proposition.

Proposition 3.2.12 (Conditions for RFEM (3.20) to be well posed)**.** *Let $\boldsymbol{U}$ be the solution of the RFEM approximation given in Definition 3.2.2 with continuous reconstruction $\mathcal{E}(\boldsymbol{U})$ given in Definition 3.2.4. We further assume that the continuous reconstruction is defined over the mesh $\mathcal{M} \equiv \mathcal{M}_0$ and the discontinuous solution is defined over the mesh $\widetilde{\mathcal{M}} \equiv \mathcal{M}_i$ for $i = 0, 1, \ldots$. Then the RFEM method is well posed if for a given p we choose*

$$q \geq p + i.$$

Furthermore, if we choose $q = p + i$ then the reconstruction operator is bijective and therefore invertible.

3.2.3 Analytic results

Here we assume that RFEM satisfies the conditions of Proposition 3.2.12 and the method is well posed. In this case we can obtain stability and convergence results in terms of the continuous reconstruction. While we do not explicitly examine the stability and convergence of the underlying solution, the underlying solution must be stable if its application under a well posed operator $\mathcal{E}$ is stable.
Through Lemma 3.2.7 we immediately obtain a nonlinear stability result when considering Hamiltonian systems. Additionally, we obtain the following results for symmetric, and skew-symmetric, linear problems.

Corollary 3.2.13 (Stability of RFEM for symmetric linear problems with forcing)**.** *Consider RFEM* (3.20)*, with $\mathcal{E}$ given by Definition 3.2.4, applied to a symmetric linear ODE, i.e., $\boldsymbol{F}(\boldsymbol{U},t) = \mathfrak{f} - A\boldsymbol{U}$ where $\mathfrak{f}$ is bounded and A is a symmetric operator. Additionally assume that $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$ and the conditions of Proposition 3.2.12 are met, then the continuous reconstruction satisfies*

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U} \right\|^2_{L_2([0,T])} + A\mathcal{E}\left(\boldsymbol{U}(T)\right) \cdot \mathcal{E}\left(\boldsymbol{U}(T)\right) \leq A\mathcal{E}\left(\boldsymbol{U}(0)\right) \cdot \mathcal{E}\left(\boldsymbol{U}(0)\right) + \|\mathfrak{f}\|^2_{L_2([0,T])}.$$

Corollary 3.2.14 (Stability of RFEM for skew-symmetric linear problems)**.** *Consider RFEM* (3.20)*, with $\mathcal{E}$ as described in Definition 3.2.4, applied to a skew-symmetric linear ODE, i.e., $\boldsymbol{F}(\boldsymbol{U},t) = -B\boldsymbol{U}$ where B is a skew-symmetric operator. Additionally assume that $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$ and the conditions of Proposition 3.2.12 are met, then the continuous reconstruction of the discrete approximation satisfies*

$$\left|\mathcal{E}\left(\boldsymbol{U}(T)\right)\right| = \left|\mathcal{E}\left(\boldsymbol{U}(0)\right)\right|.$$

Proof of Corollary 3.2.13 and Corollary 3.2.14. Corollary 3.2.13 and Corollary 3.2.14 follow identically to the proofs of Theorem 3.1.17 and Theorem 3.1.21 respectively. It is important to note that we can always choose the same test functions as in the cG case, as $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$.

□

Further to numerical stability, for linear problems we also obtain the following a priori bound for the RFEM reconstruction.

Corollary 3.2.15 (Convergence of RFEM for symmetric linear ODEs)**.** *Let $\boldsymbol{U}$ be the solution to RFEM* (3.20)*, where $\mathcal{E}$ is described by Definition 3.2.4, with the symmetric linear right hand side $\boldsymbol{F}(\boldsymbol{U},t) = \mathfrak{f}(t) - A\boldsymbol{U}$ where A is a coercive and continuous symmetric linear operator. Further let $\mathcal{M} = \mathcal{M}_0$ and $\widetilde{\mathcal{M}} = \mathcal{M}_i$ for some $i = 0, 1, ...$ as described in Definition 3.2.8, and assume that p and q satisfy the relationship*

$$q = p + i.$$

Additionally, let $\boldsymbol{u} \in C^{q+2}([0,T])$ be the exact solution of the ODE (2.1)*, then*

$$\left|\mathcal{E}(\boldsymbol{U})_n - \boldsymbol{u}(t_n)\right|^2 + \left\|\left|\Pi_{\mathbb{V}_q}\left(\mathcal{E}(\boldsymbol{U}) - \boldsymbol{u}\right)\right|_A\right\|^2_{L_2([0,t_n])} \leq C\left\|\tau^{q+2}\left|\frac{\mathrm{d}^{q+2}}{\mathrm{d}t^{q+2}}\boldsymbol{u}\right|_A\right\|^2_{L_2([0,t_n])}.$$

Proof. Follows identically to the argument outlined in §3.1.2 as $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$ with one caveat. We must redefine the interpolation operator given in Definition 3.1.23 as follows:
Let $\boldsymbol{w} \in \left(\mathcal{C}^0([0,T]) \cap L_2([0,T])\right)^D$ then define the interpolation operator
$\mathcal{I}(\boldsymbol{w}) \in \mathbb{V}^C_{q+1}\left(\mathcal{M}_0|_{I_n}\right)$ locally such that

$$\int_{I_n}\left(\mathcal{I}(\boldsymbol{w}) - \boldsymbol{w}\right)\cdot\boldsymbol{\phi}\mathrm{d}t = 0 \qquad \forall\boldsymbol{\phi} \in \mathbb{V}_{p-1}\left(\mathcal{M}_i|_{I_n}\right),$$

for $n = 0, ..., N-1$, and $\mathcal{I}(\boldsymbol{w}(t_n)) = \boldsymbol{w}(t_n)$ for $n = 0, .., N$. For this interpolation operator to be well defined we require all of its degrees of freedom to be fixed, this is the case if and only if

$$q = p + i.$$

□

Remark 3.2.16 (Kernel removal in the case where $q < p + i$)**.** *The proposed RFEM implementation, given in Definition 3.2.2, is only well posed for $q \geq p + i$, due to the kernel containing multiple elements if $q < p + i$. We can remove this problem, similarly to [74], by introducing an additional term into the method which is typically small but nonzero. Explicitly, we can redefine the method by seeking $\boldsymbol{U} \in \mathbb{V}_p\left(\widetilde{\mathcal{M}}\right)$ such that*

$$\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\boldsymbol{U})\cdot\boldsymbol{V}\mathrm{d}t + \sigma\sum_{n=0}^{N-1}[\![\boldsymbol{U}_n]\!]\cdot\boldsymbol{V}_n^+ = \int_0^T \boldsymbol{F}(\mathcal{E}(\boldsymbol{U}))\cdot\boldsymbol{V}\mathrm{d}t \qquad \forall\boldsymbol{V} \in \mathbb{V}_q\left(\widetilde{\mathcal{M}}\right),$$

for some $\sigma \neq 0$. This additional term comes at the cost of the geometric structure of the

method, i.e., for Hamiltonian ODEs the Hamiltonian function will not be conserved, but this error can be controlled by choosing σ to be small.

3.2.4 A discontinuous adaptive algorithm with a continuous reconstruction on a fixed mesh

We have now developed all of the tools we need to define a discontinuous adaptive algorithm utilising RFEM which possesses a continuous reconstruction on a fixed mesh. We define the RFEM adaptive algorithm as follows. Throughout the adaptive procedure we fix the mesh of the conforming approximation to be $\mathcal{M} \equiv \mathcal{M}_0$ and shall assume that p is fixed.

1. Set $n = 0$.
2. Set $i = 0$.
3. Consider the local RFEM approximation (3.21) over I_n with the reconstruction $\mathcal{E}$ as given in Definition 3.2.4. Seek a dG solution over $\widetilde{\mathcal{M}} \equiv \mathcal{M}_i|_{I_n}$ with $q = p + i$ and compute an appropriate local error.
4. If the local error is above a set tolerance then set $i = i + 1$ and go to 3.
5. If $n < N - 1$ then set $n = n + 1$ and go to 2.

To validate our adaptive procedure we shall investigate the RFEM approximation over the meshes $\widetilde{\mathcal{M}} \equiv \mathcal{M}_0$ and $\widetilde{\mathcal{M}} \equiv \mathcal{M}_1$ in §3.3.

3.3 Numerical experiments

Here we illustrate the numerical performance of the methods discussed within this chapter. The experimental code written for this purpose has been partially implemented in the software for automated system for solving differential equations through finite element methods "Firedrake" [153]. Here we employ a Gauss quadrature of order $2q$, where q is the degree of the polynomial space considered in the finite element approximation, to minimise the quadrature errors introduced into the approximation. Note that some quadrature error shall always enter our approximation for non-polynomial right hand sides $\boldsymbol{F}(\boldsymbol{U}, t)$. For the computation of errors we increase the quadrature degree to $2q + 4$.

When considering nonlinear right hand sides we employ the PETSc Newton line search method as our nonlinear solver with a tolerance of 10^{-15}, see [20]. For limited linear numerical experiments we assemble the underlying linear system directly and solve using the Python library Numpy. We additionally utilise the Python library Matplotlib as our primary visualisation tool.
We restrict our numerical study to Hamiltonian systems, in particular, we consider the following problems for the case $D = 2$ with the skew-symmetric matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Additionally, we shall write our solution vector in the form $\boldsymbol{u}(t) = (u_1(t), u_2(t))^{\mathrm{T}}$.

Example 3.3.1 (Harmonic oscillator)**.** *The harmonic oscillator is described via the Hamiltonian*

$$\mathcal{H}(u_1, u_2) = \frac{1}{2}u_1^2 + \frac{1}{2}u_2^2,$$

with corresponding system

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u_1 &= -u_2 \\ \frac{\mathrm{d}}{\mathrm{d}t}u_2 &= u_1. \end{aligned}$$

Through standard techniques for solving ODEs we observe that the solution of the harmonic oscillator is

$$\begin{aligned} u_1(t) &= -u_2(0)\sin(t) + u_1(0)\cos(t) \\ u_2(t) &= u_2(0)\cos(t) + u_1(0)\sin(t), \end{aligned}$$

where $u_1(0)$ and $u_2(0)$ are the prescribed initial conditions. We shall use this right hand side to benchmark our schemes.

Example 3.3.2 (Pendulum problem)**.** *The pendulum problem is described via the Hamiltonian*

$$\mathcal{H}(u_1, u_2) = \frac{1}{2}u_1^2 - \cos(u_2),$$

with corresponding system

$$\frac{\mathrm{d}}{\mathrm{d}t}u_1 = \sin(u_2)$$
$$\frac{\mathrm{d}}{\mathrm{d}t}u_2 = u_1.$$

While it is possible to obtain an exact solution through utilising Jacobi elliptic functions, see [26], we shall not press this point here.

Example 3.3.3 (Lennard-Jones oscillator)**.** *The Lennard-Jones oscillator has the Hamiltonian function*

$$\mathcal{H}(u_1, u_2) = \frac{1}{2}u_1^2 + u_2^{-12} - 2u_2^{-6},$$

and can be written as the system

$$\frac{\mathrm{d}}{\mathrm{d}t}u_1 = 12u_2^{-13} - 12u_2^{-7}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}u_2 = u_1.$$

This Hamiltonian system is very nonlinear, and as such we expect significant quadrature errors to encroach on any appropriate finite element approximation.

Throughout our numerical experiments, for clarity of exposition, we consider *uniform* time steps $\tau_n = \tau$. This restriction is not required anywhere in our analysis.
To benchmark our numerical schemes we fix the polynomial degree q and compute a sequence of solutions with $\tau = \tau(i) = 2^{-i}$ for a sequence of refinement levels, $i = l, \ldots, L$. With this in mind we construct the following definition.

Definition 3.3.4 (Experimental order of convergence)**.** *Given two sequences $a(i)$ and $\tau(i) \searrow 0$ we define the* experimental order of convergence *(EOC) to be the local slope of the $\log(a(i))$ over $\log(\tau(i))$ curve, i.e.,*

$$EOC(a, \tau; i) = \frac{\log\left(\frac{a(i+1)}{a(i)}\right)}{\log\left(\frac{\tau(i+1)}{\tau(i)}\right)}.$$

In the sequel $a(i)$ will be a sequence of errors with $\tau(i)$ the corresponding sequence of time step sizes.

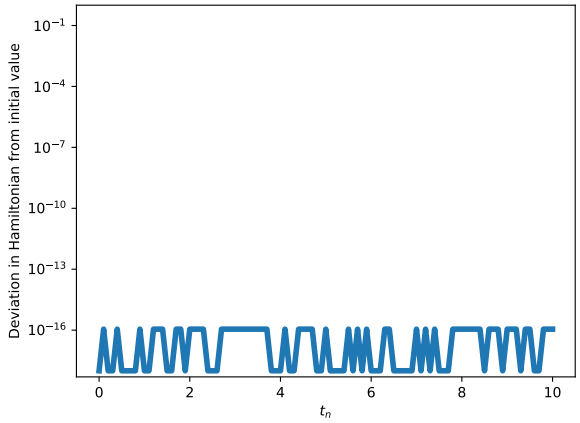
3.3.1 The cG method

Here we consider the cG method as given by Definition 3.1.3. We investigate the deviation in the Hamiltonian functions for both the pendulum problem and the Lennard-Jones oscillator in Figure 3.10 for degree $q = 0, 1$. Note that, through Theorem 3.1.6, we have that the cG method conserves the Hamiltonian function at the nodes. However, as discussed in Remark 3.1.7, for the Lennard-Jones oscillator we cannot exactly compute the finite element method due to our inability to choose a quadrature method which evaluates the method exactly. As such, we observe that the Hamiltonian function deviates for our implementation of the Lennard-Jones oscillator.
Benchmarking the cG method (3.1) with degree $q = 0, 1$ in accordance with Definition 3.3.4 for the harmonic oscillator we obtain Figure 3.11. We observe optimal convergence. Note that, due to the way we have presented the cG method optimal convergence rates are $\mathcal{O}(\tau^{q+2})$, as the numerical approximation is a degree $q + 1$ polynomial.

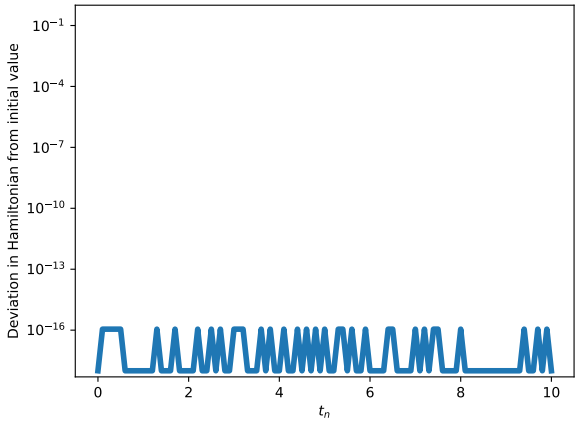
3.3.2 The upwind dG method

We consider the upwind dG method as described in Definition 3.1.12.
Investigating the nodal deviation in the Hamiltonian functions for both the pendulum problem and the Lennard-Jones oscillator for degree $q = 0, 1$ we obtain Figure 3.12. As expected from Remark 3.1.14, we observe that the Hamiltonian function dissipates over time. However, for the Lennard-Jones oscillator this dissipation is not monotonic, this is caused by the inexact quadrature approximation.
Additionally, when simulating the harmonic oscillator we observe optimal experimental convergence rates for degree $q = 0, 1$, see Figure 3.13.

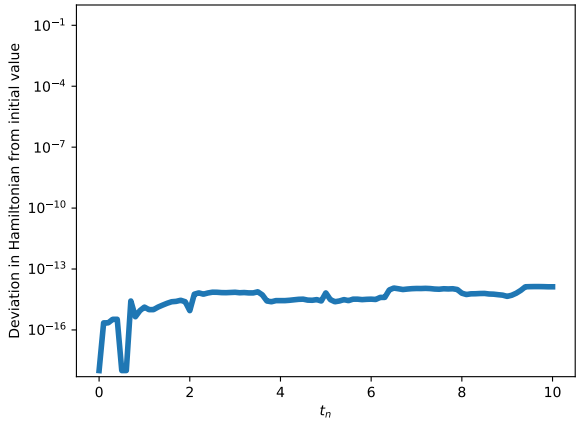
Figure 3.10: We examine the *nodal* deviation in the Hamiltonian function for the cG method (3.1) with various polynomial degrees, q, for various Hamiltonian systems. In particular we consider the pendulum problem and the Lennard-Jones oscillator as given in Example 3.3.2 and Example 3.3.3 respectively. For the pendulum problem we enforce the initial data $u_1(0) = 0.1, u_2(0) = 0.1$, and for the Lennard-Jones oscillator we enforce that $u_1(0) = 0.1, u_2(0) = -1.3$. For all simulation we employ the uniform time step $\tau = 0.1$. We observe that the Hamiltonian function is conserved for the pendulum problem, but not for the Lennard-Jones oscillator due to the inexact numerical integration of the method.



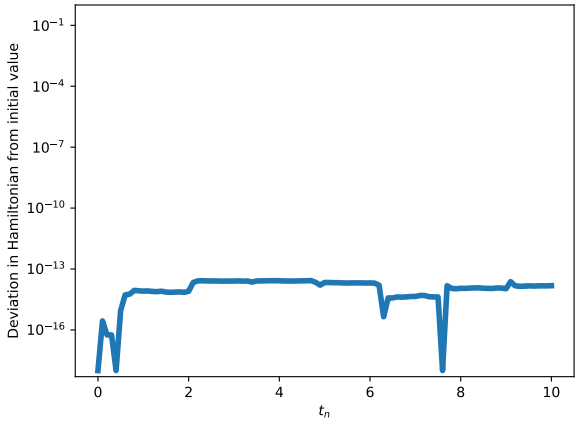
(a) We consider the pendulum problem for $q = 0$.



(b) We consider the pendulum problem for $q = 1$.

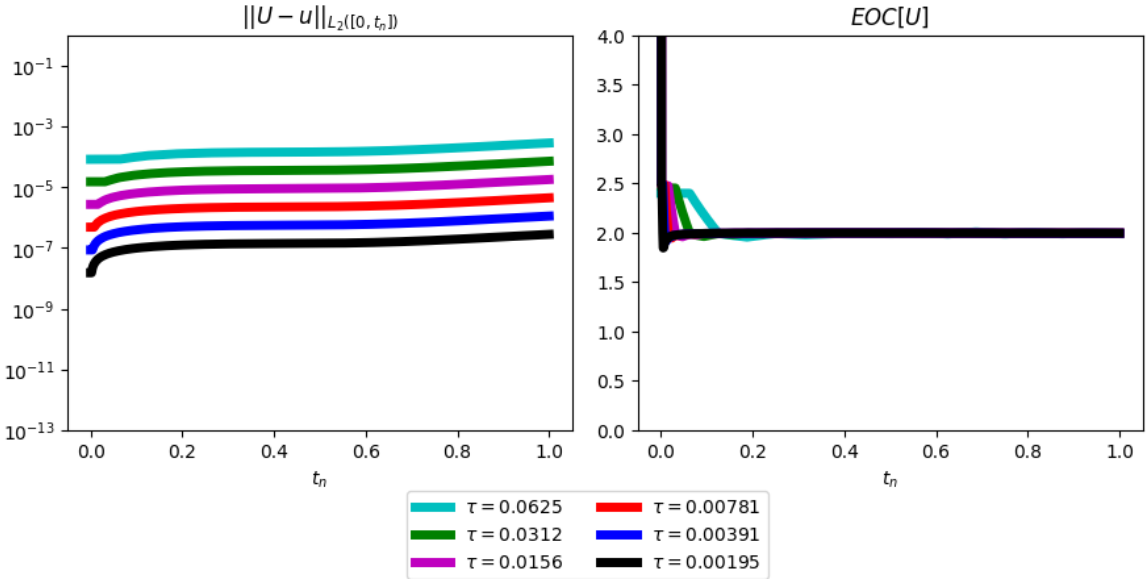


(c) We consider the Lennard-Jones oscillator for $q = 0$.

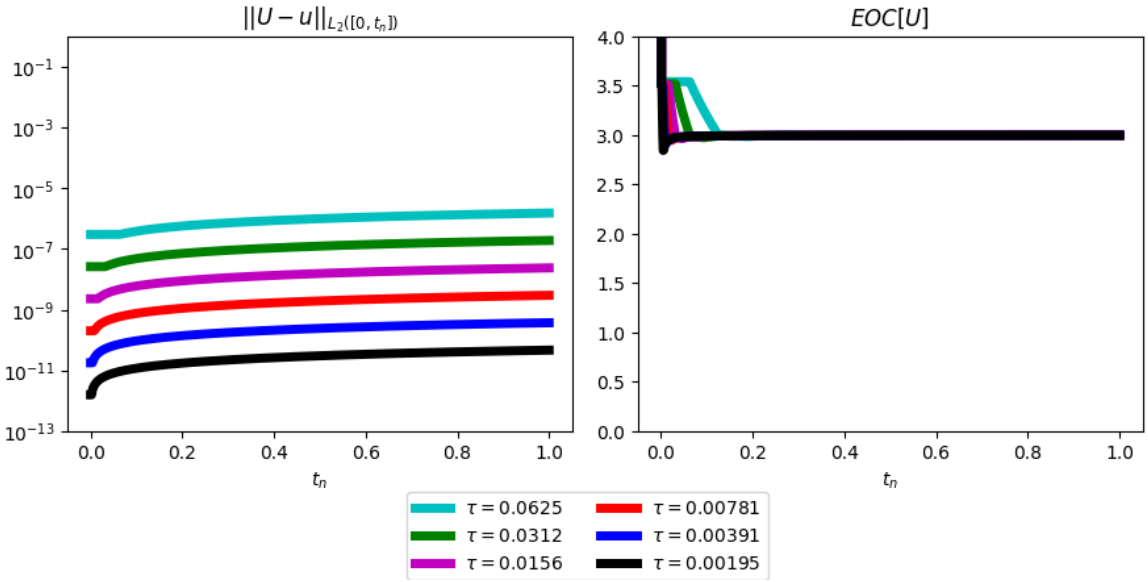


(d) We consider the Lennard-Jones oscillator for $q = 1$.

Figure 3.11: We examine the cG method (3.1) with various polynomial degrees, q, approximating harmonic oscillator discussed in Example 3.3.1 subject to the initial data $u_1(0) = 1, u_2(0) = 1$. We measure errors in the $L_2(0, t_n)$ norm and plot the corresponding EOC, and notice optimal experimental convergence rates, in the sense that our error converges as fast as the best polynomial approximation for each degree.

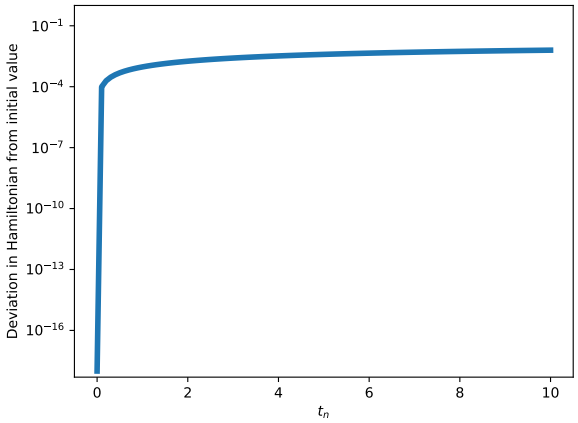


(a) Here $q = 0$.

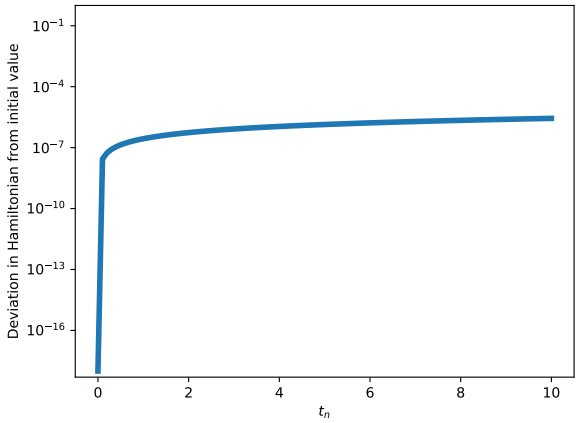


(b) Here $q = 1$.

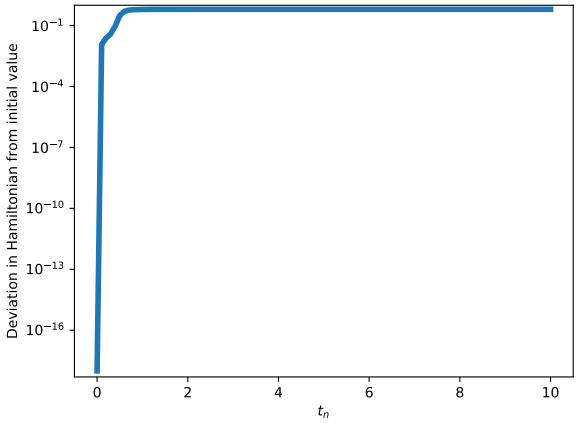
Figure 3.12: We examine the *nodal* deviation in the Hamiltonian function for the upwind dG method (3.8) with various polynomial degrees, q, for various right hand sides. In particular we consider the right hand sides which describe the pendulum problem and the Lennard-Jones oscillator as given in Example 3.3.2 and Example 3.3.3 respectively. For the pendulum problem we enforce the initial data $u_1(0) = 0.1, u_2(0) = 0.1$, and for the Lennard-Jones oscillator we enforce that $u_1(0) = 0.1, u_2(0) = -1.3$. For all simulation we employ the uniform time step $\tau = 0.1$. We observe that the Hamiltonian function deviations for both right hand sides, which we expect from the dissipative nature of the upwind dG method.



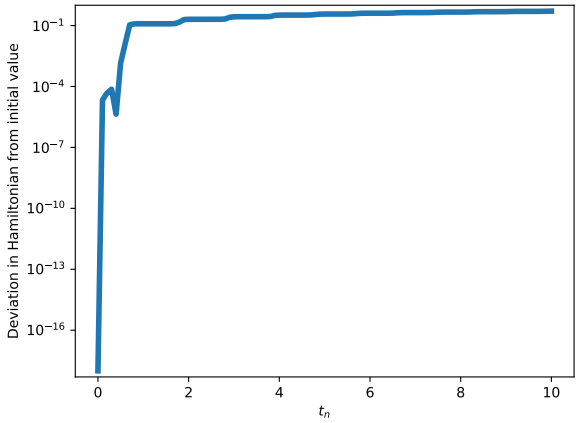
(a) We consider the pendulum problem for $q = 0$.



(b) We consider the pendulum problem for $q = 1$.

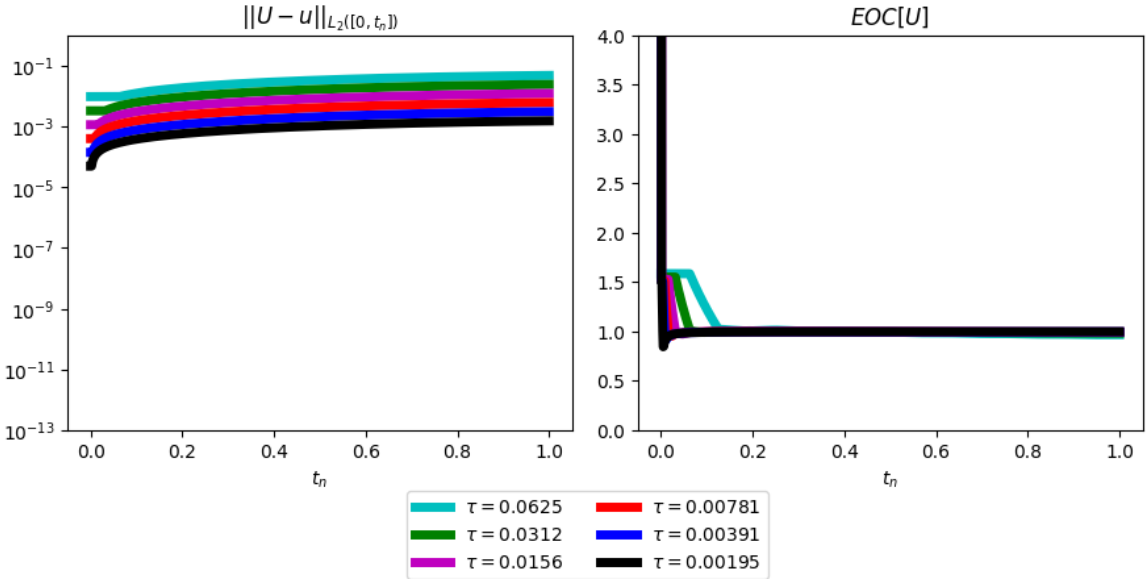


(c) We consider the Lennard-Jones oscillator for $q = 0$.

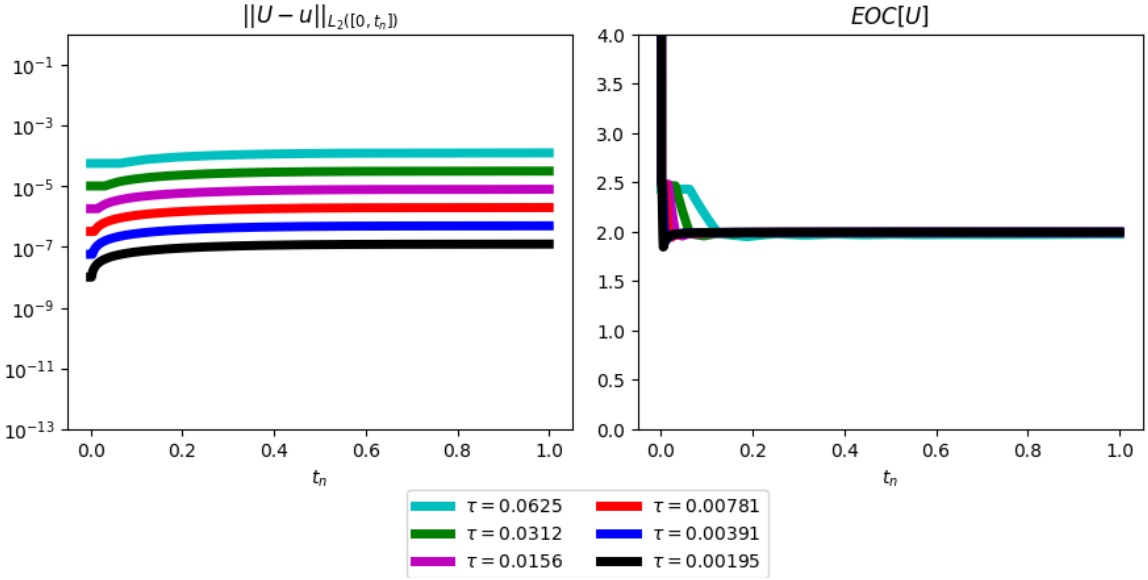


(d) We consider the Lennard-Jones oscillator for $q = 1$.

Figure 3.13: We examine the upwind dG method (3.8) with various polynomial degrees, q, approximating harmonic oscillator discussed in Example 3.3.1 subject to the initial data $u_1(0) = 1, u_2(0) = 1$. We measure errors in the $L_2(0, t_n)$ norm and plot the corresponding EOC, and notice optimal experimental convergence rates.



(a) Here $q = 0$.

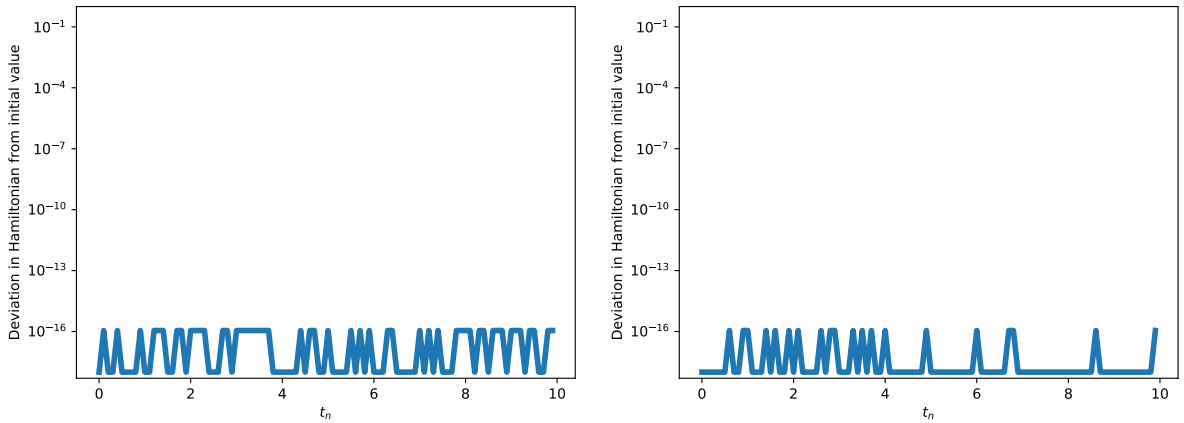


(b) Here $q = 1$.

3.3.3 The RFEM method

Here we consider the RFEM method developed in §3.2, utilising the mesh structures given in Definition 3.2.8. We begin by considering the case where $\mathcal{M} \equiv \widetilde{\mathcal{M}} \equiv \mathcal{M}_0$ with $p = q$, i.e., the case considered in §3.2.1 where our underlying solution is not a refinement of the conforming reconstruction. Note that through Lemma 3.2.10 we have that the continuous reconstruction of this implementation is equivalent to the cG method.
We consider the nodal deviation of the Hamiltonian applied to the continuous reconstruction of U_1 and U_2, i.e., $\mathcal{H}\left(\mathcal{E}\left(U_1\right), \mathcal{E}\left(U_2\right)\right)$, for the pendulum problem with $p = q = 0, 1$ in Figure 3.14.
We also consider the experimental convergence rates when simulating the harmonic oscillator in Figure 3.15. We observe that the errors for both the underlying solution $\boldsymbol{U}$ and reconstruction $\mathcal{E}\left(\boldsymbol{U}\right)$ are optimal in the sense of best approximability.

Figure 3.14: We examine the *nodal* deviation in the Hamiltonian function for the continuous reconstruction of the RFEM method (3.20) with the reconstruction given by Definition 3.2.4, $\mathcal{M} \equiv \widetilde{\mathcal{M}} \equiv \mathcal{M}_0$ and $p = q$. We consider various polynomial degrees, p, for the pendulum problem given in Example 3.3.2, subject to the initial data $u_1(0) = 0.1, u_2(0) = 0.1$. For all simulation we employ the uniform time step $\tau = 0.1$. We observe that the Hamiltonian function is conserved for the pendulum problem.

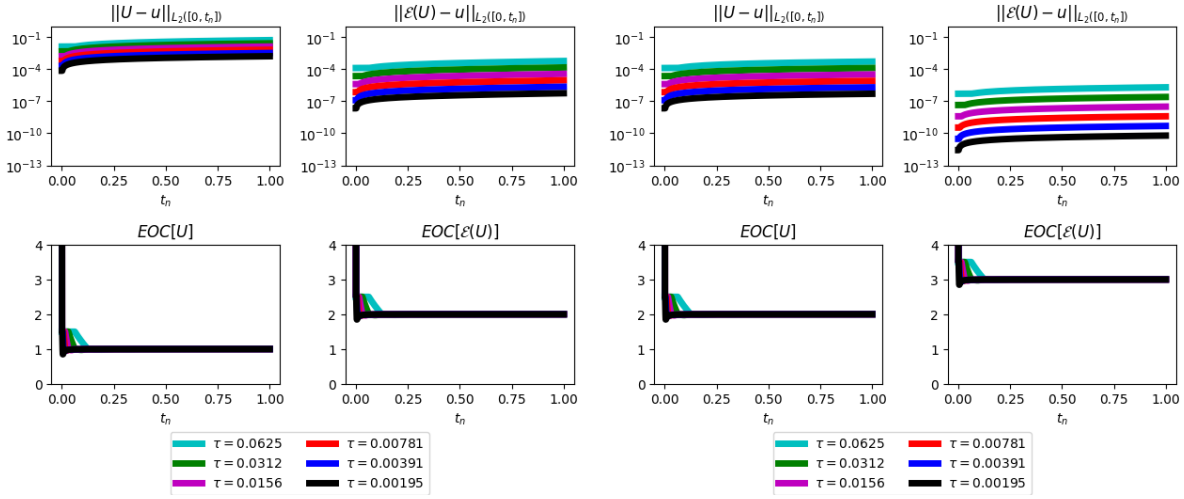


(a) Here $q = 0$.

(b) Here $q = 1$.

A key property of RFEM is the ability to define the underlying solution on a more refined mesh than that of the continuous reconstruction, leading to an adaptive discontinuous approximation with a continuous reconstruction over a fixed mesh. With this in mind, we investigate the behaviour of RFEM where the discontinuous approximation $\boldsymbol{U}$ exists on a refinement of the conforming reconstruction $\mathcal{E}\left(\boldsymbol{U}\right)$. In particular, we choose $\mathcal{M} \equiv \mathcal{M}_0$

Figure 3.15: We examine the error and EOC of the RFEM method (3.20), with $\mathcal{E}$ as given in Definition 3.2.4, $\mathcal{M} \equiv \widetilde{\mathcal{M}} \equiv \mathcal{M}_0$ and $p = q$. We consider various polynomial degrees, q, approximating harmonic oscillator discussed in Example 3.3.1 subject to the initial data $u_1(0) = 1, u_2(0) = 1$. We measure errors in the $L_2([0, t_n])$ norm and plot the corresponding EOC, and notice optimal experimental convergence rates.



(a) Here $q = 0$.

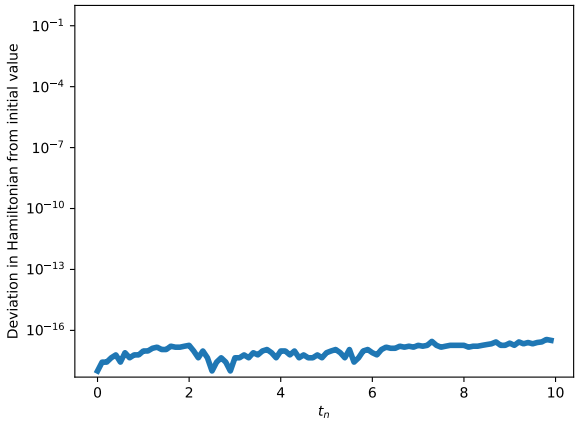
(b) Here $q = 1$.

and $\widetilde{\mathcal{M}} \equiv \mathcal{M}_1$. In view of Proposition 3.2.12, we choose $q = p+1$ and restrict ourselves to the case where $p = 0$. While we note that we can choose q larger at the cost of increasing the computational complexity, we do not do this here as it does not significantly improve the behaviour of the method.

We show the nodal deviation of the Hamiltonian for the harmonic oscillator in Figure 3.16 for $p = 0$ and $q = 1$, and observe that the Hamiltonian function is conserved nodally.

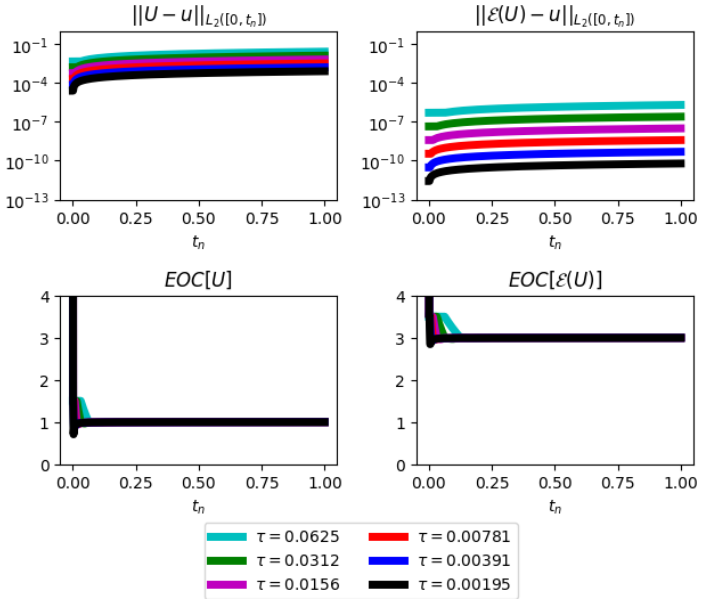
We investigate the error and experimental order of convergence in Figure 3.17, where we observe optimal behaviour for both $\boldsymbol{U}$ and $\mathcal{E}(\boldsymbol{U})$. Notice that the error for $\boldsymbol{U}$ is smaller here than in the nonadaptive case shown in Figure 3.15, but not considerably. The additional accuracy of the underlying adaptive method is best observed through the continuous reconstruction.

Figure 3.16: We examine the *nodal* deviation in the Hamiltonian function for the continuous reconstruction of RFEM (3.20) with the reconstruction given by Definition 3.2.4, $\mathcal{M} \equiv \mathcal{M}_0$, $\widetilde{\mathcal{M}} \equiv \mathcal{M}_1$, $p = 0$ and $q = 1$. We approximate numerical the harmonic oscillator given in Example 3.3.1, subject to the initial data $u_1(0) = 0.1, u_2(0) = 0.1$. For this simulation we employ the uniform time step $\tau = 0.1$. We observe that the Hamiltonian function is conserved nodally.



(a) Here $q = 0$.

Figure 3.17: We examine the error and EOC of RFEM (3.20), with $\mathcal{E}$ as given in Definition 3.2.4, $\mathcal{M} \equiv \mathcal{M}_0$, $\widetilde{\mathcal{M}} \equiv \mathcal{M}_1$, $p = 0$ and $q = 1$. We approximate the solution of the harmonic oscillator discussed in Example 3.3.1 subject to the initial data $u_1(0) = 1, u_2(0) = 1$. We measure errors in the $L_2([0, t_n])$ norm and plot the corresponding EOC, and notice optimal experimental convergence rates.



(a) Here $q = 0$.

3.4 Conclusion

We introduced the two finite element methods typically used for time integration, the cG method and upwind dG method, and discussed their geometric properties, as can be found in [105]. In addition, we conducted a stability analysis on these methods and proved an a priori error bound for the cG method.

We then introduced a new temporal finite element method, RFEM, as the first step towards a fully adaptive (discontinuous) space-time finite element method which possesses a structure preserving temporal reconstruction. We conducted a stability and a priori analysis of this method in addition to proposing a temporally adaptive algorithm which has the potential to be generalised into a space-time adaptive algorithm.

We observed that the upwind dG method was dissipative which, while useful for numerical stability, did not yield solutions which preserved any geometric structure over time. The cG method and RFEM both preserved energy over time for Hamiltonian ODEs. Further to this, RFEM generated a structure preserving solution while also possessing an underlying discontinuous solution, which lends itself to the generation of fully adaptive space-time numerical methods which preserve geometric structure over time.

Chapter 4

An introduction to Hamiltonian PDEs and their approximation

A fundamental challenge in the numerical analysis of partial differential equations is the development of consistent and conservative algorithms for Hamiltonian PDEs. These are a specific class of PDE which have physically relevant algebraic and geometric structures associated to them, see [95, 117]. They arise form a variety of areas, not least meteorological, such as the semi-geostrophic equations [155], and oceanographical equations, such as the Korteweg de-Vries (KdV) and nonlinear Schrödinger equations [144]. The KdV and Schrödinger equations are particularly special examples, in that they are bi-Hamiltonian. This means the equations have two different Hamiltonian formulations which, in turn, is one way to understand the notion of *integrability* of the problems. Regardless, the applications motivate the need for accurate long time simulations for reliable prediction of future behaviour in both meteorology and oceanography.

Within this chapter we propose a new methodology for obtaining a numerical solution which preserves the underlying structure of a Hamiltonian PDE, i.e., the "Hamiltonian functional" over time. We will then restrict our attention and investigate the design of a conservative numerical scheme for the *linearised* KdV equation, proving numerical stability and a new a priori error bound. In subsequent chapters we shall develop schemes for a multitude of more complex Hamiltonian PDEs. The method that we design for the linearised problem serendipitously falls within the framework for localised discontinuous Galerkin methods which have proven to be quite successful for linearised KdV, see [183, 182, 128, 98]. These methods have been found to superconverge at the nodes. The application of hybrid discontinuous Galerkin to this problem is currently under develop-

ment, with the first hybridised discontinuous Galerkin scheme for the stationary linear KdV problem being developed in [43] and also superconverge at the nodes.

4.1 Hamiltonian PDEs and a methodology for their discretisation

Let $u = u(t, x)$, where $t \in [0, T]$ and $x \in S^1$, where S^1 is the periodic interval $[0, 1)$, which can be thought of as a lower dimensional representation of a sphere. A Hamiltonian PDE can be written with respect to the Hamiltonian H as

$$u_t = -\mathcal{P}\frac{\delta}{\delta u}H(u) \tag{4.1}$$

where $\frac{\delta}{\delta u}$ denotes the variational derivative and $\mathcal{P}$ is a differential operator which induces a Poisson bracket, i.e., $\mathcal{P}$ defines the Poisson bracket

$$\{g_1(u), g_2(u)\}_{\mathcal{P}} := \left\langle \frac{\delta g_1}{\delta u}, \mathcal{P}\frac{\delta g_2}{\delta u} \right\rangle. \tag{4.2}$$

This bracket satisfies the skew-symmetry condition

$$\{g_1(u), g_2(u)\}_{\mathcal{P}} = -\{g_2(u), g_1(u)\}_{\mathcal{P}},$$

and distributivity, i.e. for $a, b \in \mathbb{R}$

$$\{ag_1(u) + bg_2(u), g_3(u)\}_{\mathcal{P}} = a\{g_1(u), g_3(u)\}_{\mathcal{P}} + b\{g_2(u), g_3(u)\}_{\mathcal{P}}.$$

It also satisfies the product rule

$$\{g_1(u)g_2(u), g_3(u)\}_{\mathcal{P}} = \{g_1(u), g_3(u)\}_{\mathcal{P}}g_2(u) + g_1(u)\{g_2(u), g_3(u)\}_{\mathcal{P}}$$

and the Jacobi identity

$$\{g_1(u), \{g_2(u), g_3(u)\}_{\mathcal{P}}\}_{\mathcal{P}} + \{g_2(u), \{g_3(u), g_1(u)\}_{\mathcal{P}}\}_{\mathcal{P}} + \{g_3(u), \{g_1(u), g_2(u)\}_{\mathcal{P}}\}_{\mathcal{P}} = 0.$$

Here, and in the sequel, $\langle \cdot, \cdot \rangle$ denotes the spatial L_2 inner product over S^1. Note that by skew-symmetry we have with $g_1(u) = g_2(u)$ that

$$\left\langle \frac{\delta g_1}{\delta u}, \mathcal{P} \frac{\delta g_1}{\delta u} \right\rangle = 0. \tag{4.3}$$

For clarity of exposition we have initially restricted the notion of a Hamiltonian PDE to scalar problems, but this is not always the case, in fact we shall study a *vectorial* Hamiltonian PDE in Chapter 8·

This structure is satisfied by many PDEs arising in physical situations not least, the KdV equation

$$\begin{aligned} u_t + 6uu_x + u_{xxx} &= 0, \\ u(0,x) &= u_0, \end{aligned} \tag{4.4}$$

with $u_0 = u_0(x)$ some sufficiently smooth function. This problem is *bi-Hamiltonian*, this means the problem can be written in two Hamiltonian forms, that is, there exist two distinct Hamiltonian operators $\mathcal{P}_1, \mathcal{P}_2$ and two different Hamiltonians H_1, H_2 such that

$$u_t = -\mathcal{P}_1 \frac{\delta}{\delta u} H_1(u) = -\mathcal{P}_2 \frac{\delta}{\delta u} H_2(u).$$

The KdV equation has a bi-Hamiltonian structure given by the Hamiltonians and differential operators

$$\begin{aligned} \mathcal{P}_1(\cdot) &= (\cdot)_x & H_1 &= \left\langle u^3 - \frac{u_x^2}{2}, 1 \right\rangle \\ \mathcal{P}_2(\cdot) &= (\cdot)_{xxx} + 4u(\cdot)_x + 2u_x & H_2 &= \frac{1}{2} \langle u, u \rangle . \end{aligned} \tag{4.5}$$

Note that more complex higher order Hamiltonian problems such as the Camassa-Holm (CH) equation also exist

$$\begin{aligned} m_t + um_x + 2mu_x = 0, \qquad & m = u - u_{xx}, \\ u(0,x) &= u_0, \end{aligned}$$

see [95, 13.1]. Both of these problems are *bi-Hamiltonian*, however the CH equation is a bi-Hamiltonian PDE with respect to the variable $m := u - u_{xx}$ with the Hamiltonian and

differential operators and functions

$$\mathcal{P}_1(\cdot) = (\cdot)_x - (\cdot)_{xxx} \qquad H_1 = \frac{1}{2}\left\langle u^2 + u_x^2, u\right\rangle$$
$$\mathcal{P}_2(\cdot) = m_x + m(\cdot)_x \qquad H_2 = \frac{1}{2}\left\langle u, u\right\rangle + \frac{1}{2}\left\langle u_x, u_x\right\rangle.$$

It should be pointed out that whether the problem at hand is of Hamiltonian or bi-Hamiltonian structure, the underlying Hamiltonians are conserved quantities, that is,

$$\frac{\mathrm{d}}{\mathrm{d}t}H(u) = \left\langle \frac{\delta}{\delta u}H(u), u_t \right\rangle = \{H(u), H(u)\}_{\mathcal{P}} = 0, \tag{4.6}$$

by (4.2) and (4.3).

Remark 4.1.1 (Choice of differential operators)**.** *Notice that in both KdV and CH problems one Hamiltonian is quadratic and one is cubic in the nonlinearity. We will focus on the development of schemes that conserve nonquadratic invariants. The reasons for this are twofold: Numerical schemes which preserve quadratic invariants are well developed, and our methodology on the discrete level is more succinct when the differential operator is linear, which is typically the case for nonquadratic invariants. For example, the differential operator $\mathcal{P}_2$ for KdV, as described in* (4.5) *is a nonlinear so more difficult to represent in the discrete setting. Conservation of the higher order invariants proves challenging both in the design of spatial discretisations but also for temporal discretisations since the "usual" geometric integrators that would be used in a method of lines approach typically based upon the Gauss-Radau family of Runge-Kutta schemes are* no longer conservative*, see §5.2.*

The methodology we propose for the construction of conservative numerical schemes for Hamiltonian problems is based on the observation that the argument in (4.6) requires $\frac{\delta}{\delta u}H(u)$ to be admissible as a test function, as u is an admissible test function. When performing these calculations on the PDE itself this is indeed the case, however when considering a numerical scheme $\frac{\delta}{\delta u}H(u)$ will not be admissible unless $H(u)$ is quadratic. The aforementioned methodology can be broken down as follows:

1. After the introduction of a diagnostic variable

$$v = \frac{\delta}{\delta u}H(u)$$

we can rewrite the Hamiltonian PDE (4.1) as the system

$$\begin{aligned} u_t + \mathcal{P}v &= 0 \\ v - \frac{\delta}{\delta u} H(u) &= 0. \end{aligned} \tag{4.7}$$

2. Multiplying (4.7) by v and u_t respectively and integrating over the spatial domain we find

$$\begin{aligned} \langle u_t + \mathcal{P}v, v \rangle &= 0 \\ \left\langle u_t, v - \frac{\delta}{\delta u} H(u) \right\rangle &= 0. \end{aligned} \tag{4.8}$$

Due to the skew-symmetric structure of $\mathcal{P}$ highlighted in (4.3) we have from the first equation of (4.8) that

$$\langle u_t, v \rangle = 0.$$

This permits us to simplify the second equation of (4.8) yielding

$$\begin{aligned} 0 =& \left\langle u_t, \frac{\delta}{\delta u} H(u) \right\rangle \\ =& \frac{\mathrm{d}}{\mathrm{d}t} \langle H(u), 1 \rangle . \end{aligned}$$

3. The technique may appear to add additional unnecessary complications, which at the PDE level is true, however upon discretisation of the system (4.8) we are able to mimic steps 1 and 2 on the discrete level to obtain a conservative scheme, assuming that the discrete Poisson bracket is skew-symmetric. In the sequel, we will apply this methodology to KdV with an aim to conserve the first Hamiltonian H_1 which physically represents the energy. This Hamiltonian represents the energy as it is associated with a time translation (Lie) symmetry via Noether's theorem, see [145, 147].

4.2 The linearised KdV equation

Examining the leading order asymptotic behaviour of the KdV equation (4.4) we find, after rescaling, that a linearised KdV equation can be described by the PDE

$$u_t + u_x + u_{xxx} = 0, \tag{4.9}$$

subject to some appropriate initial data $u(0, x) = u_0(x)$. Seeking a travelling wave solution, i.e., assuming that $u(t, x) = f(\zeta)$ where $\zeta = x - ct$ for some constant c representing the speed of the solution, we can rewrite (4.9) as the ODE

$$-\frac{1}{c} f'(\zeta) + f'(\zeta) + f^{(3)}(\zeta) = 0,$$

where $f'(\zeta)$ and $f^{(3)}(\zeta)$ represent the first and third derivatives with respect to ζ. Integrating allows us to write

$$\left(1 - \frac{1}{c}\right) f(\zeta) + f^{(2)}(\zeta) = C_1,$$

where C_1 is an arbitrary constant. We find, due to the enforcement of periodic boundaries, that an exact solution of the problem has the form

$$u(t, x) = C_1 + C_2 \sin\left(\alpha\left(x - (1 - \alpha^2)t\right)\right) + C_3 \cos\left(\alpha\left(x - (1 - \alpha^2)t\right)\right), \tag{4.10}$$

where $\alpha = 2\pi k$ for $k \in \mathbb{Z}$ and $c = 1 - \alpha^2$. While exact travelling wave solutions are obtainable, we will use the linearised KdV equation as a prototypical example affording us insight into the analysis of our numerical schemes for nonlinear KdV type equations and Hamiltonian PDEs in general.

4.2.1 Finite element notation

As is often the case, before introducing our numerical scheme for linearised KdV we are required to introduce a plethora of notation. Recall that our spatial domain is the periodic unit interval S^1, then we partition our domain such that $0 =: x_0 < x_1 < \cdots < x_M := 1$. We define a *spatial finite element* as $\mathcal{J}_m := (x_m, x_{m+1})$ which possesses an element length denoted $h_m := x_{m+1} - x_m$. Throughout we refer to continuous functions in lower case Roman letters, i.e. $u = u(t, x)$, and spatially discrete functions as upper case Roman letters, i.e., $U = U(t, x)$. When there is no ambiguity, for clarity of exposition, we shall

not explicitly write the dependencies of functions. Unless stated otherwise, all functions have the input arguments (t, x).

Definition 4.2.1 (Spatial finite element spaces)**.** *Let $\mathbb{P}_q(\mathcal{J}_m)$ denote the space of polynomials of degree q on the element $\mathcal{J}_m$, then the* discontinuous *finite element space is given by*

$$\mathbb{V}_q = \{U : S^1 \to \mathbb{R} : U|_{\mathcal{J}_m} \in \mathbb{P}_q(\mathcal{J}_m) \text{ for } m = 0, ..., M-1\}.$$

Further to this we define the continuous *finite element space as*

$$\mathbb{V}_q^C = \mathbb{V}_q \cap C^0\left(S^1\right),$$

where $C^0\left(S^1\right)$ denotes the space of continuous functions.

Similarly, for the temporal case presented in §3.1, we define the mesh function $h \in \mathbb{V}_0$ as the piecewise constant finite element function representing the length of an element, i.e.,

$$h|_{\mathcal{J}_m} = h_m.$$

We will additionally write the largest element as

$$h_{max} := \max_x h = \max_{m \in [0, M-1]} h_m,$$

and define the largest element within a local "patch" of elements as

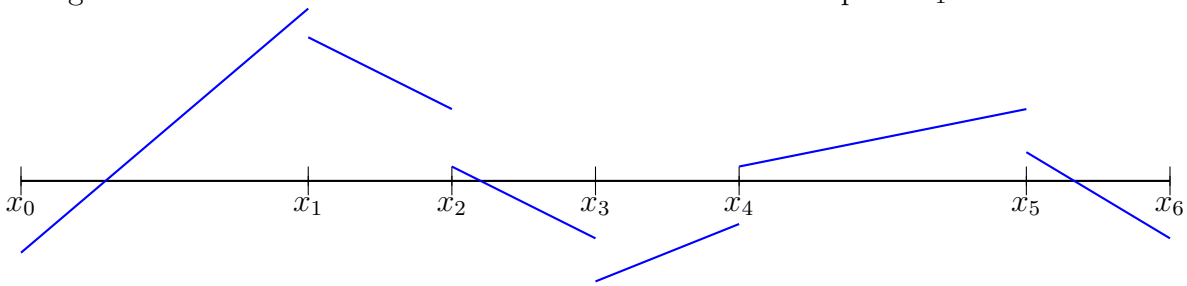
$$\widetilde{h_m} := \max_{i \in [m-1, m, m+1]} h_i. \tag{4.11}$$

In addition, throughout this work we restrict our minimal and maximal element such that

$$h_{max} \leq C h_{min},$$

for some constant C independent of h_m.
As we discussed in Chapter 3, in the development of discontinuous finite element schemes it is paramount to enforce some kind of communication between elements. This is performed through fluxes at the nodes of the finite elements as given in Definition 3.1.11.
Recall we refer to the spatial L_2 inner product as $\langle \cdot, \cdot \rangle$, we shall refer to temporal integrals using standard integral notation to avoid confusion when integrating over both space and

Figure 4.1: An illustrative finite element function in the space $\mathbb{V}_1$ where $M = 6$.



time.

Remark 4.2.2 (A clarifying remark about notation)**.** *Note that our spatial L_2 inner product $\langle \cdot, \cdot \rangle$ is defined over the whole spatial domain but does not include the endpoints of elements, i.e., for some $f, g \in L_2(S^1)$*

$$\langle f, g \rangle := \sum_{m=0}^{M-1} \int_{\mathcal{J}_m} fg \mathrm{d}x.$$

We define $\langle \cdot, \cdot \rangle$ this way as our discontinuous finite element functions do not exist at the endpoints of the elements, so when integrating over the whole of S^1 spatial derivatives of discontinuous functions are not well defined.

4.2.2 Development of a spatially discrete scheme

The linearisation of the KdV equation is in its own right a (bi-)Hamiltonian PDE and can be defined through the coupled differential operators and Hamiltonian functions

$$\mathcal{P}_1(\cdot) = (\cdot)_x \qquad H_1 = \frac{1}{2}\langle u, u \rangle - \frac{1}{2}\langle u_x, u_x \rangle$$
$$\mathcal{P}_2(\cdot) = (\cdot)_{xxx} + (\cdot)_x \qquad H_2 = \frac{1}{2}\langle u, u \rangle ,$$

respectively. In line with the methodology outlined in the prequel we introduce the first variation of the Hamiltonian H_1 as an auxiliary variable v, allowing us to rewrite linearised KdV as the system

$$u_t + v_x = 0$$
$$v - u - u_{xx} = 0.$$

Note that the first variation of the second Hamiltonian is simply u, which suggests it may be possible to design a scheme which preserves *both* of the Hamiltonians over time. The difficulty in designing a conservative scheme following the methodology in the prequel is the design of the discrete Poisson bracket. To discretise the first differential operator we need to mimic the following argument on the discrete level

$$\langle v, \mathcal{P}_1 v \rangle = \langle v, v_x \rangle = 0,$$

where $\langle \cdot, \cdot \rangle$ denotes the spatial inner product which induces the L_2 norm. That is to say we require that our discrete differential operator preserves the skew-symmetry of the continuous operator. This leads us to defining the following first derivative operator.

Definition 4.2.3 (Discrete operator for first spatial derivatives)**.** *Let $W \in \mathbb{V}_q$, then $\mathcal{G} : \mathbb{V}_q \to \mathbb{V}_q$ such that*

$$\langle \mathcal{G}(W), \phi \rangle = \langle W_x, \phi \rangle - \sum_{m=0}^{M-1} [\![W_m]\!] \{\phi_m\} \qquad \forall \phi \in \mathbb{V}_q,$$

where the jump $[\![\cdot]\!]$ and average $\{\cdot\}$ are the spatial equivalent of the jump and average given in Definition 3.1.11. Similar operators for first derivatives are defined in [31] and [78].

Lemma 4.2.4 (Integration by parts with the discrete operator for first spatial derivatives)**.** *Let $U, V \in \mathbb{V}_q$, then the finite element operator for first spatial derivatives described in Definition 4.2.3 possesses the discrete integration by parts identity*

$$\langle \mathcal{G}(U), V \rangle = - \langle U, \mathcal{G}(V) \rangle .$$

This allows us to conclude that the spatial inner product of $\mathcal{G}$ is indeed a skew-symmetric operation.

Proof. Applying Definition 4.2.3, and integration by parts, we see that

$$\begin{aligned}\langle \mathcal{G}(U), V\rangle &= \langle U_x, V\rangle - \sum_{m=0}^{M-1} [\![U_m]\!]\{V_m\} \\ &= -\langle U, V_x\rangle + \sum_{m=0}^{M-1} [\![U_m V_m]\!] - [\![U_m]\!]\{V_m\} \\ &= -\langle U, V_x\rangle + \sum_{m=0}^{M-1} \{U_m\}[\![V_m]\!] \\ &= -\langle U, \mathcal{G}(V)\rangle,\end{aligned}$$

through integration by parts, and observing that

$$\begin{aligned}[\![U_m V_m]\!] - [\![U_m]\!]\{V_m\} &= \left(U_m^- V_m^- - U_m^+ V_m^+\right) - \frac{1}{2}\left(U_m^- - U_m^+\right)\left(V_m^- + V_m^+\right) \\ &= \frac{1}{2}\left(U_m^- + U_m^+\right)\left(V_m^- - V_m^+\right) \\ &= \{U_m\}[\![V_m]\!],\end{aligned}$$

through the definitions of the jump and the average (3.6) and (3.7) respectively. □

Discretising the second differential operator is more tricky, as the operator needs to consist of both a first and third derivative, and satisfy the skew-symmetry condition, i.e.,

$$\langle v, \mathcal{P}_2 v\rangle = \langle v, v_x\rangle + \langle v, v_{xxx}\rangle = \frac{1}{2}\left\langle \left(v^2\right)_x, 1\right\rangle - \frac{1}{2}\left\langle \left(v_x^2\right)_x, 1\right\rangle = 0,$$

after integration by parts. Our methodology for the design of this discrete operator follows from the discrete design of $\mathcal{P}_1$. We will introduce an *additional* auxiliary variable representing the first spatial derivative of the solution allowing us to write the linearised KdV equation as a system of PDEs with at most one spatial derivative, i.e.,

$$\begin{aligned} u_t + v_x &= 0 \\ v - u - w_x &= 0 \\ w - u_x &= 0. \end{aligned} \tag{4.12}$$

This allows us to recursively employ $\mathcal{G}$ for the discrete version of the differential operator. With the design of discrete differential operators in mind, we introduce the following spatially discrete method.

Definition 4.2.5 (Spatially discrete scheme for the linearised KdV equation)**.** *Our spatially discrete scheme for the linearised KdV equation is given by seeking $U, V, W \in \mathbb{V}_q$ such that*

$$\begin{aligned} \langle U_t + \mathcal{G}(V), \phi \rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q \\ \langle V - U - \mathcal{G}(W), \psi \rangle &= 0 \qquad \forall \psi \in \mathbb{V}_q \\ \langle W - \mathcal{G}(U), \chi \rangle &= 0 \qquad \forall \chi \in \mathbb{V}_q, \end{aligned} \tag{4.13}$$

subject to the initial data $U(0,x) = \Pi u_0(x)$ where Π denotes the L_2 projection into the finite element space, and where $\mathcal{G}$ is given by Definition 4.2.3. Note that through the initial data for U we can obtain initial data for the auxiliary variables V, W, although we shall not delve into this intricacy until the introduction of the fully discrete scheme.

Remark 4.2.6 (A local discontinuous Galerkin method)**.** *The numerical scheme described in Definition 4.2.5 falls within the framework of local discontinuous Galerkin* (LDG) *methods, in which the auxiliary variables V and W are removed and the method is implemented in* primal form*, see [183, 98]. As such, the computational complexity of this method is not increased by orders of magnitude through the introduction of diagnostic variables. Rewriting the scheme in primal form does however increase the stencil, which we expect for a standard discretisation of higher order derivatives. In this work we do not implement the scheme within the LDG framework to keep the exposition as simple as possible to convey the main ideas.*

Theorem 4.2.7 (Conserved quantities for the spatially discrete scheme for linearised KdV)**.** *Let U, V, W be the numerical solution of the spatially discrete scheme for linearised KdV as described in Definition 4.2.5, then the discrete mass*

$$\mathcal{F}_1(U) = \langle U, 1 \rangle,$$

the discrete momentum

$$\mathcal{F}_2(U) = \frac{1}{2} \langle U, U \rangle$$

and the discrete energy

$$\mathcal{F}_3(U, W) = \frac{1}{2} \langle W, W \rangle - \frac{1}{2} \langle U, U \rangle$$

are preserved over time. That is to say that $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_1(U) = 0$, $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_2(U) = 0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_3(U, W) = 0$.

Proof. We begin by considering conservation of mass. We have that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_1(U) = \langle U_t, 1\rangle = \langle \mathcal{G}(V), 1\rangle = 0$$

through choosing $\phi = 1$ in (4.13), and then observing that $\mathcal{G}(V)$ is orthogonal to constants. We can either observe this orthogonality directly through the definition of $\mathcal{G}$ or we can view it as a consequence of Lemma 4.2.4 as $\mathcal{G}(1) = 0$.
To observe conservation of momentum and energy we will follow the methodology outlined in the previous section. However the proofs are not quite as simple due to the introduction of another auxiliary variable as a methodology of discretising the second differential operator $\mathcal{P}_2$. In view of conservation of momentum we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_2(U) &= \langle U_t, U\rangle \\ &= -\langle \mathcal{G}(V), U\rangle \\ &= \langle V, \mathcal{G}(U)\rangle\end{aligned}$$

through choosing $\phi = U$ in (4.13) and Lemma 4.2.4. Further choosing $\chi = V$, $\psi = W$ and then $\chi = U$ we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_2(U) &= \langle V, W\rangle \\ &= \langle U + \mathcal{G}(W), W\rangle \\ &= \langle U, \mathcal{G}(U)\rangle + \langle \mathcal{G}(W), W\rangle \\ &= 0,\end{aligned}$$

through the skew-symmetry of the spatial inner product induced by $\mathcal{G}$. With respect to energy conservation we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_3(U, W) &= \langle W_t, W\rangle - \langle U_t, U\rangle \\ &= \langle \mathcal{G}(U_t), W\rangle - \langle U_t, U\rangle,\end{aligned}$$

after choosing $\chi = W$ in the *temporal* derivative of (4.13). Note that as time is continuous, and our test functions are purely spatial, taking the temporal derivative of the spatial

numerical scheme is a valid operation. Applying Lemma 4.2.4 we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_3(U,W) &= \langle U_t, \mathcal{G}(W) - U\rangle \\ &= \langle U_t, V\rangle,\end{aligned}$$

after choosing $\psi = U_t$ in (4.13). Finally choosing $\phi = V$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_3(U,W) = \langle \mathcal{G}(V), V\rangle = 0,$$

and energy is conserved. □

Remark 4.2.8 (A continuous version of the spatially discrete scheme)**.** *If we enforced continuity on our trial and test finite element spaces in the spatial scheme for linearised KdV, Theorem 4.2.7 would still hold as all choices of test functions are equally valid when the finite element space is global continuous. This would not be the case if we needed to test with spatial derivatives of functions at any point. However, even though the continuous method possesses fewer degrees of freedom (through the enforcement of continuity) than its discontinuous counterpart, it cannot be implemented in primal form, see Remark 4.2.6. As such, the continuous method has a significantly higher computational complexity than the discontinuous method. In the sequel we shall focus on the discontinuous method for not only this reason, but also because the discontinuous method contains within it the continuous method, and more importantly, the a priori analysis we discuss in the sequel requires discontinuity.*

Remark 4.2.9 (Uniqueness of the spatially discrete scheme)**.** *Unfortunately our spatially discrete scheme described in Definition 4.2.5 is* not *unique for odd order polynomial degrees. This is due to the kernel of the operator $\mathcal{G}$ having dimension* 2 *for odd polynomial degrees, see [78]. Fortunately for even order polynomial degrees we have uniqueness as the dimension of the kernel of $\mathcal{G}$ is* 1*. As such, the analysis of this numerical scheme in the sequel is only valid for* even polynomial degrees*. Note that similar phenomena are observed in [31].*

It is possible to resolve this issue. Taking inspiration from [78] by introducing upwind and

downwind discrete gradients $\mathcal{G}^+, \mathcal{G}^- \in \mathbb{V}_q$ *such that*

$$\begin{aligned} \left\langle \mathcal{G}^+(W), \phi \right\rangle &= \langle W_x, \phi \rangle - \sum_{m=0}^{M-1} [\![W_m]\!] \phi_m^+ \qquad \forall \phi \in \mathbb{V}_q \\ \left\langle \mathcal{G}^-(W), \phi \right\rangle &= \langle W_x, \phi \rangle - \sum_{m=0}^{M-1} [\![W_m]\!] \phi_m^- \qquad \forall \phi \in \mathbb{V}_q, \end{aligned} \tag{4.14}$$

respectively. With these new operators in mind we can examine the spatially discrete scheme by seeking $U, V, W \in \mathbb{V}_q$ *such that*

$$\begin{aligned} \left\langle U_t + \mathcal{G}^+(V), \phi \right\rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q \\ \left\langle V - U - \mathcal{G}^-(W), \psi \right\rangle &= 0 \qquad \forall \psi \in \mathbb{V}_q \\ \left\langle W - \mathcal{G}^-(U), \chi \right\rangle &= 0 \qquad \forall \chi \in \mathbb{V}_q, \end{aligned} \tag{4.15}$$

subject to appropriate initial data. Before continuing with our analysis on the linearised scheme proposed in Definition 4.2.5 we shall investigate the stability properties of (4.15).

Lemma 4.2.10 (Stability of the alternative scheme for linearised KdV (4.15))**.** *Let* U *be the solution of* (4.15) *subject to the initial data* $U(0,x) = \Pi u_0$*, where* Π *is the* L_2 *projection into the finite element space. Then* U *is stable over time, i.e.,*

$$\|U(t,x)\|^2_{L_2(S^1)} \leq \|U(0,x)\|^2_{L_2(S^1)}.$$

Before proving Lemma 4.2.10 we shall present several useful properties of the operators $\mathcal{G}^+$ and $\mathcal{G}^-$.

Proposition 4.2.11 (Properties of $\mathcal{G}^+$ and $\mathcal{G}^-$)**.** *Let* $\mathcal{G}^+$ *and* $\mathcal{G}^-$ *be as defined in* (4.14)*. Further let* $V, W \in \mathbb{V}_q$*, then the integration by parts identity*

$$\left\langle \mathcal{G}^+(V), W \right\rangle = -\left\langle V, \mathcal{G}^-(W) \right\rangle \tag{4.16}$$

holds. While $\mathcal{G}^+$ *and* $\mathcal{G}^-$ *are not conservative operators in the sense that they are orthogonal to their arguments, we do have that for* $W \in \mathbb{V}_q$

$$\left\langle \mathcal{G}^+(W), W \right\rangle \geq 0, \tag{4.17}$$

and

$$\left\langle \mathcal{G}^-(W), W \right\rangle \leq 0. \tag{4.18}$$

Proof. To show (4.16) we let $V, W \in \mathbb{V}_q$, then through the definitions of $\mathcal{G}^+$ and $\mathcal{G}^-$ we observe that

$$\begin{aligned}\left\langle \mathcal{G}^+(V), W \right\rangle &:= \langle V_x, W\rangle - \sum_{m=0}^{M-1} [\![V_m]\!] W_m^+ = -\langle V, W_x\rangle + \sum_{m=0}^{M-1} \left([\![VW]\!] - [\![V_m]\!] W_m^+ \right) \\ &= -\langle V, W_x\rangle + \sum_{m=0}^{M-1} [\![W]\!] W_m^- =: -\left\langle V, \mathcal{G}^-(W) \right\rangle,\end{aligned}$$

through integration by parts and the definition of the jump operator (3.6). In view of (4.17) we observe, through the application of Stokes theorem, that

$$\begin{aligned}\left\langle \mathcal{G}^+(W), W \right\rangle &= \left\langle \frac{1}{2}\left(W^2\right)_x, 1 \right\rangle - \sum_{m=0}^{M-1} [\![W_m]\!] W_m^+ \\ &= \sum_{m=0}^{M-1} \left(\frac{1}{2} [\![W_m^2]\!] - [\![W_m]\!] W_m^+ \right).\end{aligned} \tag{4.19}$$

Cauchy's inequality tells us that

$$-W_m^+ W_m^- \geq -\frac{1}{2}\left(W_m^+\right) - \frac{1}{2}\left(W_m\right)^-,$$

allowing us to write that

$$-[\![W_m]\!] W_m^+ := -\left(W_m^+\right)^2 - W_m^+ W_m^- \geq \frac{1}{2}\left(W_m^+\right) - \frac{1}{2}\left(W_m^-\right) = \frac{1}{2}[\![W_m^2]\!]. \tag{4.20}$$

Applying (4.20) to (4.19) we observe that

$$\left\langle \mathcal{G}^+(W), W \right\rangle \geq 0,$$

as required. The final result (4.18) follows directly through the amalgamation of (4.16) and (4.17).

□

Proof of Lemma 4.2.10. To prove Lemma 4.2.10 we solely require that the discrete momentum *dissipates*. As such, we mimic the steps used in the proof of Theorem 4.2.7. After

choosing $\phi = U$ in (4.15) we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_2(U) &= -\left\langle \mathcal{G}^+(V), U\right\rangle \\ &= \left\langle V, \mathcal{G}^-(U)\right\rangle,\end{aligned}$$

after application of (4.16). Further choosing $\psi = \mathcal{G}^-(U)$ and then $\chi = \mathcal{G}^+(\mathcal{G}^-(U))$ we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_2(U) &= \left\langle U, \mathcal{G}^-(U)\right\rangle + \left\langle \mathcal{G}^-(W), \mathcal{G}^-(U)\right\rangle \\ &= \left\langle U, \mathcal{G}^-(U)\right\rangle - \left\langle W, \mathcal{G}^+\left(\mathcal{G}^-(U)\right)\right\rangle \\ &= \left\langle U, \mathcal{G}^-(U)\right\rangle - \left\langle \mathcal{G}^-(U), \mathcal{G}^+\left(\mathcal{G}^-(U)\right)\right\rangle.\end{aligned}$$

Through (4.17) and (4.18), as outlined in Proposition 4.2.11, we observe that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_2(U) \leq 0.$$

As such, applying the fundamental theorem of calculus over time we observe that

$$\mathcal{F}_2(U(t,x)) \leq \mathcal{F}_2(U(0,x)),$$

as required.

□

Remark 4.2.12 (Dissipation of energy in the alternative spatially discrete scheme for linearised KdV)**.** *As we found in Lemma 4.2.10, the alternative spatial finite element scheme proposed in Remark 4.2.9 does not preserve a discrete momentum, and is dissipative in nature. Due to the dissipative nature of the scheme we also do not conserve a discrete energy. However, we do* not *have that the energy functional $\mathcal{F}_3(U,W)$ is dissipative over time.*

We shall now refocus our attention on the spatially discrete scheme described in Definition 4.2.5.

Lemma 4.2.13 (Stability of the spatial scheme for linearised KdV)**.** *Let U, V, W be the solution of the spatial scheme for linearised KdV as described in Definition 4.2.5. Further assume that the initial momentum and energy are bounded, i.e., $\mathcal{F}_2(U(0,x)) < \infty$ and*

$\mathcal{F}_3(U(0,x),W(0,x)) < \infty$, *then*

$$\|U(t,x)\|^2_{L_2(S^1)} = \|U(0,x)\|^2_{L_2(S^1)}$$

and

$$\|W(t,x)\|^2_{L_2(S^1)} = \|W(0,x)\|^2_{L_2(S^1)},$$

for $t \in [0,T]$.

Proof. Recall in Theorem 4.2.7 we proved that the discrete momentum $\frac{1}{2}\langle U,U\rangle$ and discrete energy $\frac{1}{2}\langle W,W\rangle - \frac{1}{2}\langle U,U\rangle$ are preserved over time. Conservation of momentum immediately gives us control over the solution in L_2, i.e.,

$$\|U(t,x)\|^2_{L_2(S^1)} = \|U(0,x)\|^2_{L_2(S^1)}.$$

Additionally conservation of energy affords us control over the solution in a discrete H^1 norm, as

$$\|W(t,x)\|^2_{L_2(S^1)} - \|U(t,x)\|^2_{L_2(S^1)} = \|W(0,x)\|^2_{L_2(S^1)} - \|U(0,x)\|^2_{L_2(S^1)}$$

allows us to write

$$\|W(t,x)\|^2_{L_2(S^1)} = \|W(0,x)\|^2_{L_2(S^1)}$$

after application of the conservation of momentum. Note that as the energy is not sign definite we cannot conclude stability in a H^1 norm without conservation of energy *and* momentum.

□

Theorem 4.2.14 (A priori error bound for the spatially discrete scheme for linearised KdV)**.** *Let* u,v,w *be the exact solution to the linearised KdV system* (4.12) *subject to appropriate initial data and enforce the regularity* $u,w \in C^2([0,T],H^{q+1}(S^1))$ *and* $v \in C^1([0,T],H^{q+1}(S^1))$ *on the exact solution. Additionally let* U,V,W *be the finite element approximation of this system as described in Definition 4.2.5, and the polynomial degree be even, then the following bound is satisfied*

$$\|u(t,x)-U(t,x)\|^2_{L_2(S^1)} + \|v(t,x)-V(t,x)\|^2_{L_2(S^1)} + \|w(t,x)-W(t,x)\|^2_{L_2(S^1)} \leq h_{max}^{2q+2}\gamma(t),$$

where

$$\gamma(t) := C\left(\beta(t) + |v|^2_{H^{q+1}(S^1)} + \int_0^t \exp(s)\,\beta(s)\mathrm{d}s\right),$$

for $\beta(t)$ as given in (4.24)*, or more succinctly can be increased slightly to be written as,*

$$\begin{aligned}\beta(t) := C\max_{[0,t]}\left(\max_{i=0,1,2}\left(\left|z^{(i)}\right|^2_{H^{q+1}(S^1)} + \left|z^{(i)}_t\right|^2_{H^{q+1}(S^1)}\right)\right)\\ + C\int_0^t \max_{i=0,1,2}\left(\left|z^{(i)}\right|^2_{H^{q+1}(S^1)}\right) + \max_{i=0,2}\left(\left|z^{(i)}_s\right|^2_{H^{q+1}(S^1)} + \left|z^{(i)}_{ss}\right|^2_{H^{q+1}(S^1)}\right)\mathrm{d}t,\end{aligned}$$

with $z^0 = u$, $z^1 = v$ and $z^2 = w$. Further C is constant and h_{max} is the size of the largest spatial element.

We shall defer the proof of Theorem 4.2.14 to later in this section while we introduce the projection operators and lemmas needed to present the proof concisely. The remainder of §4.2.2 is dedicated to setting up and subsequently proving Theorem 4.2.14. We begin by introducing a discrete projection operator, proving that it converges with respect to the exact solution, and then showing that it converges with respect to the discrete solution through energy arguments.

Definition 4.2.15 (Projection operator)**.** *Let $u \in H^1(S^1) \oplus \mathbb{V}_q$, then we define the projection operator $\mathcal{S} : H^1(S^1) \oplus \mathbb{V}_q \to \mathbb{V}_q$ such that*

$$\begin{aligned}\langle \mathcal{S}(u), \psi\rangle &= \langle u, \psi\rangle \qquad && \forall \psi \in \mathbb{V}_{q-1}\\ \{\mathcal{S}(u)_m\} &= \{u_m\} && \text{for } m = 0, .., M.\end{aligned}$$

Note that this projection operator is similar to those defined in [31, 78], and through counting the number of degrees of freedom we observe that the projection operator is uniquely defined.

Remark 4.2.16 (Interplay between $\mathcal{S}$ and $\mathcal{G}$)**.** *The projection operator $\mathcal{S}$ acts like the identity function under application of $\mathcal{G}$, i.e.,*

$$\mathcal{G}(\mathcal{S}(u)) = \mathcal{G}(u).$$

We can observe this directly through the definition of $\mathcal{S}$, as

$$\begin{aligned}\langle \mathcal{G}(\mathcal{S}(u)), \phi\rangle &= -\langle \mathcal{S}(u), \mathcal{G}(\phi)\rangle \\ &= -\langle \mathcal{S}(u), \phi_x\rangle + \sum_{m=0}^{M-1} [\![\phi]\!] \{\mathcal{S}(u)\} \\ &= -\langle u, \phi_x\rangle + \sum_{m=0}^{M-1} [\![\phi]\!] \{u\} \\ &= -\langle u, \mathcal{G}(\phi)\rangle \\ &= \langle \mathcal{G}(u), \phi\rangle,\end{aligned} \tag{4.21}$$

after utilising Lemma 4.2.4.

Remark 4.2.17 (A priori bound for a globally *continuous* spatial finite element method)**.** *Recall in Remark 4.2.8 we stated that we cannot obtain a priori bounds if we assume that our spatially discrete finite element method is globally continuous. The reason for this is we cannot design a continuous operator satisfying* (4.21)*. In fact, if we could design such an operator the remainder of our a priori analysis would follow for the continuous method via an equivalent argument to that in the sequel.*

Lemma 4.2.18 (Error bounds for $\mathcal{S}$,[78])**.** *Let $\boldsymbol{u} \in H^1(S^1)$, and $\mathcal{S} : H^1(S^1) \oplus \mathbb{V}_q \to \mathbb{V}_q$ be the projection operator as given in Definition 4.2.15, then we have that*

$$\|\mathcal{S}(u) - u\|_{L_2(S^1)} \leq C \left| h^{q+1} u \right|_{H^{q+1}(S^1)},$$

where $h \in \mathbb{V}_0$ is the spatial step size of functions in the finite element space $\mathbb{V}_q$.

Proof. See the proof of Lemma 8 in [78]. □

A key component in proving Theorem 4.2.14 will be splitting the error with the discrete quantity $\mathcal{S}$, to be more concise we will rewrite the error in each component of the solution as

$$\begin{aligned}\mathsf{e}_u &:= u - U = (u - \mathcal{S}(u)) + (\mathcal{S}(u) - U) =: \rho^u + \theta^u \\ \mathsf{e}_v &:= v - V = (v - \mathcal{S}(v)) + (\mathcal{S}(v) - V) =: \rho^v + \theta^v \\ \mathsf{e}_w &:= w - W = (w - \mathcal{S}(w)) + (\mathcal{S}(w) - W) =: \rho^w + \theta^w.\end{aligned} \tag{4.22}$$

Notice that ρ^u, ρ^v, ρ^w is simply the error of the projection operator $\mathcal{S}$, which we have already quantified Lemma 4.2.18. We will now quantify the error between the two discrete

objects, i.e, $\theta^u, \theta^v, \theta^w$.

Theorem 4.2.19 (The error between the projection operator $\mathcal{S}$ and the spatially discrete scheme for linearised KdV)**.** *Let u, v, w be the exact solution to the linearised KdV system* (4.12) *subject to appropriate initial data and enforce the regularity $u, w \in C^2([0,T], H^{q+1}(S^1))$ and $v \in C^1([0,T], H^{q+1}(S^1))$ on the exact solution. Additionally let U, V, W be the finite element approximation of this system as described in Definition 4.2.5. Further let $\mathcal{S}$ be the projection operator as described in Definition 4.2.15, and ρ, θ as defined in* (4.22)*, and* the polynomial degree q be even (for uniqueness of $\mathcal{G}$)*, then*

$$\|\theta^u\|^2_{L_2(S^1)} + \|\theta^v\|^2_{L_2(S^1)} + \|\theta^w\|^2_{L_2(S^1)} \leq h^{2q+2}_{max} \left(\beta(t) + \int_0^t 12 \exp(s)\, \beta(s) \mathrm{d}s \right), \tag{4.23}$$

where

$$\begin{aligned} \beta(t) = C \Bigg(& |u(0,x)|^2_{H^{q+1}(S^1)} + |v(0,x)|^2_{H^{q+1}(S^1)} + |w(0,x)|^2_{H^{q+1}(S^1)} + |u_t(0,x)|^2_{H^{q+1}(S^1)} \\ & + |w_t(0,x)|^2_{H^{q+1}(S^1)} + |u|^2_{H^{q+1}(S^1)} + |v|^2_{H^{q+1}(S^1)} + |u_t|^2_{H^{q+1}(S^1)} + |w_t|^2_{H^{q+1}(S^1)} \\ & + \int_0^t |u|^2_{H^{q+1}(S^1)} + |v|^2_{H^{q+1}(S^1)} + |w|^2_{H^{q+1}(S^1)} + |u_s|^2_{H^{q+1}(S^1)} \\ & + |v_s|^2_{H^{q+1}(S^1)} + |w_s|^2_{H^{q+1}(S^1)} + |u_{ss}|^2_{H^{q+1}(S^1)} + |w_{ss}|^2_{H^{q+1}(S^1)} \, \mathrm{d}s \Bigg), \end{aligned} \tag{4.24}$$

with the constant C depending on the constant given in Lemma 4.2.18.

To prove Theorem 4.2.19 we will utilise Gronwall's integral inequality as a fundamental tool in the quantification of error propagation over time, as in [176]. For completeness we shall present this inequality before proceeding with the proof.

Lemma 4.2.20 (Gronwall's lemma, [83, 27])**.** *Suppose that $f(t)$ satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t} f(t) \leq \int_0^t g(s) f(s) \mathrm{d}s + h(t),$$

where $f(t)$ is differentiable, $g(t)$ is continuous and $h(t)$ is integrable, and all three functions are non-negative, then

$$f(t) \leq h(t) + \int_0^t \exp\left(\int_s^t g(\tau) \mathrm{d}\tau \right) g(s) h(s) \mathrm{d}s.$$

While we shall not explicitly use it here, we also present the differential version of Gronwall's inequality for use in the sequel.

Lemma 4.2.21 (Gronwall's lemma - differential version, [83, 27])**.** *Suppose that $f(t)$ satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}f(t) \leq g(t)f(t) + h(t),$$

where $f(t)$ is differentiable, $g(t)$ is continuous and $h(t)$ is integrable, and all three functions are non-negative, then

$$f(t) \leq \exp\left(\int_0^t g(s)\mathrm{d}s\right)\left(f(0) + \int_0^t h(s)\mathrm{d}s\right).$$

Proof of Theorem 4.2.19. Before proving the error bound (4.23) we first need to examine which linearised KdV-like problem $\mathcal{S}(u), \mathcal{S}(v), \mathcal{S}(w)$ solve. With this is mind we find for all $\phi \in \mathbb{V}_q$ that

$$\begin{aligned}\langle \mathcal{S}(u)_t + \mathcal{G}(\mathcal{S}(v)), \phi\rangle &= \langle \mathcal{S}(u)_t + \mathcal{G}(v), \phi\rangle \\ &= \langle \mathcal{S}(u)_t - u_t + \mathcal{G}(v) - \mathcal{G}(v), \phi\rangle \\ &= -\langle \rho_t^u, \phi\rangle,\end{aligned} \tag{4.25a}$$

through application of Remark 4.2.16 and by subtracting the first equation of the PDE (4.12). Similarly we find for all $\psi \in \mathbb{V}_q$

$$\begin{aligned}\langle \mathcal{S}(v) - \mathcal{S}(u) - \mathcal{G}(\mathcal{S}(w)), \psi\rangle &= \langle \mathcal{S}(v) - v - \mathcal{S}(u) + u - \mathcal{G}(w) + \mathcal{G}(w), \psi\rangle \\ &= \langle \rho^u - \rho^v, \psi\rangle,\end{aligned} \tag{4.25b}$$

and for all $\chi \in \mathbb{V}_q$

$$\langle \mathcal{S}(w) - \mathcal{G}(\mathcal{S}(u)), \chi\rangle = \langle \mathcal{S}(w) - w - \mathcal{G}(u) + \mathcal{G}(u), \chi\rangle = -\langle \rho^w, \chi\rangle. \tag{4.25c}$$

Subtracting the spatially discrete scheme from linearised KdV from (4.25) we find

$$\begin{aligned}\langle \theta_t^u + \mathcal{G}(\theta^v), \phi\rangle &= -\langle \rho_t^u, \phi\rangle \qquad \forall \phi \in \mathbb{V}_q \\ \langle \theta^v - \theta^u - \mathcal{G}(\theta^w), \psi\rangle &= \langle \rho^u - \rho^v, \psi\rangle \qquad \forall \psi \in \mathbb{V}_q \\ \langle \theta^w - \mathcal{G}(\theta^u), \chi\rangle &= -\langle \rho^w, \chi\rangle \qquad \forall \chi \in \mathbb{V}_q.\end{aligned} \tag{4.26}$$

In order to gain control over $\theta^u, \theta^v, \theta^w$ we shall utilise similar arguments to those given

in the proof of Theorem 4.2.7, i.e., the proof that the spatial scheme for linearised KdV preserves momentum and energy. Applying this methodology to the error equation (4.26) will allow us to rewrite our errors in terms of the residuals.
Choosing $\phi = \theta_t^u$ in (4.26) allows us to write

$$\frac{1}{2}\left\langle \theta^u, \theta^u \right\rangle_t = -\left\langle \mathcal{G}\left(\theta^u\right), \theta^v \right\rangle - \left\langle \rho_t^u, \theta^u \right\rangle,$$

after application of Lemma 4.2.4. Further choosing $\chi = \theta^v$ and $\psi = \theta^w$ tells us that

$$\begin{aligned}
\frac{1}{2}\left\langle \theta^u, \theta^u \right\rangle_t &= -\left\langle \theta^v, \theta^w \right\rangle - \left\langle \rho_t^u, \theta^u \right\rangle + \left\langle \theta^v, \rho^w \right\rangle \\
&= -\left\langle \theta^u, \theta^w \right\rangle + \left\langle \mathcal{G}\left(\theta^w\right), \theta^w \right\rangle - \left\langle \rho_t^u, \theta^u \right\rangle + \left\langle \theta^v, \rho^w \right\rangle + \left\langle \rho^v, \theta^w \right\rangle - \left\langle \rho^u, \theta^w \right\rangle \\
&= -\left\langle \mathcal{G}\left(\theta^u\right), \theta^u \right\rangle - \left\langle \rho_t^u, \theta^u \right\rangle + \left\langle \theta^v, \rho^w \right\rangle + \left\langle \rho^v, \theta^w \right\rangle - \left\langle \rho^u, \theta^w \right\rangle + \left\langle \rho^w, \theta^u \right\rangle \\
&= -\left\langle \rho_t^u, \theta^u \right\rangle + \left\langle \theta^v, \rho^w \right\rangle + \left\langle \rho^v, \theta^w \right\rangle - \left\langle \rho^u, \theta^w \right\rangle + \left\langle \rho^w, \theta^u \right\rangle,
\end{aligned} \tag{4.27}$$

after choosing $\chi = \theta^u$ and applying the skew-symmetry of $\mathcal{G}$. We ultimately wish to apply Gronwall's inequality to obtain error bounds for θ^u, but we cannot achieve an error bound without first gaining control over the auxiliary variables θ^v, θ^w. With this in mind take the temporal derivative of the latter two error equations (4.26) yielding

$$\begin{aligned}
\left\langle \theta_t^v - \theta_t^u - \mathcal{G}\left(\theta_t^w\right), \psi \right\rangle &= \left\langle \rho_t^u - \rho_t^v, \psi \right\rangle \qquad \forall \psi \in \mathbb{V}_q \\
\left\langle \theta_t^w - \mathcal{G}\left(\theta_t^u\right), \chi \right\rangle &= -\left\langle \rho_t^w, \chi \right\rangle \qquad \forall \chi \in \mathbb{V}_q.
\end{aligned} \tag{4.28}$$

Choosing $\psi = \theta^v$ in (4.28) we see that

$$\frac{1}{2}\left\langle \theta^v, \theta^v \right\rangle_t = \left\langle \theta_t^u, \theta^v \right\rangle + \left\langle \mathcal{G}\left(\theta_t^w\right), \theta^v \right\rangle - \left\langle \rho_t^v, \theta^v \right\rangle + \left\langle \rho_t^u, \theta^v \right\rangle. \tag{4.29}$$

Through choosing $\phi = \theta^v$ in (4.26)

$$\left\langle \theta_t^u, \theta^v \right\rangle = -\left\langle \mathcal{G}\left(\theta^v\right), \theta^v \right\rangle - \left\langle \rho_t^u, \theta^v \right\rangle = -\left\langle \rho_t^u, \theta^v \right\rangle. \tag{4.30}$$

Additionally choosing $\phi = \theta_t^w$ in (4.26) and then $\psi = \theta_t^u$ in (4.28)

$$\begin{aligned}
\left\langle \mathcal{G}\left(\theta_t^w\right), \theta^v \right\rangle &= -\left\langle \theta_t^w, \mathcal{G}\left(\theta^v\right) \right\rangle = \left\langle \theta_t^u, \theta_t^w \right\rangle + \left\langle \rho_t^u, \theta_t^w \right\rangle \\
&= \left\langle \theta_t^u, \mathcal{G}\left(\theta_t^u\right) \right\rangle + \left\langle \rho_t^u, \theta_t^w \right\rangle - \left\langle \rho_t^w, \theta_t^u \right\rangle = \left\langle \rho_t^u, \theta_t^w \right\rangle - \left\langle \rho_t^w, \theta_t^u \right\rangle.
\end{aligned} \tag{4.31}$$

Applying (4.30) and (4.31) to (4.29) yields

$$\frac{1}{2}\left\langle \theta^v, \theta^v \right\rangle_t = -\left\langle \rho_t^v, \theta^v \right\rangle + \left\langle \rho_t^u, \theta_t^w \right\rangle - \left\langle \rho_t^w, \theta_t^u \right\rangle. \tag{4.32}$$

To gain control over θ^w we choose $\chi = \theta^w$ in (4.28) allowing us to write

$$\begin{aligned}
\frac{1}{2}\left\langle \theta^w, \theta^w \right\rangle_t &= \left\langle \mathcal{G}\left(\theta_t^u\right), \theta^w \right\rangle - \left\langle \rho_t^w, \theta^w \right\rangle \\
&= -\left\langle \theta_t^u, \mathcal{G}\left(\theta^w\right) \right\rangle - \left\langle \rho_t^w, \theta^w \right\rangle \\
&= -\left\langle \theta_t^u, \theta^v \right\rangle + \left\langle \theta_t^u, \theta^u \right\rangle - \left\langle \rho_t^w, \theta^w \right\rangle - \left\langle \rho^v, \theta_t^u \right\rangle + \left\langle \rho^u, \theta_t^u \right\rangle \\
&= \left\langle \mathcal{G}\left(\theta^v\right), \theta^v \right\rangle + \frac{1}{2}\left\langle \theta^u, \theta^u \right\rangle_t - \left\langle \rho_t^w, \theta^w \right\rangle - \left\langle \rho^v, \theta_t^u \right\rangle + \left\langle \rho^u, \theta_t^u \right\rangle + \left\langle \rho_t^u, \theta^v \right\rangle \\
&= \frac{1}{2}\left\langle \theta^u, \theta^u \right\rangle_t - \left\langle \rho_t^w, \theta^w \right\rangle - \left\langle \rho^v, \theta_t^u \right\rangle + \left\langle \rho^u, \theta_t^u \right\rangle + \left\langle \rho_t^u, \theta^v \right\rangle,
\end{aligned} \tag{4.33}$$

after choosing $\psi = \theta_t^u$ and $\phi = \theta^v$ in (4.26). Notice that (4.33) also depends on $\left\langle \theta^u, \theta^u \right\rangle_t$. We now sum (4.27), (4.32) and (4.33), but we double the contribution of (4.27) in the summation to avoid the $\left\langle \theta^u, \theta^u \right\rangle_t$ terms in (4.27) and (4.33) from cancelling each other out, yielding

$$\begin{aligned}
\frac{1}{2}\left\langle \theta^u, \theta^u \right\rangle_t + \frac{1}{2}\left\langle \theta^v, \theta^v \right\rangle_t + \frac{1}{2}\left\langle \theta^w, \theta^w \right\rangle_t = &-2\left\langle \rho_t^u, \theta^u \right\rangle + 2\left\langle \theta^v, \rho^w \right\rangle + 2\left\langle \rho^v, \theta^w \right\rangle - 2\left\langle \rho^u, \theta^w \right\rangle \\
&+ 2\left\langle \rho^w, \theta^u \right\rangle - \left\langle \rho_t^v, \theta^v \right\rangle + \left\langle \rho_t^u, \theta_t^w \right\rangle - \left\langle \rho_t^w, \theta_t^u \right\rangle \\
&- \left\langle \rho_t^w, \theta^w \right\rangle - \left\langle \rho^v, \theta_t^u \right\rangle + \left\langle \rho^u, \theta_t^u \right\rangle + \left\langle \rho_t^u, \theta^v \right\rangle.
\end{aligned} \tag{4.34}$$

Before we can apply Gronwall's inequality notice that we have θ_t terms on the right hand side of (4.34) which need to be removed. Note that through integration by parts *in time* we have

$$\begin{aligned}
\int_0^t \left\langle \rho_s^w, \theta_s^u \right\rangle \mathrm{d}s &= -\int_0^t \left\langle \rho_{ss}^w, \theta^u \right\rangle \mathrm{d}s + \left\langle \rho_t^w, \theta^u \right\rangle - \left\langle \rho_t^w(0,x), \theta^u(0,x) \right\rangle \\
\int_0^t \left\langle \rho_s^u, \theta_s^w \right\rangle \mathrm{d}s &= -\int_0^t \left\langle \rho_{ss}^u, \theta^w \right\rangle \mathrm{d}s + \left\langle \rho_t^u, \theta^w \right\rangle - \left\langle \rho_t^u(0,x), \theta^w(0,x) \right\rangle \\
\int_0^t \left\langle \rho^v, \theta_s^u \right\rangle \mathrm{d}s &= -\int_0^t \left\langle \rho_s^v, \theta^u \right\rangle \mathrm{d}s + \left\langle \rho^v, \theta^u \right\rangle - \left\langle \rho^v(0,x), \theta^u(0,x) \right\rangle \\
\int_0^t \left\langle \rho^u, \theta_s^u \right\rangle \mathrm{d}s &= -\int_0^t \left\langle \rho_s^u, \theta^u \right\rangle \mathrm{d}s + \left\langle \rho^u, \theta^u \right\rangle - \left\langle \rho^u(0,x), \theta^u(0,x) \right\rangle.
\end{aligned} \tag{4.35}$$

Integrating (4.34) over time we obtain

$$\begin{aligned}\frac{1}{2}\langle\theta^u,\theta^u\rangle+\frac{1}{2}\langle\theta^v,\theta^v\rangle+\frac{1}{2}\langle\theta^w,\theta^w\rangle=&\frac{1}{2}\langle\theta^u(0,x),\theta^u(0,x)\rangle+\frac{1}{2}\langle\theta^v(0,x),\theta^v(0,x)\rangle\\&+\frac{1}{2}\langle\theta^w(0,x),\theta^w(0,x)\rangle\\&+\int_0^t-2\langle\rho^u_s,\theta^u\rangle-2\langle\theta^v,\rho^w\rangle-2\langle\rho^v,\theta^w\rangle+2\langle\rho^u,\theta^w\rangle\\&+2\langle\rho^w,\theta^u\rangle-\langle\rho^v_s,\theta^v\rangle+\langle\rho^u_s,\theta^w_s\rangle+\langle\rho^w_s,\theta^u_s\rangle\\&-\langle\rho^w_s,\theta^w\rangle-\langle\rho^v,\theta^u_s\rangle-\langle\rho^u,\theta^u_s\rangle+\langle\rho^u_s,\theta^v\rangle\,\mathrm{d}s.\end{aligned}\tag{4.36}$$

through the application of the fundamental theorem of calculus. Utilising (4.35) we can remove the temporal derivatives of θ in (4.36) allowing us to write

$$\begin{aligned}\frac{1}{2}\langle\theta^u,\theta^u\rangle+\frac{1}{2}\langle\theta^v,\theta^v\rangle+\frac{1}{2}\langle\theta^w,\theta^w\rangle=&\frac{1}{2}\langle\theta^u(0,x),\theta^u(0,x)\rangle+\frac{1}{2}\langle\theta^v(0,x),\theta^v(0,x)\rangle\\&+\frac{1}{2}\langle\theta^w(0,x),\theta^w(0,x)\rangle\\&+\int_0^t2\langle\rho^u_s,\theta^u\rangle+2\langle\theta^v,\rho^w\rangle+2\langle\rho^v,\theta^w\rangle-2\langle\rho^u,\theta^w\rangle\\&-2\langle\rho^w,\theta^u\rangle+\langle\rho^v_s,\theta^v\rangle-\langle\rho^u_{ss},\theta^w\rangle+\langle\rho^w_{ss},\theta^u\rangle\\&+\langle\rho^w_s,\theta^w\rangle+\langle\rho^v_s,\theta^u\rangle-\langle\rho^u_s,\theta^u\rangle-\langle\rho^u_s,\theta^v\rangle\,\mathrm{d}s\\&-\langle\rho^w_t,\theta^u\rangle+\langle\rho^w_t(0,x),\theta^u(0,x)\rangle+\langle\rho^u_t,\theta^w\rangle\\&-\langle\rho^u_t(0,x),\theta^w(0,x)\rangle-\langle\rho^v,\theta^u\rangle+\langle\rho^v(0,x),\theta^u(0,x)\rangle\\&+\langle\rho^u,\theta^u\rangle-\langle\rho^u(0,x),\theta^u(0,x)\rangle\,.\end{aligned}$$

Applying Hölder's inequality, and for the temporally integrated terms and terms evaluated

at $t = 0$ Cauchy's inequality, we find

$$\begin{aligned}
\frac{1}{2}\left\|\theta^u\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\theta^v\right\|^2_{L_2(S^1)} & \\
+\frac{1}{2}\left\|\theta^w\right\|^2_{L_2(S^1)} \leq\ & 2\left\|\theta^u(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\theta^v(0,x)\right\|^2_{L_2(S^1)} + \left\|\theta^w(0,x)\right\|^2_{L_2(S^1)} \\
& + \frac{1}{2}\left\|\rho^w_t(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\rho^u_t(0,x)\right\|^2_{L_2(S^1)} \\
& + \frac{1}{2}\left\|\rho^v(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\rho^u(0,x)\right\|^2_{L_2(S^1)} \\
& + 3\int_0^t \left\|\theta^u\right\|^2_{L_2(S^1)} + \left\|\theta^v\right\|^2_{L_2(S^1)} + \left\|\theta^w\right\|^2_{L_2(S^1)} + \left\|\rho^u\right\|^2_{L_2(S^1)} \\
& + \left\|\rho^v\right\|^2_{L_2(S^1)} + \left\|\rho^w\right\|^2_{L_2(S^1)} + \left\|\rho^u_s\right\|^2_{L_2(S^1)} + \left\|\rho^v_s\right\|^2_{L_2(S^1)} \\
& + \left\|\rho^w_s\right\|^2_{L_2(S^1)} + \left\|\rho^u_{ss}\right\|^2_{L_2(S^1)} + \left\|\rho^w_{ss}\right\|^2_{L_2(S^1)} \mathrm{d}s \\
& - \left\langle \rho^w_t, \theta^u \right\rangle + \left\langle \rho^u_t, \theta^w \right\rangle - \left\langle \rho^v, \theta^u \right\rangle + \left\langle \rho^u, \theta^u \right\rangle .
\end{aligned}$$

Through Cauchy's inequality with ϵ we an write

$$\begin{aligned}
& - \left\langle \rho^w_t, \theta^u \right\rangle + \left\langle \rho^u_t, \theta^w \right\rangle \\
& \quad - \left\langle \rho^v, \theta^u \right\rangle + \left\langle \rho^u, \theta^u \right\rangle \leq \frac{1}{4\epsilon_1}\left\|\rho^w_t\right\|^2_{L_2(S^1)} + \epsilon_1\left\|\theta^u\right\|^2_{L_2(S^1)} + \frac{1}{4\epsilon_2}\left\|\rho^u_t\right\|^2_{L_2(S^1)} + \epsilon_2\left\|\theta^w\right\|^2_{L_2(S^1)} \\
& \qquad\qquad + \frac{1}{4\epsilon_1}\left\|\rho^v\right\|^2_{L_2(S^1)} + \epsilon_1\left\|\theta^u\right\|^2_{L_2(S^1)} + \frac{1}{4\epsilon_1}\left\|\rho^u\right\|^2_{L_2(S^1)} + \epsilon_1\left\|\theta^u\right\|^2_{L_2(S^1)} .
\end{aligned}$$

Choosing $\epsilon_1 = \frac{1}{12}$ and $\epsilon_2 = \frac{1}{4}$ allows us to write

$$\begin{aligned}
\frac{1}{4}\left\|\theta^u\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\theta^v\right\|^2_{L_2(S^1)} & \\
+\frac{1}{4}\left\|\theta^w\right\|^2_{L_2(S^1)} \leq\ & 2\left\|\theta^u(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\theta^v(0,x)\right\|^2_{L_2(S^1)} + \left\|\theta^w(0,x)\right\|^2_{L_2(S^1)} \\
& + \frac{1}{2}\left\|\rho^w_t(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\rho^u_t(0,x)\right\|^2_{L_2(S^1)} \\
& + \frac{1}{2}\left\|\rho^v(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\rho^u(0,x)\right\|^2_{L_2(S^1)} \\
& + 3\left\|\rho^w_t\right\|^2_{L_2(S^1)} + \left\|\rho^u_t\right\|^2_{L_2(S^1)} + 3\left\|\rho^v\right\|^2_{L_2(S^1)} + 3\left\|\rho^u\right\|^2_{L_2(S^1)} \\
& + 3\int_0^t \left\|\theta^u\right\|^2_{L_2(S^1)} + \left\|\theta^v\right\|^2_{L_2(S^1)} + \left\|\theta^w\right\|^2_{L_2(S^1)} + \left\|\rho^u\right\|^2_{L_2(S^1)} \\
& + \left\|\rho^v\right\|^2_{L_2(S^1)} + \left\|\rho^w\right\|^2_{L_2(S^1)} + \left\|\rho^u_s\right\|^2_{L_2(S^1)} + \left\|\rho^v_s\right\|^2_{L_2(S^1)} \\
& + \left\|\rho^w_s\right\|^2_{L_2(S^1)} + \left\|\rho^u_{ss}\right\|^2_{L_2(S^1)} + \left\|\rho^w_{ss}\right\|^2_{L_2(S^1)} \mathrm{d}s
\end{aligned} \tag{4.37}$$

In preparation for the application of Gronwall's inequality we define

$$\begin{aligned}\alpha(t) = {} & 2\left\|\theta^u(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\theta^v(0,x)\right\|^2_{L_2(S^1)} + \left\|\theta^w(0,x)\right\|^2_{L_2(S^1)} \\ & + \frac{1}{2}\left\|\rho^w_t(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\rho^u_t(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\rho^v(0,x)\right\|^2_{L_2(S^1)} + \frac{1}{2}\left\|\rho^u(0,x)\right\|^2_{L_2(S^1)} \\ & + 3\left\|\rho^w_t\right\|^2_{L_2(S^1)} + \left\|\rho^u_t\right\|^2_{L_2(S^1)} + 3\left\|\rho^v\right\|^2_{L_2(S^1)} + 3\left\|\rho^u\right\|^2_{L_2(S^1)} \\ & + 3\int_0^t \left\|\rho^u\right\|^2_{L_2(S^1)} + \left\|\rho^v\right\|^2_{L_2(S^1)} + \left\|\rho^w\right\|^2_{L_2(S^1)} + \left\|\rho^u_s\right\|^2_{L_2(S^1)} + \left\|\rho^v_s\right\|^2_{L_2(S^1)} + \left\|\rho^w_s\right\|^2_{L_2(S^1)} \\ & + \left\|\rho^u_{ss}\right\|^2_{L_2(S^1)} + \left\|\rho^w_{ss}\right\|^2_{L_2(S^1)} \mathrm{d}s.\end{aligned}$$

Notice that $\alpha(t)$ is an a priori known quantity, recall that in Lemma 4.2.18 we showed that

$$\left\|\rho^u\right\|_{L_2(S^1)} \leq C\left|h^{q+1}u\right|_{H^{q+1}(S^1)},$$

and similarly for v and w. Additionally recall that at the initial time $U(0)$ is simply the L_2 projection of the initial condition into the finite element space we have that

$$\begin{aligned}\left\|\theta^u(0)\right\|_{L_2(S^1)} &:= \left\|\mathcal{S}\left(u\right)(0) - U(0)\right\|_{L_2(S^1)} = \left\|\mathcal{S}\left(u\right)(0) - \Pi u(0)\right\|_{L_2(S^1)} \\ &\leq \left\|\mathcal{S}\left(u\right)(0) - u(0)\right\|_{L_2(S^1)} + \left\|u(0) - \Pi u(0)\right\|_{L_2(S^1)} \\ &\leq \left\|\rho^u\right\|_{L_2(S^1)} \leq C\left|h^{q+1}u(0)\right|_{H^{q+1}(S^1)},\end{aligned}$$

by Lemma 4.2.18 and the optimally of the L_2 projection, and again the same argument holds with respect to v and w. With this in mind we can bound $\alpha(t)$ by

$$\begin{aligned}\alpha(t) \leq {} & Ch^{2q+2}_{max}\Bigg(\left|u(0,x)\right|^2_{H^{q+1}(S^1)} + \left|v(0,x)\right|^2_{H^{q+1}(S^1)} + \left|w(0,x)\right|^2_{H^{q+1}(S^1)} + \left|u_t(0,x)\right|^2_{H^{q+1}(S^1)} \\ & + \left|w_t(0,x)\right|^2_{H^{q+1}(S^1)} + \left|u\right|^2_{H^{q+1}(S^1)} + \left|v\right|^2_{H^{q+1}(S^1)} + \left|u_t\right|^2_{H^{q+1}(S^1)} + \left|w_t\right|^2_{H^{q+1}(S^1)} \\ & + \int_0^t \left|u\right|^2_{H^{q+1}(S^1)} + \left|v\right|^2_{H^{q+1}(S^1)} + \left|w\right|^2_{H^{q+1}(S^1)} + \left|u_s\right|^2_{H^{q+1}(S^1)} \\ & + \left|v_s\right|^2_{H^{q+1}(S^1)} + \left|w_s\right|^2_{H^{q+1}(S^1)} + \left|u_{ss}\right|^2_{H^{q+1}(S^1)} + \left|w_{ss}\right|^2_{H^{q+1}(S^1)} \mathrm{d}s\Bigg) =: h^{2q+2}_{max}\beta(t).\end{aligned}$$

We can rewrite (4.37) in the condensed form

$$\frac{1}{4}\left\|\theta^u\right\|^2_{L_2(S^1)}+\frac{1}{4}\left\|\theta^v\right\|^2_{L_2(S^1)} +\frac{1}{4}\left\|\theta^w\right\|^2_{L_2(S^1)}\leq h_{max}^{2q+2}\beta(t) + 12\int_0^t \frac{1}{4}\left\|\theta^u\right\|^2_{L_2(S^1)}+\frac{1}{4}\left\|\theta^v\right\|^2_{L_2(S^1)}+\frac{1}{4}\left\|\theta^w\right\|^2_{L_2(S^1)}\mathrm{d}s$$

after decreasing the lower bound. Applying Gronwall's inequality, see Lemma 4.2.20, we find

$$\frac{1}{4}\left\|\theta^u\right\|^2_{L_2(S^1)}+\frac{1}{4}\left\|\theta^v\right\|^2_{L_2(S^1)}+\frac{1}{4}\left\|\theta^w\right\|^2_{L_2(S^1)}\leq h_{max}^{2q+2}\left(\beta(t)+\int_0^t 12\exp(s)\,\beta(s)\mathrm{d}s\right),$$

concluding the proof.

□

Now we have bounded the error between the discrete projection operator $\mathcal{S}$ and our numerical solution we can prove the a priori error bound for our spatial scheme for linearised KdV.

Proof of Theorem 4.2.14. For convenience we will again write the error as described in (4.22). Through the triangle inequality we have that

$$\left\|\mathsf{e}_u\right\|^2_{L_2(S^1)}+\left\|\mathsf{e}_v\right\|^2_{L_2(S^1)}+\left\|\mathsf{e}_w\right\|^2_{L_2(S^1)}\leq \frac{3}{2}\left(\left\|\rho^u\right\|^2_{L_2(S^1)}+\left\|\rho^v\right\|^2_{L_2(S^1)}+\left\|\rho^w\right\|^2_{L_2(S^1)}\right) + \frac{3}{2}\left(\left\|\theta^u\right\|^2_{L_2(S^1)}+\left\|\theta^v\right\|^2_{L_2(S^1)}+\left\|\theta^w\right\|^2_{L_2(S^1)}\right).$$

Applying Lemma 4.2.18 and Theorem 4.2.19 we obtain the a priori error bound for our spatial scheme for linearised KdV.

□

Remark 4.2.22 (A priori error bounds for the alternative spatially discrete scheme (4.15))**.** *The alternative spatially discrete scheme* (4.15) *also satisfies an a priori bound of order* $\mathcal{O}\left(h_{max}^{q+1}\right)$*. However, for* (4.15) *an optimal bound exists for* all *polynomial degree* q*. Observe through Lemma 4.2.10 we obtain numerical stability, and the operators* $\mathcal{G}^+$ *and* $\mathcal{G}^-$ *are unique, see Remark 4.2.9. We obtain these a priori bounds through mimicking the arguments made in this section with one caveat, instead of using the projection operator given in Definition 4.2.15 to split the continuous and discrete errors we utilise two new*

projection operators. These projection operators are described as follows: Let $u \in H^1(S^1)$*, then we define* $\mathcal{S}^+ : H^1(S^1) \oplus \mathbb{V}_q \to \mathbb{V}_q$ *such that*

$$\left\langle \mathcal{S}^+(u), \psi \right\rangle = \langle u, \psi \rangle \qquad \forall \psi \in \mathbb{V}_{q-1}$$

and

$$\mathcal{S}^+(u)_m = u_m^+ \qquad \text{for } m = 0, .., M.$$

Additionally we define $\mathcal{S}^- : H^1(S^1) \oplus \mathbb{V}_q \to \mathbb{V}_q$ *such that*

$$\left\langle \mathcal{S}^-(u), \psi \right\rangle = \langle u, \psi \rangle \qquad \forall \psi \in \mathbb{V}_{q-1}$$

and

$$\mathcal{S}^-(u)_m = u_m^- \qquad \text{for } m = 0, .., M.$$

While the alternative scheme can be bounded a priori for more polynomial degrees than the scheme we primarily focus on this section, is allow dissipates momentum. *As we are interested in the long time dynamical behaviour of solutions we shall not focus on this alternative scheme, as we do not want to dampen our dynamics over long time. As such, we now refocus our attention on the spatially discrete scheme described in Definition 4.2.5. We shall return to examine this modified scheme in our numerical experiments.*

4.2.3 Fully discrete scheme and numerical experiments

We will discretise our temporal interval such that $0 =: t_0 < t_1 < \cdots < t_N =: T$ with a step size $\tau_n := t_{n+1} - t_n$. In this subsection, and throughout the remainder of this work, when our temporal discretisation is within the finite difference framework we shall denote temporally discrete functions with superscripts, i.e., $u^n(x)$ is counterpart to $u(t, x)$ at the point $t = t_n$.

We wish to discretise time in the spatially discrete scheme for linearised KdV (4.13) such that we preserve the mass, momentum, and energy of the scheme over time, i.e., a fully discrete version of Theorem 4.2.7 holds. Due to the quadratic, or lower order, nature of the invariants we can choose a temporal discretisation which preserves quadratic invariants, which are discussed in Chapter 2. For compatibility with our temporal discretisations in the subsequent chapters we choose the Crank-Nicholson method.

Definition 4.2.23 (Fully discrete scheme for the linearised KdV equation)**.** *Let* $U^j, W^j \in$

$\mathbb{V}_q$ *be given for* $j = 0, .., n$. *Then we seek* $U^{n+1}, V^{n+1}, W^{n+1} \in \mathbb{V}_q$ *such that*

$$\begin{aligned} \left\langle \frac{U^{n+1} - U^n}{\tau_n} + \mathcal{G}\left(V^{n+1}\right), \phi \right\rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q \\ \left\langle V^{n+1} - U^{n+\frac{1}{2}} - \mathcal{G}\left(W^{n+\frac{1}{2}}\right), \psi \right\rangle &= 0 \qquad \forall \psi \in \mathbb{V}_q \\ \left\langle W^{n+1} - \mathcal{G}\left(U^{n+1}\right), \chi \right\rangle &= 0 \qquad \forall \chi \in \mathbb{V}_q, \end{aligned} \tag{4.40}$$

where $U^{n+\frac{1}{2}} := \frac{1}{2}\left(U^{n+1} + U^n\right)$ *for* $n = 0, .., N-1$. *Further we define the initial data* $U^0 = \Pi u_0(x)$ *where* Π *denotes the* L_2 *projection into the finite element space, we initialise* W *such that* $W^0 = \mathcal{G}\left(U^0\right)$, *and* $\mathcal{G}$ *is given by Definition 4.2.3.*

Remark 4.2.24 (Temporal discretisation of auxiliary variables)**.** *Through rewriting the scheme* (4.40) *in primal form we observe that* U *is discretised temporally by Crank-Nicholson. However, we do not discretise the auxiliary variables explicitly as* $V^{n+\frac{1}{2}}$ *and* $W^{n+\frac{1}{2}}$. *First consider the auxiliary variable* V^{n+1}, *as this variable is diagnostic, and does not evolve in time, we do not require information about it from the previous time step. If, for example, we evaluated the auxiliary variable at* $V^{n+\frac{1}{2}}$ *then it would not alter any subsequent analytic results, or change the rate of convergence in our numerical experiments. It would however introduce additional numerical artefacts, as we would need to enforce initial data on* all three *variables instead of just* U^0 *and* W^0, *so small errors in* V *could propagate over time. We have discretised in* W *such that we are permitted to take a* discrete temporal derivative *in the third equation of* (4.40). *The ability to do this was fundamental to the proof of energy conservation in the spatially discrete case, so we expect it to help us in the fully discrete case.*

Proposition 4.2.25 (Conservation of invariants in the fully discrete scheme)**.** *As we are employing a quadratic invariant preserving temporal method, and our spatial method preserves a discrete mass, momentum and energy of linearised KdV, our fully discrete scheme preserves appropriate discrete invariants, i.e.,*

$$\begin{aligned} \mathcal{F}_1\left(U^{n+1}\right) &= \mathcal{F}_1(U^n) \\ \mathcal{F}_2\left(U^{n+1}\right) &= \mathcal{F}_2(U^n) \\ \mathcal{F}_3\left(U^{n+1}, W^{n+1}\right) &= \mathcal{F}_3(U^n, W^n). \end{aligned}$$

Proof. The proof for conservation of these invariants follows from the same methodology as the proof of Theorem 4.2.7, but we shall present it here for completeness.

To show conservation of mass we need only choose $\phi = 1$ in (4.40) finding

$$\mathcal{F}_1\left(U^{n+1}\right) - \mathcal{F}_1(U^n) = \left\langle U^{n+1} - U^n, 1\right\rangle = \tau_n \left\langle \mathcal{G}\left(V^{n+1}\right), 1\right\rangle = 0.$$

To show momentum conservation we require compatibility between the terms for W in the second and third equations of (4.40). We obtain this by summing the third equation on the current step, and on the previous step, allowing us to write

$$\left\langle W^{n+\frac{1}{2}} - \mathcal{G}\left(U^{n+\frac{1}{2}}\right), \chi\right\rangle = 0. \tag{4.41}$$

Choosing $\phi = U^{n+\frac{1}{2}}$ in (4.40), and subsequently $\chi = V^{n+1}$ in (4.41), $\psi = W^{n+\frac{1}{2}}$ in (4.40), and $\chi = U^{n+\frac{1}{2}}$ in (4.41) we have that

$$\begin{aligned}
\mathcal{F}_2\left(U^{n+1}\right) - \mathcal{F}_2(U^n) &= \frac{1}{2}\left\langle U^{n+1}, U^{n+1}\right\rangle - \frac{1}{2}\left\langle U^n, U^n\right\rangle = \left\langle U^{n+1} - U^n, U^{n+\frac{1}{2}}\right\rangle \\
&= -\tau_n \left\langle \mathcal{G}\left(V^{n+1}\right), U^{n+\frac{1}{2}}\right\rangle = \tau_n \left\langle V^{n+1}, \mathcal{G}\left(U^{n+\frac{1}{2}}\right)\right\rangle \\
&= \tau_n \left\langle V^{n+1}, W^{n+\frac{1}{2}}\right\rangle \\
&= \tau_n \left\langle U^{n+\frac{1}{2}} + \mathcal{G}\left(W^{n+\frac{1}{2}}\right), W^{n+\frac{1}{2}}\right\rangle = \tau_n \left\langle U^{n+\frac{1}{2}}, W^{n+\frac{1}{2}}\right\rangle \\
&= \tau_n \left\langle U^{n+\frac{1}{2}}, \mathcal{G}\left(U^{n+\frac{1}{2}}\right)\right\rangle \\
&= 0,
\end{aligned}$$

and the discrete momentum is conserved.

Recall in the spatially discrete proof of conservation of energy we were required to take the temporal derivative of the third equation in the numerical scheme. In the fully discrete case the equivalent operation is taking the difference of the third equation of (4.40) between the temporal nodes t_{n+1} and t_n, this yields

$$\left\langle W^{n+1} - W^n - \mathcal{G}\left(U^{n+1}\right) + \mathcal{G}\left(U^n\right), \chi\right\rangle = 0. \tag{4.42}$$

After choosing $\chi = W^{n+\frac{1}{2}}$ in (4.42), $\psi = \frac{U^{n+1}-U^n}{\tau_n}$ in (4.40) and then $\phi = V^{n+1}$ in (4.40)

we have

$$\begin{aligned}\mathcal{F}_3\left(U^{n+1},W^{n+1}\right)-\mathcal{F}_3(U^n,W^n) &= \frac{1}{2}\left\langle W^{n+1},W^{n+1}\right\rangle-\frac{1}{2}\left\langle U^{n+1},U^{n+1}\right\rangle\\ &\quad-\frac{1}{2}\left\langle W^n,W^n\right\rangle+\frac{1}{2}\left\langle U^n,U^n\right\rangle\\ &= \left\langle W^{n+1}-W^n,W^{n+\frac{1}{2}}\right\rangle-\left\langle U^{n+1}-U^n,U^{n+\frac{1}{2}}\right\rangle\\ &= \left\langle \mathcal{G}\left(U^{n+1}\right)-\mathcal{G}\left(U^n\right),W^{n+\frac{1}{2}}\right\rangle-\left\langle U^{n+1}-U^n,U^{n+\frac{1}{2}}\right\rangle\\ &= -\left\langle U^{n+1}-U^n,\mathcal{G}\left(W^{n+\frac{1}{2}}\right)\right\rangle-\left\langle U^{n+1}-U^n,U^{n+\frac{1}{2}}\right\rangle\\ &= -\left\langle V^{n+1},U^{n+1}-U^n\right\rangle\\ &= \tau_n\left\langle V^{n+1},\mathcal{G}\left(V^{n+1}\right)\right\rangle=0,\end{aligned}$$

confirming that the discrete energy is conserved.

□

Lemma 4.2.26 (Stability of the fully discrete scheme for linearised KdV)**.** *Let U^n,V^n,W^n be the numerical solution of the fully discrete scheme for linearised KdV as described in Definition 4.2.23 for $n=0,..,N$. Further assume that the initial momentum and energy are bounded, i.e., $\mathcal{F}_2(U^0)<\infty$ and $\mathcal{F}_3(U^0,W^0)<\infty$, then*

$$\|U^n\|^2_{L_2(S^1)}=\left\|U^0\right\|^2_{L_2(S^1)}$$

and

$$\|W^n\|^2_{L_2(S^1)}=\left\|W^0\right\|^2_{L_2(S^1)}.$$

Proof. Similarly to the spatially discrete case given in Lemma 4.2.13 the result is obtained as a consequence of the conservation of momentum and energy of the fully discrete scheme discussed in Proposition 4.2.25.

□

Now we have developed the fully discrete scheme we shall conduct numerical experiments. Our spatial discretisation has been built through Firedrake, see [153], utilising Unified Form Language, see [8]. We have implemented an order $2q$ Gauss quadrature approximation, so the finite element method is computed *exactly*, i.e., we do not have additional quadrature errors contributing to our scheme, with the exception of the L_2 projection of the initial condition u_0. We employ a direct LU decomposition solver from the PETSc library [20] to solve the assembled linear system. Additionally, we will use a combination of Paraview and Matplotlib as visualisation tools.

For simplicity we assume that both the time step and spatial element length are uniform. When conducting numerical experiments with a spatial component we shall also stretch our spatial domain from $S^1(0,1) \to S^1(0,40)$. The reason for this stretching at present is solely for compatibility with numerical experiments in the sequel where solution dynamics are more difficult to visualise on small domains. With this stretching in mind we can redefine our exact solution to the linearised KdV equation, (4.10) as

$$u(t,x) = C_1 + C_2 \sin\left(\alpha\left(x - (1-\alpha^2)t\right)\right) + C_3 \cos\left(\alpha\left(x - (1-\alpha^2)t\right)\right), \tag{4.43}$$

where $\alpha = \frac{2\pi k}{40}$ for $k \in \mathbb{Z}$.
While we present the experimental order of convergence in the previous chapter, see Definition 3.3.4, we recall it here to clarify the definition in the PDE setting. For each benchmark test we fix the polynomial degree q and compute a sequence of solutions with $h = h(i) = 2^{-i}$ and τ chosen either so $\tau \ll h$, to make the temporal discretisation error negligible, so $\tau = Ch$ so temporal discretisation error dominates. This is done for a sequence of refinement levels, $i = l, \ldots, L$.

Definition 4.2.27 (Experimental order of convergence)**.** *Given two sequences $a(i)$ and $h(i) \searrow 0$ we define the* experimental order of convergence *(EOC) to be the local slope of the $\log(a(i))$ vs. $\log(h(i))$ curve, i.e.,*

$$EOC(a,h;i) = \frac{\log\left(\frac{a(i+1)}{a(i)}\right)}{\log\left(\frac{h(i+1)}{h(i)}\right)}.$$

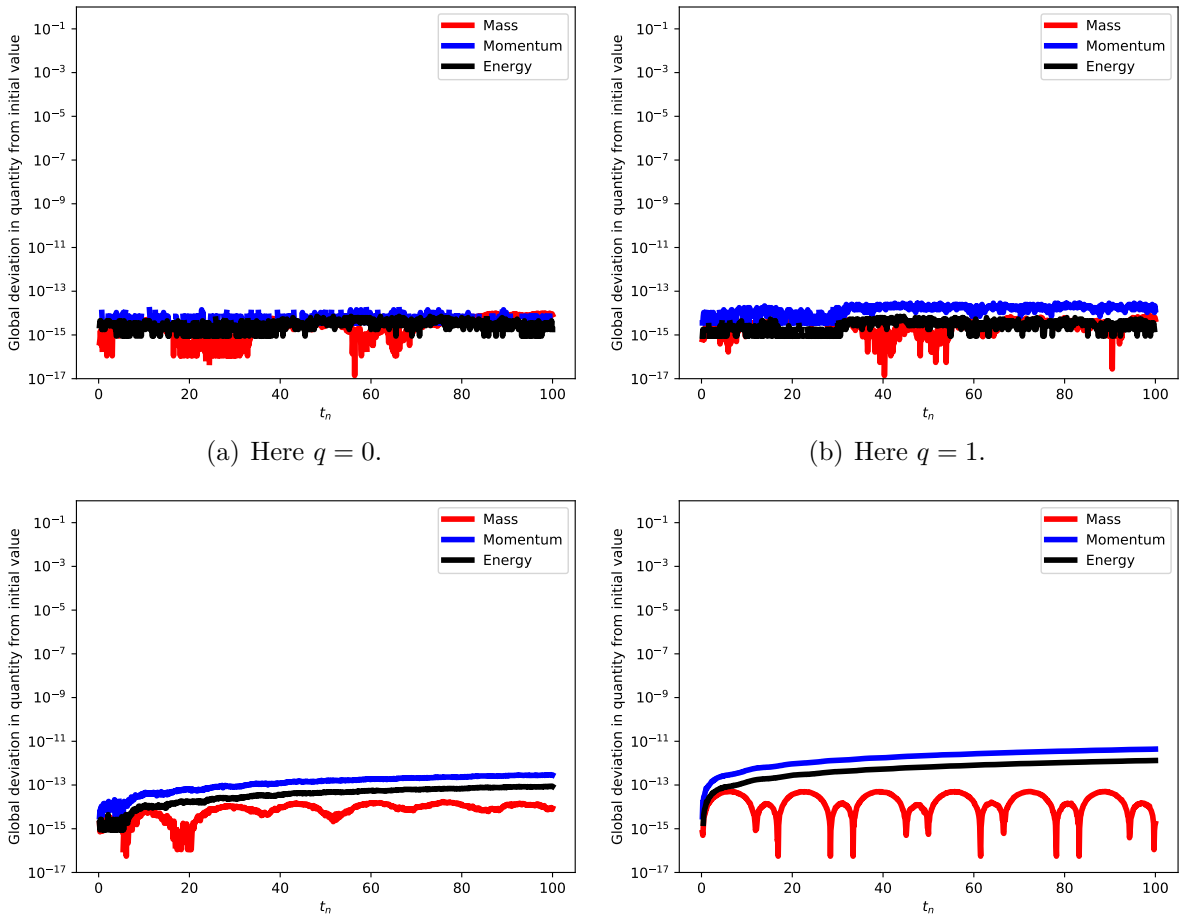
As in Definition 3.3.4 $a(i)$ represents a sequence of errors and $h(i)$ the corresponding sequence of element sizes. In practice, we assume that the time step size is either coupled with the element sizes or that it is sufficiently small.

4.2.3.1 A trigonometric test case

Here we shall investigate how well our fully discrete scheme approximates the exact solution to the linearised KdV equation (4.43) where we have chosen $C_1 = C_3 = 0$ and $C_2 = 1$. We observe in Figure 4.2 and Figure 4.3 the numerical deviation in conserved quantities. We also obtain the EOC for the aforementioned two cases, where τ is fixed to be small and we vary only h in Figure 4.4, and where $\tau = Ch$ in Figure 4.5.
We shall also perform numerical experiments on the finite element scheme defined by (4.15) as defined in Remark 4.2.9. Note that we have only described the *spatial* discretisation

here, the temporal discretisation is equivalent to that described for our fully discrete scheme as seen in (4.40). We plot the global deviation in the "conserved quantities" in Figure 4.6, and the EOC in Figure 4.7 and Figure 4.8.

Figure 4.2: Here we examine the conservative discretisation scheme with various polynomial degrees, q, approximating the exact solution (4.43). We show the *global* deviation in the three invariants mass, momentum and energy. In each test we take a fixed spatial discretisation parameter of $h = 1$ and fixed time step of $\tau = 0.1$. The simulations are run for long time to test conservativity with $T = 100$ in each case.



(a) Here $q = 0$.

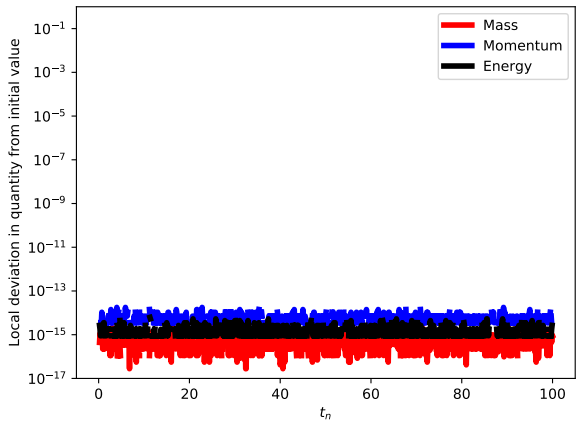
(b) Here $q = 1$.

(c) Here $q = 2$.

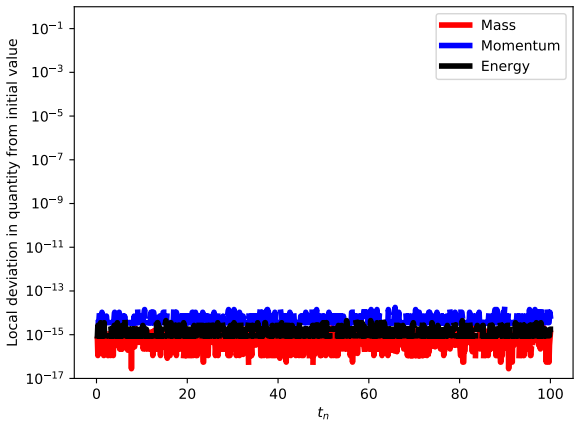
(d) Here $q = 3$.

Remark 4.2.28 (Propagation of machine error)**.** *Analytically, we have proven that the mass, momentum, and energy are preserved. Numerically when examining the global change in these quantities in Figure 4.2 we observe the deviation in these quantities, at least for the cases where $q = 2, 3$, propagate over time. This is likely due to small local errors in the implementation of our numerical scheme, such as only being able to store*

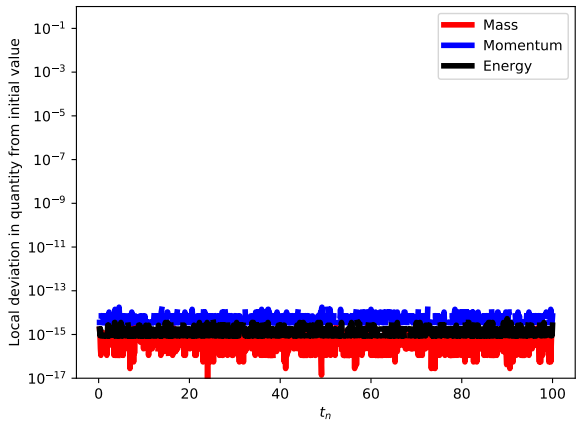
Figure 4.3: Here we examine the conservative discretisation scheme with various polynomial degrees, q, approximating the exact solution (4.43). We show the *local* deviation in the three invariants mass, momentum and energy. In each test we take a fixed spatial discretisation parameter of $h = 1$ and fixed time step of $\tau = 0.1$. The simulations are run for long time to test conservativity with $T = 100$ in each case.



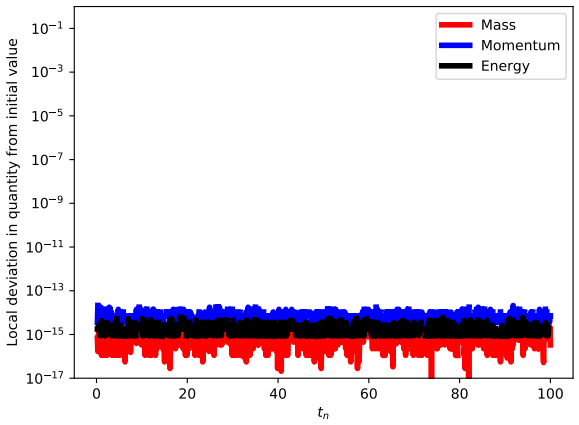
(a) Here $q = 0$.



(b) Here $q = 1$.



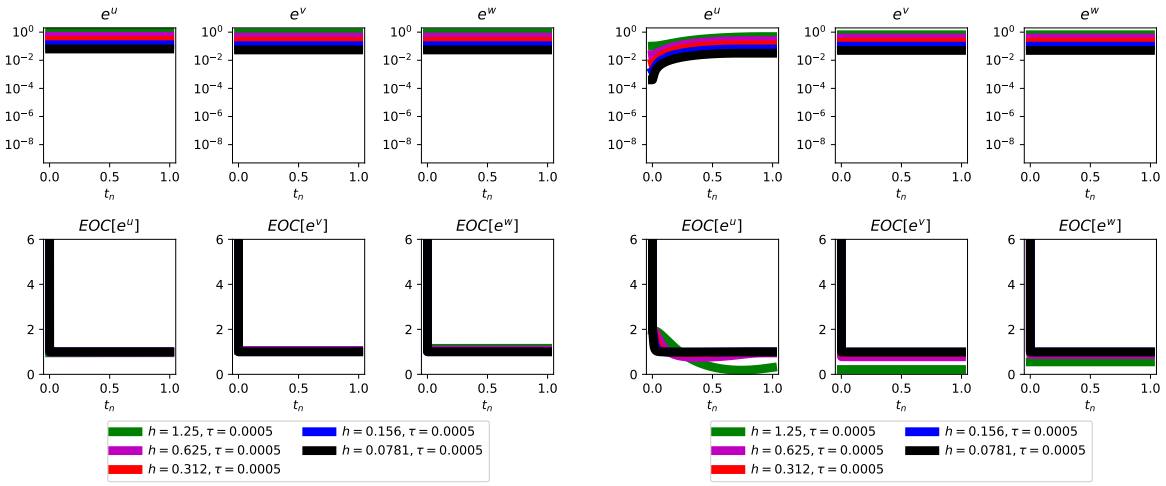
(c) Here $q = 2$.



(d) Here $q = 3$.

numbers to finite precision, which are committed on every *time step. There is no reason that these errors should cancel each other out, but sometimes they do, particularly for lower order polynomial degrees. In Figure 4.3 we observe that the deviation in our conserved quantities locally, i.e., the difference between the invariant on the current time step and the previous time step, is sufficiently small for all time.*

Figure 4.4: Here we examine the conservative discretisation scheme with various polynomial degrees, q, approximating the exact solution (4.43) with $C_1 = C_3 = 0$ and $C_2 = 1$. We measure errors in the $L_\infty(0, t_n; L_2(S^1(0, 40)))$ norm for each variable in the system and plot the EOC for test runs that benchmark both the spatial and temporal discretisation. Here we fix $\tau = 0.0005$ such that the spatial error always dominates. We denote the error in U as $\mathsf{e}_u := \|u - U\|_{L_\infty(0,t_n;L_2(S^1(0,40)))}$, and similarly for v and w.

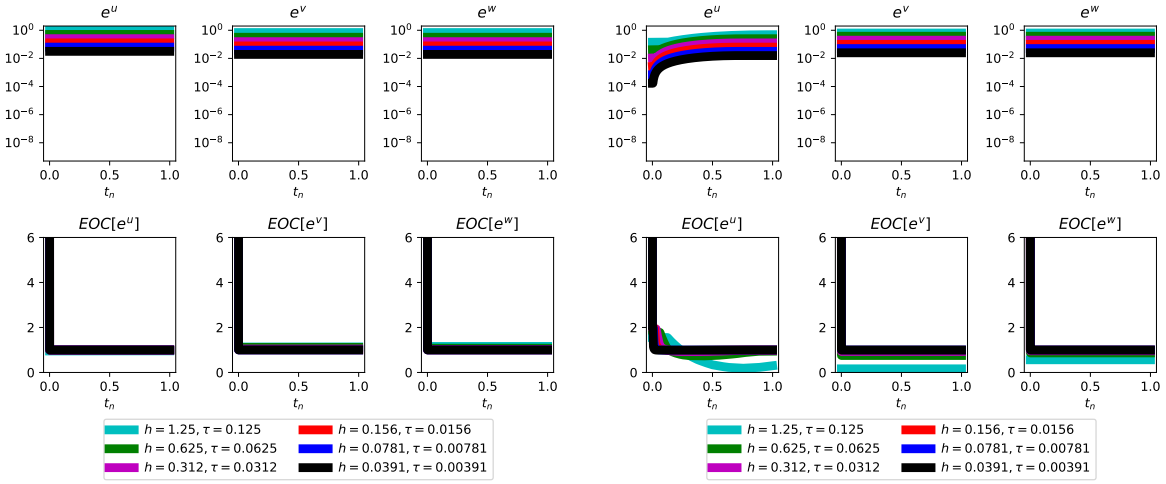


(a) Here $q = 0$.

(b) Here $q = 1$.

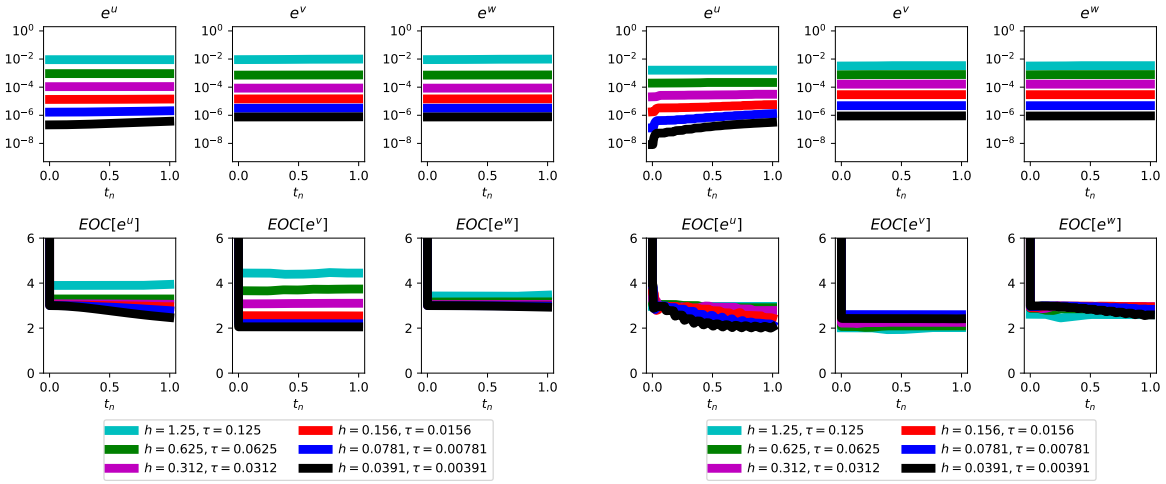
Remark 4.2.29 (Suboptimal spatial convergence for odd polynomial degrees)**.** *For odd polynomial degree we numerically observe suboptimal convergence by one order of the conservative numerical scheme described in Definition 4.2.23. This does not contradict our analytical results as these only hold for even polynomial degrees as our scheme is not unique, see Remark 4.2.9. We observe that by applying the fix discussed in the aforementioned remark we achieve a numerical scheme with EOC in agreement with best approximation, however we do not preserve a discrete momentum or mass over long time.*

Figure 4.5: Here we examine the conservative discretisation scheme with various polynomial degrees, q, approximating the exact solution (4.43) with $C_1 = C_3 = 0$ and $C_2 = 1$. We measure errors in the $L_\infty(0, t_n; L_2(S^1(0, 40)))$ norm for each variable in the system and plot the EOC for test runs that benchmark both the spatial and temporal discretisation. Here $\tau = C\tau$, so we expect different errors to dominate for different polynomials. We denote the error in U as $\mathsf{e}_u := \|u - U\|_{L_\infty(0,t_n;L_2(S^1(0,40)))}$, and similarly for v and w.



(a) Here $q = 0$. Note that the spatial and temporal errors are converging at the same rate.

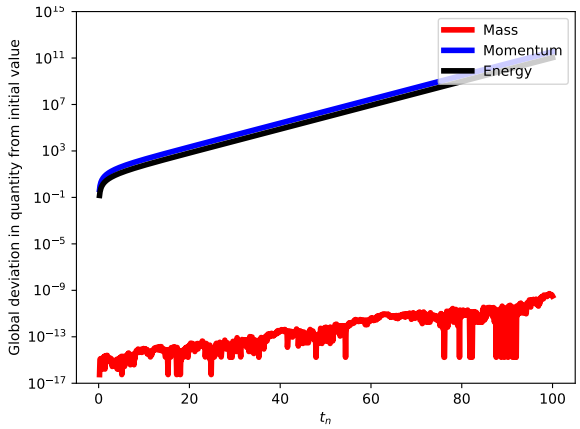
(b) Here $q = 1$. Note that the spatial error dominates.



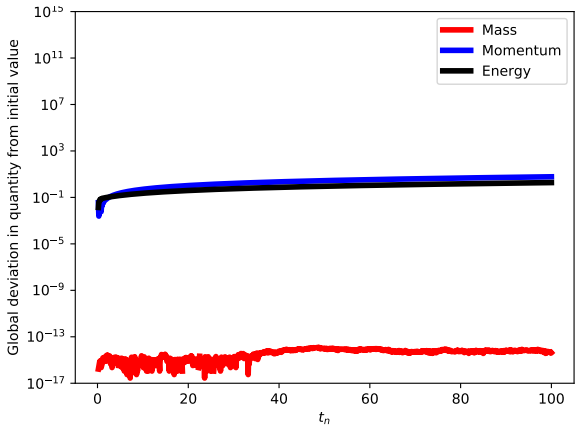
(c) Here $q = 2$. Note that the temporal error dominates.

(d) Here $q = 3$. Note that the temporal error dominates.

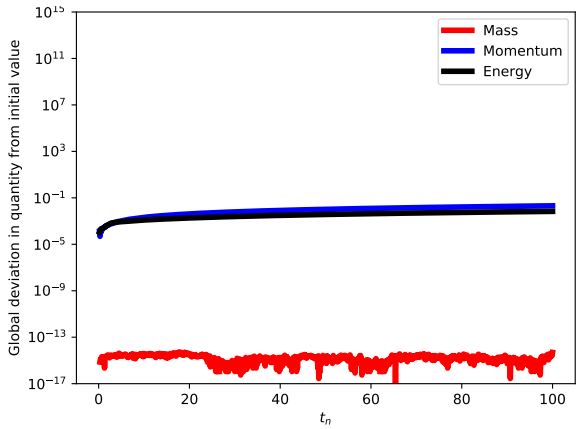
Figure 4.6: Here we examine the *nonconservative* scheme given by the amalgamation of the spatially discrete scheme (4.15) with the temporal discretisation employed for (4.40) for various polynomial degrees, q, approximating the exact solution (4.43). We show the *global* deviation in the three invariants mass, momentum and energy. In each test we take a fixed spatial discretisation parameter of $h = 1$ and fixed time step of $\tau = 0.1$. The simulations are run for long time to test conservativity with $T = 100$ in each case. We observe that only mass is preserved numerically for $q > 0$. When $q = 0$ the numerical scheme is so unstable that the increase in magnitude of the solution causes large deviations in the mass.



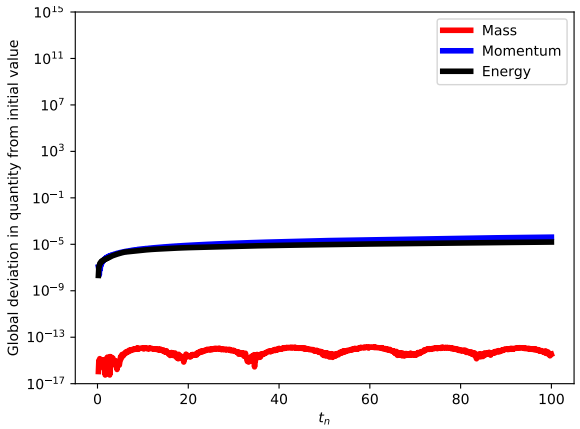
(a) Here $q = 0$.



(b) Here $q = 1$.

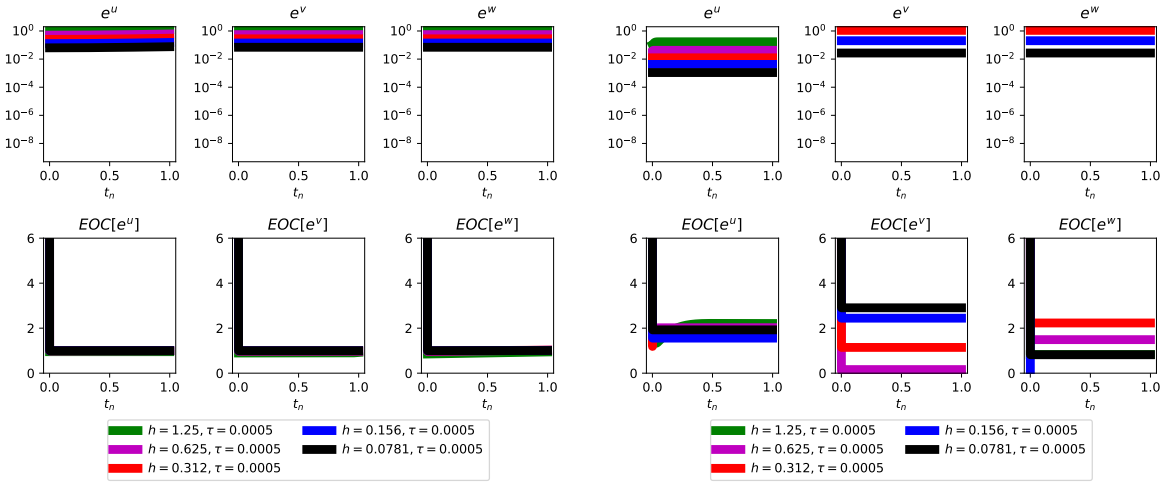


(c) Here $q = 2$.



(d) Here $q = 3$.

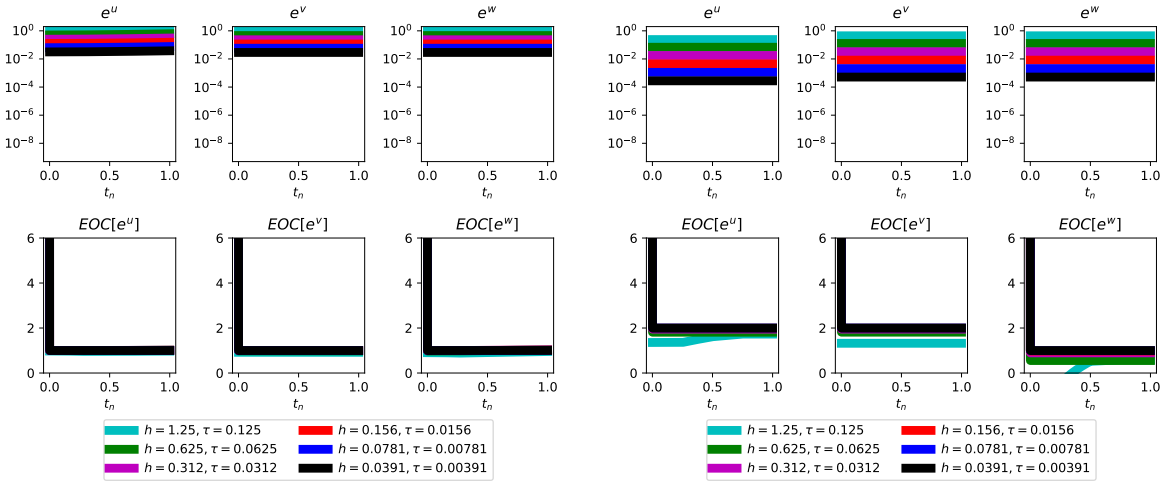
Figure 4.7: Here we examine the *nonconservative* scheme given by the amalgamation of the spatially discrete scheme (4.15) with the temporal discretisation employed for (4.40) for various polynomial degrees, q, approximating the exact solution (4.43) with $C_1 = C_3 = 0$ and $C_2 = 1$. We measure errors in the $L_\infty(0, t_n; L_2(S^1(0, 40)))$ norm for each variable in the system and plot the EOC for test runs that benchmark both the spatial and temporal discretisation. Here we fix $\tau = 0.0005$ such that the spatial error always dominates. We denote the error in U as $\mathsf{e}_u := \|u - U\|_{L_\infty(0,t_n;L_2(S^1(0,40)))}$, and similarly for v and w.



(a) Here $q = 0$.

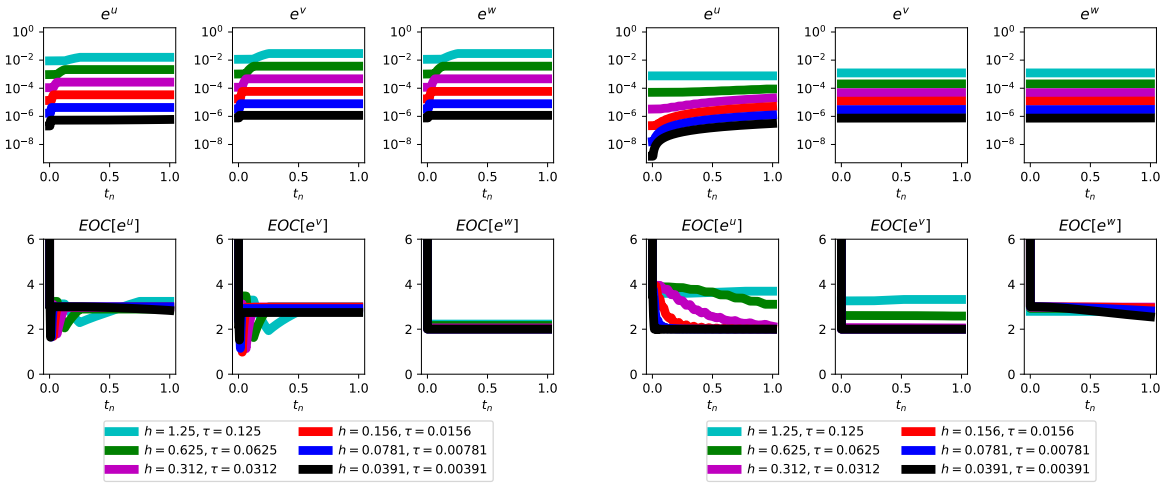
(b) Here $q = 1$.

Figure 4.8: Here we examine the *nonconservative* scheme given by the amalgamation of the spatially discrete scheme (4.15) with the temporal discretisation employed for (4.40) for various polynomial degrees, q, approximating the exact solution (4.43) with $C_1 = C_3 = 0$ and $C_2 = 1$. We measure errors in the $L_\infty(0, t_n; L_2(S^1(0, 40)))$ norm for each variable in the system and plot the EOC for test runs that benchmark both the spatial and temporal discretisation. Here $\tau = C\tau$, so we expect different errors to dominate for different polynomials. We denote the error in U as $\mathsf{e}_u := \|u - U\|_{L_\infty(0,t_n;L_2(S^1(0,40)))}$, and similarly for v and w.



(a) Here $q = 0$. Note that the spatial errors dominate.

(b) Here $q = 1$. Note that the spatial errors and temporal errors converge at the same rate with respect to u and v. With respect to w the spatial error dominates. The error with respect to w is heuristically a H^1 error with respect to u so the slower convergence rate is somewhat expected.



(c) Here $q = 2$. Note that the temporal error dominates.

(d) Here $q = 3$. Note that the temporal error dominates.

4.2.3.2 An initial condition which solves the KdV equation

Here we initialise our numerical scheme with an exact solution to the standard KdV equation. This exact solution, as given in Chapter 5, is a *solitary wave* in the KdV equation known as a *soliton*, but of course in the linearised KdV equation it is not a supported solution. We describe the initial data as

$$U^0(x) = \Pi\left(\frac{1}{2}\text{sech}\left(\frac{1}{2}(x-20)\right)^2\right). \tag{4.44}$$

This test case is motivated by [125] where an exact solution of the linearised KdV equation is enforced as the initial condition for a numerical discretisation of the KdV equation. It is observed here, as well as in Chapter 5 and Chapter 8, that the initial condition decomposes into solitons, a natural solution for the KdV equation. Here we investigate if the converse is true.

We display the deviation in conserved quantities in Figure 4.10 and examine the numerical solution dynamics in Figure 4.9.

Figure 4.9: Here we show the dynamics of the approximation generated by our fully discrete scheme for linearised KdV with polynomial degree $q = 1$ and $\tau = h = 0.1$ approximating the solution to (4.12) with initial conditions given by (4.44). Notice that initially, dispersive waves emanate from the soliton, although no significant structure is visible over long time.

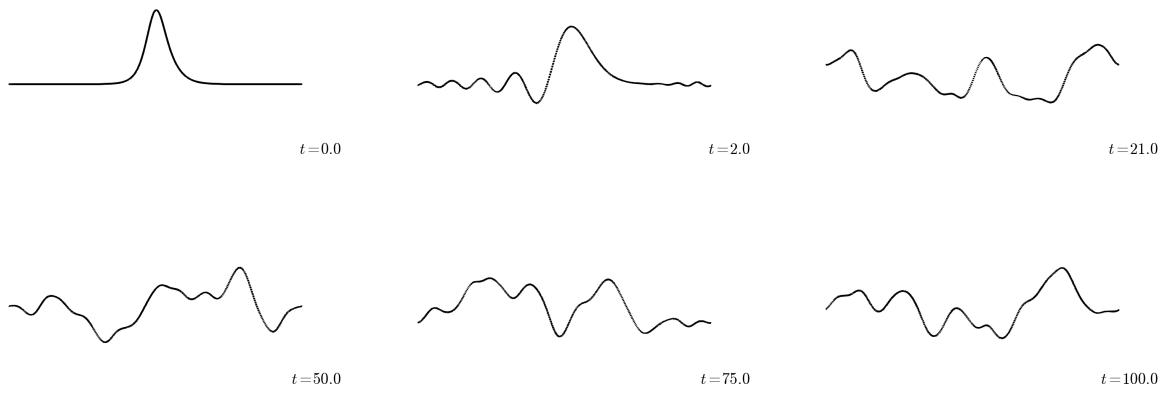
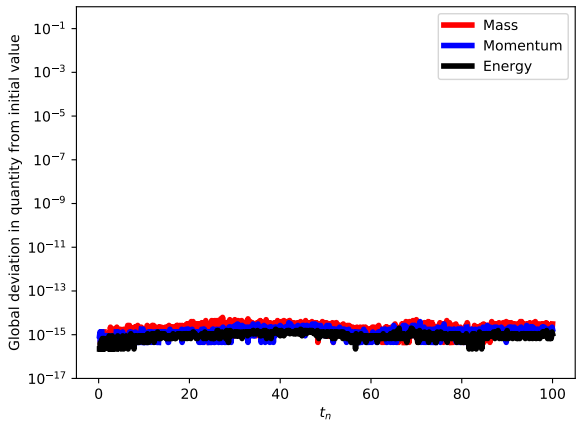
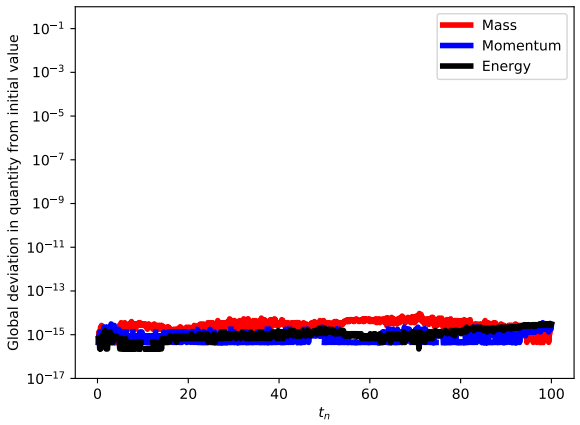


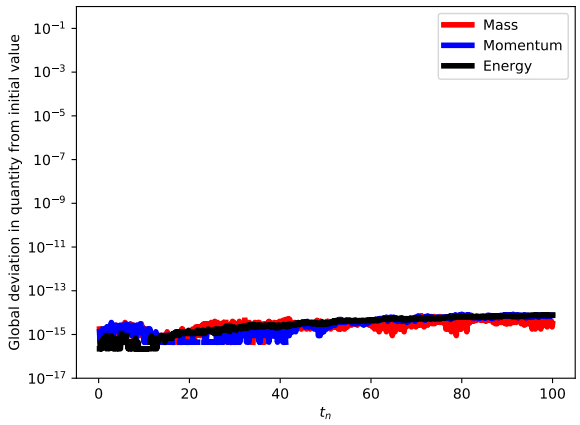
Figure 4.10: Here we examine the conservative discretisation scheme with various polynomial degrees, q, with the initial condition (4.44). We show the *global* deviation in the three invariants mass, momentum and energy. In each test we take a fixed spatial discretisation parameter of $h = 1$ and fixed time step of $\tau = 0.1$. The simulations are run for long time to test conservativity with $T = 100$ in each case.



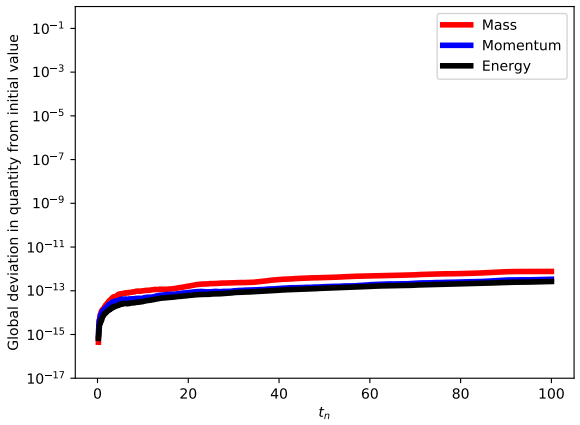
(a) Here $q = 0$.



(b) Here $q = 1$.



(c) Here $q = 2$.



(d) Here $q = 3$.

4.2.3.3 Discontinuous initial data

We will now enforce a discontinuous initial condition to our scheme for the linearised KdV equation. It is important to remember that this initial condition lacks the smoothness to be a solution of the continuous problem. We can describe the initial data by

$$U^0(x) = \begin{cases} 1 \text{ for } x \in [10, 20] \\ 0 \text{ otherwise.} \end{cases} \tag{4.45}$$

We display the deviation in mass, momentum and energy globally in Figure 4.12, as well as examining the numerical solution dynamics in Figure 4.11.

Figure 4.11: Here we show the dynamics of the approximation generated by our fully discrete scheme for linearised KdV with polynomial degree $q = 1$ and $\tau = h = 0.1$ approximating the solution to (4.12) with initial conditions given by (4.45) simulated over long time. We note that similar phenomena have been observed numerically in [148, 118, 107]. While this test case violates the conditions for our analysis we observe that the solution remains stable.

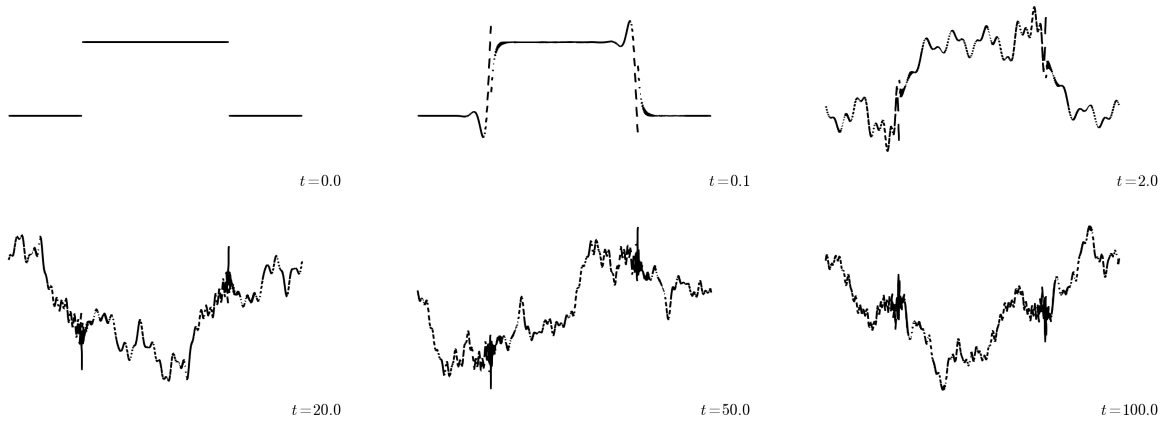
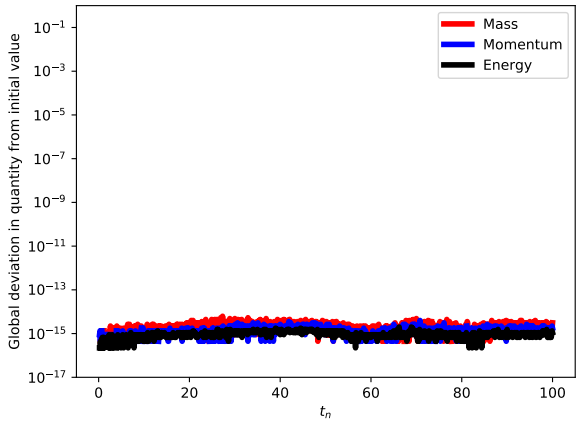
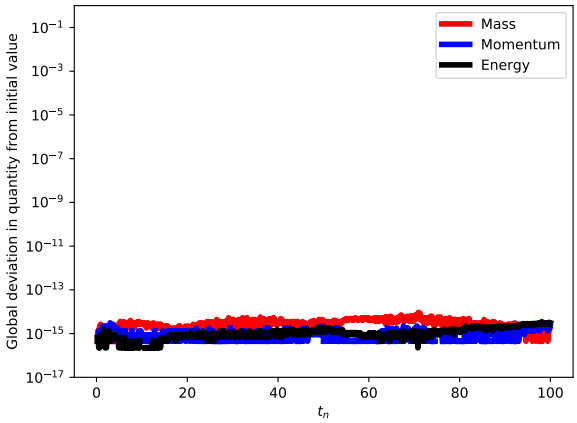


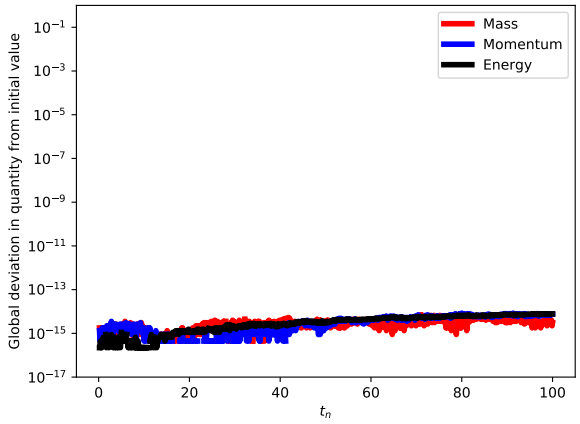
Figure 4.12: Here we examine the conservative discretisation scheme with various polynomial degrees, q, with the initial condition (4.45). We show the *global* deviation in the three invariants mass, momentum and energy. In each test we take a fixed spatial discretisation parameter of $h = 1$ and fixed time step of $\tau = 0.1$. The simulations are run for long time to test conservativity with $T = 100$ in each case.



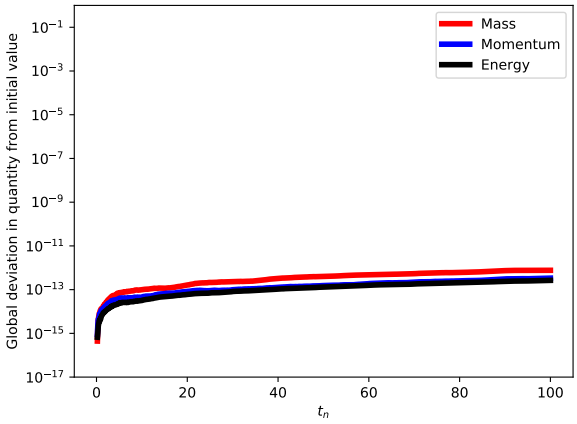
(a) Here $q = 0$.



(b) Here $q = 1$.



(c) Here $q = 2$.



(d) Here $q = 3$.

4.3 Conclusion

In this chapter we introduced a methodology for the design of conservative finite element methods for Hamiltonian PDEs. We go on to design such a scheme for the linearised KdV equation, show that for this linear problem we preserve the momentum and the energy, and utilise these conservative properties to prove a priori error bounds when the scheme is uniquely defined. We observe numerically that even when the scheme is not unique it is still conservative.

Chapter 5

Invariant preserving schemes for the KdV equation

Here we give a numerical comparison of two discontinuous Galerkin methods for the Korteweg-de Vries *(KdV)* equation, one of which has been developed following the methodology outlined in §4.1. Similarly to the previous chapter, in view of the importance of conservation properties for this problem and other related Hamiltonian problems for long time simulations, these methods are constructed such that different discrete invariants of the problem are conserved. Due to the nonlinearity of the problem it does not seem possible to construct discrete schemes that preserve more than two invariants following our discretisation methodology. As such we look at two schemes. One scheme conserves the mass and momentum (which is a quadratic invariant), and the other the mass and energy (which is a cubic invariant). We summarise with numerical experiments aimed at testing the robustness, long time accuracy and computational speed of these methods.

5.1 The KdV equation and two spatial discretisations

5.1.1 The continuous problem

Similarly to the previous chapter, let $u = u(t, x)$, where $t \in [0, T]$ and $x \in S^1$ with S^1 the periodic unit interval. For brevity when there is no ambiguity we will not explicitly write the dependencies of u. Recall that the KdV equation is given by

$$u_t + 6uu_x + u_{xxx} = 0$$
$$u(0, x) = u_0,$$

in (4.4) for $u_0 = u_0(x)$ sufficiently smooth, where subscripts represent partial derivatives. Notice that the KdV equation can be written in conservative form, i.e.,

$$u_t = -\left(3u^2 + u_{xx}\right)_x . \tag{5.1}$$

In view of the PDE (5.1) we observe mass conservation, that is,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle u, 1\rangle = \langle u_t, 1\rangle = -\left\langle \left(3u^2 + u_{xx}\right)_x, 1\right\rangle = 0,$$

utilising Stokes theorem and the periodic boundary conditions. We will refer to the mass in the sequel as

$$\mathcal{F}_1(u) := \langle u, 1\rangle . \tag{5.2}$$

Similarly, again by (4.4) and integration by parts we find that a momentum is also conserved, that is,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\langle u, u\rangle\right) = \langle u_t, u\rangle = -\langle 6uu_x + u_{xxx}, u\rangle = -\left\langle 2\left(u^3\right)_x - \frac{1}{2}\left(u_x^2\right)_x, 1\right\rangle = 0.$$

We will denote the momentum by

$$\mathcal{F}_2(u) = \frac{1}{2}\langle u, u\rangle . \tag{5.3}$$

Further the problem conserves an energy, that is,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\langle u_x, u_x\rangle - \left\langle u^2, u\right\rangle\right) &= \langle u_{xt}, u_x\rangle - 3\left\langle u_t, u^2\right\rangle \\ &= -\left\langle u_t, u_{xx} + 3u^2\right\rangle \\ &= \left\langle \left(u_{xx} + 3u^2\right)_x, u_{xx} + 3u^2\right\rangle = 0,\end{aligned}$$

through integration by parts and Stokes' theorem utilising the periodic boundary conditions. We will write the energy as

$$\mathcal{F}_3(u) = \frac{1}{2}\langle u_x, u_x\rangle - \left\langle u^2, u\right\rangle . \tag{5.4}$$

These are only the first three invariants of the problem, in fact, infinitely many conservation laws can be constructed through Bäcklund transformations, see [2]. Notice that our momentum $\mathcal{F}_2(u)$ and energy $\mathcal{F}_3(u)$ are exactly the Hamiltonian operators for the KdV

equation described in (4.5).

Remark 5.1.1 (Invariant induced norms)**.** *Notice that the only invariant which induces a norm without further interpolation arguments is momentum. Both the mass* (5.2) *and energy* (5.4) *are* not signed*. This will add significant complications to the analysis of finite element schemes which do not preserve the momentum. For example, through momentum conservation we have*

$$\|u(t,x)\|^2_{L_2(S^1)} = \|u(0,x)\|^2_{L_2(S^1)}$$

immediately, whereas energy conservation yields the nonlinear notion of stability

$$\mathcal{F}_3(u(t,x)) = \mathcal{F}_3(u(0,x)),$$

which requires the Gagliardo-Nirenberg interpolation inequality to be related to a norm, see Proposition 8.1.3.

5.1.2 Spatial finite element notation

Before introducing our invariant preserving spatial finite element schemes we direct the reader to recall the notation found in §4.2.1. With this notation in mind we now introduce our invariant conserving schemes.

5.1.3 Momentum conserving spatial discretisation

The momentum conserving scheme we consider was developed and analysed in [31, 114]. Note that this scheme is actually described for a general class of KdV-type equations, but here we restrict our study to the KdV equation for clarity of exposition. This numerical scheme arises through defining two discrete operators: Let $W \in \mathbb{V}_q$, then the first operator deals with the nonlinear term, define $\mathcal{N} : \mathbb{V}_q \to \mathbb{V}_q$ such that

$$\langle \mathcal{N}(W), \phi \rangle = -\left\langle W^2, \phi_x \right\rangle + \frac{1}{3}\sum_{m=0}^{M-1} \left(W_m^{+\,2} + W_m^+ W_m^- + W_m^{-\,2} \right) [\![\phi_m]\!] \qquad \forall \phi \in \mathbb{V}_q, \quad (5.5)$$

where $W_m^+ = \lim_{x \searrow x_m} W(x)$ and $W_m^- = \lim_{x \nearrow x_m} W(x)$. This operator is conservative in the sense that $\langle \mathcal{N}(W), W \rangle = 0 \ \forall W \in \mathbb{V}_q$, as we will discuss in the proof of Proposition

5.1.3. We also define the dispersion operator $\mathcal{D} : \mathbb{V}_q \to \mathbb{V}_q$ for $W \in \mathbb{V}_q$ such that

$$\langle \mathcal{D}(W), \phi \rangle = \langle W_x, \phi_{xx} \rangle + \sum_{m=0}^{M-1} W_{xx}^{+} [\![\phi_m]\!] - [\![W_m]\!] \phi_{xx}^{+} - \{W_{xm}\} [\![\phi_{xm}]\!] \qquad \forall \phi \in \mathbb{V}_q. \quad (5.6)$$

Note that the dispersion operator is similarly conservative in the sense that $\langle \mathcal{D}(W), W \rangle = 0$, but this operator is only well defined for $q \geq 2$, for $q = 1$ it lacks consistency due to loss of information in the flux terms.

Definition 5.1.2 (Momentum conserving spatially discrete scheme,[31])**.** *Let $\mathcal{N}$ and $\mathcal{D}$ be the operators defined in* (5.5) *and* (5.6) *respectively. Then seek $U \in \mathbb{V}_q$ (for $q \geq 2$) such that*

$$\begin{aligned} \langle U_t + 3\mathcal{N}(U) + \mathcal{D}(U), \phi \rangle = 0 \qquad \forall \phi \in \mathbb{V}_q \\ U(0, x) = \Pi u_0, \end{aligned} \quad (5.7)$$

where Π is the L_2 projection into the finite element space.

Proposition 5.1.3 (Conservation of momentum by (5.7), [31])**.** *Let $U \in \mathbb{V}_q$ be the solution of the spatial finite element approximation given in Definition 5.1.2 for $2 \leq q \in \mathbb{Z}$. The discrete mass $\mathcal{F}_1(U)$ and momentum $\mathcal{F}_2(U)$ are conserved, that is to say that*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}_1(U) = 0$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}_2(U) = 0.$$

Proof. We observe discrete mass conservation through choosing $\phi = 1$ in (5.7), as

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}_1(U) = \langle U_t, 1 \rangle = -\langle 3\mathcal{N}(U) + \mathcal{D}(U), 1 \rangle = 0,$$

through orthogonality of $\mathcal{N}$ and $\mathcal{D}$ with constants. Note that we observe this orthogonality directly through the operators respective definitions (5.5) and (5.6).
Conservation of momentum can be seen through choosing $\phi = U$ in (5.7) as

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}_2(U) = \langle U_t, U \rangle = -\langle 3\mathcal{N}(U) + \mathcal{D}(U), U \rangle = 0,$$

utilising orthogonality of $\mathcal{N}(U)$ and $\mathcal{D}(U)$ with U. Note that this orthogonality is observed through direct calculation. For $\mathcal{N}$ we have

$$\begin{aligned}\langle \mathcal{N}(U), U\rangle &= -\left\langle U^2, U_x\right\rangle + \frac{1}{3}\sum_{m=0}^{M-1}\left(U_m^{+2} + U_m^+U_m^- + U_m^{-2}\right)[\![U_m]\!] \\ &= -\frac{1}{3}\left\langle \left[U^3\right]_x, 1\right\rangle + \frac{1}{3}\sum_{m=0}^{M-1}[\![U_m^3]\!] \\ &= 0,\end{aligned}$$

through the application of the fundamental theorem of calculus applied elementwise. For $\mathcal{D}$ we have

$$\begin{aligned}\langle \mathcal{D}(U), U\rangle &= \langle U_x, U_{xx}\rangle - \sum_{m=0}^{M-1}\{U_{xm}\}[\![U_{xm}]\!] \\ &= \frac{1}{2}\left\langle \left(U_x^2\right)_x, 1\right\rangle - \sum_{m=0}^{M-1}\frac{1}{2}[\![U_{xm}^{\,2}]\!] \\ &= 0,\end{aligned}$$

through application of the definitions of the jumps and averages, as well as the fundamental theorem of calculus elementwise. Note that these operators have been designed such that they both respect the aforementioned orthogonality conditions as a methodology of designing a conservative scheme.

□

As the momentum conserving scheme (5.7) preserves $\mathcal{F}_2(U)$ over time from Remark 5.1.1 we observe that we have numerical stability in the L_2 norm. This numerical stability allows for the development of the following suboptimal error bound.

Theorem 5.1.4 (An a priori bound for the momentum conserving scheme,[31])**.** *Let $U \in \mathbb{V}_q$ be the spatial finite element approximation as given in Definition 5.1.2 for $2 \leq q \in \mathbb{Z}$, and let u be the corresponding exact solution to* (4.4)*. Further* assume that the polynomial degree q and the number of nodes N are both even*, then*

$$\|u(t,x) - U(t,x)\|_{L_2(S^1)} \leq C_1 \exp(C_2 t)\, h_{max}^q,$$

where the constants C_1 and C_2 depend on p and the magnitude of the solution u and its derivatives.

5.1.4 Energy conserving spatial discretisation

We now design a new energy conserving scheme utilising the methodology outlined in §4.1, as the energy functional corresponds to one of the Hamiltonian formulations of the KdV equation. We begin by introducing the variational derivative of the energy as an auxiliary variable, allowing us to rewrite the KdV equation as the system

$$\begin{aligned} u_t + v_x &= 0 \\ v - 3u^2 - u_{xx} &= 0, \end{aligned}$$

on the continuous level. Then to prove energy conservation on the continuous level we show

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_3(u) &= \langle u_{xt}, u_x \rangle - 3\left\langle u_t, u^2 \right\rangle = \langle u_t, u_{xx} \rangle - \left\langle u_t, 3u^2 \right\rangle \\ &= \langle u_t, v \rangle = \langle v_x, v \rangle = \frac{1}{2}\left\langle \left(v^2\right)_x, 1 \right\rangle = 0, \end{aligned}$$

through integration by parts and the fundamental theorem of calculus. It is this argument we wish to mimic on the discrete level. Before discretising this system we define the following spatial operators for the first and second spatial derivatives.

Definition 5.1.5 (Discrete operator for first spatial derivatives)**.** *Let $W \in \mathbb{V}_q$, then recall that $\mathcal{G} : \mathbb{V}_q \to \mathbb{V}_q$ such that*

$$\langle \mathcal{G}(W), \phi \rangle = \langle W_x, \phi \rangle - \sum_{m=0}^{M-1} [\![W_m]\!] \{\phi_m\} \qquad \forall \phi \in \mathbb{V}_q.$$

Note that the operator for first derivatives is given in Definition 4.2.3, we recall it here for the readers convenience.

Definition 5.1.6 (Discrete operator for second spatial derivatives: Symmetric interior penalty, [11])**.** *Let $\gamma = 1$ and $W \in \mathbb{V}_q$, then $\mathcal{A}_h : \mathbb{V}_q \times \mathbb{V}_q \to \mathbb{V}_q$ such that*

$$\begin{aligned} \mathcal{A}_h(W, \boldsymbol{\phi}) = \langle W_x, \boldsymbol{\phi}_x \rangle + \sum_{m=0}^{M-1} \Big(&- \gamma [\![W_m]\!] \{\boldsymbol{\phi}_{xm}\} \\ &- \gamma [\![\boldsymbol{\phi}_m]\!] \{W_{xm}\} + \frac{\sigma}{h_m} [\![W_{xm}]\!] [\![\boldsymbol{\phi}_{xm}]\!] \Big) \qquad \forall \boldsymbol{\phi} \in \mathbb{V}_q, \end{aligned} \tag{5.8}$$

where σ is a sufficiently large constant to guarantee coercivity. Numerically we choose it to be at least 10, see [5]. While $\gamma = 1$ we refer to the method as the consistent interior

penalty method, however, we also consider the case where $\gamma = 0$. If $\gamma = 0$ we refer to the method as the inconsistent interior penalty method. Unless stated otherwise, we shall always consider the consistent method. The operator $\mathcal{A}_h$ is a symmetric *bilinear form which induces a norm. See [12] for more details and an overview of similar operators.*

The energy conserving spatial finite element scheme is then described as follows.

Definition 5.1.7 (Energy conserving spatially discrete scheme)**.** *Let the operators $\mathcal{G}$ and $\mathcal{A}_h$ be as described in Definition 4.2.3 and Definition 5.1.6, then seek $U, V \in \mathbb{V}_q$ such that*

$$\begin{aligned} \langle U_t + \mathcal{G}(V), \phi \rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q \\ \left\langle V - 3U^2, \psi \right\rangle + \mathcal{A}_h(U, \psi) &= 0 \qquad \forall \psi \in \mathbb{V}_q \\ U(0, x) &= \Pi u_0, \end{aligned} \tag{5.9}$$

where Π is the L_2 projection into the finite element space.

Proposition 5.1.8 (Conservation of energy by (5.9))**.** *Let $U, V \in \mathbb{V}_q$ be the solution to the spatial finite element scheme described by Definition 5.1.7. The discrete mass $\mathcal{F}_1(U)$ and the discrete energy*

$$\widetilde{\mathcal{F}_3}(U) := \frac{1}{2}\mathcal{A}_h(U, U) - \left\langle U^2, U \right\rangle, \tag{5.10}$$

are conserved over time, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_1(U) = 0$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathcal{F}_3}(U) = 0.$$

Proof. In view of mass conservation we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_1(U) = \langle U_t, 1 \rangle = \langle \mathcal{G}(V), 1 \rangle = 0,$$

through choosing $\phi = 1$ in (5.9) and recalling that $\mathcal{G}$ is orthogonal to constants, which can be seen through combining Lemma 4.2.4 with the definition of $\mathcal{G}$, Definition 4.2.3.
To show energy conservation we depend heavily on Lemma 4.2.4, that is that $\mathcal{G}$ is a skew-symmetric operator. Choosing $\psi = U_t$ in (5.9)

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathcal{F}_3}(U) = \mathcal{A}_h(U_t, U) - \left\langle U_t, 3U^2 \right\rangle = \langle U_t, V \rangle,$$

as $\mathcal{A}_h$ is symmetric. Further choosing $\phi = V$ in (5.9) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathcal{F}_3}(U) = -\langle \mathcal{G}(V), V\rangle = 0,$$

through the skew-symmetry of $\mathcal{G}$.

□

Remark 5.1.9 (Choice of operator for the second derivative)**.** *We do not necessarily need to choose $\mathcal{A}_h$ to be described by* (5.8) *with $\gamma = 1$ for an energy conserving method. In fact, all we need is that the bilinear form $\mathcal{A}_h(\cdot,\cdot)$ is* symmetric*. For our numerical method to be well posed we additionally need the bilinear form to be coercive. For example, choosing $\gamma = 0$ the bilinear form*

$$\mathcal{A}_h(W, \boldsymbol{\phi}) = \langle W_x, \boldsymbol{\phi}_x\rangle + \sum_{m=0}^{M-1} \frac{\sigma}{h_m} [\![W_{xm}]\!] [\![\boldsymbol{\phi}_{xm}]\!] \qquad \forall \boldsymbol{\phi} \in \mathbb{V}_q,$$

is inconsistent in the sense of [12] but would be an equally valid choice of operator. This modified choice of $\mathcal{A}_h$ leads the scheme to preserve a modified discrete energy.

Remark 5.1.10 (A remark on the uniqueness of the energy conserving scheme)**.** *Recall in §4.2 we introduced a scheme which utilised the $\mathcal{G}$ operator and we found in Remark 4.2.9 that this scheme was not uniquely defined due to multiple elements in the kernel of $\mathcal{G}$. While utilising $\mathcal{G}$ the energy conserving scheme* (5.9) *is uniquely defined, as $\mathcal{A}_h$ includes a kernel removing stabilisation term.*

5.1.5 A preliminary comparison of the numerical schemes

While both the momentum and energy conserving schemes fall within the discontinuous finite element framework there are some fundamental differences between the momentum and energy conserving schemes. The momentum conserving scheme requires a polynomial degree of $q \geq 2$, due to the weak handling of the third derivative, whereas the energy conserving scheme is valid for $q \geq 1$ as the third derivative is being "hidden" by an auxiliary variable. While the auxiliary variable allows us to use lower order polynomials it also increases the computational complexity of a naive implementation of the scheme, as our finite element approximation is a system of equations. However, it is possible to rewrite the energy conserving scheme in *primal form* and implement it as one would for local dG as is discussed in Remark 4.2.6 leading to both schemes having a comparable computational complexity.

While the momentum conserving scheme is required to be discontinuous, due to the handling of the third derivative, for the energy conserving scheme we could restrict our scheme to the continuous finite element space. We shall not consider the continuous restriction of our scheme here, but we implement a scheme following the same methodology as the energy conserving scheme in Chapter 8 which utilises continuous finite element spaces.
We note that, in view of Theorem 5.1.4, the momentum has proven a priori bounds under the assumption that the degree of the finite element space q has *even parity* and an *even number of nodes*. We do not have such a priori bounds for the energy conserving scheme, primarily because, as we discussed in Remark 5.1.1, the conservation of energy does not lead us to numerical stability in a norm.

5.2 Temporal discretisations

We now discuss the temporal discretisations which we shall couple with the momentum conserving and energy conserving schemes. It is important to note that the temporal discretisation for conservation of each invariant is distinct as *the invariants have different order nonlinearities*. Unlike for linearised KdV, see §4.2, a single temporal discretisation cannot be conservative for both momentum and energy. Note that both of these temporal discretisations are mass conserving.
Recall that we define the temporal partition $0 =: t_0 < t_1 < \cdots < t_N := T$ with a step size $\tau_n := t_{n+1} - t_n$, additionally recall that we denote a temporally discrete function of the function $u(t, x)$ at t_n as $u^n(x)$. The temporal discretisations we discuss will be fixed at second order, as can be seen through Taylor's theorem. While higher order conservative temporal methods are available, for example, through composition of the methods discussed in this work, we shall not press this point here.

5.2.1 Momentum conserving temporal discretisation

In §4.2 we discussed how the Crank-Nicholson temporal discretisation preserved quadratic invariants. As momentum is quadratic in order we again choose a Crank-Nicholson discretisation, i.e., we define our temporal discretisation as follows.

Definition 5.2.1 (Momentum conserving temporal discretisation)**.** *Let* $u_0 = u(0, x)$*, then*

for $n = 0, ..., N-1$ find u^{n+1} such that

$$\frac{u^{n+1} - u^n}{\tau_n} + 6u_x^{n+\frac{1}{2}} u^{n+\frac{1}{2}} + u_{xxx}^{n+\frac{1}{2}} = 0, \tag{5.11}$$

where $u^{n+\frac{1}{2}} = \frac{u^{n+1}+u^n}{2}$.

Remark 5.2.2 (Crank-Nicholson timestepping)**.** *For a temporal problem of the form*

$$u_t = f(u),$$

the Crank-Nicolson temporal discretisation typically takes the form

$$\frac{u^{n+1} - u^n}{\tau_n} = \frac{f\left(u^{n+1}\right) + f\left(u^n\right)}{2}.$$

For our temporally discrete momentum conserving scheme this may not initially appear to be the case, due to the handling of the nonlinear term $6u_x^{n+\frac{1}{2}} u^{n+\frac{1}{2}}$, however as we are spatially continuous we can rewrite this term as the total derivative $3\left(\left(u^{n+\frac{1}{2}}\right)^2\right)_x$ confirming that our temporal discretisation is Crank-Nicholson in nature.

Proposition 5.2.3 (Nodal conservation of momentum by (5.11))**.** *Let u^n for $n = 0, .., N$ be as described in Definition 5.2.1, then the mass and momentum are conserved nodally, i.e.,*

$$\mathcal{F}_1\left(u^{n+1}\right) = \mathcal{F}_1(u^n), \qquad \mathcal{F}_2\left(u^{n+1}\right) = \mathcal{F}_2(u^n).$$

Proof. In view of mass conservation we integrate (5.11) spatially and observe that

$$\left\langle u^{n+1} - u^n, 1 \right\rangle = -\left\langle 3\left(\left(u^{n+\frac{1}{2}}\right)^2\right)_x + u_{xxx}^{n+\frac{1}{2}}, 1 \right\rangle = 0,$$

by the fundamental theorem of calculus similar to the continuous argument. For conservation of momentum we multiply (5.11) by $u^{n+\frac{1}{2}}$, then integrating over the spatial domain we yield

$$\begin{aligned}
\mathcal{F}_2\left(u^{n+1}\right) - \mathcal{F}_2(u^n) &= \frac{1}{2}\left\langle u^{n+1}, u^{n+1} \right\rangle - \frac{1}{2}\left\langle u^n, u^n \right\rangle = \left\langle u^{n+1} - u^n, u^{n+\frac{1}{2}} \right\rangle \\
&= -\tau_n \left\langle 6u^{n+\frac{1}{2}} u_x^{n+\frac{1}{2}} + u_{xxx}^{n+\frac{1}{2}}, u^{n+\frac{1}{2}} \right\rangle \\
&= -\tau_n \left\langle 2\left(\left(u^{n+\frac{1}{2}}\right)^3\right)_x - \frac{1}{2}\left(\left(u_x^{n+\frac{1}{2}}\right)^2\right)_x, 1 \right\rangle \\
&= 0,
\end{aligned}$$

after integration by parts, the fundamental theorem of calculus and the periodic spatial boundary conditions.

□

5.2.2 Energy conserving temporal discretisation

Due to the cubic nature of the energy we cannot employ the Crank-Nicholson temporal discretisation to yield an energy conserving scheme. Instead, we introduce a temporal discretisation designed specifically such that the energy is conserved. This temporal discretisation can be viewed as a second order perturbation of Crank-Nicholson, through Taylor's theorem, with respect to the variable u, see Remark 5.2.8.

Definition 5.2.4 (Energy conserving temporal discretisation)**.** *Let $u_0 = u(0,x)$, then for $n = 0, ..., N-1$ find u^{n+1} and v^{n+1} such that*

$$\begin{aligned} \frac{u^{n+1} - u^n}{\tau_n} + v_x^{n+1} &= 0 \\ v^{n+1} - \mathcal{K}\left(u^{n+1}, u^n\right) - u_{xx}^{n+\frac{1}{2}} &= 0, \end{aligned} \tag{5.12}$$

where $u^{n+\frac{1}{2}} = \frac{u^{n+1}+u^n}{2}$ and

$$\mathcal{K}\left(u^{n+1}, u^n\right) := \left(u^{n+1}\right)^2 + u^{n+1}u^n + \left(u^n\right)^2.$$

Remark 5.2.5 (Temporal treatment of the auxiliary variable)**.** *Note that we design the temporal discretisation of our auxiliary variable v such that it is diagnostic, i.e., we evaluate it at t_{n+1} to bypass the need to provide initial data for v which in practice introduces additional errors. This* does not reduce the order of accuracy in time *as temporally we can rewrite the scheme in primal form eliminating the auxiliary variable.*

Proposition 5.2.6 (Nodal conservation of energy by (5.12))**.** *Let u^n for $n = 0, ..., N$ be as described in Definition 5.2.4, then the mass and energy are conserved nodally, i.e.,*

$$\mathcal{F}_1\left(u^{n+1}\right) = \mathcal{F}_1(u^n), \qquad \mathcal{F}_3\left(u^{n+1}\right) = \mathcal{F}_3(u^n).$$

Proof. In view of mass conservation we have that

$$\mathcal{F}_1\left(u^{n+1}\right) - \mathcal{F}_1(u^n) = \left\langle u^{n+1} - u^n, 1\right\rangle = -\tau_n \left\langle v_x^{n+1}, 1\right\rangle = 0,$$

by the fundamental theorem of calculus and the periodic boundary conditions.
In view of energy conservation, through algebraic manipulation and integration by parts, we see that

$$\begin{aligned}\mathcal{F}_3\left(u^{n+1}\right) - \mathcal{F}_3(u^n) &= \frac{1}{2}\left\langle u_x^{n+1}, u_x^{n+1}\right\rangle - \frac{1}{2}\left\langle u_x^n, u_x^n\right\rangle - \left\langle \left(u^{n+1}\right)^2, u^{n+1}\right\rangle + \left\langle (u^n)^2, u^n\right\rangle \\ &= \left\langle \left(u^{n+1}\right)^2 + u^{n+1}u^n + (u^n)^2, u^{n+1} - u^n\right\rangle - \left\langle u_{xx}^{n+\frac{1}{2}}, u^{n+1} - u^n\right\rangle.\end{aligned}$$

Choosing $\psi = \frac{u^{n+1}-u^n}{\tau_n}$ in (5.12) allows us to write that

$$\mathcal{F}_3\left(u^{n+1}\right) - \mathcal{F}_3(u^n) = \tau_n \left\langle u^{n+1} - u^n, v^{n+1}\right\rangle.$$

Note that this choice of ψ is similar to the choice made to show energy conservation in the spatial case, but instead of a time derivative we choose a discrete time derivative, realised as a difference quotient, as a test function. Choosing $\phi = v^{n+1}$ in (5.12) we see that

$$\mathcal{F}_3\left(u^{n+1}\right) - \mathcal{F}_3(u^n) = \tau_n^2 \left\langle v_x^{n+1}, v^{n+1}\right\rangle = \tau_n^2 \left\langle \left(\left(v^{n+1}\right)^2\right)_x, 1\right\rangle = 0,$$

i.e., the energy is conserved nodally.

□

Remark 5.2.7 (The design of temporal discretisations that preserve higher order invariants)**.** *The key to designing a temporal discretisation that preserves a nonquadratic invariant is solely in the discretisation of the nonlinear term of the scheme. Notice that for the momentum conserving scheme when we multiply our difference quotient with* $u^{n+\frac{1}{2}}$ *we obtain a difference of squares, i.e.,*

$$\left(u^{n+1} - u^n\right) u^{n+\frac{1}{2}} = \left(u^{n+1}\right)^2 - (u^n)^2.$$

To obtain a difference of cubes, as is required of the energy conserving scheme, we define our nonlinear term $\mathcal{K}\left(u^{n+1}, u^n\right)$ *such that we obtain a difference of cubes, i.e.,*

$$\mathcal{K}\left(u^{n+1}, u^n\right) u^{n+\frac{1}{2}} = \left(u^{n+1}\right)^3 - (u^n)^3.$$

Following this methodology we can design temporal discretisations for problems with high order nonlinearities, such as the modified KdV equation, as we will discuss later in this chapter, or the vectorial *modified KdV equation, see Chapter 8. Note that this idea is*

similar to that used in the design of discrete gradient methods which are discussed in §2.2.3.

Remark 5.2.8 (Deviation of the nonconserved invariant)**.** *Note that, after rewriting the energy conserving temporal scheme in primal form, the only difference between the momentum conserving scheme and the energy conserving scheme is the discretisation of the nonlinear term. We can explicitly write the nonlinear term of the momentum conserving scheme* (5.11) *as*

$$3\left(\left(u^{n+\frac{1}{2}}\right)^2\right)_x,$$

and for the energy conserving scheme (5.12) *as*

$$\left(\left(u^{n+1}\right)^2 + u^{n+1}u^n + (u^n)^2\right)_x.$$

Through Taylor expanding we observe that the difference between these nonlinear terms is second order, i.e.,

$$\left(u^{n+1}\right)^2 + u^{n+1}u^n + (u^n)^2 = 3\left(u^{n+\frac{1}{2}}\right)^2 + \frac{1}{4}\tau_n^2\frac{\mathrm{d}}{\mathrm{d}\xi}u(\xi,x) \quad \text{for some } \xi \in [t_n, t_{n+1}]\,. \tag{5.13}$$

In view of (5.13) *we have that for the momentum conserving scheme* (5.11) *the local deviation in energy is* $\mathcal{O}(\tau_n^2)$*, and similarly for the energy conserving scheme* (5.12) *the local deviation in momentum is* $\mathcal{O}(\tau_n^2)$*.*

Remark 5.2.9 (Order of the temporal discretisations)**.** *We can observe, through direct application of Taylor's theorem, that both* (5.11) *and* (5.12) *are* second *order accurate in time.*

5.3 Full discretisations

We will now combine the spatial and temporal schemes for both momentum conserving and energy conserving discretisations, and show that the fully discrete methods preserve their respective invariants at the temporal nodes. These proofs will be an amalgamation of the spatially discrete and temporally discrete proofs.

Recall that we define spatially discrete finite element functions through capitalisation, and temporally discrete functions through superscripts. With this in mind, we write $U^n(x)$ as the fully discrete function approximating $u(t,x)$.

Unlike in the prequel, here we allow our *spatial mesh to change over time.* This requires us to redefine our spatial notation with dependence on time as follows. We refine our spatial partition as $0 =: x_0^n < x_1^n < \cdots < x_M^n := 1$ with elements $\mathcal{J}_m^n := (x_m^n, x_{m+1}^n)$ possessing length $h_m^n := x_{m+1}^n - x_m^n$. Note that here superscripts of spatial points do not necessarily represent a sequence of points, as we have not assumed that the number of points are constant. It is possible that the number of points changes over time.

Definition 5.3.1 (Fully discrete finite element space)**.** *Let $\mathbb{P}_q^n(\mathcal{J}_m^n)$ denote the space of polynomials of degree q on an interval $\mathcal{J}_m^n \subset \mathbb{R}$ at time $t = t_n$, then the discontinuous finite element space is given by*

$$\mathbb{V}_q^n = \{U^n : S^1 \to \mathbb{R} : U^n|_{\mathcal{J}_m^n} \in \mathbb{P}_q^n(\mathcal{J}_m^n) \text{ for } m = 0, ..., M-1\},$$

where M depends on the time step, and for $n = 0, ..., N$.

Additionally, as our spatial mesh can change over time, for our numerical methods to be well defined we are required to introduce an mesh change operator which maps from the old mesh at time n to the new mesh at time $n+1$, denoted $\mathcal{P}^{n+1} : \mathbb{V}_q^n \to \mathbb{V}_q^{n+1}$. We shall define this operator more concisely in the sequel, but in the literature this operator is typically either the Lagrange interpolation operator onto the new mesh or the L_2 projection. In the case that $\mathbb{V}_q^{n+1} \equiv \mathbb{V}_q^n$ then $\mathcal{P}^{n+1}$ is the identity as the mesh does not change.
Before proceeding to introduce our fully discrete numerical schemes we first briefly discuss different potential methods of adaptivity.

Remark 5.3.2 (Different methods of adaptivity)**.** *In the literature three types of adaptivity are typically considered for finite element methods. The first of which is r-adaptivity, see [35]. In r-adaptivity the number of nodal points is* constant*, and the degrees of freedom of the method are simply redistributed, or* relocated*. A key benefit to this method of adaptivity is that the dimension of the finite element space does not change in time, and as such we have a predetermined computational complexity. This type of adaptivity is often driven by a Monge-Ampere type equation, see [36, 150], which in one spatial dimension is equivalent to the Laplacian. This type of adaptivity is also referred to as a moving mesh method, typically where the mesh is subject to Lagrangian flow, see [41, 18, 126].*
The second type of adaptivity is known as h-adaptivity, as the spatial step, typically denoted h, is adapted locally over time, see [57]. Ultimately this kind of adaptivity can be viewed as coarsening/refining the elements of the spatial mesh for a finite element method. This is a popular method of adaptivity within the finite element framework primarily as the analysis

of such methods falls within the finite element framework, as such a posteriori bounds can be constructed and their magnitude can be controlled locally through h-adaptivity, see [6]. Is it possible, if desired, to combine r-adaptivity with h-adaptivity, see [14].
The final type of adaptivity, which is implemented within the finite element framework, is known as p-adaptivity. Here the polynomial *degree of the finite element approximation is adapted locally over time, see [25, 61, 64, 112]. While p-adaptivity yields very fast convergence of solutions for smooth solutions, it is not ideal for representaing discontinuities. This has led to the coupling of p-adaptivity with h-adaptivity which handle discontinuities relatively well, see [54, 97, 169, 55, 53, 52] on hp-adaptive schemes.*
The methods that we propose allow for r-adaptivity and h-adaptivity. Throughout this work we assume that the polynomial degree of the finite element method is fixed. Note that p-adaptivity can in fact be considered, but we do not accommodate it here for clarity of exposition.

5.3.1 Fully discrete momentum conserving scheme

Definition 5.3.3 (Momentum conserving fully discrete scheme, [31])**.** *Let $U^n \in \mathbb{V}_q^n$ and $U^0 = \Pi^0 u_0$ where Π^0 is the L_2 projection into the initial finite element space. Then we seek $U^{n+1} \in \mathbb{V}_q^{n+1}$ such that*

$$\left\langle \frac{U^{n+1} - \mathcal{P}^{n+1}U^n}{\tau_n} + 3\mathcal{N}\left(U^{n+\frac{1}{2}}\right) + \mathcal{D}\left(U^{n+\frac{1}{2}}\right), \phi \right\rangle = 0 \qquad \forall \phi \in \mathbb{V}_q^{n+1}, \tag{5.14}$$

where $U^{n+\frac{1}{2}} = \frac{1}{2}\left(U^{n+1} + \mathcal{P}^{n+1}U^n\right)$, and $\mathcal{N}$ and $\mathcal{D}$ are defined by (5.5) *and* (5.6)*.*

Proposition 5.3.4 (Conservative properties of (5.14))**.** *Let U^n for $n = 0, ..., N-1$ be as described in Definition 5.3.3, then we have that*

$$\mathcal{F}_1\Big(U^{n+1}\Big) = \mathcal{F}_1\Big(\mathcal{P}^{n+1}U^n\Big),$$

and

$$\mathcal{F}_2\Big(U^{n+1}\Big) = \mathcal{F}_2\Big(\mathcal{P}^{n+1}U^n\Big),$$

where $\mathcal{F}_1$ and $\mathcal{F}_2$ are given in (5.2) *and* (5.3) *respectively. Ultimately this tells us that the deviation in mass and momentum of the momentum conserving scheme is controlled by the interpolation of the solution from the spatial mesh at t_n to the spatial mesh at t_{n+1}.*

Remark 5.3.5 (A remark on Proposition 5.3.4)**.** *If our mesh is not adaptive then the mass*

and momentum of the "momentum conserving scheme" will *be conserved, as the mesh change operator $\mathcal{P}^{n+1}$ be the identity operator. In general, over an adaptive mesh, to preserve invariants we need to design the mesh change operator $\mathcal{P}^{n+1}$ such that $\mathcal{F}_i(\mathcal{P}^{n+1}U^n) = \mathcal{F}_i(U^n)$ for a given invariant. We focus on the design of such operators in §6.2 and §7.4.*

Proof of Proposition 5.3.4. In view of mass conservation add and subtract $\mathcal{F}_1(\mathcal{P}^{n+1}U^n)$, choose $\phi = 1$ in (5.14) and utilise the orthogonality of the nonlinear and dispersion operators to the constants to find

$$\begin{aligned}\mathcal{F}_1\left(U^{n+1}\right) - \mathcal{F}_1(U^n) &= \left\langle U^{n+1} - U^n, 1\right\rangle = \left\langle U^{n+1} - \mathcal{P}^{n+1}U^n, 1\right\rangle + \left\langle \mathcal{P}^{n+1}U^n - U^n, 1\right\rangle \\ &= -\tau_n \left\langle 3\mathcal{N}\left(U^{n+\frac{1}{2}}\right) + \mathcal{D}\left(U^{n+\frac{1}{2}}\right), 1\right\rangle + \left\langle \mathcal{P}^{n+1}U^n - U^n, 1\right\rangle \\ &= \left\langle \mathcal{P}^{n+1}U^n - U^n, 1\right\rangle.\end{aligned}$$

In the same spirit as the spatially discrete case we choose $\phi = U^{n+\frac{1}{2}}$ to show that

$$\begin{aligned}\mathcal{F}_2\left(U^{n+1}\right) - \mathcal{F}_2(U^n) &= \frac{1}{2}\left\langle U^{n+1}, U^{n+1}\right\rangle - \frac{1}{2}\left\langle U^n, U^n\right\rangle \\ &= \frac{1}{2}\left\langle U^{n+1}, U^{n+1}\right\rangle - \frac{1}{2}\left\langle \mathcal{P}^{n+1}U^n, \mathcal{P}^{n+1}U^n\right\rangle \\ &\quad + \frac{1}{2}\left\langle \mathcal{P}^{n+1}U^n, \mathcal{P}^{n+1}U^n\right\rangle - \frac{1}{2}\left\langle U^n, U^n\right\rangle \\ &= \left\langle U^{n+1} - U^n, U^{n+\frac{1}{2}}\right\rangle + \frac{1}{2}\left\langle \mathcal{P}^{n+1}U^n, \mathcal{P}^{n+1}U^n\right\rangle - \frac{1}{2}\left\langle U^n, U^n\right\rangle \\ &= -\tau_n \left\langle 3\mathcal{N}\left(U^{n+\frac{1}{2}}\right) + \mathcal{D}\left(U^{n+\frac{1}{2}}\right), U^{n+\frac{1}{2}}\right\rangle \\ &\quad + \frac{1}{2}\left\langle \mathcal{P}^{n+1}U^n, \mathcal{P}^{n+1}U^n\right\rangle - \frac{1}{2}\left\langle U^n, U^n\right\rangle \\ &= \frac{1}{2}\left\langle \mathcal{P}^{n+1}U^n, \mathcal{P}^{n+1}U^n\right\rangle - \frac{1}{2}\left\langle U^n, U^n\right\rangle,\end{aligned}$$

by the orthogonality of the non-linear and dispersion operators to their argument discussed in §5.1.3.

□

Corollary 5.3.6 (Stability of the momentum conserving scheme)**.** *Assume that $\mathbb{V}_q^n \subseteq \mathbb{V}_q^{n+1}$ and that $\mathcal{P}^{n+1}$ is consistent, then the momentum conserving scheme described in Definition 5.3.3 is stable in $L_2(S^1)$ over time, i.e.,*

$$\|U^n\|_{L_2(S^1)} = \|U^0\|_{L_2(S^1)}.$$

Proof. This result follows directly from conservation of momentum discussed in Proposition 5.3.4.

□

5.3.2 Energy conserving scheme

Definition 5.3.7 (Energy conserving fully discrete scheme)**.** *Let $U^n \in \mathbb{V}_q^n$ be given, then seek $U^{n+1}, V^{n+1} \in \mathbb{V}_q^{n+1}$ such that*

$$\begin{aligned} \left\langle \frac{U^{n+1} - \mathcal{P}^{n+1}U^n}{\tau_n} + \mathcal{G}\left(V^{n+1}\right), \phi \right\rangle = 0 \qquad & \forall \phi \in \mathbb{V}_q^{n+1} \\ \left\langle V^{n+1} - \mathcal{K}\left(U^{n+1}, \mathcal{P}^{n+1}U^n\right), \psi \right\rangle + \mathcal{A}_h\left(U^{n+\frac{1}{2}}, \psi\right) = 0 \qquad & \forall \psi \in \mathbb{V}_q^{n+1}, \end{aligned} \tag{5.15}$$

where

$$\mathcal{K}\left(U^{n+1}, U^n\right) = \left(U^{n+1}\right)^2 + U^{n+1}\mathcal{P}^{n+1}U^n + \left(\mathcal{P}^{n+1}U^n\right)^2, \tag{5.16}$$

$U^{n+\frac{1}{2}} = \frac{1}{2}\left(U^{n+1} + \mathcal{P}^{n+1}U^n\right)$, $\mathcal{A}_h(\cdot,\cdot)$ and $\mathcal{G}$ are described by Definition 5.1.6 and Definition 4.2.3 respectively. Additionally U^0 is given by the L_2 projection of u_0 into the initial finite element space.

Proposition 5.3.8 (Conservative properties of (5.15))**.** *Let U^n for $n = 0, ..., N-1$ be as described by Definition 5.3.7, then we have that*

$$\mathcal{F}_1\Big(U^{n+1}\Big) = \mathcal{F}_1\Big(\mathcal{P}^{n+1}U^n\Big),$$

and

$$\widetilde{\mathcal{F}_3}\left(U^{n+1}\right) = \widetilde{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right),$$

where $\mathcal{F}_1$ and $\widetilde{\mathcal{F}_3}$ are given in (5.2) *and* (5.10) *respectively. Similarly to in Proposition 5.3.4, for the momentum conserving scheme, we have that the deviation in mass and energy for our fully discrete scheme are controlled by the interpolation of the numerical solution between the spatial mesh at time t_n to the spatial mesh at time t_{n+1}. If we do not adapt the spatial mesh over time we conserve the discrete mass and energy exactly.*

Before proving Proposition 5.3.8 we first need to introduce a discrete identity for $\mathcal{A}_h$.

Lemma 5.3.9 (Discrete identity for the bilinear form)**.** *Let $U^{n+1} \in \mathbb{V}_q^{n+1}$ and $U^n \in \mathbb{V}_q^n$,*

then

$$\mathcal{A}_h\left(U^{n+\frac{1}{2}}, \frac{U^{n+1}-U^n}{\tau_n}\right) = \frac{1}{2\tau_n}\left(\mathcal{A}_h\left(U^{n+1}, U^{n+1}\right) - \mathcal{A}_h\left(U^n, U^n\right)\right).$$

Proof. Applying the definition of $U^{n+\frac{1}{2}}$ and bilinearity we find

$$\begin{aligned}\mathcal{A}_h\left(U^{n+\frac{1}{2}}, \frac{U^{n+1}-U^n}{\tau_n}\right) &= \frac{1}{2\tau_n}\mathcal{A}_h\left(U^{n+1}+U^n, U^{n+1}-U^n\right)\\ &= \frac{1}{2\tau_n}\left(\mathcal{A}_h\left(U^{n+1}, U^{n+1}\right) - \mathcal{A}_h\left(U^n, U^n\right)\right)\end{aligned}$$

after utilising symmetry of the bilinear form. □

Proof of Proposition 5.3.8. In view of energy conservation we add and subtract the energy of $\mathcal{P}^{n+1}U^n$ allowing us to write

$$\begin{aligned}\widetilde{\mathcal{F}_3}\left(U^{n+1}\right) - \widetilde{\mathcal{F}_3}\left(U^n\right) &= \widetilde{\mathcal{F}_3}\left(U^{n+1}\right) - \widetilde{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right) + \widetilde{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right) - \widetilde{\mathcal{F}_3}\left(U^n\right)\\ &= \frac{1}{2}\left(\mathcal{A}_h\left(U^{n+1}, U^{n+1}\right) - \mathcal{A}_h\left(\mathcal{P}^{n+1}U^n, \mathcal{P}^{n+1}U^n\right)\right)\\ &\quad - \left\langle\left(U^{n+1}\right)^2, U^{n+1}\right\rangle + \left\langle\left(\mathcal{P}^{n+1}U^n\right)^2, \mathcal{P}^{n+1}U^n\right\rangle\\ &\quad + \widetilde{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right) - \widetilde{\mathcal{F}_3}\left(U^n\right).\end{aligned}$$

Applying Lemma 5.3.9 and then choosing $\psi = \frac{U^{n+1}-U^n}{\tau_n}$ in (5.15) we have find that

$$\begin{aligned}\widetilde{\mathcal{F}_3}\left(U^{n+1}\right) - \widetilde{\mathcal{F}_3}\left(U^n\right) &= \tau_n\mathcal{A}_h\left(U^{n+\frac{1}{2}}, \frac{U^{n+1}-\mathcal{P}^{n+1}U^n}{\tau_n}\right)\\ &\quad - \tau_n\left\langle\mathcal{K}\left(U^{n+1}, \mathcal{P}^{n+1}\mathcal{P}^{n+1}U^n\right), \frac{U^{n+1}-\mathcal{P}^{n+1}U^n}{\tau_n}\right\rangle\\ &\quad + \widetilde{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right) - \widetilde{\mathcal{F}_3}\left(U^n\right)\\ &= \tau_n\left\langle V^{n+1}, U^{n+1}-U^n\right\rangle + \widetilde{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right) - \widetilde{\mathcal{F}_3}\left(U^n\right)\end{aligned}$$

where $\mathcal{K}$ is defined by (5.16). Note that choosing $\phi = V^{n+1}$ in (5.15) we have

$$\begin{aligned}\widetilde{\mathcal{F}_3}\left(U^{n+1}\right) - \widetilde{\mathcal{F}_3}\left(U^n\right) &= \tau_n^2\left\langle V^{n+1}, \mathcal{G}\left(V^{n+1}\right)\right\rangle + \widetilde{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right) - \widetilde{\mathcal{F}_3}\left(U^n\right)\\ &= \widetilde{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right) - \widetilde{\mathcal{F}_3}\left(U^n\right),\end{aligned}$$

through the skew-symmetry of $\mathcal{G}$, see Lemma 4.2.4, as required.

□

Remark 5.3.10 (Preservation of invariants and adaptivity)**.** *As we observe from Proposition 5.3.4 and Proposition 5.3.8, we cannot freely adapt our mesh over time and expect to preserve the higher order invariants of the problem. In fact, we cannot even expect to preserve the mass. This presents a significant problem with the employment of adaptivity, as numerical stability in both the momentum and energy conserving schemes arises from conservation of their respective invariants. As such, we cannot expect an adaptive algorithm to be well behaved. This motivates the design of mesh change operators $\mathcal{P}^{n+1}$ such that the invariants are still preserved to regain some of this numerical stability. We shall elaborate on this point in Chapter 6.* For the remainder of this chapter we shall assume a uniform mesh in both space and time.

Remark 5.3.11 (Conservative schemes for the modified KdV equation)**.** *Throughout this chapter we have restricted our study to the KdV equation, but neither the momentum or the energy conserving schemes need to be restricted to such cases. Consider, for example, the modified KdV equation*

$$u_t + 24u^2u_x + u_{xxx} = 0,$$

which possesses the mass, momentum and energy

$$\langle u, 1\rangle, \qquad \frac{1}{2}\langle u, u\rangle, \qquad \frac{1}{8}\langle u_x, u_x\rangle - \frac{1}{2}\left\langle u^2, u^2\right\rangle,$$

respectively. The corresponding momentum conserving scheme in this case is given by letting $U^n \in \mathbb{V}_q^n$ be given, where $U^0 = \Pi^0 u_0$ with Π^0 the L_2 projection into the initial finite element space. Then, we seek $U^{n+1} \in \mathbb{V}_q^{n+1}$ such that

$$\left\langle \frac{U^{n+1} - \mathcal{P}^{n+1}U^n}{\tau_n} + 24\mathcal{N}\left(U^{n+\frac{1}{2}}\right) + \mathcal{D}\left(U^{n+\frac{1}{2}}\right), \phi \right\rangle = 0 \qquad \forall \phi \in \mathbb{V}_q^{n+1},$$

where $U^{n+\frac{1}{2}} = \frac{1}{2}\left(U^{n+1} + \mathcal{P}^{n+1}U^n\right)$, and we redefine our nonlinear operator such that for $W \in \mathbb{V}_q^{n+1}$

$$\langle \mathcal{N}(W), \phi\rangle = -\left\langle W^3, \phi_x\right\rangle + \frac{1}{4}\sum_{m=0}^{M-1}\left(\left(W_m^+\right)^3 + \left(W_m^+\right)^2 W_m^- + W_m^+\left(W_m^-\right)^2 + \left(W_m^-\right)^3\right)[\![\phi_m]\!].$$

and $\mathcal{D}$ is as described in (5.6)*, as can be seen in [31]. Mimicking the proof of Proposition 5.3.8 we find that a discrete mass and momentum are still preserved for this scheme in*

the sense that

$$\mathcal{F}_1\Big(U^{n+1}\Big) = \mathcal{F}_1\Big(\mathcal{P}^{n+1}U^n\Big)$$
$$\mathcal{F}_2\Big(U^{n+1}\Big) = \mathcal{F}_2\Big(\mathcal{P}^{n+1}U^n\Big).$$

Equivalently, we can introduce the first variation of the energy $v = 8u^3 + u_{xx}$ *as an auxiliary variable yielding the following scheme. Let* $U^n \in \mathbb{V}_q^n$ *be given, where* $U^0 = \Pi^0 u_0$ *where* Π^0 *is the initial* L_2 *projection into the finite element space. Further let* $\mathcal{G}$ *and* $\mathcal{A}_h$ *be as described in Definition 4.2.3 and Definition 5.1.6, then we seek* $U^{n+1} \in \mathbb{V}_q^{n+1}$ *such that*

$$\left\langle \frac{U^{n+1} - \mathcal{P}^{n+1}U^n}{\tau_n} + \mathcal{G}\left(V^{n+1}\right), \phi \right\rangle = 0 \qquad \forall \phi \in \mathbb{V}_q^{n+1}$$
$$\left\langle V^{n+1} - 8\mathcal{K}\left(U^{n+1}, \mathcal{P}^{n+1}U^n\right), \psi \right\rangle \mathcal{A}_h\left(U^{n+\frac{1}{2}}, \psi\right) = 0 \qquad \forall \psi \in \mathbb{V}_q^{n+1},$$

where $U^{n+\frac{1}{2}} = \frac{1}{2}\left(U^{n+1} + \mathcal{P}^{n+1}U^n\right)$ *and we redefine our nonlinear operator* $\mathcal{K}$ *as*

$$\mathcal{K}\left(U^{n+1}, \mathcal{P}^{n+1}U^n\right) = \frac{1}{4}\left(\left(U^{n+1}\right)^3 + \left(U^{n+1}\right)^2 \mathcal{P}^{n+1}U^n + U^{n+1}\left(\mathcal{P}^{n+1}U^n\right)^2 + \left(\mathcal{P}^{n+1}U^n\right)^3\right).$$

Mimicking the proof of Proposition of 5.3.8 we see that we conserve mass and the discrete energy

$$\widehat{\mathcal{F}_3}\left(U^n\right) = \frac{1}{8}\mathcal{A}_h\left(U^n, U^n\right) - \frac{1}{2}\left(U^n\right)^3,$$

in the sense that

$$\mathcal{F}_1\Big(U^{n+1}\Big) = \mathcal{F}_1\Big(\mathcal{P}^{n+1}U^n\Big)$$
$$\widehat{\mathcal{F}_3}\left(U^{n+1}\right) = \widehat{\mathcal{F}_3}\left(\mathcal{P}^{n+1}U^n\right).$$

5.4 Numerical experiments

Here we run numerical experiments with the aim of comparing the momentum conserving scheme (5.14) and energy conserving scheme (5.15). Similarly to previous chapters we implement our schemes using the automated system for solving finite element methods Firedrake [153]. We employ a Gauss quadrature of degree $3q$, which is exact for finite element functions. When computing errors we shall employ a degree $3q+4$ Gauss quadrature. We approximately solve the nonlinear component of our finite element scheme using the PETSc Newton line search method, see [20], to a tolerance of 10^{-12}.

We restrict ourselves here to a *uniform* mesh. That is we assume that τ_n is constant for all n, that h_m is constant for all m, and that our finite element mesh does not change over time.

We benchmark our schemes in the L_∞ norm in time and for the spatial component we will consider two different norms: the $L_2(S^1)$ norm and an appropriate spatial energy norm which can be written as

$$e_u := \max_n \|U^n - u(t_n)\|_{L_2(S^1)} \tag{5.17}$$

and

$$e_{u_{dG}} := \max_n \sqrt{\|(U^n - u(t_n))_x\|^2_{L_2(S^1)} + \sum_{m=0}^{M} \frac{\sigma}{h_m} [\![U^n_m]\!]^2} \tag{5.18}$$

respectively, where σ is as given in Definition 5.1.6.

While our analysis is conducted over $S^1(0,1)$ we shall stretch our periodic spatial domain to $S^1(0,40)$ as the visualisation of solution dynamics is more difficult over smaller domains.

5.4.1 One soliton simulation

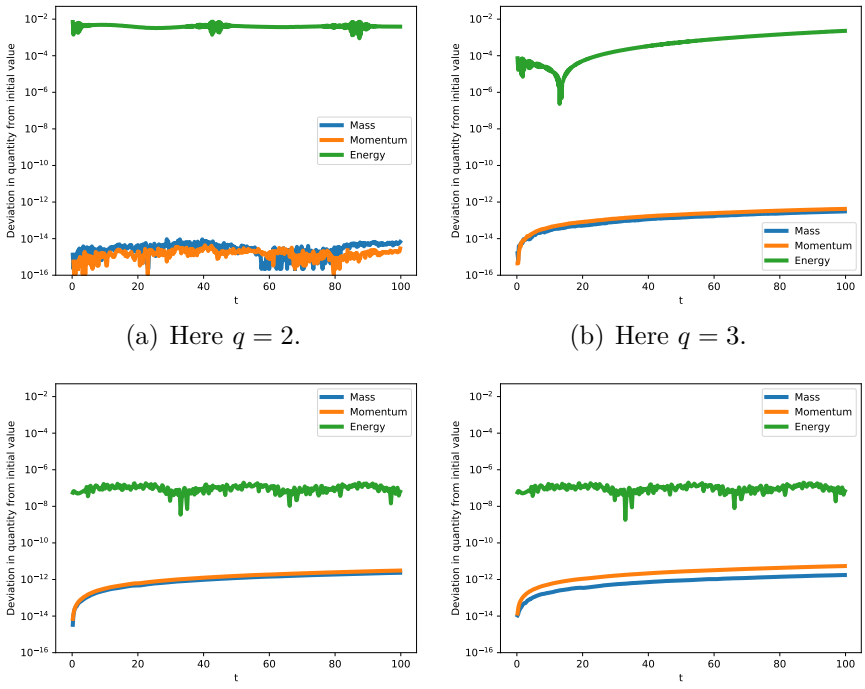
Consider the one soliton solution

$$u(t,x) = \frac{1}{2}\text{sech}\left(\frac{1}{2}\xi\right)^2, \tag{5.19}$$

where $\xi = (x - t + 20 \mod 40) - 20$.

We begin by computing the deviation in mass, momentum and energy for both of our schemes with degree $q = 2, 3, 4, 5$ over long time in Figure 5.1 and Figure 5.2, additionally we specify a nonlinear solver tolerance of 10^{-12}. Note that for the energy conserving simulation we can also obtain a degree $q = 1$ simulation which behaves similarly to the higher degree cases. In the simulation for the momentum conserving scheme we observe that the deviation in mass and momentum are below solver tolerance at each time step, and similarly the deviation in mass and energy for the energy conserving simulations are below solver tolerance at each time step. Globally however for the degree $q = 4, 5$ simulations the deviation in the conserved quantities increases above the solver tolerance globally for both schemes. This is due to the propagation of solver errors over time, see Remark 4.2.28. We notice that the deviation in energy for the momentum conserving scheme decreases as we increase the polynomial degree until the deviation is approximately 10^{-7}, whereas the deviation in momentum for the energy conserving scheme is approximately 10^{-7} regardless of spatial polynomial degree.

Figure 5.1: The deviation in mass, momentum and energy with $T = 100$ and initial condition (5.19) for the momentum conserving scheme (5.14). Further we choose $\tau_n = 0.2$, $h_m = 0.4$ and vary the polynomial degree q.



(a) Here $q = 2$.

(b) Here $q = 3$.

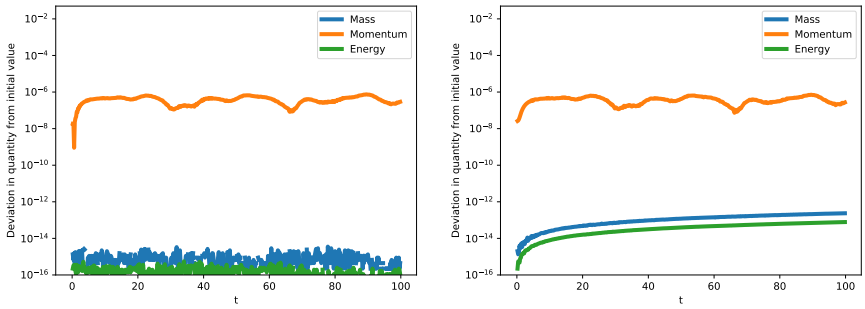
(c) Here $q = 4$.

(d) Here $q = 5$.

We benchmark our approximations against the exact solution (5.19) in Figure 5.3 for spatial degree $q = 2, 3$ with higher nonlinear solver tolerance of 10^{-8}. This higher tolerance is chosen here as it significantly decreases computational time. Note in these simulations we have fixed the time step to be sufficiently small so that it does not make a leading order contribution to the error. In all cases the EOC for the dG error (5.18) is one order lower than the L_2 error (5.17) so we shall not discuss them separately. We can also see that for the momentum conserving scheme the EOC in L_2 is 3 for both $q = 2$ and $q = 3$. This suggests a spatial error of $\mathcal{O}\left(h_m^{q+1}\right)$ for q even and $\mathcal{O}\left(h_m^{q}\right)$ for q odd. For the energy conserving scheme the spatial error appears to behave like $\mathcal{O}\left(h_m^{q+1}\right)$ for all q. Allowing the time step to vary in proportionately with the spatial step, i.e., $\tau_n = h_m$, we observe in Figure 5.4 that the temporal EOC is $\mathcal{O}\left(\tau_n^2\right)$ as we expect from our analysis.

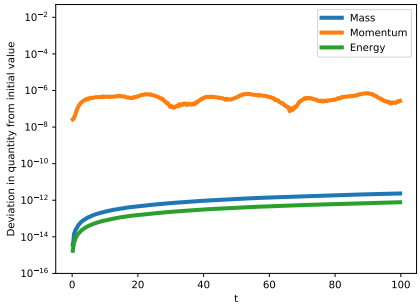
We shall also investigate how well the qualitative structure of the solution is captured. In the spirit of [32] we investigate the amplitude error, phase error and shape error of a single soliton over time.

Figure 5.2: The deviation in mass, momentum and energy with $T = 100$ and initial condition (5.19) for the energy conserving scheme (5.15). Further we choose $\tau_n = 0.2$, $h_m = 0.4$ and vary the polynomial degree q.

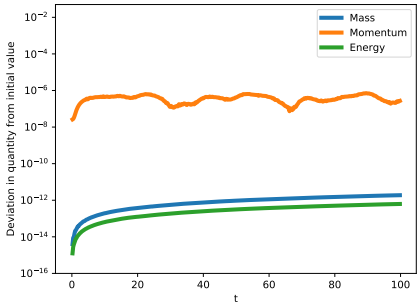


(a) Here $q = 2$.

(b) Here $q = 3$.



(c) Here $q = 4$.



(d) Here $q = 5$.

The amplitude error for U_i is then given by

$$\max_{\boldsymbol{X}} U_i - \max_{\boldsymbol{X}} u_i.$$

If the amplitude error is positive then the numerical soliton is larger than the exact solution, and vice versa. Similarly we define the phase error as

$$e_{p_i} = \operatorname*{\texttt{argmax}}_{\boldsymbol{X}} U_i - \operatorname*{\texttt{argmax}}_{\boldsymbol{X}} u_i,$$

where `argmax` represents the spatial coordinate associated to the maximum over $\boldsymbol{X}$. If the phase error is positive then the numerical approximation is moving faster than the exact solution, and vice versa. Note that this discrete measure of the error cannot detect shifts in phase which are smaller than the distance between degrees of freedom.
In addition to the amplitude and phase error we introduce the "shape error". In [32] the shape error is defined to be $\min_{y\in S^1} \|u(x+y,t_n) - U(x,t_n)\|_{L_2(S^1)}$. Numerically we approximate this error by shifting the exact solution by the distance of the phase error and computing the L_2 error at fixed times, i.e., the discrete shape error is

$$\left\|u_i(x+e_{p_i},t_n) - U_i(x,t_n)\right\|_{L_2(S^1([0,40]))}.$$

We tabulate the amplitude, phase and shape errors for the momentum and energy conserving discretisations initialised by the single soliton initial condition (5.19) in Table 5.1.

Table 5.1: Here we tabulate the phase, amplitude and shape errors committed by the momentum and energy conserving schemes (5.14) and (5.15) respectively, approximating the smooth solution (5.19). We display the minimal and maximal errors over the time interval $t \in [0, 100]$ for the phase and amplitude errors. As the shape error is signed we only display its maximal value over the interval. We show these errors for various coupled temporal and spatial discretisations and various polynomial degrees. We notice through inspecting the phase errors that both of our numerical solitons travel slower than their exact counterparts. Note that when the phase error is measured to be zero this *does not* mean that the phase of the numerical scheme is exact, only that the phase error is smaller than the distance between the degrees of freedom, $\frac{h_m}{q}$, of our numerical approximation.

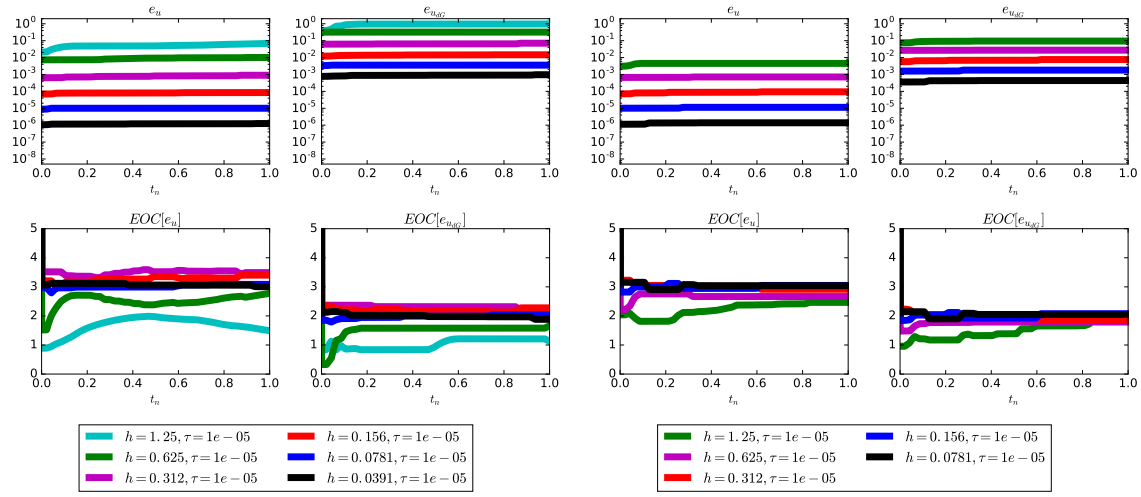
τ	h	Degree	Phase error		Amplitude error		Shape error
			min	max	min	max	max
2.5e-01	3.2e-01	2	-4.8e-01	0.0e+00	-2.6e-04	1.5e-03	3.3e-02
		3	-4.3e-01	0.0e+00	-3.4e-04	6.9e-04	2.1e-02
		4	-4.0e-01	0.0e+00	-2.4e-04	3.8e-04	1.5e-02
1.2e-01	1.6e-01	2	-1.6e-01	0.0e+00	-1.5e-04	3.7e-04	1.6e-02
		3	-1.1e-01	0.0e+00	-9.3e-05	1.4e-04	1.0e-02
		4	-1.2e-01	0.0e+00	-6.3e-05	1.1e-04	7.3e-03
6.2e-02	8.0e-02	2	-4.0e-02	0.0e+00	-5.2e-05	6.7e-05	7.4e-03
		3	-2.7e-02	0.0e+00	-2.3e-05	3.8e-05	5.1e-03
		4	-4.0e-02	0.0e+00	-2.1e-05	2.9e-05	3.7e-03

(a) Here we consider the momentum conserving scheme (5.14).

τ	h	Degree	Phase error		Amplitude error		Shape error
			min	max	min	max	max
2.5e-01	3.2e-01	1	-6.4e-01	0.0e+00	-3.7e-03	3.8e-03	5.8e-02
		2	-3.2e-01	0.0e+00	-1.1e-03	7.5e-04	2.9e-02
		3	-3.2e-01	0.0e+00	-6.8e-04	2.3e-04	1.9e-02
		4	-3.2e-01	0.0e+00	-6.2e-04	1.6e-04	1.5e-02
1.2e-01	1.6e-01	1	-1.6e-01	0.0e+00	-1.0e-03	1.1e-03	2.9e-02
		2	-8.0e-02	0.0e+00	-2.8e-04	1.9e-04	1.5e-02
		3	-1.1e-01	0.0e+00	-1.9e-04	8.0e-05	9.7e-03
		4	-1.2e-01	0.0e+00	-1.4e-04	4.7e-05	7.3e-03
6.2e-02	8.0e-02	1	-8.0e-02	0.0e+00	-2.4e-04	2.4e-04	1.5e-02
		2	-4.0e-02	0.0e+00	-7.4e-05	4.5e-05	7.3e-03
		3	-2.7e-02	0.0e+00	-4.6e-05	2.5e-05	4.9e-03
		4	-4.0e-02	0.0e+00	-3.7e-05	1.2e-05	3.6e-03

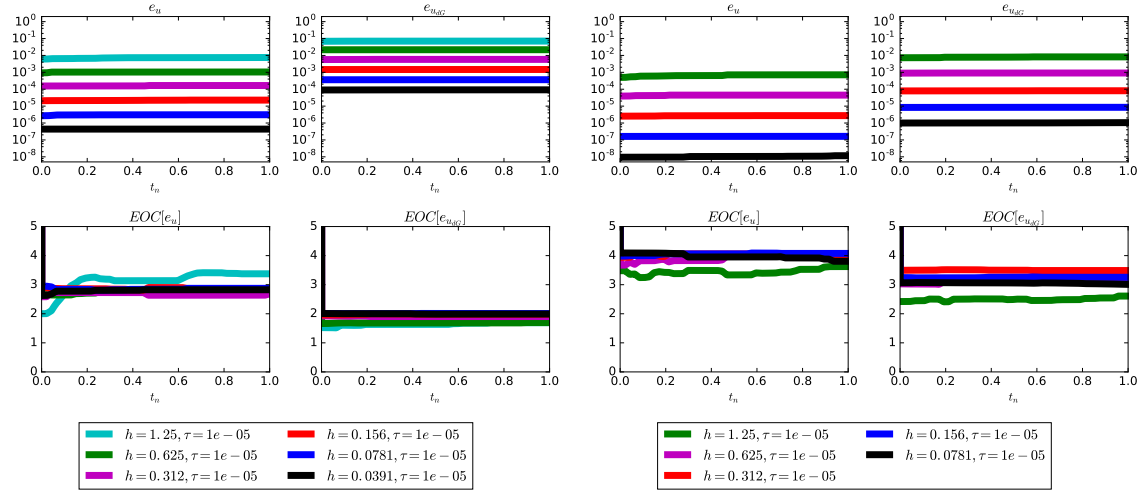
(b) Here we consider the energy conserving scheme (5.15).

Figure 5.3: The errors, as described in (5.17) and (5.18), and experimental order of convergence of the single soliton solution (5.19) for the momentum conserving scheme (5.14) or the energy conserving scheme (5.15) with polynomial degrees $q = 2, 3$. Here we fix $\tau_n = 0.00001$ and vary h_m.



(a) Momentum conserving scheme with spatial degree $q = 2$.

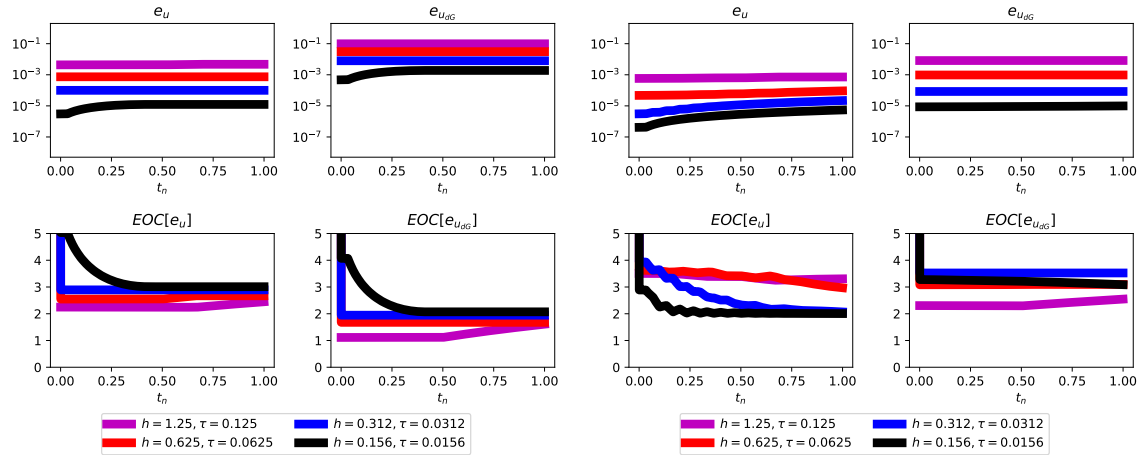
(b) Momentum conserving scheme with spatial degree $q = 3$.



(c) Energy conserving scheme with spatial degree $q = 2$.

(d) Energy conserving scheme with spatial degree $q = 3$.

Figure 5.4: The errors, as described in (5.17) and (5.18), and experimental order of convergence of the single soliton solution (5.19) for the momentum conserving scheme (5.14) or the energy conserving scheme (5.15) with polynomial degree $q = 3$. Here we vary τ_n and h_m such that $\tau_n = Ch_m$.



(a) Momentum conserving scheme with spatial degree $q = 3$. Here the temporal error dominates.

(b) Energy conserving scheme with spatial degree $q = 3$. Here the temporal error dominates.

5.4.2 Two soliton simulation

Consider the two soliton solution over the *real line* $\mathbb{R}$ of

$$u(t,x) = \frac{F}{G}, \tag{5.20}$$

where

$$F = 2(c_1 - c_2)\left(c_1 \cosh\left(\frac{\sqrt{c_2}}{2}(x - p_2 - t)\right)^2 + c_2 \sinh\left(\frac{\sqrt{c_1}}{2}(x - p_1 - t)\right)^2\right)$$

and

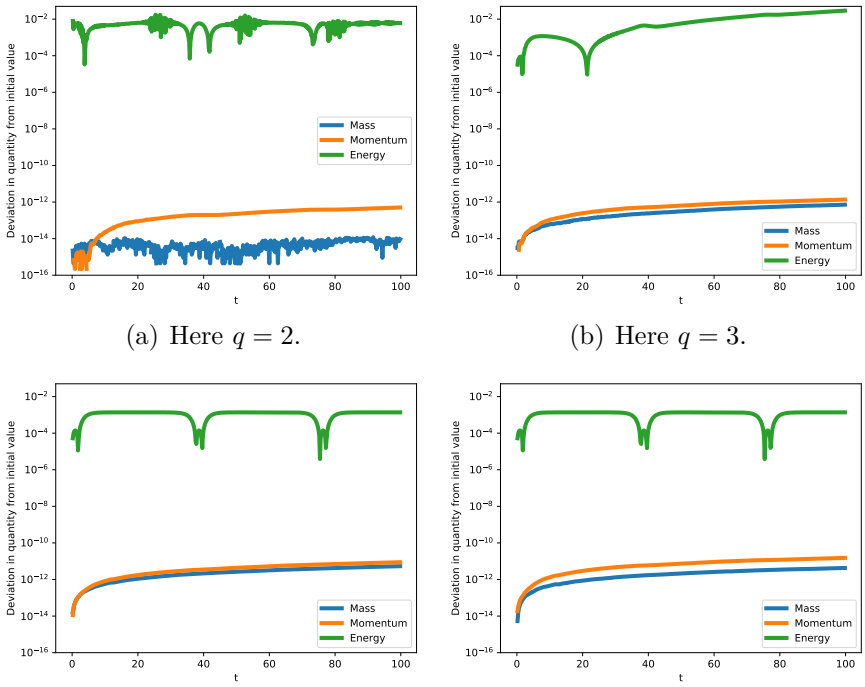
$$\begin{aligned} G = (\sqrt{c_1} - \sqrt{c_2})\cosh\left(\frac{\sqrt{c_1}}{2}(x - p_1 - t) + \frac{\sqrt{c_2}}{2}(x - p_2 - t)\right) \\ + (\sqrt{c_1} + \sqrt{c_2})\cosh\left(\frac{\sqrt{c_1}}{2}(x - p_1 - t) - \frac{\sqrt{c_2}}{2}(x - p_2 - t)\right). \end{aligned}$$

In the sequel we specify $c_1 = 1.5, c_2 = 0.6, p_1 = 20, p_2 = 21$. Note that numerically we are implementing this solution over a periodic domain, as such once our numerical solution wraps around the periodic domain the exact solution is no longer valid.

We compute the deviation in mass, momentum and energy for the two soliton solution (5.20) of both of our numerical schemes in Figure 5.5 and Figure 5.6 where our nonlinear solver tolerance is specified as 10^{-12}. We notice a similar behaviour to the 1 soliton case. Both mass and momentum are conserved over each time step for the momentum conserving scheme and mass and energy are conserved over each time step for the energy conserving scheme. In addition for $q = 2, 3$ the respective conserved quantities are conserved globally, as the deviation remains below solver tolerance. We can again observe that for the momentum conserving scheme the deviation in energy decreases as polynomial degree q increases until it reaches 10^{-3} globally, whereas the deviation in momentum for the energy conserving scheme is 10^{-3} for all polynomial degrees we consider. Note in both figures the deviation of the quantity which is *not* conserved decreases around $t \approx 40, 80$, this decrease coincides temporally with soliton interactions.

We benchmark our schemes against the exact solution (5.20) in Figure 5.7 with spatial degrees $q = 2, 3$. Similarly to the 1 soliton case we notice that the spatial L_2 error (5.17) for the momentum conserving scheme appears to be $\mathcal{O}\left(h_m^{q+1}\right)$ for q even and $\mathcal{O}\left(h_m^{q}\right)$ for q odd. Our experiments suggest that the error for the energy conserving scheme behaves

Figure 5.5: The deviation in mass, momentum and energy with $T = 100$ and initial condition (5.20) for the momentum conserving scheme (5.14). Further we choose $\tau_n = 0.2$, $h_m = 0.4$ and vary the polynomial degree q.



(a) Here $q = 2$.

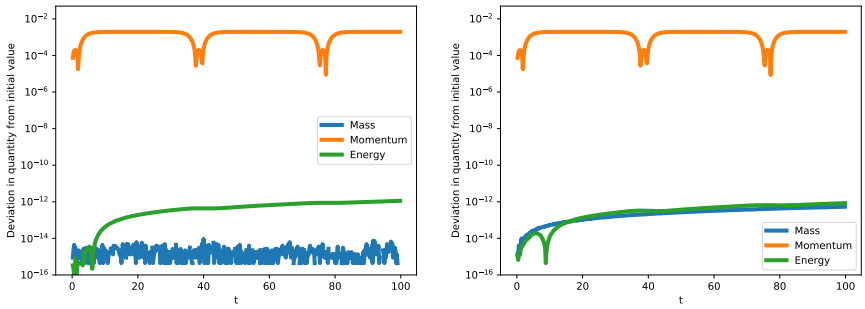
(b) Here $q = 3$.

(c) Here $q = 4$.

(d) Here $q = 5$.

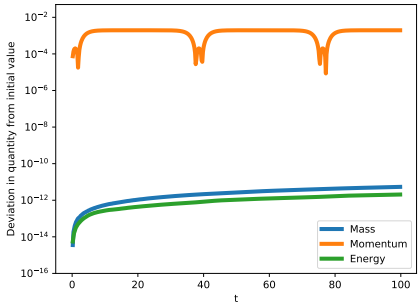
like $\mathcal{O}\left(h_m^{q+1}\right)$ for all spatial polynomial degrees. We also note that from Figure 5.8 we observe that the temporal error of two soliton solution the converges at least $\mathcal{O}\left(\tau_n^2\right)$.

Figure 5.6: The deviation in mass, momentum and energy with $T = 100$ and initial condition (5.20) for the energy conserving scheme (5.15). Further we choose $\tau_n = 0.2$, $h_m = 0.4$ and vary the polynomial degree q.

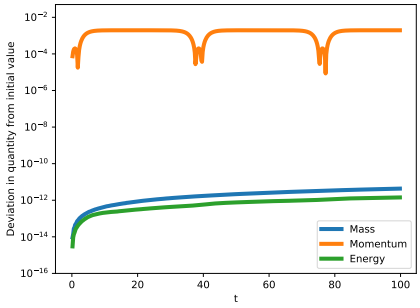


(a) Here $q = 2$.

(b) Here $q = 3$.

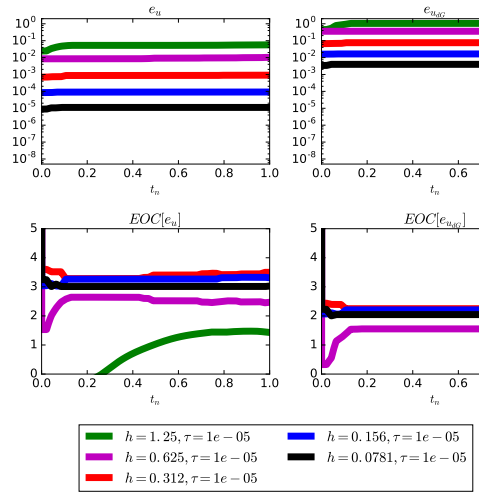


(c) Here $q = 4$.

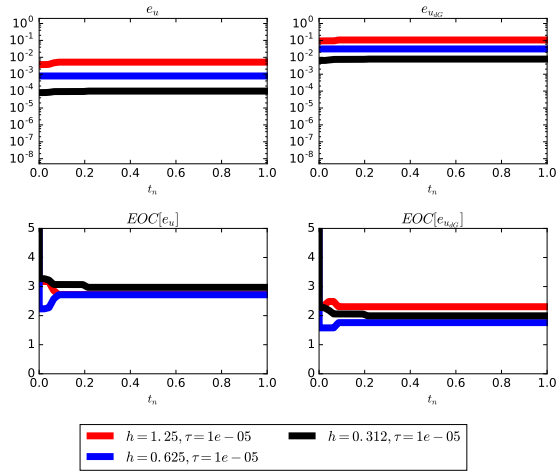


(d) Here $q = 5$.

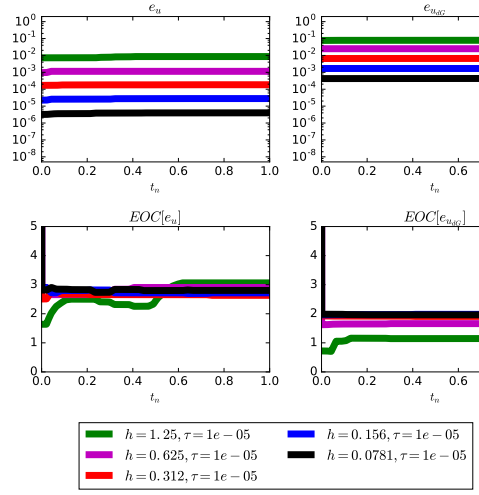
Figure 5.7: The errors, as described in (5.17) and (5.18), and experimental order of convergence of the two soliton solution (5.20) for the momentum conserving scheme (5.14) and energy conserving scheme (5.15) with polynomial degrees $q = 2, 3$. Here we fix $\tau_n = 0.00001$ and vary h_m.



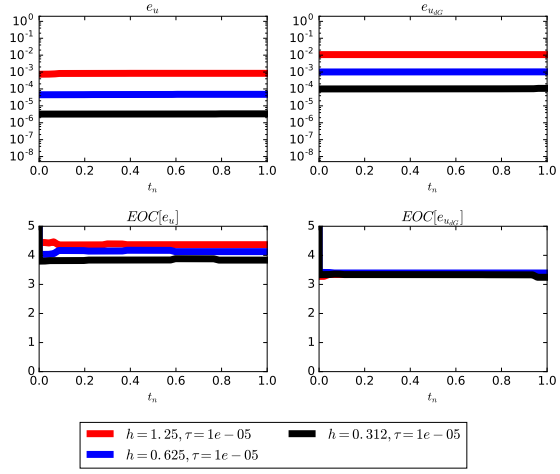
(a) Momentum conserving scheme with spatial degree $q = 2$.



(b) Momentum conserving scheme with spatial degree $q = 3$.

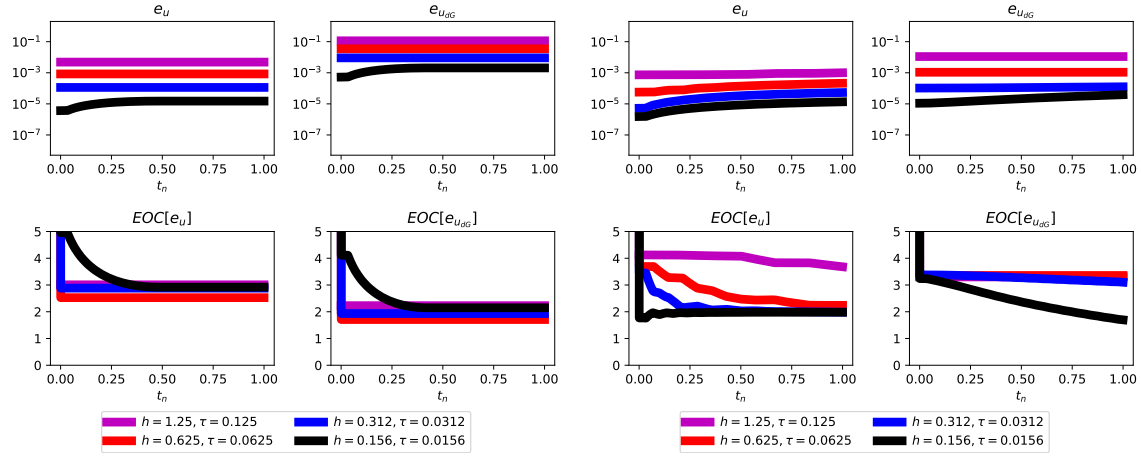


(c) Energy conserving scheme with spatial degree $q = 2$.



(d) Energy conserving scheme with spatial degree $q = 3$.

Figure 5.8: The errors, as described in (5.17) and (5.18), and experimental order of convergence of the two soliton solution (5.20) for the momentum conserving scheme (5.14) or the energy conserving scheme (5.15) with polynomial degree $q = 3$. Here we vary $\tau_n = 0.00001$ and h_m such that $\tau_n = Ch_m$.



(a) Momentum conserving scheme with spatial degree $q = 3$. Here the temporal error dominates.

(b) Energy conserving scheme with spatial degree $q = 3$. Here the temporal error dominates.

5.5 Conclusion

We introduced a momentum conserving scheme and an energy conserving scheme for the KdV equation and presented a detailed numerical comparison of both schemes. While the dynamical behaviour of both schemes is similar, the energy conserving typically outperformed the momentum conserving scheme with respect to phase error and shape error. With respect to the error in amplitude the momentum conserving scheme slightly outperformed the energy conserving scheme. We observed that the error for the energy conserving scheme is smaller, and has a faster experimental convergence rate for odd polynomial degree. While for the momentum conserving scheme a detailed error analysis exists in [31, 114], for the energy conserving scheme we could not show error bounds with existing techniques, as the energy of KdV does not induce a norm.

Chapter 6

An introduction to conservative finite element schemes as adaptive algorithms

As we found in Chapter 5, the conservation of invariants does *not* immediately extend to adaptive algorithms. Here we shall focus on a h-adaptive implementation of our scheme for linearised KdV given by Definition 4.2.23, which we discussed at length for non-adaptive problems in §4.2. For clarity of exposition, we refer the reader to our notation for spatially adaptive finite element meshes and spaces introduced in §5.3.

Before introducing our h-adaptive implementation, we first recall the notion of a *mesh change operator*, which maps a finite element function of $\mathbb{V}_q^n$ to $\mathbb{V}_q^{n+1}$. We write the mesh change function as $\mathcal{P}^{n+1}$. We shall not explicitly choose this function forthwith, but it shall be either an interpolation or projection operator. We modify our scheme for linearised KdV to allow for adaptivity as follows.

Definition 6.0.1 (An adaptive scheme for linearised KdV)**.** *Let $U^j \in \mathbb{V}_q^j$ be given for $j = 0, .., n$. Then we seek $U^{n+1}, V^{n+1}, W^{n+1} \in \mathbb{V}_q^{n+1}$ such that*

$$\begin{aligned} \left\langle \frac{U^{n+1} - \mathcal{P}^{n+1}U^n}{\tau_n} + \mathcal{G}\left(V^{n+1}\right), \phi \right\rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q^{n+1} \\ \left\langle V^{n+1} - U^{n+\frac{1}{2}} - \mathcal{G}\left(W^{n+\frac{1}{2}}\right), \psi \right\rangle &= 0 \qquad \forall \psi \in \mathbb{V}_q^{n+1} \\ \left\langle W^{n+1} - \mathcal{G}\left(U^{n+1}\right), \chi \right\rangle &= 0 \qquad \forall \chi \in \mathbb{V}_q^{n+1}, \end{aligned} \tag{6.1}$$

with $W^n = \mathcal{G}\left(\mathcal{P}^{n+1}U^n\right)$ for $\mathcal{G}$ given by Definition 4.2.3, and where

$U^{n+\frac{1}{2}} := \frac{1}{2}\left(U^{n+1} + \mathcal{P}^{n+1}U^n\right)$ *for* $n = 0, .., N-1$. *Further we define the initial data* $U^0 = \Pi^0 u_0(x)$ *where* Π^0 *denotes the* L_2 *projection into the initial finite element space.*

We shall describe how we adapt the aforementioned scheme in §6.1.

Remark 6.0.2 (The temporal discretisation of auxiliary variables in the adaptive scheme)**.** *Notice that our handling of the auxiliary variable in the adaptive scheme differs from the standard case seen in Definition 4.2.23. In our adaptive algorithm we no longer update* W^n *in a time stepping fashion. Due to the auxiliary nature of* W^n *applying a mesh change operator causes inconsistencies in the resultant numerical scheme. We instead redefine* W^n *for every time step such that it is compatible with* $\mathcal{P}^{n+1}U^n$.
When our spatial mesh is not adaptive the adaptive implementation (6.1) *is equivalent to the standard implementation* (4.40)*. Computationally however, there is additional overhead in the form of recomputing* W^n *on every time step.*

Proposition 6.0.3 (Deviation of invariants in the adaptive scheme for linearised KdV)**.** *Let* U^n *be the solution of the adaptive scheme for linearised KdV described in Definition 6.0.1 subject to some mesh change operator* $\mathcal{P}^{n+1}$*. Recall from Theorem 4.2.7 that the mass and momentum are described by*

$$\mathcal{F}_1(U^n) = \langle U^n, 1\rangle$$

and

$$\mathcal{F}_2(U^n) = \frac{1}{2}\left\langle U^n, U^n\right\rangle,$$

respectively. Additionally, we rewrite the energy in the form

$$\mathcal{F}_3(U^n) = \frac{1}{2}\left\langle \mathcal{G}\left(U^n\right), \mathcal{G}\left(U^n\right)\right\rangle - \frac{1}{2}\left\langle U^n, U^n\right\rangle$$

which is equivalent, but more illuminating in the adaptive case.
The deviation in mass, momentum and energy can then be quantified as

$$\mathcal{F}_1\Big(U^{n+1}\Big) = \mathcal{F}_1\Big(\mathcal{P}^{n+1}U^n\Big),$$

$$\mathcal{F}_2\Big(U^{n+1}\Big) = \mathcal{F}_2\Big(\mathcal{P}^{n+1}U^n\Big),$$

and

$$\mathcal{F}_3\Big(U^{n+1}\Big) = \mathcal{F}_3\Big(\mathcal{P}^{n+1}U^n\Big)$$

respectively.

Remark 6.0.4 (Conservation of invariants in the adaptive scheme for linearised KdV)**.** *If, for a given invariant $\mathcal{F}_i$, we choose a mesh change operator such that*

$$\mathcal{F}_i\Big(\mathcal{P}^{n+1}U^n\Big) = \mathcal{F}_i(U^n),$$

then the invariant is conserved but this is not guaranteed by standard interpolators*. In the sequel we shall investigate the design of mesh change operators which preserve invariants under adaptivity. For any consistent mesh change operator, we find that all invariants are preserved if our adaptive routine only allows for refinement, i.e., if $\mathbb{V}_q^n \subseteq \mathbb{V}_q^{n+1}$.*

Proof of Proposition 6.0.3. Proposition 6.0.3 follows the argument outlined in the proof of Proposition 4.2.25 with the caveat that we must write

$$\mathcal{F}_i(U^n) = \mathcal{F}_i\Big(\mathcal{P}^{n+1}U^n\Big) + \mathcal{F}_i(U^n) - \mathcal{F}_i\Big(\mathcal{P}^{n+1}U^n\Big),$$

for $i = 1, 2, 3$, similarly to in the proofs of Proposition 5.2.3 and Proposition 5.2.6.

□

Remark 6.0.5 (Lagrange multipliers for conservative adaptivity)**.** *One methodology for the design of a conservative mesh change operator is by enforcing conservation of an invariant through Lagrange multipliers, see [16]. This has proven very successful in the literature, see [63, 143], however we cannot expect mass to be conserved when constructing a Lagrange multiplier which conserves a nonlinear invariant.*

6.1 Adaptive algorithm

Before we discuss the conservative properties of the adaptive linearised KdV scheme under different mesh change operators we must first discuss how we adapt the spatial mesh over time.

While adapting the spatial mesh subject to predetermined hierarchical structure is often considered, see [165, 186, 56, c.f.], here we shall not explicitly take this approach. While a hierarchical mesh has major practical benefits for adaptivity in two dimensions and higher, for one dimensional simulations it is not paramount as a hierarchical mesh structure always exists even if we do not explicitly structure our algorithm this way.

To implement our adaptive algorithm there are multiple parameters we must specify. First we define `coarsen` and `refine` to be the percentage of elements we attempt to coarsen and refine respectively. Note that for the adaptive routine to make sense we require that $0 \leq \texttt{coarsen}$ and $0 \leq \texttt{refine}$, and $\texttt{coarsen} + \texttt{refine} \leq 100$ as we cannot simultaneously refine and coarsen an element. Additionally we define $h_{\texttt{min}}$ and $h_{\texttt{max}}$ to be the minimal and maximal sizes of any given element $\mathcal{J}_m^n$. Of course, if we attempt to coarsen or refine an element beyond these tolerances the elements shall not change, and no additional mesh change shall be implemented to compensate for this.
In addition to these parameters our adaptive algorithm is also subject to a mesh change indicator function, $\texttt{indicator}\left(U^n\right)$, in this chapter we heuristically choose $\texttt{indicator}\left(U^n\right) \in \mathbb{V}_0^n$ such that

$$\left\langle \texttt{indicator}\left(U^n\right), \phi \right\rangle = \sum_{m=0}^{M-1} \left\langle \left\{h_m\right\}^{-1} \left[\!\left[U_m^n\right]\!\right], \phi|_{\mathcal{J}_m^n} \right\rangle \qquad \forall \phi \in \mathbb{V}_0^n.$$

In the adaptive algorithm we also formally define `mesh` to describe the collection of nodal values of a given mesh.
Note that in practice, we employ the adaptivity on the initial condition to more accurately initialise the simulation.

6.2 Mesh change operators

As we found in Proposition 6.0.3, the choice of mesh change operator is paramount to the numerical conservation of invariants. Throughout this section we discuss different potential mesh change operators and their respective properties, in particular we focus on the *Lagrange interpolation* of the solution from an old mesh to a new mesh, and the *L_2 projection*. All results presented here depend on Proposition 6.0.3. First we consider Lagrange interpolation as our mesh change operator.

Definition 6.2.1 (The Lagrange interpolation operator)**.** *For a function $U^n \in \mathbb{V}_q^n$ its interpolant onto the new mesh $\mathcal{I}^{n+1}U^n \in \mathbb{V}_q^{n+1}$ is described uniquely by its values at the degrees of freedom $\boldsymbol{X}$ as*

$$\mathcal{I}^{n+1}U^n\left(\xi\right) = U^n\left(\xi\right) \qquad \forall \xi \in \boldsymbol{X}. \tag{6.2}$$

Note that our degrees of freedom are located at the zeroes of the Lagrange basis functions.

Algorithm 6.1 Adaptive algorithm

Require:

The number of time steps N
An initial mesh `mesh`
An initial finite element solution U^0
A coarsening percentage `coarsen`
A refining percentage `refine`
A maximum possible element size $h_{\texttt{max}}$
A minimum possible element size $h_{\texttt{min}}$
An indicator function `indicator`

Ensure:

An adaptive finite element solution U^n for $n = 0, ..., N$

1: **for** $n = 0 : N - 1$ **do** ▷ Loop over time
2: $\texttt{ind} = \texttt{indicator}\,(U^n)$ ▷ Compute the indicator function on the current time step
3: $\texttt{marker} = \texttt{zeroes}(\texttt{length}(\texttt{mesh}) - 1)$ ▷ Initialise a vector of zeroes the same length as the number of elements on the mesh
4: **for** $m = 0 : \texttt{length}(\texttt{marker})$ **do** ▷ Loop over elements of the mesh
5: **if** $\|\texttt{ind}\|_{L_2(\mathcal{J}_m^n)}$ is within `coarsen` percent of the smallest value and the resulting element size is small than $h_{\texttt{max}}$ **then**
6: $\texttt{marker}[m] = -1$ ▷ Mark element for coarsening
7: **else if** $\|\texttt{ind}\|_{L_2(\mathcal{J}_m^n)}$ is within `refine` percent of the largest value and the resulting element sizes are both larger than $h_{\texttt{min}}$ **then**
8: $\texttt{marker}[m] = 1$ ▷ Mark element for refinement
9: $\texttt{mesh} = \texttt{mesh\_change}(\texttt{mesh}, \texttt{marker})$ ▷ Change the mesh in accordance with `marker`
10: $U^{n+1} = \texttt{solve}(\mathcal{P}^{n+1}U^n, \texttt{mesh})$ ▷ Solve the finite element method over one time step

Proposition 6.2.2 (Conservative properties of the interpolant)**.** *Let U^n be the solution of the adaptive scheme for linearised KdV given in Definition 6.0.1, and $\mathcal{I}^{n+1}$ be the Lagrange interpolation operator given in Definition 6.2.1. Under refining the mesh all invariants are preserved, i.e., if $\mathbb{V}_q^n \subseteq \mathbb{V}_q^{n+1}$ we have that*

$$\mathcal{F}_i\Big(\mathcal{I}^{n+1}U^n\Big) = \mathcal{F}_i(U^n),$$

for $i = 1, 2, 3$. However, under an adaptive algorithm which includes coarsening we do not preserve any invariants.

Remark 6.2.3 (The Lagrange interpolation operator and adaptivity)**.** *The Lagrange interpolant* (6.2) *does preserve the mass, momentum and energy under refinement, however if the algorithm involves coarsening all of these invariants are lost, therefore it is* not *an ideal mesh change operator for a conservative adaptive algorithm. For an adaptive algorithm to be practically useful the ability to coarsen is essential, otherwise adaptivity increases the computational complexity of the algorithm.*

Proof of Proposition 6.2.2. If $\mathbb{V}_q^n \subseteq \mathbb{V}_q^{n+1}$ then the Lagrange interpolation operator is exact, i.e., $\mathcal{I}^{n+1}U^n = U^n$, and as such all invariants are conserved through Proposition 6.0.3. However, if this is not the case then we may construct counter examples demonstrating that the invariants are not conserved, such as the examples explicitly computed in §6.3. □

We now consider the L_2 projection as our mesh change operator. While we have regularly utilised this projection throughout we write it explicitly here to add clarity to the behaviour of the L_2 projection as a mesh change operator.

Definition 6.2.4 (The L_2 projection)**.** *For a function $U^n \in \mathbb{V}_q^n$ the L_2 projection into $\mathbb{V}_q^{n+1}$ is given by seeking $\Pi^{n+1}U^n \in \mathbb{V}_q^{n+1}$ such that*

$$\Big\langle \Pi^{n+1}U^n, \phi \Big\rangle = \langle U^n, \phi \rangle \qquad \forall \phi \in \mathbb{V}_q^{n+1}. \tag{6.3}$$

Proposition 6.2.5 (Conservative properties of the L_2 projection)**.** *Let U^n be the solution of the adaptive scheme for linearised KdV given in Definition 6.0.1, and Π^{n+1} be the L_2 projection described in Definition 6.2.4. Under mesh refinement all invariants are preserved, i.e., if $\mathbb{V}_q^n \subseteq \mathbb{V}_q^{n+1}$ then*

$$\mathcal{F}_i\Big(\Pi^{n+1}U^n\Big) = \mathcal{F}_i(U^n),$$

for $i = 1, 2, 3$. Further, if we also allow for coarsening, i.e., $\mathbb{V}_q^n \nsubseteq \mathbb{V}_q^{n+1}$, then the L_2 projection conserves mass, and dissipates momentum, i.e.,

$$\mathcal{F}_1\Big(\Pi^{n+1}U^n\Big) = \mathcal{F}_1(U^n)$$

and

$$\mathcal{F}_2\Big(\Pi^{n+1}U^n\Big) \leq \mathcal{F}_2(U^n).$$

Remark 6.2.6 (The L_2 projection as a mesh change operator)**.** *Through the amalgamation of Proposition 6.2.5 with Proposition 6.0.3, we find that our adaptive algorithm* (6.1) *with the L_2 projection as its mesh change operator not only conserves the mass over all time, but also either conserves or dissipates momentum (depending on whether the algorithm coarsens). As momentum can be written in terms of the L_2 norm, we can conclude that the adaptive algorithm is numerically stable with the L_2 projection as a mesh change operator.*

Proof of Proposition 6.2.5. If $\mathbb{V}_q^n \subseteq \mathbb{V}_q^{n+1}$ then the L_2 projection (6.3) is exact, and through Proposition 6.0.3 all three invariants are conserved. Relaxing this assumption to allow for mesh coarsening we observe, through choosing $\phi = 1$ in (6.3), that

$$\mathcal{F}_1\Big(\Pi^{n+1}U^n\Big) = \mathcal{F}_1(U^n),$$

i.e., mass is conserved. Additionally choosing $\phi = \Pi^{n+1}U^n$ we find that

$$\begin{aligned}
\frac{1}{2}\left\|\Pi^{n+1}U^n\right\|_{L_2(S^1)}^2 &= \mathcal{F}_2\Big(\Pi^{n+1}U^n\Big) \\
&= \frac{1}{2}\left\langle \Pi^{n+1}U^n, \Pi^{n+1}U^n \right\rangle \\
&= \frac{1}{2}\left\langle U^n, \Pi^{n+1}U^n \right\rangle \\
&\leq \frac{1}{2}\left\|U^n\right\|_{L_2(S^1)} \left\|\Pi^{n+1}U^n\right\|_{L_2(S^1)},
\end{aligned}$$

through Cauchy's inequality. Dividing both sides by $\|\Pi^{n+1}U^n\|_{L_2(S^1)}$ and squaring the resulting identity we observe that

$$\mathcal{F}_2\Big(\Pi^{n+1}U^n\Big) \leq \mathcal{F}_2(U^n),$$

i.e., the momentum dissipates.

□

Remark 6.2.7 (A momentum *conserving* mesh change operator)**.** *While the L_2 projection dissipates momentum it is possible to construct a mesh change operator which* exactly *preserves momentum. Consider the mesh change operator $\mathcal{P}^{n+1} : \mathbb{V}_q^n \to \mathbb{V}_q^{n+1}$ defined by seeking $\mathcal{P}^{n+1}U^n \in \mathbb{V}_q^{n+1}$ such that*

$$\left\langle \left(\mathcal{P}^{n+1}U^n\right)^2, \phi \right\rangle = \left\langle (U^n)^2, \phi \right\rangle \qquad \forall \phi \in \mathbb{V}_q^{n+1},$$

where $U^n \in \mathbb{V}_q^{n+1}$. Through choosing $\phi = 1$ we observe the momentum is exactly conserved, however this Lagrange interpolation operator is not mass conserving and has a higher computational complexity than the L_2 projection.

Remark 6.2.8 (Adaptive change in energy)**.** *Unfortunately, for the numerical scheme under consideration, the energy does not induce a norm as it is not signed. As such we cannot obtain stability in a norm for our adaptive scheme. However, in §7.4 we investigate a numerical scheme where the energy induces a norm and propose an energy dissipating mesh change operator.*

Remark 6.2.9 (Mesh change operators for *hp*-adaptive algorithms)**.** *Often, in the literature, h-adaptivity is combined with p-adaptivity, see Remark 5.3.2. With this in mind we observe that all results for the L_2 projection hold when our spatial mesh is also adaptive in polynomial degree.*

6.3 Numerical experiments

Here we shall conduct numerical experiments on our adaptive scheme for linearised KdV (6.1). As discussed in §4.2.3, at its core our code utilises Firedrake with a $2q$ order Gauss quadrature. The mesh adaptive algorithm and mesh change operators have been implemented in Python 3 and utilise Numpy linear solvers where appropriate. Additionally we used Matplotlib for visualisation. Similarly to previous chapters, throughout our numerical experiments we stretch our periodic spatial domain from $S^1(0,1)$ to $S^1(0,40)$, primarily for consistency with results in the prequel.
Here we approximate numerically the exact solution of linearised KdV

$$u(t,x) = \sin\left(\alpha\left(x - (1-\alpha^2)t\right)\right), \tag{6.4}$$

where $\alpha = \frac{2\pi}{40}$. Unless stated otherwise, we shall choose the adaptive parameters discussed

in §6.1 as

$$\begin{aligned} \texttt{coarsen} &= 10 & h_{\texttt{max}} &= 1 \\ \texttt{refine} &= 60 & h_{\texttt{min}} &= 0.2. \end{aligned} \tag{6.5}$$

Additionally we choose a fixed time step of $\tau_n = 0.1$, and an initial uniform spatial mesh of $h_m = 0.4$. Notice our spatial mesh is deliberately taken to be very coarse.

First we examine the solution dynamics for different mesh change operators. We choose the Lagrange interpolation operator in Figure 6.1, and the L_2 projection operator in Figure 6.2. We do not observe significant differences in the solution dynamics between mesh change operators, which is likely due to the linear nature of the underlying problem.

Additionally, we examine the values of each invariant over time adapting the mesh subject to the Lagrange interpolation operator in Figure 6.3 and subject to the L_2 projection in Figure 6.4. We observe that the L_2 projection outperforms the Lagrange interpolant as a mesh change operator with respect to the invariants, as expected from our analysis in §6.2.

While we cannot visually differentiate between the accuracy of the solution dynamics for different mesh change operators, we quantify the respective errors in Figure 6.5. We observe that the L_2 projector marginally outperforms the Lagrange interpolation operator. Through increasing the percentage of elements which are coarsened to $\texttt{coarsen} = 30$, we observe that the L_2 projection significantly outperforms the Lagrange interpolation operator in Figure 6.6

Figure 6.1: Here we examine the dynamics of the adaptive algorithm for linearised KdV (6.1) with the *Lagrange interpolant* (6.2) as the mesh change operator. The mesh coordinates are represented by vertical blue lines at the bottom of each solution snapshot. We initialise the simulation with the L_2 projection of (6.4) at $t = 0$ and employ the adaptive parameters (6.5). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We allow for the initial mesh to be adapted.

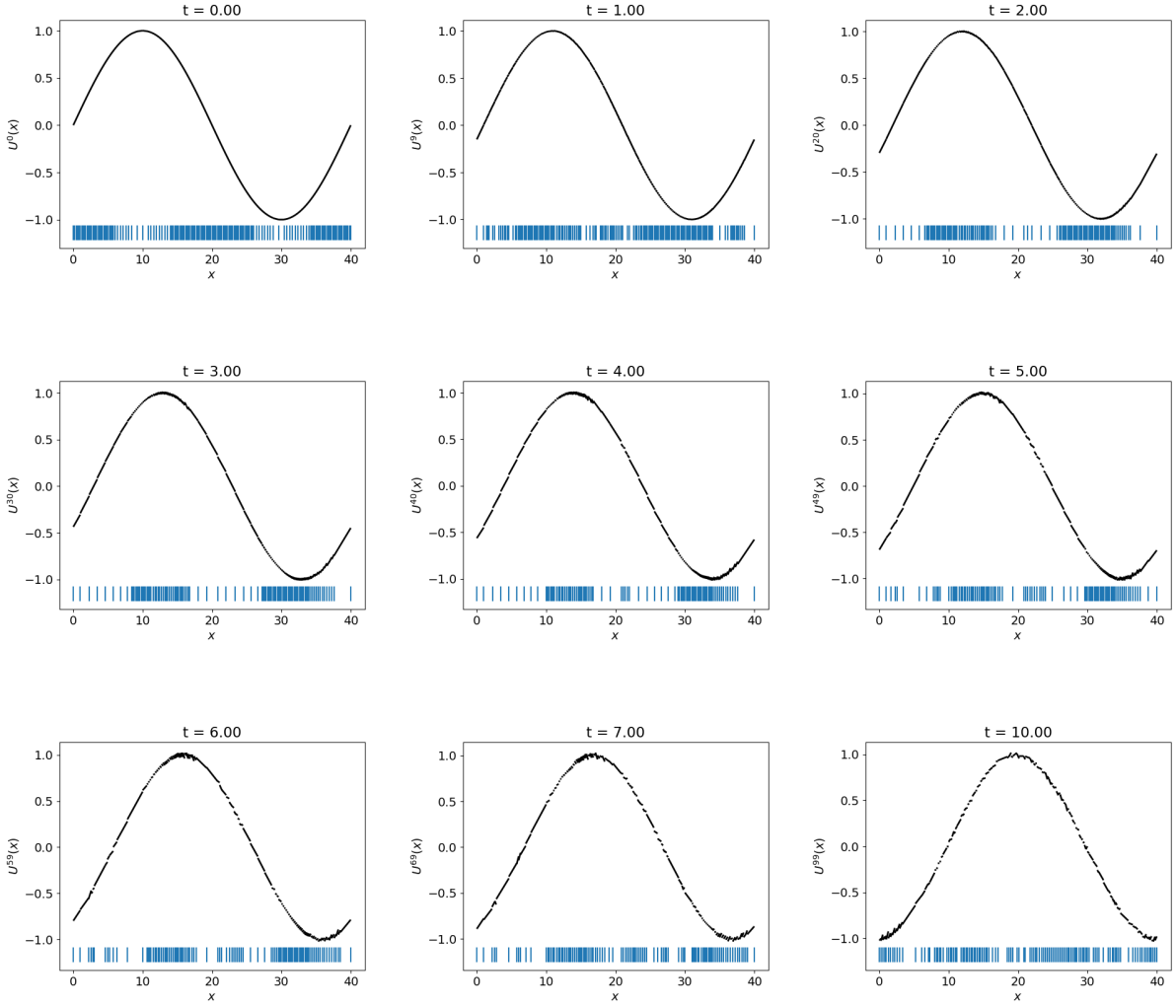


Figure 6.2: Here we examine the dynamics of the adaptive algorithm for linearised KdV (6.1) with the L_2 *projection* (6.3) as the mesh change operator. The mesh coordinates are represented by vertical blue lines at the bottom of each solution snapshot. We initialise the simulation with the L_2 projection of (6.4) at $t = 0$ and employ the adaptive parameters (6.5). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We allow for the initial mesh to be adapted.

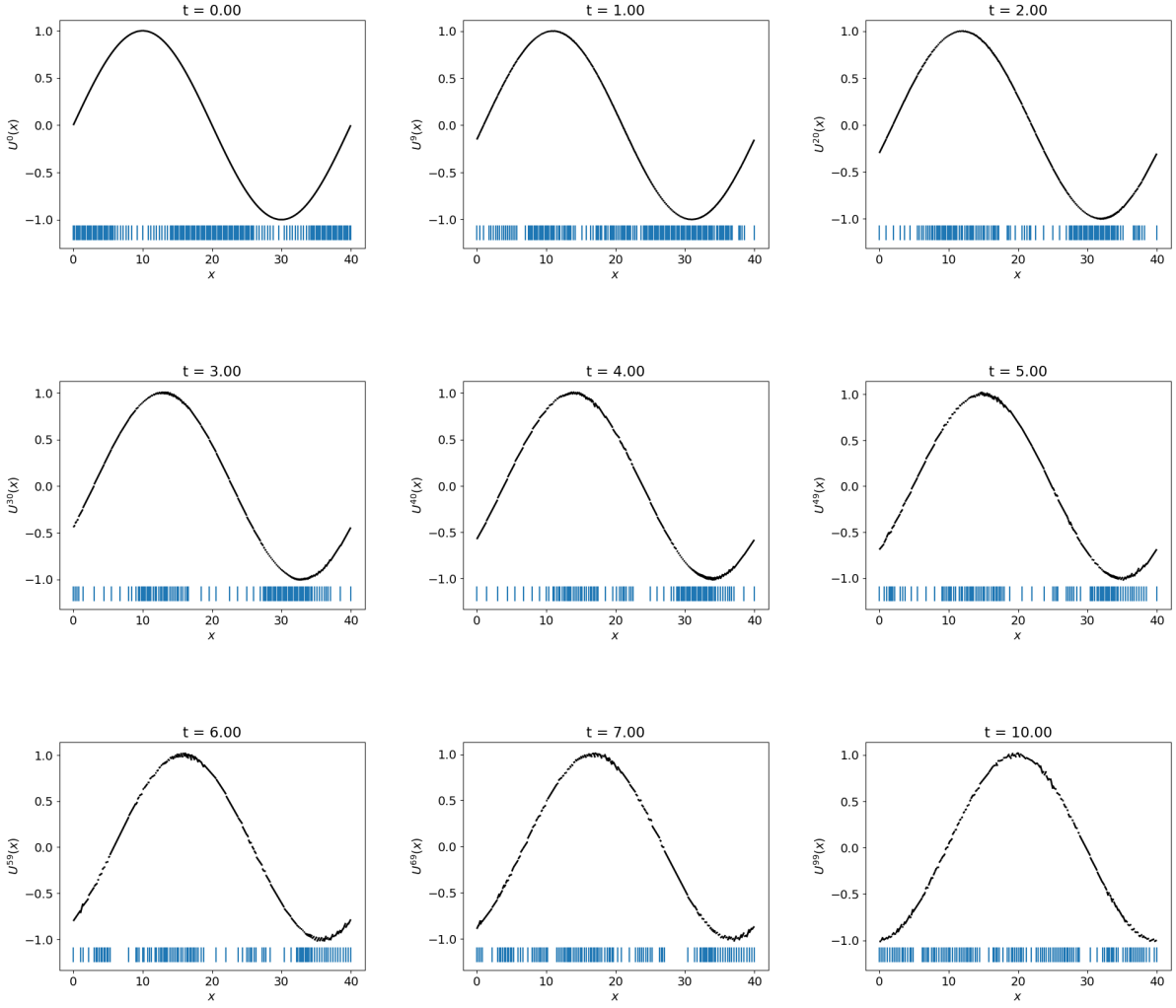
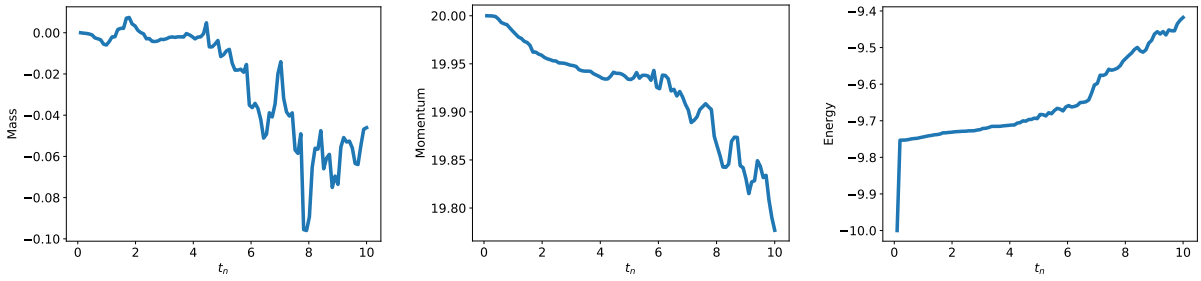
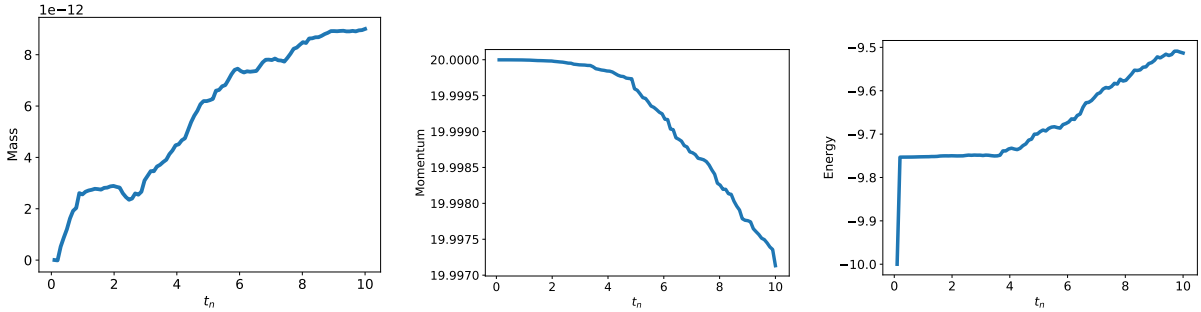


Figure 6.3: Here we examine the values of the invariants (mass, momentum and energy) at the temporal nodes in the adaptive algorithm for linearised KdV (6.1) with the *Lagrange interpolant* (6.2) as the mesh change operator. We initialise the simulation with the L_2 projection of (6.4) into the initial finite element space and employ the adaptive parameters (6.5). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. All invariants deviate non-monotonically.



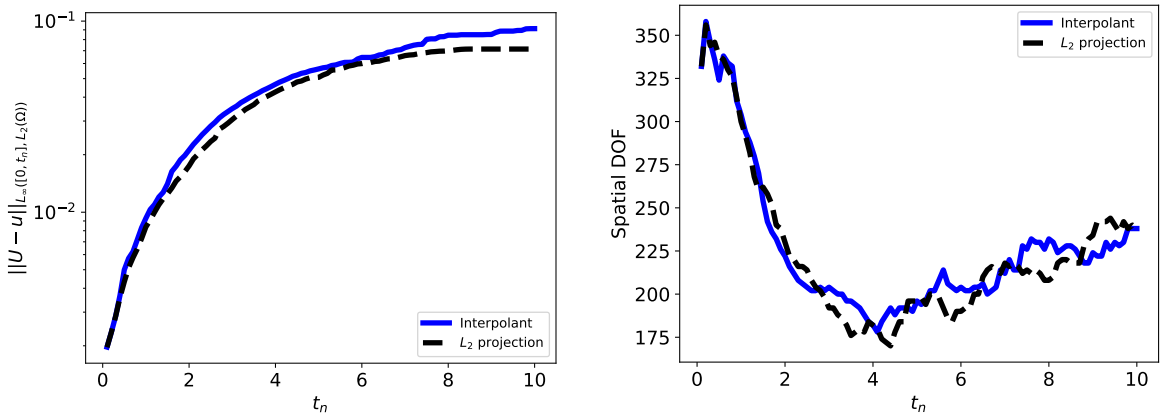
(a) The mass. (b) The momentum. (c) The energy.

Figure 6.4: Here we examine the values of the invariants (mass, momentum and energy) at the temporal nodes in the adaptive algorithm for linearised KdV (6.1) with the L_2 *projection* (6.3) as the mesh change operator. We initialise the simulation with the L_2 projection of (6.4) into the initial finite element space and employ the adaptive parameters (6.5). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We notice that mass is conserved and momentum decreases monotonically. The deviation in momentum is also significantly smaller than in the case where the mesh change operator is chosen to be the Lagrange interpolant.



(a) The mass. (b) The momentum. (c) The energy.

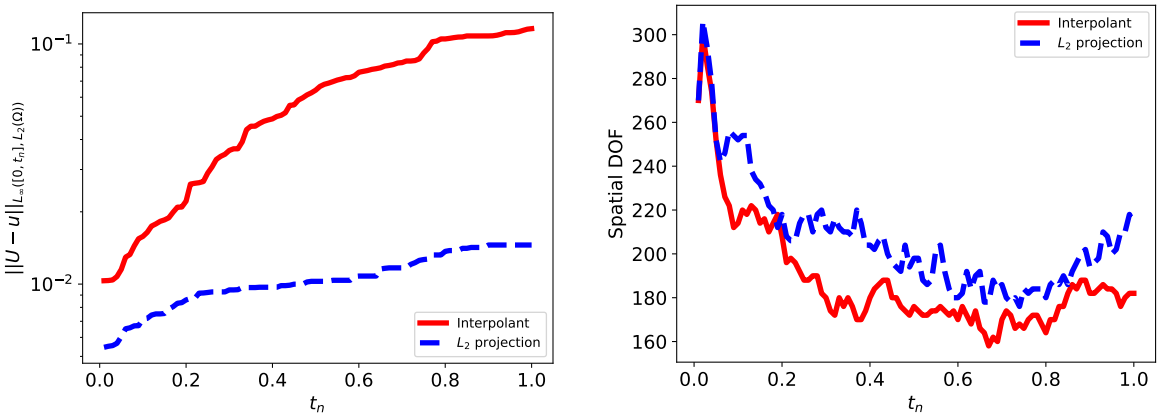
Figure 6.5: Here we examine the error measured in the Bochner norm $L_\infty([0,T], L_2(S^1))$ for linearised KdV (6.1) with the Lagrange interpolant (6.2) or the L_2 projection (6.3). We additionally plot the error the number of degrees of freedom used in each simulation on each time step. The simulations are initialised by L_2 projection of (6.4) into the initial finite element space and employ the adaptive parameters (6.5). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We observe that both simulations use a comparable number of degrees of freedom, with the L_2 projection marginally outperforming the Lagrange interpolation operator.



(a) The error.

(b) The number of degrees of freedom.

Figure 6.6: Here we examine the error measured in the Bochner norm $L_\infty([0,T], L_2(S^1))$ for linearised KdV (6.1) with the Lagrange interpolant (6.2) or the L_2 projection (6.3). We additionally plot the error the number of degrees of freedom used in each simulation on each time step. The simulations are initialised by L_2 projection of (6.4) into the initial finite element space and employ the adaptive parameters (6.5) with `coarsen` $= 30$. Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We observe that the L_2 projection significantly outperforms the Lagrange interpolation operator. Note that the number of degrees of freedom the algorithm uses with the L_2 projection as a mesh change operator is slightly higher, this is likely due to the spatial location of the leading order error changing and the algorithm attempting to refine nodes which are not already at the maximal refinement level. The increase in the number of degrees of freedom is *not* the reason that the L_2 projection outperforms the Lagrange interpolation operator.



(a) The error.

(b) The number of degrees of freedom.

6.4 Conclusion

We presented an introduction to an adaptive routine for conservative problems through the conservative numerical scheme for linearised KdV introduced in §4.2. We found that for conservativity, and therefore numerical stability, the mesh change operator of an adaptive scheme needs to be carefully chosen such that the invariants are bounded. It is likely that we did not encounter significant numerical difficulties with the mesh change operators which did not fall within this framework due to the linearity of the problem considered.

Chapter 7

Conservative Galerkin methods for dispersive Hamiltonian PDEs

In this chapter we focus on the design of conservative discontinuous Galerkin schemes for generalised third order KDV type equations. The techniques we employ allow for the derivation of optimal a priori and a posteriori bounds, the latter of which is crucial for the development of adaptive algorithms. We specify our family of problems such that *their energy defines a norm*. It is through conserving a discrete version of this energy that optimal a priori and a posteriori bounds can be constructed. Further, we use our a posteriori bounds to drive an adaptive algorithm, similar to the adaptive algorithm developed in Chapter 6. Recall that interpolation based adaptive methods do not preserve any of the underlying structure of the problem, and are not necessarily stable. We develop a new mesh change projection which is stable in an appropriate energy norm of the problem. While we set up the framework to prove nonlinear error bounds here we only explicitly show bounds for the linear case.

7.1 Necessary definitions and the continuous problem

Here we formulate the model problem, fix notation and give some basic assumptions. We describe some known results and history of the defocusing generalised Korteweg-de Vries equation, highlighting the Hamiltonian structure of the equation. We show that the underlying Hamiltonian structure naturally yields an induced stability of the solutions to the PDE system and give a summary of some exact solutions for specific nonlinearities.

Throughout this work we consider the dispersive KdV type problem

$$u_t - f'(u)_x + u_{xxx} = 0, \tag{7.1}$$

where

$$f(u) = \frac{\alpha}{\beta} u^\beta \text{ for } \beta \in 2\mathbb{N}, \alpha > 0,$$

and $f'(u)$ represents the first Frechet derivative. An example of this is given through choosing $\beta = 4$ and $\alpha = 1$, i.e.,

$$u_t - u^2 u_x + u_{xxx} = 0.$$

Notice the sign in front of the nonlinearity. This problem is sometimes referred to as the defocusing mKdV equation, with the focusing mKdV equation, which we briefly discussed in Remark 5.3.11, having the opposing sign on the nonlinearity.

Proposition 7.1.1 (Invariants of the continuous problem)**.** *The dispersive problem* (7.1) *has the invariants*

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle u, 1 \rangle = \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{2} \langle u, u \rangle \right) = \frac{\mathrm{d}}{\mathrm{d}t} \left(\left\langle \frac{1}{2} u_x^2 + f(u), 1 \right\rangle \right) = 0, \tag{7.2}$$

as can be seen through a similar argument to that conducted in §5.1.1. Physically the invariants presented in (7.2) *represent the mass, momentum and energy of the PDE* (7.1) *respectively.*

Before continuing with our investigation of defocusing KdV type equations we first introduce standard Sobolev space and norm notation, as discussed in [69]. Recall that we denote the standard Lebesgue spaces by $L_p(S^1)$, $1 \leq p \leq \infty$, with corresponding norms $\|\cdot\|_{L_p(S^1)}$. Let also $H^s(S^1)$, be the Hilbertian Sobolev space of index $s \in \mathbb{R}$ of real-valued functions defined on S^1, constructed via standard interpolation and/or duality procedures, along with the corresponding norm and seminorm

$$\|u\|_{W^{k,p}(S^1)} := \begin{cases} \left(\sum_{|\alpha| \leq k} \|\mathrm{D}^\alpha u\|_{L_p(S^1)}^p \right)^{1/p} & \text{if } p \in [1, \infty) \\ \sum_{|\alpha| \leq k} \|\mathrm{D}^\alpha u\|_{L_\infty(S^1)} & \text{if } p = \infty \end{cases}$$

$$|u|_{W^{k,p}(S^1)} := \left\| \mathrm{D}^k u \right\|_{L_p(S^1)}$$

respectively. We also make use of the following notation for time dependent Sobolev

(Bochner) spaces:

$$\mathcal{C}^i(0,T;H^k(S^1)) := \left\{ u : [0,T] \to H^k(S^1) : \ u \text{ and } \frac{\mathrm{d}^j}{\mathrm{d}^j t} u \text{ for } 0 < j \leq i \text{ is continuous} \right\},$$
$$L_\infty(0,T;H^k(S^1)) := \left\{ u : [0,T] \to H^k(S^1) : \ \operatorname{ess\,sup}_{t\in[0,T]} \|u(t)\|_{H^k(S^1)} < \infty \right\}.$$

Remark 7.1.2 (Representation of the energy of (7.1))**.** *The energy of* (7.1)*,*

$$\left\langle \frac{1}{2} u_x^2, + f(u), 1 \right\rangle,$$

defines a norm, which can be seen more clearly when $f(u)$ is written in terms of a norm, i.e.,

$$f(u) = \frac{\alpha}{\beta} \|u\|^\beta_{L_\beta(S^1)}. \tag{7.3}$$

Throughout the sequel we shall freely apply (7.3)*.*

Remark 7.1.3 (Pointwise solution control)**.** *As the energy of the problem induces a norm, we obtain stability of the continuous problem. More concisely, energy conservation*

$$\frac{\mathrm{d}}{\mathrm{d}t} \left\langle \frac{1}{2} u_x^2 + f(u), 1 \right\rangle = 0$$

immediately shows that,

$$\|u\|_{\mathcal{C}^i(0,T;H^1(S^1))} \leq C \|u(0)\|_{H^1(S^1)},$$

as $f(u) \geq 0$. Since $H^1(S^1) \subset L_\infty(S^1)$ we see

$$\sup_{t\in[0,\infty]} \|u(t)\|_{L_\infty(S^1)} \leq C \|u(0)\|_{H^1(S^1)},$$

i.e., the solution to the continuous problem remains in a bounded set. It is this argument we shall later mimic on the discrete level to demonstrate numerical stability.

We shall now briefly discuss the exact solution of (7.1) in both a linear and nonlinear case.

Example 7.1.4 (Exact solution to the linear problem)**.** *Let $f(u) = \frac{1}{2}u^2$, then under the ansatz that $u(t,x) = u(\xi)$, with $\xi = c\,(x + (1+c^2)t)$ we find that*

$$u(t,x) = C_1 \sin(\xi) + C_2 \cos(\xi), \tag{7.4}$$

solves (7.1) *where* $c = 2l\pi$ *for* $l \in \mathbb{Z}$ *and* C_1, C_2 *denote real constants. Due to the linear nature of the problem any possible linear combination of* (7.4) *for various attainable parameter values is also a solution.*

Example 7.1.5 (Exact solution to the nonlinear problem)**.** *With* $f(u) = \frac{1}{2}u^4$*, then under the ansatz that* $u(t,x) = u(x+ct)$ *it can be shown that the positon solution*

$$u(x,t) = \frac{1}{2}\operatorname{csch}\left(c^{\frac{1}{2}}\frac{(x-ct)}{2}\right)^2$$

formally solves (7.1)*. It is well-known that one can bijectively map solutions from the defocusing mKdV equation to solutions to the KdV equation employing the Miura transform, see [142, 3]. Although, it is worth noting that it is not possible to get smooth, nonsingular position solutions of the defocusing mKdV through inverse scattering techniques because of the singularity that is inherent in its Darboux transformation. For* $f(u) = \frac{1}{4}u^4$ *we can, however find kink*

$$u(x,t) = (3c)^{1/2}\tanh\left(\frac{(2c)^{\frac{1}{2}}}{2}(x+ct)\right)$$

and anti-kink solutions

$$u(x,t) = -(3c)^{1/2}\tanh\left(\frac{(2c)^{\frac{1}{2}}}{2}(x+ct)\right),$$

that are smooth, but are not periodic. To establish periodic, smooth exact solutions, one must examine Jacobi elliptic functions [149]. Let $sn(x,k)$ *denote that Jacobi elliptic function with modulus* $k \in [0,1)$*, then a solution is given by [51]*

$$u(x,t) = k\,\mathrm{sn}(x + (k^2+1)t, k), \tag{7.5}$$

after employing a spatial rescaling to satisfy the periodic boundary conditions.

7.2 Discretisation and a priori analysis

Here we approximate (7.1) by a semi-discrete discontinuous Galerkin method. We suggest the reader recalls the finite element notion and discretisation parameters introduced in §4.2.1, and in particular the definition of the spatial finite element space given by Definition 4.2.1.

Following the methodology for the design of conservative Galerkin schemes outlined in §4.1, which we additionally implemented in Chapter 5, we introduce the *first variation* of the energy as an auxiliary variable. This allows us to rewrite (7.1) as

$$\begin{aligned} u_t + v_x &= 0 \\ v + f'(u) - u_{xx} &= 0. \end{aligned}$$

Definition 7.2.1 (Spatially discrete scheme for defocusing KdV type equations)**.** *Let $\mathcal{G} : \mathbb{V}_q \to \mathbb{V}_q$ be the spatial first derivative operator given in Definition 4.2.3. Additionally let $\mathcal{A}_h : \mathbb{V}_q \times \mathbb{V}_q \to \mathbb{V}_q$ be a* symmetric *bilinear form representing the weak formulation of a second spatial derivative. Our spatially discrete scheme for defocusing KdV type equations* (7.1) *is given by seeking $U, V \in \mathbb{V}_q$ such that*

$$\begin{aligned} \langle U_t + \mathcal{G}(V), \phi \rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q \\ \langle V + f'(U), \psi \rangle + \mathcal{A}_h(U, \psi) &= 0 \qquad \forall \psi \in \mathbb{V}_q, \end{aligned} \tag{7.6}$$

with initial condition $U(0, x) = \Pi u_0$.

While we formally defined the dG type norm in (5.18) as the spatial component of a Bochner norm, we shall concisely define it here as

$$\|W\|_{dG}^2 := \sum_{m=0}^{M-1} \left(\|W_x\|_{L_2(\mathcal{J}_m)}^2 + \{h_m\}^{-1} [\![W_m]\!]^2 \right).$$

We impose that the bilinear form $\mathcal{A}_h$ is coercive and continuous with respect to the dG norm, that is, there exists a $C_A, c_A > 0$ such that for all $U, V \in \mathbb{V}_q$

$$\begin{aligned} \mathcal{A}_h(U, V) &\leq C_A \|U\|_{dG} \|V\|_{dG}, \\ c_A \|U\|_{dG}^2 &\leq \mathcal{A}_h(U, U). \end{aligned}$$

When examining the method (7.6) numerically we shall choose $\mathcal{A}_h$ to be the interior penalty method described in Definition 5.1.6 where $\gamma = 0$.

Proposition 7.2.2 (Conservativity of discrete invariants)**.** *Recall that we write the mass as $\mathcal{F}_1(U) := \langle U, 1 \rangle$. Additionally, we redefine the discrete energy of* (7.6) *as*

$$F_3(U) = \frac{1}{2} \mathcal{A}_h(U, U) + \frac{\alpha}{\beta} \|U\|_{L_\beta(S^1)}^\beta$$

Let $U \in \mathbb{V}_q$ be the solution of the discrete scheme (7.6)*. Then U conserves the mass and the discrete energy, i.e.,*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_1(U) &= 0\\ \frac{\mathrm{d}}{\mathrm{d}t}F_3(U) &= 0.\end{aligned}$$

Proof. The desired results follow from an identical argument to that made in the proof of Proposition 5.1.8.

□

We now present an a priori analysis of the spatially discrete scheme (7.6). We begin by introducing a perturbed error equation allowing us to compute the difference between the numerical solution and two yet undetermined discrete functions.

Lemma 7.2.3 (Perturbed error equation and its deviation in energy)**.** *Let U, V be a solution of* (7.6) *and let $\widetilde{U}, \widetilde{V} \in \mathbb{V}_q$ be a solution to the perturbed problem*

$$\begin{aligned}\left\langle \widetilde{U}_t, +\mathcal{G}\left(\widetilde{V}\right), \phi\right\rangle &= -\left\langle E^u, \phi\right\rangle \quad \forall \phi \in \mathbb{V}_q\\ \left\langle \widetilde{V} + f'(\widetilde{U}), \psi\right\rangle + \mathcal{A}_h\left(\widetilde{U}, \psi\right) &= -\left\langle E^v, \psi\right\rangle \quad \forall \psi \in \mathbb{V}_q,\end{aligned} \tag{7.7}$$

for some $E^u, E^v \in \mathbb{V}_q$. Then, with $\theta^u = U - \widetilde{U}$ and $\theta^v = V - \widetilde{V}$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\mathcal{A}_h\left(\theta^u, \theta^u\right) + \frac{\alpha}{\beta}\left\|\theta^u\right\|^{\beta}_{L_\beta(S^1)}\right) = \mathfrak{I}_1 + \mathfrak{I}_2, \tag{7.8}$$

where

$$\mathfrak{I}_1 = \left\langle f'(U) - f'(\widetilde{U}), E^u\right\rangle - \left\langle E^v, \mathcal{G}\left(\theta^v\right)\right\rangle + \mathcal{A}_h\left(\theta^u, E^u\right)$$

and

$$\begin{aligned}\mathfrak{I}_2 = &\left\langle \Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right), E^u\right\rangle\\ &+ \left\langle \mathcal{G}\left(\Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right)\right), \left(E^v - f'(U) + f'(\widetilde{U})\right)\right\rangle\\ &- \mathcal{A}_h\left(\theta^u, \mathcal{G}\left(\Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right)\right)\right),\end{aligned}$$

with Π representing the L_2 projection into the finite element space.

Proof. To begin we note that a discrete error equation is given by taking the difference of

(7.6) and (7.7) yielding

$$\begin{aligned} \langle \theta_t^u + \mathcal{G}\left(\theta^v\right), \phi \rangle &= \langle E^u, \phi \rangle \\ \left\langle \theta^v + f'(U) - f'(\widetilde{U}), \psi \right\rangle + \mathcal{A}_h\left(\theta^u, \psi\right) &= \langle E^v, \phi \rangle . \end{aligned} \tag{7.9}$$

Explicitly computing the time derivative

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F_3\left(\theta^u\right) := \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\mathcal{A}_h\left(\theta^u, \theta^u\right) + \frac{\alpha}{\beta}\left\|\theta^u\right\|_{L_\beta(S^1)}^\beta\right) &= \mathcal{A}_h\left(\theta^u, \theta_t^u\right) + \langle f'(\theta^u), \theta_t^u \rangle \\ &= \mathcal{A}_h\left(\theta^u, \theta_t^u\right) + \left\langle f'(U) - f'(\widetilde{U}), \theta_t^u \right\rangle \\ &\quad - \left\langle f'(U) - f'(\widetilde{U}) - f'(\theta^u), \theta_t^u \right\rangle , \end{aligned}$$

through adding and subtracting appropriately. Now making use of (7.9) we see

$$\frac{\mathrm{d}}{\mathrm{d}t} F_3\left(\theta^u\right) = \langle E^v - \theta^v, \theta_t^u \rangle - \left\langle f'(U) - f'(\widetilde{U}) - f'(\theta^u), \theta_t^u \right\rangle =: \mathfrak{I}_1 + \mathfrak{I}_2.$$

Again using (7.9) we see

$$\begin{aligned} \mathfrak{I}_1 &= \langle E^v - \theta^v, \theta_t^u \rangle \\ &= \langle E^v - \theta^v, E^u - \mathcal{G}\left(\theta^v\right) \rangle \\ &= \langle E^v, E^u \rangle - \langle \theta^v, E^u \rangle - \langle E^v, \mathcal{G}\left(\theta^v\right) \rangle , \end{aligned}$$

where we have used skew-symmetry of $\mathcal{G}$. Further, again by (7.9), we have

$$\mathfrak{I}_1 = \left\langle \left(f'(U) - f'(\widetilde{U})\right), E^u \right\rangle - \langle E^v, \mathcal{G}\left(\theta^v\right) \rangle + \mathcal{A}_h\left(\theta^u, E^u\right) . \tag{7.10}$$

For the other term, making use of (7.9) analogously to the previous argument we see

$$\begin{aligned} \mathfrak{I}_2 &= \left\langle \left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right), \theta_t^u \right\rangle \\ &= \left\langle \Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right), \theta_t^u \right\rangle \\ &= \left\langle \Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right), E^u - \mathcal{G}\left(\theta^v\right) \right\rangle \\ &= \left\langle \Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right), E^u \right\rangle + \left\langle \mathcal{G}\left(\Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right)\right), \theta^v \right\rangle \\ &= \left\langle \Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right), E^u \right\rangle \\ &\quad + \left\langle \mathcal{G}\left(\Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right)\right), E^v - f'(U) + f'(\widetilde{U}) \right\rangle \\ &\quad - \mathcal{A}_h\left(\theta^u, \mathcal{G}\left(\Pi\left(f'(U) - f'(\widetilde{U}) - f'(\theta^u)\right)\right)\right), \end{aligned}$$

concluding the proof. □

Now we have quantified the error between the numerical solution of (7.6) and two discrete objects, we fix these discrete objects, and shall quantify the discrete error (7.8) in the linear case.

Lemma 7.2.4. *Suppose $v \in H^{q+2}(S^1)$ and modify the projection operator $\mathcal{S}(v) \in \mathbb{V}_q$ given in Definition 4.2.15 such that*

$$\begin{aligned} \langle \mathcal{S}(v), \phi \rangle &= \langle v, \phi \rangle \qquad \forall \phi \in \mathbb{V}_{q-1} \\ \{\mathcal{S}(v)_m\} &= v(x_m). \end{aligned}$$

Then

$$\|v - \mathcal{S}(v)\|_{L_2(S^1)} + \|v_x - \mathcal{G}(\mathcal{S}(v))\|_{L_2(S^1)} \leq Ch^{q+1} |v|_{H^{q+2}(S^1)}.$$

Proof. To show the L_2 bound, it suffices to notice that $\mathcal{S}(v)$ is exact when $v \in \mathbb{V}_q$, allowing the use of Bramble-Hilbert. For the gradient bound, note through the definition of $\mathcal{G}$ we have

$$\begin{aligned} \|\Pi(v_x) - \mathcal{G}(\mathcal{S}(v))\|_{L_2(S^1)} &= \sup_{\phi \in L_2, \|\phi\| \leq 1} \langle (\Pi(v_x) - \mathcal{G}(\mathcal{S}(v))), \phi \rangle \\ &= \sup_{\phi \in L_2, \|\phi\| \leq 1} \langle v_x - \mathcal{G}(\mathcal{S}(v)), \Pi\phi \rangle \\ &= \sup_{\phi \in L_2, \|\phi\| \leq 1} -\langle v - \mathcal{S}(v), (\Pi\phi)_x \rangle + \sum_{m=0}^{M-1} \{v_m - \mathcal{S}(v)_m\} [\![\Pi\phi_m]\!] \\ &= 0, \end{aligned}$$

by the definition of $\mathcal{S}$. Hence $\Pi(v_x) = \mathcal{G}(\mathcal{S}(v))$ and the result follows through standard approximation properties of the L_2 projection.

□

Lemma 7.2.5 (Inconsistent Ritz projectors and its error control)**.** *For $u \in H^{q+1}(S^1), v \in H^{q+2}(S^1)$, let $\mathcal{R}(u) \in \mathbb{V}_q$ satisfy*

$$\mathcal{A}_h(\mathcal{R}(u), \phi) + \langle \mathcal{R}(u), \phi \rangle = \mathcal{A}_h(u, \phi) + \langle u + v - \mathcal{S}(v), \phi \rangle. \tag{7.11}$$

Then we have for h_{max} small enough

$$\|u - \mathcal{R}(u)\|_{L_2(S^1)} + h_{max} \|u - \mathcal{R}(u)\|_{dG} \leq Ch_{max}^{q+1} \left(|u|_{H^{q+1}(S^1)} + |v|_{H^{q+2}(S^1)} \right). \tag{7.12}$$

Proof. To show (7.12) we note that through the definition (7.11) we have the orthogonality result

$$\mathcal{A}_h\left(\mathcal{R}\left(u\right)-u,\phi\right)+\left\langle \mathcal{R}\left(u\right)-u,\phi\right\rangle =\left\langle \left(v-\mathcal{S}\left(v\right)\right),\phi\right\rangle \qquad \forall\phi\in\mathbb{V}_q.$$

Hence we have, for any $W\in\mathbb{V}_q$

$$\begin{aligned}
c_A\left\|W-\mathcal{R}\left(u\right)\right\|_{dG}^2 & \\
+\left\|W-\mathcal{R}\left(u\right)\right\|_{L_2(S^1)}^2 &\leq \mathcal{A}_h\left(W-\mathcal{R}\left(u\right),W-\mathcal{R}\left(u\right)\right)+\left\langle W-\mathcal{R}\left(u\right),W-\mathcal{R}\left(u\right)\right\rangle \\
&= \mathcal{A}_h\left(W-u,W-\mathcal{R}\left(u\right)\right)+\left\langle W-u,W-\mathcal{R}\left(u\right)\right\rangle \\
&\quad +\mathcal{A}_h\left(u-\mathcal{R}\left(u\right),W-\mathcal{R}\left(u\right)\right)+\left\langle u-\mathcal{R}\left(u\right),W-\mathcal{R}\left(u\right)\right\rangle \\
&= \mathcal{A}_h\left(W-u,W-\mathcal{R}\left(u\right)\right)+\left\langle W-u,W-\mathcal{R}\left(u\right)\right\rangle \\
&\quad +\left\langle \mathcal{S}\left(v\right)-v,W-\mathcal{R}\left(u\right)\right\rangle \\
&\leq \frac{1}{2}\Bigg(C_A^2\left\|W-u\right\|_{dG}^2+\left\|W-u\right\|_{L_2(S^1)}^2+C_A^2\left\|W-\mathcal{R}\left(u\right)\right\|_{dG}^2 \\
&\quad +2\left\|W-\mathcal{R}\left(u\right)\right\|_{L_2(S^1)}^2+\left\|\mathcal{S}\left(v\right)-v\right\|_{L_2(S^1)}^2\Bigg).
\end{aligned}$$

Thus, choosing $W=\Pi u$, using standard approximation properties of the L_2 projector as well as the bound from Lemma 7.2.4

$$\left\|W-\mathcal{R}\left(u\right)\right\|_{dG}^2\leq C\left(h_{max}^{2q}\left|u\right|_{H^{q+1}(S^1)}^2+h_{max}^{2q+2}\left|v\right|_{H^{q+2}(S^1)}^2\right),$$

and hence the dG norm bound follows from the triangle inequality. To show the L_2 norm, let $z\in H^2$ solve the dual problem

$$-z_{xx}+z=u-\mathcal{R}\left(u\right),$$

then standard elliptic regularity, see [80], shows that

$$\left|z\right|_{H^2(S^1)}\leq C\left\|u-\mathcal{R}\left(u\right)\right\|_{L_2(S^1)}. \tag{7.13}$$

Hence, for any $Z \in \mathbb{V}_q$

$$\begin{aligned}
\|u - \mathcal{R}(u)\|^2_{L_2(S^1)} &= \langle u - \mathcal{R}(u), u - \mathcal{R}(u)\rangle \\
&= \langle -z_{xx} + z, u - \mathcal{R}(u)\rangle \\
&= \mathcal{A}_h(z, u - \mathcal{R}(u)) + \langle z, u - \mathcal{R}(u)\rangle \\
&= \mathcal{A}_h(z - Z, u - \mathcal{R}(u)) + \langle z - Z, u - \mathcal{R}(u)\rangle + \langle \mathcal{S}(v) - v, Z\rangle \\
&\leq C_A \|z - Z\|_{dG} \|u - \mathcal{R}(u)\|_{dG} + \|z - Z\|_{L_2(S^1)} \|u - \mathcal{R}(u)\|_{L_2(S^1)} \\
&\quad + \|\mathcal{S}(v) - v\|_{L_2(S^1)} \|Z\|_{L_2(S^1)},
\end{aligned}$$

using Cauchy-Schwarz. Choosing $Z = \Pi z$, we have

$$\begin{aligned}
\|u - \mathcal{R}(u)\|^2_{L_2(S^1)} &\leq Ch_{max} |z|_{H^2(S^1)} \|u - \mathcal{R}(u)\|_{dG} + Ch^2_{max} |z|_{H^2(S^1)} \|u - \mathcal{R}(u)\|_{L_2(S^1)} \\
&\quad + \|\mathcal{S}(v) - v\|_{L_2(S^1)} \|z\|_{L_2(S^1)}, \\
&\leq Ch_{max} \|u - \mathcal{R}(u)\|_{L_2(S^1)} \|u - \mathcal{R}(u)\|_{dG} + Ch^2_{max} \|u - \mathcal{R}(u)\|^2_{L_2(S^1)} \\
&\quad + Ch^{q+1}_{max} |v|_{H^{q+2}(S^1)},
\end{aligned}$$

using the elliptic regularity result (7.13) and Lemma 7.2.4. Hence

$$\left(1 - Ch^2_{max}\right) \|u - \mathcal{R}(u)\|_{L_2(S^1)} \leq Ch_{max} \|u - \mathcal{R}(u)\|_{dG} + Ch^{q+1}_{max} |v|_{H^{q+2}(S^1)},$$

as required for h_{max} small enough. □

Now we have quantified the difference between the numerical solution (7.6) and appropriate discrete functions we obtain an a priori bound for (7.6) in the linear case.

Theorem 7.2.6 (A priori bound - linear case)**.** *Suppose $f(u) = \frac{1}{2}u^2$, in this case the PDE* (7.1) *is linear and given by*

$$u_t - u_x + u_{xxx} = 0.$$

Let U solve (7.6)*) and the conditions of Lemma 7.2.3, Lemma 7.2.4 and Lemma 7.2.5 hold. Then, for $t \in [0, T]$,*

$$\begin{aligned}
&\|(u - U)(t)\|^2_{dG} \\
&\quad + \|(u - U)(t)\|^2_{L_2(S^1)} \leq C_1 \exp(C_2 t) \Bigg(\|(u - U)(0)\|^2_{dG} + \|(u - U)(0)\|^2_{L_2(S^1)} \\
&\qquad + h^{2q}_{max} \int_0^t |u_t(s, x)|^2_{H^{q+1}(S^1)} + |v(s, x)|^2_{H^{q+2}(S^1)} \, \mathrm{d}s \Bigg).
\end{aligned} \tag{7.14}$$

Proof. We begin by noting that, since $f'(u) = u$, in Lemma 7.2.3 $\mathfrak{I}_2 = 0$, hence we see that

$$\frac{\mathrm{d}}{\mathrm{d}t}F_3\left(\theta^u\right) = \left\langle \theta^u, E^u \right\rangle - \left\langle E^v, \mathcal{G}\left(\theta^v\right)\right\rangle + \mathcal{A}_h\left(\theta^u, E^u\right), \tag{7.15}$$

through (7.10). Observe that the term $\mathcal{G}\left(\theta^v\right)$ is not controllable in $\mathcal{G}\left(\theta^u\right)$ and also will not be of an optimal order. It is prudent for fixed U, V to choose $\widetilde{U}, \widetilde{V}$ such that $E^v = 0$. This then constrains choices for the pair $\widetilde{U}, \widetilde{V}$. We pick $\widetilde{V} = \mathcal{S}\left(v\right)$ and then choose $\widetilde{U} = \mathcal{R}\left(u\right)$. This choice ensures that the perturbed equations

$$\begin{aligned}\left\langle \widetilde{U}_t + \mathcal{G}\left(\widetilde{V}\right), \phi\right\rangle &= -\left\langle E^u, \phi\right\rangle \quad \forall\ \Phi \in \mathbb{V}_q \\ \left\langle \widetilde{V} + f'(\widetilde{U}), \psi\right\rangle + \mathcal{A}_h\left(\widetilde{U}, \psi\right) &= -\left\langle E^v, \psi\right\rangle \quad \forall\ \psi \in \mathbb{V}_q,\end{aligned}$$

are satisfied with

$$\begin{aligned}E^u &= u_t - \widetilde{U}_t + v_x - \mathcal{G}\left(\widetilde{V}\right) \\ E^v &= 0.\end{aligned}$$

Substituting this into (7.15) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}F_3\left(\theta^u\right) = \left\langle \theta^u, E^u \right\rangle + \mathcal{A}_h\left(\theta^u, E^u\right).$$

Now, through Cauchy's inequality we see

$$\frac{\mathrm{d}}{\mathrm{d}t}F_3\left(\theta^u\right) \leq \frac{1}{2}\left(\left\|\theta^u\right\|^2_{L_2(S^1)} + C_A^2\left\|\theta^u\right\|^2_{dG} + \left\|E^u\right\|^2_{L_2(S^1)} + C_A^2\left\|E^u\right\|^2_{dG}\right).$$

Hence Gronwall's inequality, see Lemma 4.2.21, implies that

$$F_3\left(\theta^u(t)\right) \leq \exp\left(C_A^2 t\right)\left(F_3\left(\theta^u(0)\right) + \int_0^t \left\|E^u(s)\right\|^2_{L_2(S^1)} + C_A^2\left\|E^u(s)\right\|^2_{dG}\,\mathrm{d}s\right).$$

It remains to bound the term E^u. We do this by splitting into two components and controlling them individually. First note that since we are in a semi discrete setting, Lemma 7.2.5 yields

$$\left\|u_t - \mathcal{R}\left(u\right)_t\right\|_{L_2(S^1)} \leq C h_{max}^{q+1}\left(\left|u_t\right|_{H^{q+1}(S^1)} + \left|v\right|_{H^{q+2}(S^1)}\right).$$

Further, Lemma 7.2.4 immediately gives

$$\|v_x - \mathcal{G}(\mathcal{S}(v))\|_{L_2(S^1)} \leq C h_{max}^{q+1} |v|_{H^{q+2}(S^1)},$$

hence

$$\|E^u\|^2_{L_2(S^1)} + C_A^2 \|E^u\|^2_{dG} \leq C h_{max}^{2q} \left(|u_t|^2_{H^{q+1}(S^1)} + |v|^2_{H^{q+2}(S^1)} \right),$$

as required.

□

7.3 A posteriori analysis

Here we give an a posteriori analysis of the semi discrete scheme posed in §7.2. We proceed along similar lines to the a priori analysis in that we examine solutions of perturbed equations, taking account of different effects errors induced will have. The difference being, in this section we make use of the stability framework of the underlying PDE.

Lemma 7.3.1 (Perturbed error equation and its deviation in energy)**.** *Let* $u \in C^1([0,T], H^3(S^1))$ *be a strong solution to* (7.1) *and suppose* $\widetilde{u} \in C^1([0,T], H^3(S^1))$ *satisfies*

$$\widetilde{u}_t - f'(\widetilde{u})_x + \widetilde{u}_{xxx} = -\mathfrak{R},$$

for some $\mathfrak{R} \in L_2(S^1)$*. Additionally assume that the bilinear form* $\mathcal{A}_h$ *is consistent in the sense that its action on a continuous function is the same as the corresponding continuous operator. Then, with* $\rho := u - \widetilde{u}$

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{2} \mathcal{A}_h(\rho, \rho) + \frac{\alpha}{\beta} \|\rho\|^\beta_{L_\beta(S^1)} \right) = \mathcal{J}_1 + \mathcal{J}_2,$$

where

$$\begin{aligned}
\mathcal{J}_1 &= \langle -\rho_{xx} + f'(u) - f'(\widetilde{u}), \mathfrak{R} \rangle \\
\mathcal{J}_2 &= \langle (f'(\rho) - f'(u) + f'(\widetilde{u})), \mathfrak{R} \rangle + \langle (f'(\rho) - f'(u) + f'(\widetilde{u}))_x, \rho_{xx} \rangle \\
&\quad - \langle f'(\rho), (f'(u) - f'(\widetilde{u}))_x \rangle .
\end{aligned}$$

Proof. To begin, we note that $\rho = u - \widetilde{u}$ satisfies the discrete error equation

$$\rho_t - (f'(u)_x - f'(\widetilde{u})_x) + \rho_{xxx} = \mathfrak{R}. \tag{7.16}$$

Note that the bilinear form $\mathcal{A}_h$ is consistent, so explicitly computing the time derivative of the energy we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F_3(\rho) &= \langle \rho_x, \rho_{xt}\rangle + \langle f'(\rho), \rho_t\rangle \\ &= -\langle \rho_{xx}, \rho_t\rangle + \langle f'(u) - f'(\widetilde{u}), \rho_t\rangle + \langle f'(\rho) - f'(u) - f'(\widetilde{u}), \rho_t\rangle =: \mathcal{J}_1 + \mathcal{J}_2.\end{aligned}$$

Making use of (7.16) we see

$$\begin{aligned}\mathcal{J}_1 &= -\langle \rho_{xx}, \rho_t\rangle + \langle f'(u) - f'(\widetilde{u}), \rho_t\rangle \\ &= \langle -\rho_{xx} + f'(u) - f'(\widetilde{u}), \mathfrak{R} - \rho_{xxx} + f'(u)_x - f'(\widetilde{u})_x\rangle \\ &= \langle -\rho_{xx} + f'(u) - f'(\widetilde{u}), \mathfrak{R}\rangle .\end{aligned}$$

Further,

$$\begin{aligned}\mathcal{J}_2 &= \langle f'(\rho) - f'(u) + f'(\widetilde{u}), \rho_t\rangle \\ &= \langle f'(\rho) - f'(u) + f'(\widetilde{u}), \mathfrak{R} - \rho_{xxx} + f'(u)_x - f'(\widetilde{u})_x\rangle \\ &= \langle f'(\rho) - f'(u) + f'(\widetilde{u}), \mathfrak{R}\rangle + \langle (f'(\rho) - f'(u) + f'(\widetilde{u}))_x, \rho_{xx}\rangle \\ &\quad - \langle f'(\rho) - f'(u) + f'(\widetilde{u}), (f'(u) - f'(\widetilde{u}))_x\rangle \\ &= \langle f'(\rho) - f'(u) + f'(\widetilde{u}), \mathfrak{R}\rangle + \langle (f'(\rho) - f'(u) + f'(\widetilde{u}))_x, \rho_{xx}\rangle \\ &\quad - \langle f'(\rho), (f'(u) - f'(\widetilde{u}))_x\rangle ,\end{aligned}$$

as required. □

Remark 7.3.2 (The linear case)**.** *For the sake of exposition, similarly to the a priori case, we have divided the contributions of the energy identity into two components, $\mathcal{J}_1$ and $\mathcal{J}_2$, where $\mathcal{J}_2 = 0$ in the case the problem is linear.*
Here we conduct the a posteriori analysis for the linear problem, hence in this section we take $f(u) = \frac{1}{2}u^2$, we leave the analysis of the nonlinear problem for future work.

Definition 7.3.3 (Discrete reconstruction operator $\mathcal{D}$)**.** *We define $\mathcal{D} : \mathbb{V}_q \to \mathbb{V}_{q+1}$ to be the* discrete reconstruction operator *satisfying for $W \in \mathbb{V}_q$*

$$\langle \mathcal{D}(W)_x - \mathcal{G}(W), \phi\rangle = 0 \qquad \forall \phi \in \mathbb{V}_q,$$

where $\mathcal{G}$ is given in Definition 4.2.3, and

$$\mathcal{D}(W)_m = \{W_m\},$$

for $m = 0, ..., M$.

Remark 7.3.4 (Continuity of the discrete reconstruction operator)**.** *Note that* $\mathcal{D}$ *is constructed such that for any* $W \in \mathbb{V}_q$ *we have that* $\mathcal{D}(W) \in \mathbb{V}_{q+1} \cap \mathcal{C}^0(S^1)$*. In addition we have the approximation properties, proofs of which can be found in [133],*

$$\|W - \mathcal{D}(W)\|^2_{L_2(S^1)} \leq C \left\| h^{1/2} [\![W]\!] \right\|^2_{L_2}$$
$$\|W - \mathcal{D}(W)\|^2_{dG} \leq \left\| h^{-1/2} [\![\Psi]\!] \right\|^2_{L_2}.$$

Remark 7.3.5 (Orthogonality)**.** *Note that* $\mathcal{D}$ *is constructed such that for any* $W \in \mathbb{V}_q$ *and* $\phi \in \mathbb{V}_{q-1}$ *we have that*

$$\langle \mathcal{D}(W) - W, \phi \rangle = 0.$$

A proof can be found in [75]

Definition 7.3.6 (Elliptic reconstruction)**.** *Let* $\mathscr{R} : \mathbb{V}_q \to H^1(S^1)$*, we define the elliptic reconstruction* $\mathscr{R}(U)$ *as the solution of*

$$-\mathscr{R}(U)_{xx} + f'(\mathscr{R}(U)) = -\mathcal{D}(V) \tag{7.17}$$

with average value matching the discrete solution, that is

$$\langle \mathscr{R}(U) - U, 1 \rangle = 0$$

Proposition 7.3.7 (Regularity bound for the reconstruction)**.** *The elliptic problems defining the reconstruction operators in Definition 7.3.6 are well posed, moreover, thanks to elliptic regularity (see [80]), we have*

$$\|\mathscr{R}(U)\|_{H^{k+1}(S^1)} \leq C \|\mathcal{D}(V)\|_{H^{k-1}(S^1)} \qquad \text{for } k = 0, 1, 2.$$

Lemma 7.3.8 (Reconstructed PDE)**.** *The reconstruction given in Definition 7.3.6 satisfies*

$$\mathscr{R}(U)_t - f'(\mathscr{R}(U))_x + \mathscr{R}(U)_{xxx} = \mathfrak{R}, \tag{7.18}$$

with

$$\mathfrak{R} = (\mathscr{R}(U) - U)_t.$$

Proof. Since $\mathscr{R}(U)$ satisfies (7.17) and the problem data $\mathcal{D}(V) \in H^1(S^1)$ it is clear that

$\mathscr{R}(U) \in H^3(S^1)$ and satisfies

$$-\mathscr{R}(U)_{xxx} + f'(\mathscr{R}(U))_x = -\mathcal{D}(V)_x. \tag{7.19}$$

Further, the first equation of the semi discrete scheme (7.6) states

$$\begin{aligned} 0 &= \langle U_t + \mathcal{G}(V), \phi \rangle \\ &= \langle U_t + \mathcal{D}(V)_x, \phi \rangle, \end{aligned} \tag{7.20}$$

using the discrete reconstruction given in Definition 7.3.3. Since $U_t, \mathcal{D}(V)_x \in \mathbb{V}_q$ (7.20) can be written pointwise as

$$U_t + \mathcal{D}(V)_x = 0. \tag{7.21}$$

Substituting (7.19) and (7.21) into (7.18) we see

$$\begin{aligned} \mathscr{R}(U)_t - f'(\mathscr{R}(U))_x + \mathscr{R}(U)_{xxx} &= \mathscr{R}(U)_t + \mathcal{D}(V)_x \\ &= (\mathscr{R}(U) - U)_t, \end{aligned}$$

as required.

□

Proposition 7.3.9 (A posteriori control for the elliptic problem)**.** *The reconstruction $\mathscr{R}(U)$ is the elliptic reconstruction of U [132]. There exists an optimal order elliptic a posteriori estimate controlling $\|U - \mathscr{R}(U)\|_{L_2(S^1)}$ and $\|U - \mathscr{R}(U)\|_{dG}$, that is, there exist functionals $\eta_{0,1}$ depending only upon U and the problem data such that*

$$\begin{aligned} \|U - \mathscr{R}(U)\|_{L_2(S^1)} &\leq \eta_0(U, g) \sim O(h_{max}^{q+1}) \\ \|U - \mathscr{R}(U)\|_{dG} &\leq \eta_1(U, g) \sim O(h_{max}^{q}), \end{aligned}$$

where g represents the right hand side of the elliptic reconstruction. Indeed, with $g := -\mathcal{D}(V)$ in (7.17), for $\mathcal{A}_h(\cdot,\cdot)$ given by the interior penalty discretisation (5.8) an estimate of the form

$$\eta_0(U, g) = C \sum_{m=0}^{M-1} \left(h_m^4 \|g + U_{xx} - f'(U)\|_{L_2(\mathcal{J}_m)}^2 + \widetilde{h_m^{-3}} [\![U_{xm}]\!]^2 + \sigma \widetilde{h_m} [\![U_m]\!]^2 + \widetilde{h_m} [\![V_m]\!]^2 \right),$$

$$\eta_1(U, g) = C \sum_{m=0}^{M-1} \left(h_m^2 \|g + U_{xx} - f'(U)\|_{L_2(\mathcal{J}_m)}^2 + \widetilde{h_m} [\![U_{xm}]\!]^2 + \sigma \widetilde{h_m^{-1}} [\![U_m]\!]^2 + \widetilde{h_m} [\![V_m]\!]^2 \right),$$

where $\widetilde{h_m}$ is the maximal element size in a local patch of elements as given in (4.11).

Proof. The result is an extension of those found in [178, 116] whilst noting that, due to the definition of the elliptic reconstruction $\mathscr{R}(U)$, U satisfies the orthogonality condition

$$\mathcal{A}_h\left(\mathscr{R}(U) - U, \phi\right) + \left\langle f'(\mathscr{R}(U)) - f'(U), \phi\right\rangle = \left\langle V - \mathcal{D}(V), \phi\right\rangle \qquad \forall \phi \in \mathbb{V}_q.$$

This induces an inconsistency error that can be controlled by the results in Remark 7.3.4. □

Remark 7.3.10 (Alternative estimators)**.** *One of the strengths of the elliptic reconstruction methodology is the ability to use other types of estimator that are not residual based. Indeed, recovery based a posteriori estimators have been widely used since their introduction by the engineering community in the 1980s. Their success in applications is due to their simplicity of implementation, milder dependence of problem data than other estimators and certain superconvergence properties. Work carried out on recovery estimators has reached a state of maturity for elliptic problems, see [6, 23, 187, 123] and subsequent references.*

Theorem 7.3.11 (A posteriori bound - linear case)**.** *Suppose $f(u) = \frac{1}{2}u^2$. Further, let U solve* (7.6) *and the conditions of Lemma 7.3.1 and Lemma 7.3.8 hold. Then, for $t \in [0, T]$,*

$$\begin{aligned}\left\|(u-U)(t)\right\|^2_{L_2(S^1)} &\\ + \left\|(u-U)(t)\right\|^2_{dG} \leq \exp(t)\Bigg(&\left\|(u-U)(0)\right\|^2_{dG} + \left\|(u-U)(0)\right\|^2_{L_2(S^1)} \\ &+ \int_0^t \eta_1(U_t,(s)g_t(s))^2 + \eta_0(U_t(s),g_t(s))^2\,\mathrm{d}s\Bigg).\end{aligned} \tag{7.22}$$

Proof. Since $f'(u) = u$, in Lemma 7.3.1 $\mathfrak{R} = \mathscr{R}(U)_t - U_t$, hence

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F_3(\rho) &= \left\langle -\rho_{xx} + \rho, \mathfrak{R}\right\rangle \\ &= \mathcal{A}_h\left(\rho, \mathscr{R}(U)_t\right) - \mathcal{A}_h\left(\rho, U_t\right) + \left\langle \rho, \mathscr{R}(U)_t - U_t\right\rangle \\ &\leq \frac{1}{2}\left(C_A^2\left\|\rho_x\right\|^2_{L_2(S^1)} + C_A^2\left\|\mathscr{R}(U)_t - U_t\right\|^2_{dG} + \left\|\rho\right\|^2_{L_2(S^1)} + \left\|\mathscr{R}(U)_t - U_t\right\|^2_{L_2(S^1)}\right).\end{aligned}$$

Gronwall's inequality, given in Lemma 4.2.21, implies

$$F_3\left(\rho(t)\right) \leq \exp\left(C_A^2 t\right)\Big(F_3\left(\rho(0)\right) + \int_0^t C_A^2 \left\|\left(\mathscr{R}\left(U\right)_t - U_t\right)(s)\right\|_{dG}^2 + \left\|\left(\mathscr{R}\left(U\right)_t - U_t\right)(s)\right\|_{L_2(S^1)}^2 \,\mathrm{d}s.\Big)$$

It remains to computationally bound $\mathscr{R}\left(U\right)_t - U_t$ for which we can invoke the results of Proposition 7.3.9, concluding the proof.

□

Remark 7.3.12 (Suboptimality in L_2)**.** *The bound for the pointwise in time L_2 error, appearing on the left-hand side of* (7.22)*, is tight only for very short times. As we will observe in §7.5 on a uniform mesh of size $h \to 0$ the gradient term $\|u - U\|_{dG} = \mathcal{O}\left(h_{max}^q\right)$, while $\|(u-U)(t)\|_{L_2(S^1)} = \mathcal{O}\left(h_{max}^{q+1}\right)$.*

7.4 Temporal discretisation and the design of a stable adaptive algorithm

Practically, a fully discrete approximation scheme is required for implementation. Here we present an argument for designing a fully discrete scheme similar to that in §4.2 and §5.3. For brevity, we suggest that the reader recalls the required notation for the temporal discretisation as presented in §5.3.

Definition 7.4.1 (Fully discrete scheme for defocusing KdV type equations)**.** *Given $U^0 \in \mathbb{V}_q^0$, for $n \in [0, N-1]$ find $U^{n+1} \in \mathbb{V}_q^{n+1}$ such that*

$$\begin{aligned} \left\langle \frac{U^{n+1} - \mathcal{P}^{n+1}U^n}{\tau_n} + \mathcal{G}\left(V^{n+1}\right), \phi \right\rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q^{n+1} \\ \left\langle V^{n+1} + \frac{f(U^{n+1}) - f(\mathcal{P}^{n+1}U^n)}{U^{n+1} - \mathcal{P}^{n+1}U^n}, \psi \right\rangle + \mathcal{A}_h\left(U^{n+\frac{1}{2}}, \psi\right) &= 0 \qquad \forall \psi \in \mathbb{V}_q^{n+1} \\ U^0 &= \Pi^0 u_0 \end{aligned} \tag{7.23}$$

where $U^{n+\frac{1}{2}} := U^{n+1} + \mathcal{P}^{n+1}U^n$ and Π^0 denotes the L_2 orthogonal projector into the initial finite element space, $\mathcal{G}$ is described by Definition 4.2.3 and $\mathcal{A}_h$ is described by Definition 5.1.6.

Remark 7.4.2 (Conserving invariants in mesh dependent invariant)**.** *It is not possible to conserve invariants which depend adaptively on the underlying spatial mesh, as this allows the invariant to change in time. As such, we must remove the spatial dependency of* $\mathcal{A}_h$ *by replacing* h *with* h_{min} *in* (5.8)*, where* h_{min} *is the minimal obtainable spatial element size in the adaptive simulation.*

Proposition 7.4.3 (Conservativity of the fully discrete scheme)**.** *Let* $\{U^n\}_{n=0}^N$ *be the fully discrete scheme generated by* (7.23)*, then we have that*

$$\mathcal{F}_1\Big(U^{n+1}\Big) = \mathcal{F}_1\Big(\mathcal{P}^{n+1}U^n\Big)$$

and

$$F_3\left(U^{n+1}\right) = F_3\left(\mathcal{P}^{n+1}U^n\right)$$

Proof. The proof of Proposition 7.4.3 follows the same methodology as the proof of Proposition 5.3.8, noting the symmetric of the bilinear form $\mathcal{A}_h$.

□

Remark 7.4.4 (Conservation over spatially adapting meshes)**.** *Assuming a non-adaptive spatial mesh our fully discrete scheme through Proposition 7.4.3 we have that a discrete mass and energy are conserved. However, over an adaptive mesh this is not necessarily the case and is highly dependent on the mesh change operator* $\mathcal{P}^{n+1}$*. Recall, from Chapter 6, that through choosing the mesh change operator to be the Lagrange interpolant* (6.2) *no invariants are conserved for arbitrary adaptations. We also found that the* L_2 *projection* (6.3) *conserves the mass, and dissipates momentum yielding numerical stability for the adaptive algorithm* (6.1)*. Unfortunately the scheme under consideration is not momentum conserving in the non-adaptive setting, and as such the* L_2 *projection does not lead to numerical stability here. This leads us to propose a new mesh change operator with an aim of dissipating the energy.*

In view of Remark 7.4.4, we define the Ritz projection as the following mesh change operator.

Definition 7.4.5 (The Ritz projection)**.** *For a function* $U^n \in \mathbb{V}_q^n$*, we define the Ritz*

projection $\mathscr{L}^{n+1} : \mathbb{V}_q^n \to \mathbb{V}_q^{n+1}$ *by seeking* $\mathscr{L}^{n+1}U^n \in \mathbb{V}_q^{n+1}$ *such that*

$$\begin{aligned}\frac{1}{\beta^*}\mathcal{A}_h\left(\mathscr{L}^{n+1}U^n, \phi\right) \\ + \frac{\alpha}{\beta}\left\langle \left(\mathscr{L}^{n+1}U^n\right)^{\beta-1}, \phi \right\rangle = \frac{1}{\beta^*}\mathcal{A}_h\left(U^n, \phi\right) + \frac{\alpha}{\beta}\left\langle \left(U^n\right)^{\beta-1}, \phi \right\rangle \qquad \forall \phi \in \mathbb{V}_q^{n+1},\end{aligned} \tag{7.24}$$

where $\frac{1}{\beta} + \frac{1}{\beta^*} = 1$.

Proposition 7.4.6 (Conservative properties of the Ritz projection)**.** *Let* U^n *be the solution of the adaptive scheme* (7.23)*, and* $\mathscr{L}^{n+1}$ *be the Ritz projection described in Definition 7.4.5. Under an adaptive mesh the energy is stable, i.e.,*

$$F_3\left(\mathscr{L}^{n+1}U^n\right) \leq F_3\left(U^n\right).$$

In addition for the linear problem, so $f(u) = \frac{1}{2}u^2$*, then mass is also conserved as long as* $\mathcal{A}_h$ *is consistent, i.e.,*

$$\mathcal{F}_1\Big(\mathscr{L}^{n+1}U^n\Big) = \mathcal{F}_1(U^n).$$

Remark 7.4.7 (The Ritz projection as a mesh change operator)**.** *Through amalgamating Proposition 7.4.6 with Proposition 7.4.3, we find that our adaptive algorithm for linear problems, i.e.,* $f(u) = \frac{1}{2}u^2$*, is mass conservative. Further to this, we find that the energy dissipates. As the energy induces a norm, we immediately obtain stability of the adaptive algorithm with the Ritz projection as the mesh change operator in the natural energy norm for the problem.*

Proof of Proposition 7.4.6. If we restrict ourselves to the linear case where $f(u) = \frac{1}{2}u$, then through choosing $\phi = 1$ in (7.24) we observe that mass is conserved for any consistent $\mathcal{A}_h$, i.e.,

$$\left\langle \mathscr{L}^{n+1}U^n, 1 \right\rangle = \left\langle U^n, 1 \right\rangle.$$

Now let us consider the stability properties of the Ritz projector. Through choosing $\phi = \mathscr{L}^{n+1}U^n$ we observe that

$$\frac{1}{\beta^*}\mathcal{A}_h\left(\mathscr{L}^{n+1}U^n - U^n, \mathscr{L}^{n+1}U^n\right) + \frac{\alpha}{\beta}\left\langle \left(\mathscr{L}^{n+1}U^n\right)^{\beta-1} - \left(U^n\right)^{\beta-1}, \mathscr{L}^{n+1}U^n \right\rangle = 0. \tag{7.25}$$

For clarity of exposition we may the two terms components in (7.25) independently, as

the two terms do not interact. Examining the first term we find that

$$\begin{aligned}\frac{1}{\beta^*}\mathcal{A}_h\left(\mathscr{L}^{n+1}U^n, \mathscr{L}^{n+1}U^n\right) &= \frac{1}{\beta^*}\mathcal{A}_h\left(U^n, \mathscr{L}^{n+1}U^n\right) \\ &\leq \frac{1}{\beta^*}\mathcal{A}_h\left(U^n, U^n\right)^{\frac{1}{2}}\mathcal{A}_h\left(\mathscr{L}^{n+1}U^n, \mathscr{L}^{n+1}U^n\right)^{\frac{1}{2}},\end{aligned}$$

as $\mathcal{A}_h(\cdot,\cdot)$ defines an inner product. Additionally, through Cauchy's inequality with ϵ where $\epsilon = \frac{1}{2}$ we find that

$$\frac{1}{2\beta^*}\mathcal{A}_h\left(\mathscr{L}^{n+1}U^n, \mathscr{L}^{n+1}U^n\right) \leq \frac{1}{2\beta^*}\mathcal{A}_h\left(U^n, U^n\right). \tag{7.26}$$

Through application of Hölder's inequality to the second form of (7.25) we can write

$$\begin{aligned}\frac{\alpha}{\beta}\left\|\mathscr{L}^{n+1}U^n\right\|^{\beta}_{L_\beta(S^1)} &= \frac{\alpha}{\beta}\left\langle (U^n)^{\beta-1}, \mathscr{L}^{n+1}U^n\right\rangle \\ &\leq \frac{\alpha}{\beta}\left\|(U^n)^{\beta-1}\right\|_{L_{\beta^*}(S^1)}\left\|\mathscr{L}^{n+1}U^n\right\|_{L_\beta(S^1)}\end{aligned}$$

where $\frac{1}{\beta} + \frac{1}{\beta^*} = 1$. Further, applying Young's inequality we have that

$$\frac{\alpha}{\beta}\left\|\mathscr{L}^{n+1}U^n\right\|^{\beta}_{L_\beta(S^1)} \leq \frac{\alpha}{\beta}\left(\frac{1}{\beta^*}\left\|(U^n)^{\beta-1}\right\|^{\beta^*}_{L_{\beta^*}(S^1)} + \frac{1}{\beta}\left\|\mathscr{L}^{n+1}U^n\right\|^{\beta}_{L_\beta(S^1)}\right).$$

Noting that through the definition of the Lebesgue norms we have

$$\left\|(U^n)^{\beta-1}\right\|_{L_{\beta^*}(S^1)} = \|U^n\|^{\beta}_{L_\beta(S^1)},$$

which allows us to conclude that

$$\frac{\alpha}{\beta\beta^*}\left\|\mathscr{L}^{n+1}U^n\right\|^{\beta}_{L_\beta(S^1)} \leq \frac{\alpha}{\beta\beta^*}\|U^n\|^{\beta}_{L_\beta(S^1)}, \tag{7.27}$$

after observing that $1 - \frac{1}{\beta} = \frac{1}{\beta^*}$. Combining (7.26) with (7.27) completes the proof.

□

7.5 Numerical experiments

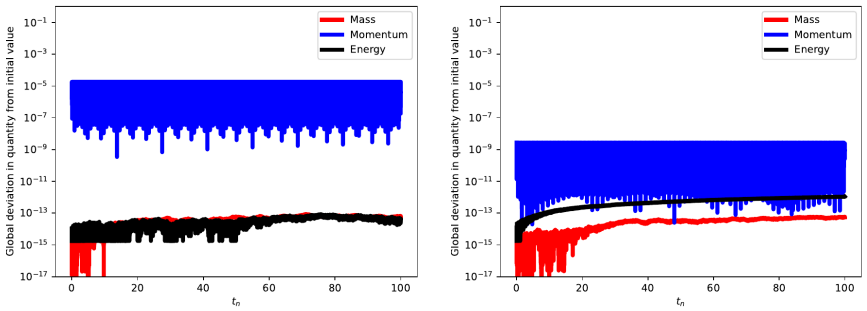
Here we conduct numerical experiments for the proposed fully discrete scheme (7.23) in both the linear and nonlinear case, with particular emphasis on the linear case. In the

uniform setting we present the experimental order of convergence and plots displaying the deviation in mass, momentum and energy, similarly to those seen in §4.2.3 and §5.4. We go on to present results in the adaptive setting, and investigate solution dynamics along with the deviation in invariants over adaptive spatial meshes, similarly to §6.3.
The non-adaptive components of our code are implemented in Firedrake [153], and depending on the nature of the problem, utilise either a direct solver or Newton line search method with a tolerance of 10^{-12} in PETSc [20]. We utilise a Gauss quadrature of high enough degree that the finite element method is evaluated exactly, and when integrating continuous functions we ensure that our quadrature approximation does not introduce leading order errors. The adaptive components of our code are implemented in Python 3 with mesh change operators utilising Numpy linear solvers.
Similarly to the prequel, in our numerical experiments we stretch our periodic domain such that $S^1(0,1) \to S^1(0,40)$ for consistency. For the linear problem, i.e., when $f(u) = \frac{1}{2}u^2$ in (7.1), we numerically simulate a spatially stretched version of the exact solution (7.4) with $C_1 = 1$, $C_2 = 0$ and $l = 1$. For the nonlinear problem $f(u) = \frac{1}{4}u^4$ we simulate the exact solution (7.5) with modulus $k = 0.9$ and rescale the spatial domain by $x \to 1.031236845339269070 37\tilde{x}$ to numerically enforce the solution is periodic up to a tolerance of 10^{-15}.

7.5.1 Uniform experiments

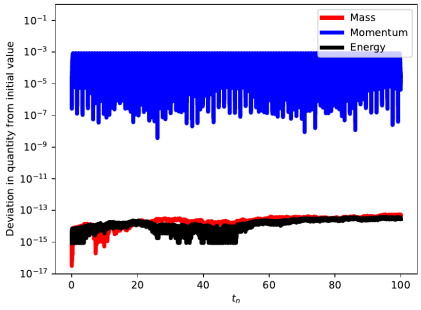
We begin by examining the global deviation in invariants for both the linear and nonlinear problems. We observe, in Figure 7.1, that both problems conserve the expected invariants. Notice that often these deviations propagate in time, this is due to the propagation of errors below either machine precision or our solver tolerance.
We plot the experimental order of convergence for the linear problem in Figure 7.2. We observe that the dG norm sharply obtains the same convergence rate observed in the a priori bound (7.14). However, the L_2 component of the error superconverges with respect to the a priori bound, in fact the error here agrees with best approximation results in the L_2 norm.

Figure 7.1: The deviation in mass, momentum and energy with $T = 100$ for the scheme (7.23) with either $f(u) = \frac{1}{2}u^2$ or $f(u) = \frac{1}{4}u^4$. If $f(u) = \frac{1}{2}u^2$ then we initialise the scheme with (7.4), otherwise if $f(u) = \frac{1}{4}u^4$ we initialise the scheme with (7.5). Further we choose $\tau_n = 0.2$, $h_m = 0.4$ and vary the polynomial degree q.

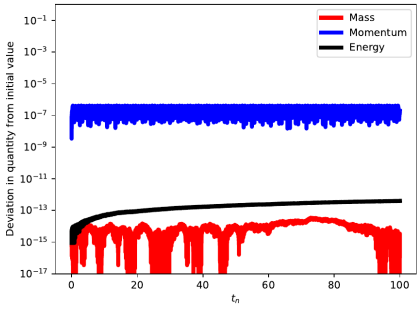


(a) Here $f(u) = \frac{1}{2}u^2$ and $q = 1$.

(b) Here $f(u) = \frac{1}{2}u^2$ and $q = 2$.

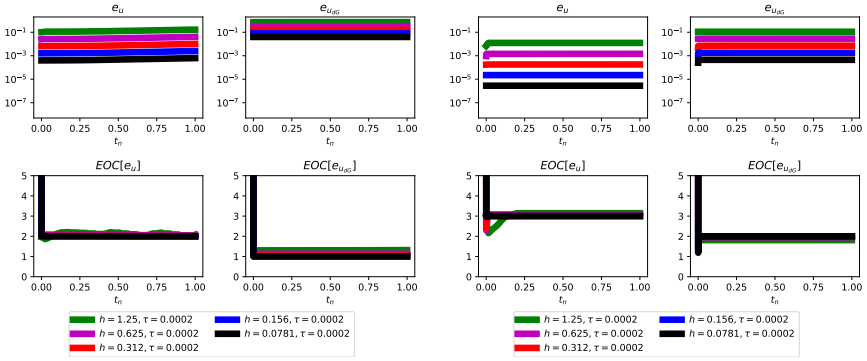


(c) Here $f(u) = \frac{1}{4}u^4$ and $q = 1$.



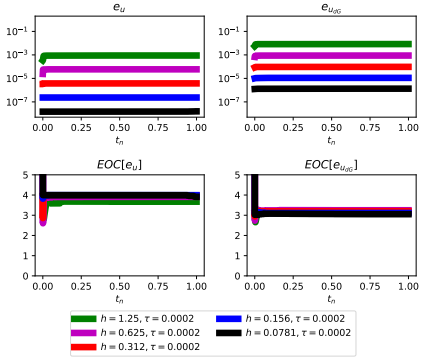
(d) Here $f(u) = \frac{1}{4}u^4$ and $q = 2$.

Figure 7.2: The errors of (7.23) in both the L_2 and dG norm (5.18), and associated experimental order of convergence with the corresponding exact solution (7.4) with polynomial degrees $q = 1, 2, 3$. Here we fix $\tau_n = 0.0002$ and varying h_m. We observe the a priori bound (7.14) is attained, however the L_2 component superconverges with respect to the a priori bound.



(a) Here $q = 1$.

(b) Here $q = 2$.



(c) Here $q = 3$.

7.5.2 Adaptive experiments

Here we implement the adaptive algorithm (7.23) following the methodology outlined in §6.1, subject to the mesh change parameters

$$\begin{aligned} \texttt{coarsen} &= 10 & h_{\texttt{max}} &= 1 \\ \texttt{refine} &= 60 & h_{\texttt{min}} &= 0.2. \end{aligned} \tag{7.28}$$

We shall also consider the effect of increasing the percentage of elements coarsened from $\texttt{coarsen} = 10$ to $\texttt{coarsen} = 30$. Throughout we assume an initial uniform spatial mesh with element size $h_m = 0.4$.

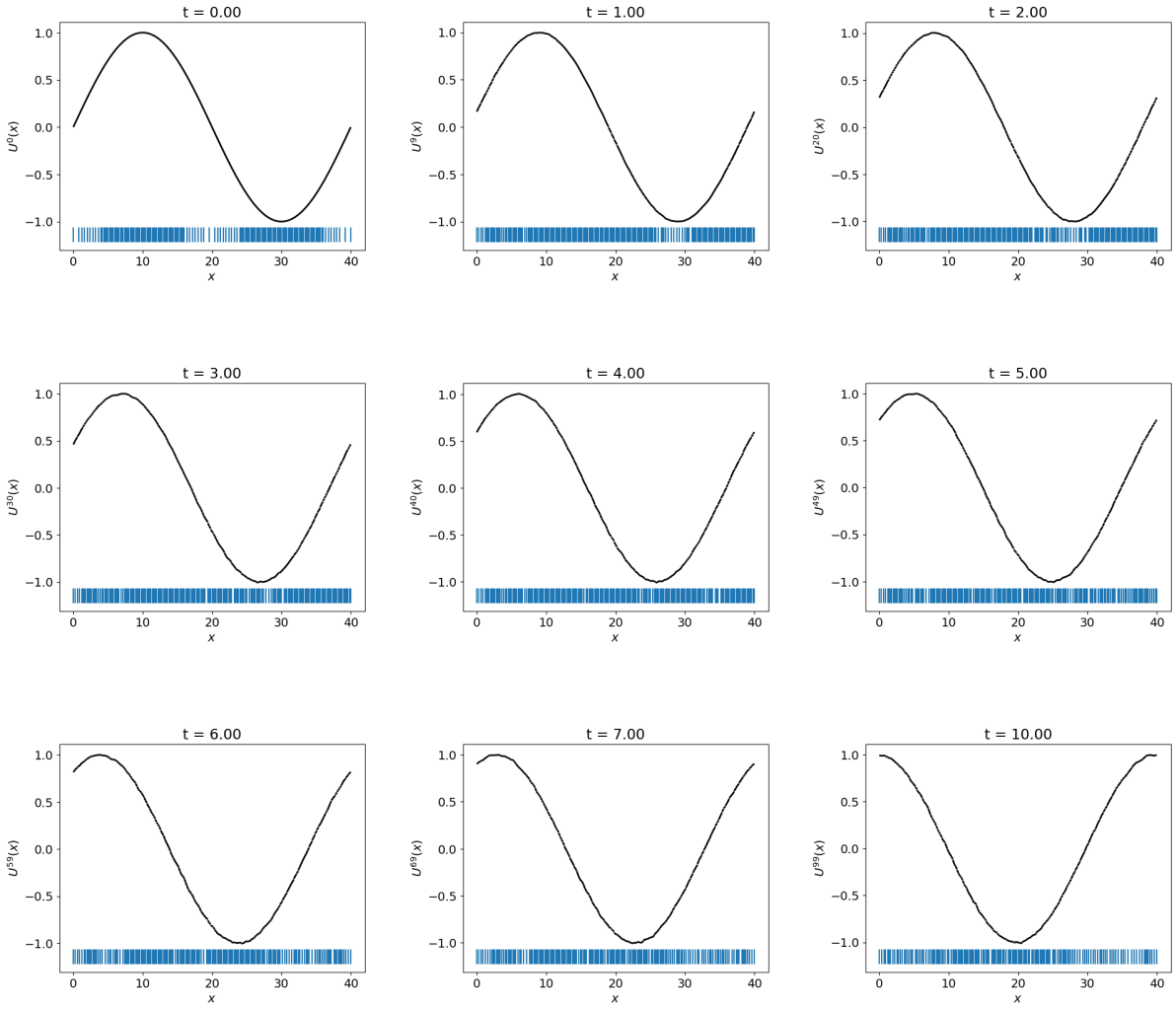
We focus our attention on not only the Ritz projection (7.24) as a mesh change operator, but also the Lagrange interpolant (6.2) and L_2 projection (6.3).

We begin by examining solution dynamics of the linear problems for different mesh change operators. In Figure 7.3, Figure 7.4 and Figure 7.5 we choose the Lagrange interpolant, L_2 projection and Ritz projection respectively as mesh change operators for the linear problem $f(u) = \frac{1}{2}u$. We observe that while all three solutions look similar in this case the mesh adaptivity is highly dependent on the mesh change operator that we consider, however a similar overall number of degrees of freedom are used. We observe that over time the L_2 projection leads to a nonsmooth approximation. While the Lagrange interpolant is significantly smoother the smoothest approximation is obtained by the Ritz projector, likely due to its stability in the energy norm.

While we cannot explicitly see significant differences through visualising the solution dynamics for different choices of mesh change operators, through examining the deviation in the "invariants" of the scheme we obtain a measure of how well the adaptive schemes are performing. We consider the invariants of the adaptive scheme with Lagrange interpolant, L_2 projection and Ritz projection as mesh change operators in Figure 7.6, Figure 7.7 and Figure 7.8 respectively. The Lagrange interpolation operator appears to dissipate the momentum and energy in this case, but it does not conserve the mass. The L_2 projection, while conserving the mass, is not stable with respect to the energy. We observe that for the Ritz projection the mass is conserved, and the energy dissipates as we expect from Proposition 7.4.6.

We can further compare the mesh change operators through their respective errors, see Figure 7.9. We observe that in the energy norm the Ritz projector slightly outperforms the interpolant, however, in the L_2 norm the L_2 projector and interpolant outperform the

Figure 7.3: Here we examine the dynamics of the adaptive algorithm (7.23) with the *Lagrange interpolant* (6.2) as the mesh change operator. The mesh coordinates are represented by vertical blue lines at the bottom of each solution snapshot. We initialise the simulation with the L_2 projection of (7.4) at $t = 0$ and employ the adaptive parameters (7.28). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We allow for the initial mesh to be adapted.



Ritz projector. By increasing the percentage of elements we coarsen to `coarsen` $= 30$, we obtain Figure 7.10 which exaggerates this point.

Figure 7.4: Here we examine the dynamics of the adaptive algorithm (7.23) with the L_2 *projection* (6.3) as the mesh change operator. The mesh coordinates are represented by vertical blue lines at the bottom of each solution snapshot. We initialise the simulation with the L_2 projection of (7.4) at $t = 0$ and employ the adaptive parameters (7.28). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We allow for the initial mesh to be adapted.

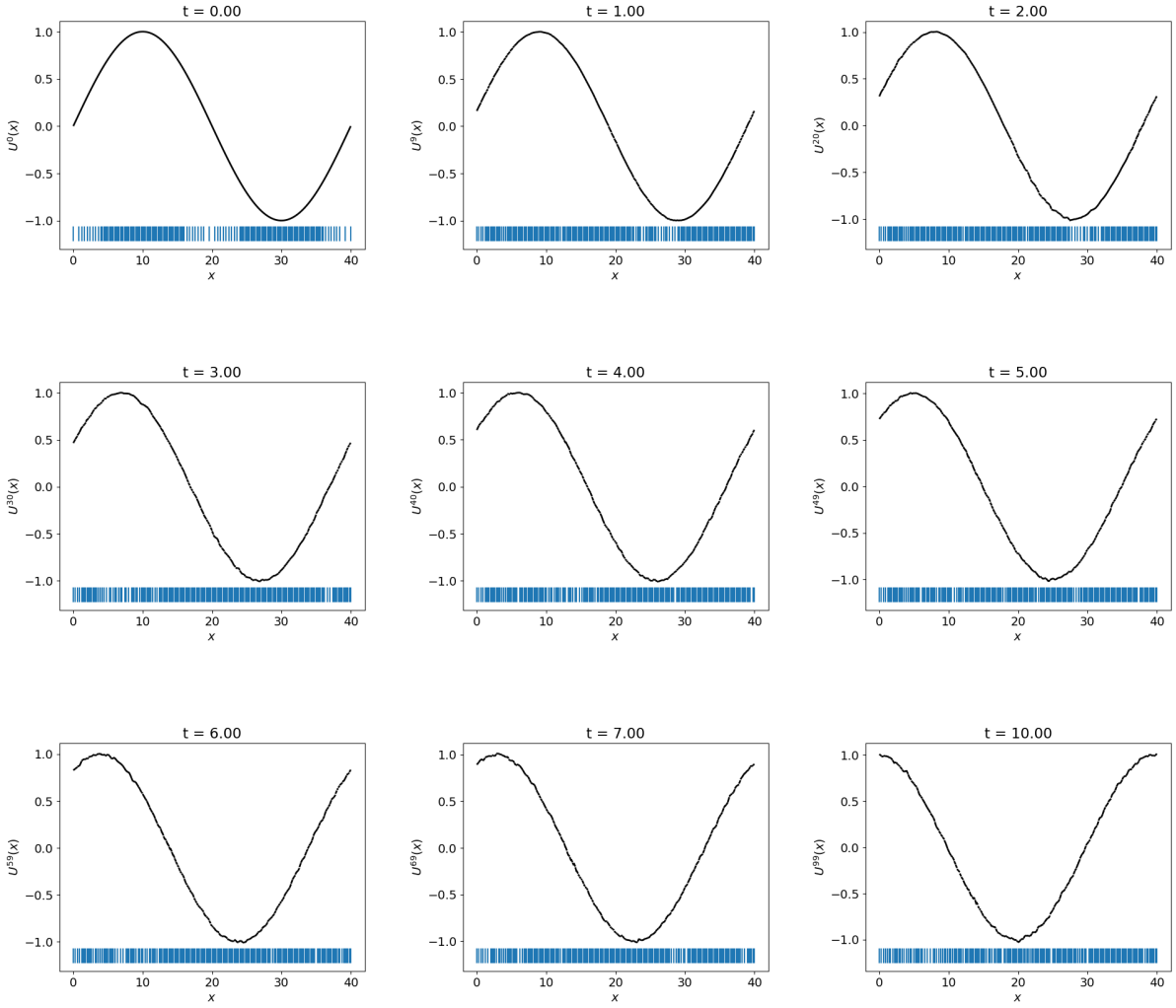


Figure 7.5: Here we examine the dynamics of the adaptive algorithm (7.23) with the *Ritz projection* (7.24) as the mesh change operator. The mesh coordinates are represented by vertical blue lines at the bottom of each solution snapshot. We initialise the simulation with the L_2 projection of (7.4) at $t = 0$ and employ the adaptive parameters (7.28). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We allow for the initial mesh to be adapted.

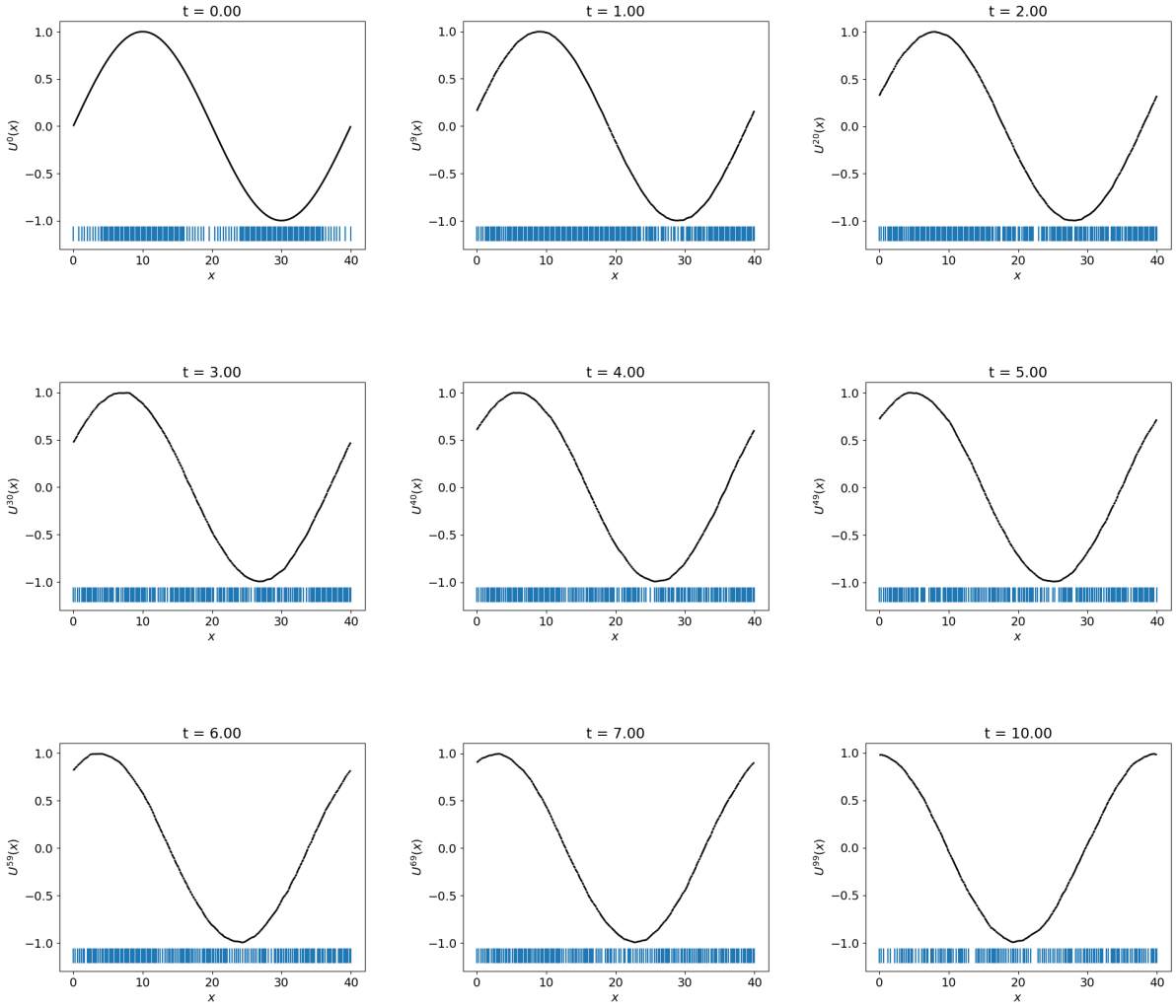
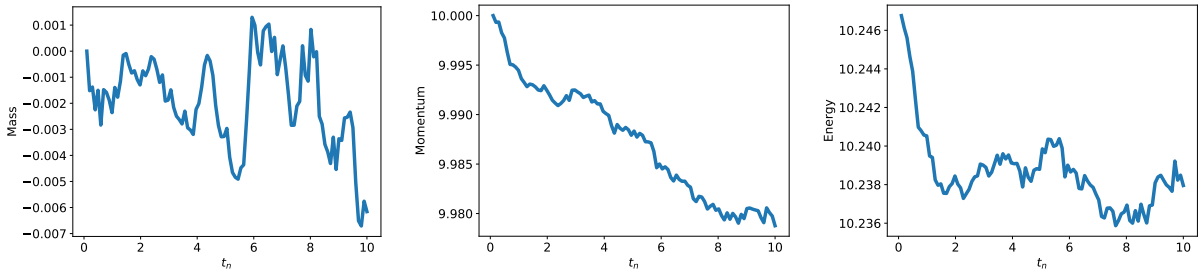
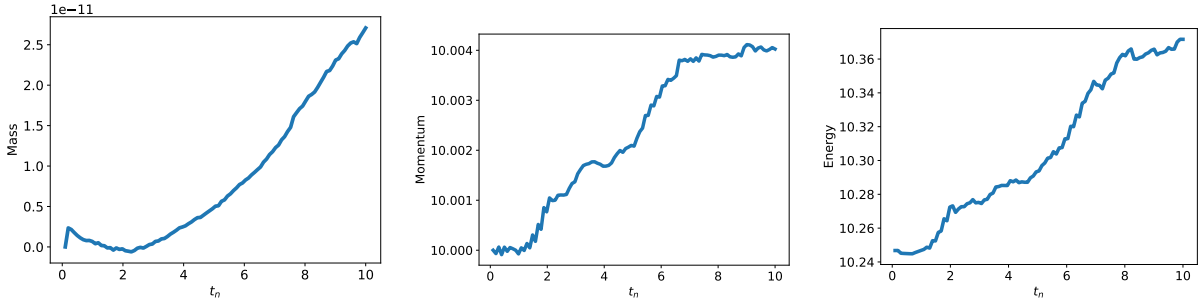


Figure 7.6: Here we examine the values of the invariants (mass, momentum and energy) at the temporal nodes in the adaptive algorithm (7.23) for $f(u) = \frac{1}{2}u^2$ with the *Lagrange interpolant* (6.2) as the mesh change operator. We initialise the simulation with the L_2 projection of (7.4) into the initial finite element space and employ the adaptive parameters (7.28). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We observe that the mass deviates non-monotonically, however the momentum and energy dissipate globally.



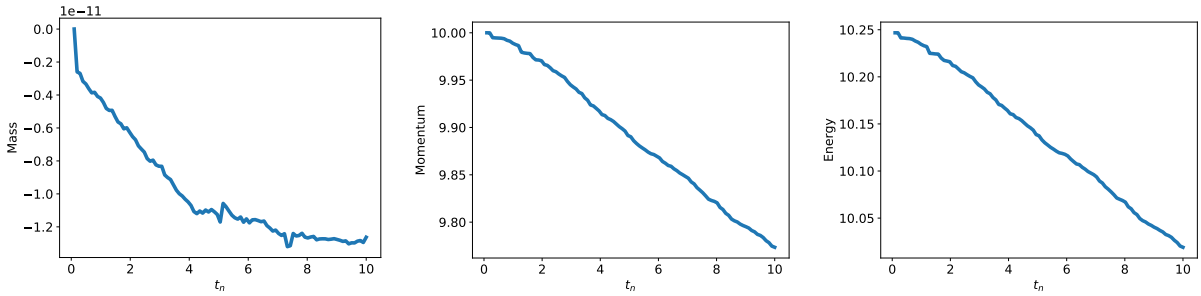
(a) The mass. (b) The momentum. (c) The energy.

Figure 7.7: Here we examine the values of the invariants (mass, momentum and energy) at the temporal nodes in the adaptive algorithm (7.23) for $f(u) = \frac{1}{2}u^2$ with the L_2 *projection* (6.3) as the mesh change operator. We initialise the simulation with the L_2 projection of (7.4) into the initial finite element space and employ the adaptive parameters (7.28). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We observe that the mass in conserved, however the energy increases globally indicating an instability in the algorithm.



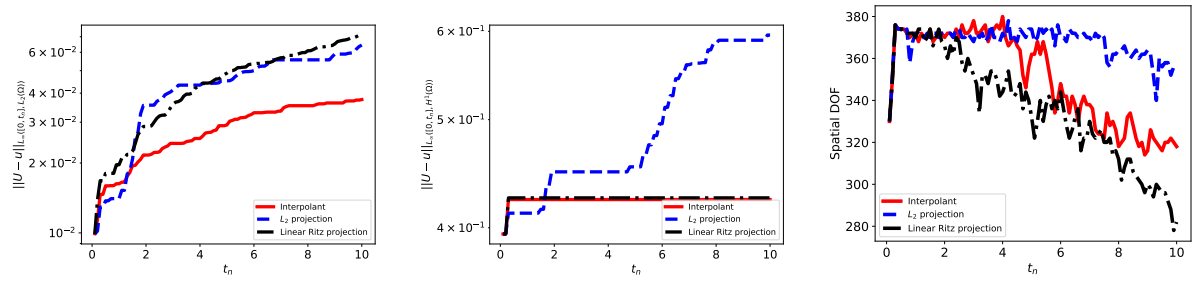
(a) The mass. (b) The momentum. (c) The energy.

Figure 7.8: Here we examine the values of the invariants (mass, momentum and energy) at the temporal nodes in the adaptive algorithm (7.23) for $f(u) = \frac{1}{2}u^2$ with the *Ritz projection* (7.24) as the mesh change operator. We initialise the simulation with the L_2 projection of (7.4) into the initial finite element space and employ the adaptive parameters (7.28). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We observe that the mass is conserved, and the momentum and energy monotonically dissipate.



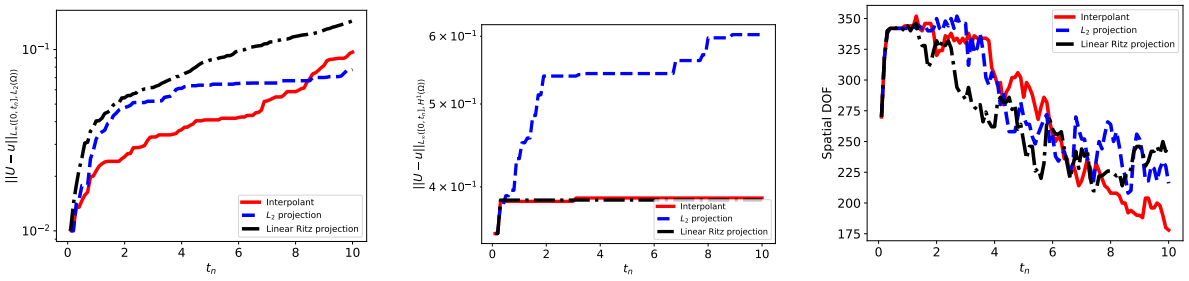
(a) The mass. (b) The momentum. (c) The energy.

Figure 7.9: Here we examine the error measured both the L_2 and dG norm (5.18) for the adaptive algorithm (7.23) with the Lagrange interpolant (6.2), the L_2 projection (6.3) or the Ritz projection (7.24) as the mesh change operator. We plot with the error the number of degrees of freedom used in each simulation on each time step. The simulations are initialised by L_2 projection of (7.4) into the initial finite element space and employ the adaptive parameters (7.28). Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We observe that all simulations use a comparable number of degrees of freedom with the L_2 projection using slightly more. In the L_2 norm for this simulation the Lagrange interpolant has the smallest error. In the dG norm the Ritz projector and interpolant behave comparably, although for the Ritz projector fewer degrees of freedom are used.



(a) The error in L_2. (b) The error the dG norm (5.18). (c) The number of degrees of freedom.

Figure 7.10: Here we examine the error measured both the L_2 and dG norm (5.18) for the adaptive algorithm (7.23) with the Lagrange interpolant (6.2), the L_2 projection (6.3) or the Ritz projection (7.24) as the mesh change operator. We display alongside the error the number of degrees of freedom used in each simulation on each time step. The simulations are initialised by L_2 projection of (7.4) into the initial finite element space and employ the adaptive parameters (7.28) with `coarsen` $= 30$. Further to this we consider the uniform time step $\tau_n = 0.1$, and initialise our spatial mesh as a uniform mesh with the spatial element size $h_m = 0.4$. We observe that all simulations use a comparable number of degrees of freedom, the Ritz projector marginally outperforms the Lagrange interpolant in the energy norm. In the L_2 norm the Lagrange interpolant and L_2 projector behave comparably and outperform the Ritz projector. We also notice that the error in the L_2 projection blows up in the dG norm.



(a) The error in L_2. (b) The error the dG norm (5.18). (c) The number of degrees of freedom.

7.6 Conclusion

In this chapter we designed a new numerical scheme for a general dissipative KdV type equation. We proved optimal a priori and a posteriori error bounds in the spatially discrete linear case. We also proposed a fully discrete adaptive algorithm, and designed a mesh change operator which guarantees numerical stability of this adaptive algorithm. We have left the study of the nonlinear case for future work.

Chapter 8

A conservative discretisation for the vectorial modified KdV equation

The work in this chapter has been published, see [107], and was conducted in collaboration with Georgios Papamikos during his time at the University of Reading.

We design a consistent Galerkin scheme for the approximation of the vectorial modified Korteweg-de Vries *(vmKdV)* equation with periodic boundary conditions. We demonstrate that the scheme conserves energy up to solver tolerance. In this sense the method is consistent with the energy balance of the continuous system. This energy balance ensures there is no numerical dissipation allowing for extremely accurate long time simulations free from numerical artefacts. Various numerical experiments are shown demonstrating the asymptotic convergence of the method with respect to the discretisation parameters. Some simulations are also presented that correctly capture the unusual interactions between solitons in the vectorial setting.

To the best of the author's knowledge, the discretisation discussed here is the *first* designed to conserve the energy of the vmKdV equation. Similarly to previous chapters, as the vmKdV falls within the framework of Hamiltonian PDEs, we employ the methodology discussed in §4.1 for the discretisation of Hamiltonian operators. However, the vectorial case possesses a significantly different structure to the scalar Hamiltonian PDEs in the prequel. For example, there is no mass conservation in the vectorial case. Moreover, the corresponding Hamiltonian operator is a nonlocal differential operator in the vectorial case which presents additional complexity in its discrete representation.

8.1 The continuous problem

In this section we formulate the model problem, fix notation and give some basic assumptions. We describe some known results and history of the vmKdV equation, highlighting the Hamiltonian structure of the equation. We show that the underlying Hamiltonian structure naturally yields an induced stability of the solutions to the PDE system and give a brief description of how to construct some exact solutions using a dressing method. We then show how the system can be written through induced auxiliary variables which are the basis of the design of our numerical scheme.

Recall that throughout we denote the standard Lebesgue spaces by $L_p(I)$ for $I \subseteq \mathbb{R}$, $p \in [1, \infty]$, equipped with corresponding norms $\|u\|_{L_p(I)}$. In addition, we denote $H^k(I)$ to be the Hilbert Sobolev space of order k of real-valued functions defined over $I \subseteq \mathbb{R}$ with norm $\|u\|_{H^k(I)}$.

The vmKdV equation is an evolutionary PDE for a real, D-vector valued function

$$\begin{array}{rccl} \boldsymbol{u}: & \mathbb{R}^2 & \to & \mathbb{R}^D \\ & (x,t) & \mapsto & \boldsymbol{u}(x,t) = (u_1, \ldots, u_d)^{\mathrm{T}} \end{array}$$

and is given by

$$\boldsymbol{u}_t + \frac{3}{2}\boldsymbol{u} \cdot \boldsymbol{u}\boldsymbol{u}_x + \boldsymbol{u}_{xxx} = \boldsymbol{0}. \tag{8.1}$$

Here we are using "$\boldsymbol{x} \cdot \boldsymbol{y}$" as the Euclidean inner product between two vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$. In the sequel we shall also write "$|\boldsymbol{x}|$" as the induced Euclidean norm of $\boldsymbol{x}$.

A particular case of the vmKdV system occurs when $d = 2$, $\boldsymbol{u} = (u_1, u_2)^{\mathrm{T}}$ when (8.1) can be identified with the complex modified KdV *(mKdV)* equation

$$y_t + \frac{3}{2}|y|^2 y_x + y_{xxx} = 0$$

for the complex dependent variable $y = u_1 + iu_2$. Sometimes this is also called the Hirota mKdV equation [92]. When $d = 1$ we obtain the famous mKdV equation which has been studied numerically in the context of Galerkin methods in [31] and Remark 5.3.11.

Equation (8.1) admits both Lie and discrete point symmetries and has infinitely many conservation laws. Indeed, under the action of the orthogonal group $O_d(\mathbb{R})$, which is the

group of real $d \times d$ matrices such that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I}$,

$$\widetilde{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{u}, \quad \text{for } \boldsymbol{A} \in O_d(\mathbb{R})$$

equation (8.1) remains invariant. Moreover, vmKdV is invariant under the translations

$$\widetilde{x} = x + \epsilon, \quad \widetilde{t} = t + \gamma$$

and under the scaling transformation

$$\left(\widetilde{x}, \widetilde{t}, \widetilde{\boldsymbol{u}}\right) = \left(e^{\epsilon}x, e^{3\epsilon}t, e^{-\epsilon}\boldsymbol{u}\right).$$

Proposition 8.1.1 (Conservative properties of solutions)**.** *The vmKdV equation admits the following conservation laws:*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}f_2(\boldsymbol{u}) &= \frac{\mathrm{d}}{\mathrm{d}x}g_2(\boldsymbol{u})\\ \frac{\mathrm{d}}{\mathrm{d}t}f_4(\boldsymbol{u}) &= \frac{\mathrm{d}}{\mathrm{d}x}g_4(\boldsymbol{u}),\end{aligned}$$

with $\frac{\mathrm{d}}{\mathrm{d}t}$ *and* $\frac{\mathrm{d}}{\mathrm{d}x}$ *representing the total derivatives with respect to* t *and* x *respectively, where the conserved densities are given by*

$$\begin{aligned}f_2(\boldsymbol{u}) &= \frac{1}{2}\left|\boldsymbol{u}\right|^2\\ f_4(\boldsymbol{u}) &= \frac{1}{2}\left|\boldsymbol{u}_x\right|^2 - \frac{1}{8}\left|\boldsymbol{u}\right|^4\end{aligned}$$

and the corresponding fluxes are

$$\begin{aligned}g_2(\boldsymbol{u}) &= \left|\boldsymbol{u}_x\right|^2 - 2\boldsymbol{u}\cdot\boldsymbol{u}_{xx} - \frac{3}{4}\left|\boldsymbol{u}\right|^4\\ g_4(\boldsymbol{u}) &= \frac{1}{8}\left|\boldsymbol{u}\right|^6 - \left|\boldsymbol{u}\right|^2\left|\boldsymbol{u}_x\right|^2 - \frac{1}{2}\left(\boldsymbol{u}\cdot\boldsymbol{u}_x\right)^2 - \boldsymbol{u}_x\cdot\boldsymbol{u}_{xxx} + \frac{1}{2}\left|\boldsymbol{u}_{xx}\right|^2 + \frac{1}{2}\left|\boldsymbol{u}\right|^2\boldsymbol{u}\cdot\boldsymbol{u}_{xx}.\end{aligned}$$

Proof. To prove that the total time derivative of $f_2(\mathbf{u})$ and $f_4(\mathbf{u})$ are in the image of $\frac{\mathrm{d}}{\mathrm{d}x}$ we use the Euler operator $\boldsymbol{E} = (E_1, ..., E_d)$, where

$$E_i(f) = \sum_{k=0}^{\infty}\left(-\frac{\mathrm{d}}{\mathrm{d}x}\right)^k \partial_{u_{i_{kx}}} f, \quad u_{i_{kx}} = u_{i\underbrace{x\ldots x}_{k}}$$

and the fact that $\mathrm{Ker}\boldsymbol{E} = \mathrm{Im}\frac{\mathrm{d}}{\mathrm{d}x}$, see [147] for a proof. On the other hand in order to calculate the corresponding fluxes g_2 and g_4 we apply the homotopy operator [147] to $\frac{\mathrm{d}}{\mathrm{d}t}f_2(\mathbf{u})$ and $\frac{\mathrm{d}}{\mathrm{d}t}f_4(\mathbf{u})$ respectively. The homotopy operator is given by

$$\mathtt{H}(f(\mathbf{u})) = \int_0^1 \sum_{i=1}^{d} \mathrm{I}_i(f)(\lambda\mathbf{u})\frac{d\lambda}{\lambda}$$

where

$$\mathrm{I}_i(f) = \sum_{k=1}^{\infty}\left(\sum_{s=0}^{k-1} u_{i_{sx}}\left(-\frac{\mathrm{d}}{\mathrm{d}x}\right)^{k-s-1}\right) f_{u_{i_{kx}}}.$$

□

Corollary 8.1.2 (Conservative properties of solutions on S^1.)**.** *Let S^1 be the unitary circle, i.e., $[0,1]$ with matching endpoints and recall that $\langle\cdot,\cdot\rangle$ denotes the spatial L_2 inner product over S^1. Then from Proposition 8.1.1 it follows that, upon defining*

$$F_2(\boldsymbol{u}) := \langle f_2(\boldsymbol{u}), 1\rangle$$

as the momentum *functional and*

$$F_4(\boldsymbol{u}) := \langle f_4(\boldsymbol{u}), 1\rangle$$

as the energy *functional for periodic solutions, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}F_2(\boldsymbol{u}) = \frac{\mathrm{d}}{\mathrm{d}t}F_4(\boldsymbol{u}) = 0.$$

Moreover this does not just hold for periodic solutions over S^1. Indeed, one can consider the equation (8.1) *over $\mathbb{R}$ and require that solutions decay at infinity, for example Schwartz functions, and the result holds. A particular example of such solutions are the much celebrated soliton and breather solutions. Note that we refer to $F_2(\boldsymbol{u})$ as the momentum as it is associated with the space translation Lie symmetry, and similarly the energy $F_4(\boldsymbol{u})$ is associated with the time translation Lie symmetry.*

Proposition 8.1.3 (A continuous stability bound)**.** *Let the vmKdV system* (8.1)*, defined over S^1, be coupled with initial conditions $\boldsymbol{u}_0$ satisfying $F_2(\boldsymbol{u}_0) = C_2 < \infty$ and $F_4(\boldsymbol{u}_0) =$*

$C_4 < \infty$ then $\boldsymbol{u}$ satisfies

$$\|\boldsymbol{u}_x(t)\|_{L_2(S^1)} \leq \left(4C_4 + \frac{C_{GN}^8 C_2^3}{2}\right)^{1/2},$$

where C_{GN} is a constant appearing from the Gagliardo-Nirenberg interpolation inequality.

Proof. In view of the definition of $F_4(\boldsymbol{u})$ we have that

$$\begin{aligned} \|\boldsymbol{u}_x\|_{L_2(S^1)}^2 &= 2F_4(\boldsymbol{u}) + \frac{1}{4}\|\boldsymbol{u}\|_{L_4(S^1)}^4 \\ &= 2F_4(\boldsymbol{u}_0) + \frac{1}{4}\|\boldsymbol{u}\|_{L_4(S^1)}^4, \end{aligned} \tag{8.2}$$

through the conservativity of $F_4(\boldsymbol{u})$ given in Theorem 8.1.1. Now making use of the Gagliardo-Nirenberg interpolation inequality there exists a constant C_{GN} such that

$$\|\boldsymbol{u}\|_{L_4(S^1)} \leq C_{GN}\|\boldsymbol{u}\|_{L_2(S^1)}^{3/4}\|\boldsymbol{u}_x\|_{L_2(S^1)}^{1/4},$$

hence

$$\begin{aligned} \frac{1}{4}\|\boldsymbol{u}\|_{L_4(S^1)}^4 &\leq \frac{1}{4}C_{GN}^4\|\boldsymbol{u}\|_{L_2(S^1)}^3\|\boldsymbol{u}_x\|_{L_2(S^1)} \\ &\leq \frac{1}{32}C_{GN}^8\|\boldsymbol{u}\|_{L_2(S^1)}^6 + \frac{1}{2}\|\boldsymbol{u}_x\|_{L_2(S^1)}^2, \end{aligned} \tag{8.3}$$

through Young's inequality. Substituting (8.3) into (8.2) we see

$$\begin{aligned} \frac{1}{2}\|\boldsymbol{u}_x\|_{L_2(S^1)}^2 &\leq 2F_4(\boldsymbol{u}_0) + \frac{C_{GN}^8}{32}\|\boldsymbol{u}\|_{L_2(S^1)}^6 \\ &\leq 2F_4(\boldsymbol{u}_0) + \frac{C_{GN}^8}{4}F_2(\boldsymbol{u})^3 \\ &\leq 2F_4(\boldsymbol{u}_0) + \frac{C_{GN}^8}{4}F_2(\boldsymbol{u}_0)^3 \\ &\leq 2C_4 + \frac{C_{GN}^8 C_2^3}{4}, \end{aligned}$$

using the conservativity of $F_2(\boldsymbol{u})$, concluding the proof. □

Remark 8.1.4 (Hierarchy of conservation laws)**.** *Note that the vmKdV equation* (8.1) *admits an infinite hierarchy of conserved quantities. For example, after $F_2(\boldsymbol{u})$ and $F_4(\boldsymbol{u})$*

the next member of the hierarchy is

$$F_6(\boldsymbol{u}) = \frac{1}{2}\left\langle |\boldsymbol{u}|^3, |\boldsymbol{u}|^3 \right\rangle + 10\left\langle \boldsymbol{u}\cdot\boldsymbol{u}_x, \boldsymbol{u}\cdot\boldsymbol{u}_x \right\rangle + \left\langle |\boldsymbol{u}|^2, |\boldsymbol{u}_x|^2 \right\rangle + 7\left\langle |\boldsymbol{u}|^2, \boldsymbol{u}\cdot\boldsymbol{u}_{xx} \right\rangle + 4\left\langle |\boldsymbol{u}_{xx}|, |\boldsymbol{u}_{xx}| \right\rangle.$$

A generating function of the conserved densities for the vmKdV is constructed using its Lax representation in [4].
Together with the Gagliardo-Nirenberg interpolation inequality one may derive a priori bounds of a similar form to that given in Theorem 8.1.3 but in higher order norms. Indeed, for $s \in \mathbb{N}$ the conservation law F_{2s} naturally gives rise to a stability bound in H^{s-1}.

8.1.1 Exact solutions to the vmKdV system

The vmKdV equation (8.1) is integrable and has already drawn some attention [4, 10]. Its integrability properties were derived using the structure equation for the evolution of a curve embedded in an n-dimensional Riemannian manifold with constant curvature [158, 9, 134]. The associated Cauchy problem can be studied analytically using the inverse scattering transform [2, 146, 70]. As it admits a zero curvature representation (or a Lax representation, see [125]), i.e., it can be written in the following form:

$$U_t - V_x + [U, V] = 0,$$

where $U = U(\boldsymbol{u}; \lambda)$ and $V = V(\boldsymbol{u}; \lambda)$ are appropriate matrices in a Lie algebra having a polynomial dependence on a spectral parameter $\lambda \in \mathbb{C}$. One can construct, see [4], a Darboux matrix M [154, 156] that maps the pair (U, V) to

$$(U, V) \mapsto (\widetilde{U}, \widetilde{V}) = (MUM^{-1} + M_x M^{-1}, MUM^{-1} + M_t M^{-1}) \tag{8.4}$$

and $\widetilde{U} = U(\widetilde{\boldsymbol{u}}; \lambda)$ and $\widetilde{V} = V(\widetilde{\boldsymbol{u}}; \lambda)$. In other words $\widetilde{U}$ and $\widetilde{V}$ have the same structure as U and V respectively. The transformation (8.4) implies a nonlocal symmetry $\boldsymbol{u} \mapsto \widetilde{\boldsymbol{u}}$ of the vmKdV, known as a Bäcklund transformation. Such transformations that have applications in geometry [154] are characteristic of integrable equations. Starting with the trivial background solution $\boldsymbol{u} = \boldsymbol{0}$ one can then recursively and algebraically construct the soliton solutions of vmKdV equation (8.1). For example, when $d = 2$ a 1-soliton solution

is given by

$$\boldsymbol{u} = \frac{2\mu}{\cosh(\xi_\mu)}\boldsymbol{E}, \tag{8.5}$$

where $\mu \in \mathbb{R}$, $\xi_\mu = \mu(x - c_\mu) - \mu^3 t$, for some shift $c_\mu \in \mathbb{R}$ and $\boldsymbol{E}$ is a constant unit vector. A 2-soliton solution is given by

$$\boldsymbol{u} = \frac{F_{\mu,\nu}}{G}\boldsymbol{E}_1 + \frac{F_{\nu,\mu}}{G}\boldsymbol{E}_2, \tag{8.6}$$

where $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are constant unit vectors, $\mu,\ \nu \in \mathbb{R}$ with $\mu \neq \pm\nu$ and

$$F_{k,l} = 2(l^2 - k^2)l\cosh(\xi_k) \tag{8.7}$$

and

$$G = (\mu^2 + \nu^2)\cosh(\xi_\mu)\cosh(\xi_\nu) - 2\mu\nu\sinh(\xi_\mu)\sinh(\xi_\nu) - 2\mu\nu\boldsymbol{E}_1 \cdot \boldsymbol{E}_2. \tag{8.8}$$

The 1-soliton (8.5) and 2-soliton (8.6) solutions, while elegant, are not the most general of their kind, see [4] for details. Nevertheless, the exact solutions (8.5) and (8.6) are both perfectly adequate for benchmarking our scheme which we shall use them for in §8.4. Such solutions can also be derived using Hirota's bilinear form [93, 10].

Solitons are, however, a special class of solution for this problem with a very particular structure. In general one cannot write down closed form solutions for this problem motivating the need for long time accurate numerical schemes. We shall proceed by describing the Hamiltonian structure of the vmKdV problem which forms the basis for the design of our numerical scheme.

Remark 8.1.5 (Hamiltonian formulation of vmKdV)**.** *The vmKdV system is Hamiltonian and thus it can be written as*

$$\boldsymbol{u}_t = \mathcal{P}(\boldsymbol{u})\frac{\delta F_4(\boldsymbol{u})}{\delta \boldsymbol{u}},$$

where $\mathcal{P}(\boldsymbol{u})$ is a Hamiltonian operator, $F_4(\boldsymbol{u})$ is the corresponding Hamiltonian and $\frac{\delta \cdot}{\delta \boldsymbol{u}}$ denotes the first variation with respect to $\boldsymbol{u}$, as discussed in Chapter 4 for scalar problems. For this specific problem the Hamiltonian operator acts on a real, d-vector function $\boldsymbol{y}$ and takes the form

$$\mathcal{P}(\boldsymbol{u})\boldsymbol{y} := \boldsymbol{y}_x - \boldsymbol{u}\lrcorner\left[\frac{\mathrm{d}}{\mathrm{d}x}^{-1}(\boldsymbol{y}\otimes\boldsymbol{u} - \boldsymbol{u}\otimes\boldsymbol{y})\right], \tag{8.9}$$

as described in [9], where $\frac{\mathrm{d}}{\mathrm{d}x}^{-1}$ is the formal inverse operator of $\frac{\mathrm{d}}{\mathrm{d}x}$, $\otimes$ is the tensor product

between vectors and $\lrcorner$ *is an interior product defined through*

$$\boldsymbol{x} \lrcorner (\boldsymbol{y} \otimes \boldsymbol{z}) = (\boldsymbol{x} \cdot \boldsymbol{y})\, \boldsymbol{z}.$$

This then induces a Poisson bracket

$$\{F, G\} := \left\langle \frac{\delta F}{\delta u}, \mathcal{P}(\boldsymbol{u}) \frac{\delta G}{\delta u} \right\rangle,$$

a skew-symmetric bilinear form satisfying the Jacobi identity. In view of the skew-symmetry of $\mathcal{P}(\boldsymbol{u})$ *we have*

$$\frac{\mathrm{d}}{\mathrm{d}t} F_4(\boldsymbol{u}) = \{F_4(\boldsymbol{u}), F_4(\boldsymbol{u})\} = 0.$$

Notice also that the vmKdV system can also be written as

$$\boldsymbol{u}_t = \{\boldsymbol{u}, F_4(\boldsymbol{u})\}.$$

The main idea behind the discretisation we propose is to correctly represent the Hamiltonian operator in the finite element space whilst preserving the skew-symmetry property of the underlying bracket. *Indeed, the proof of Proposition 8.1.1 motivates rewriting the vmKdV system by introducing auxiliary variables to represent different components of the Hamiltonian operator. We consider seeking the tuple* $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ *such that*

$$\begin{aligned} \mathbf{0} &= \boldsymbol{u}_t + \boldsymbol{v}_x + \boldsymbol{w} \\ \mathbf{0} &= \boldsymbol{v} - \frac{1}{2} |\boldsymbol{u}|^2 \boldsymbol{u} - \boldsymbol{u}_{xx} \\ \mathbf{0} &= \boldsymbol{w} - |\boldsymbol{u}|^2 \boldsymbol{u}_x + (\boldsymbol{u}_x \cdot \boldsymbol{u})\, \boldsymbol{u}. \end{aligned} \tag{8.10}$$

Notice that $\boldsymbol{v} = \frac{\delta F_4(\boldsymbol{u})}{\delta u}$ *and* $\boldsymbol{w} = \boldsymbol{u} \lrcorner \left[\frac{\mathrm{d}}{\mathrm{d}x}^{-1} (\boldsymbol{v} \otimes \boldsymbol{u} - \boldsymbol{u} \otimes \boldsymbol{v}) \right]$. *This form of* $\boldsymbol{w}$ *is extremely important as the Hamiltonian operator given in* (8.9) *is nonlocal. The fact that it can be "localised" by removing the* $\frac{\mathrm{d}}{\mathrm{d}x}^{-1}$ *allows for the efficient approximation by Galerkin methods.*

This reformulation also means that in the case both arguments of the Poisson bracket are the Hamiltonian we may write

$$0 = \{F_4(\boldsymbol{u}), F_4(\boldsymbol{u})\} = \langle \boldsymbol{v}, \boldsymbol{v}_x + \boldsymbol{w} \rangle\,. \tag{8.11}$$

It is exactly this structure that we try to exploit.

Remark 8.1.6 (Relation to the scalar case)**.** *As already mentioned when $d = 1$, the problem reduces to the mKdV equation. In this case $w \equiv 0$ and the mixed system coincides with that discussed in Remark 5.3.11. Energy conservative schemes can be derived, and a priori bounds can be proven for a* linearised *version of the problem, see Chapter 4. For $d > 1$, $\boldsymbol{w}$ is not necessarily zero and represents the additional contribution arising from the Hamiltonian operator described in Remark 8.1.5.*

Proposition 8.1.7 (The mixed system is conservative)**.** *Let $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ be given by* (8.10) *then we have that*

$$\frac{\mathrm{d}}{\mathrm{d}t} F_4(\boldsymbol{u}) = \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{2} \langle \boldsymbol{u}_x, \boldsymbol{u}_x \rangle - \frac{1}{8} \left\langle |\boldsymbol{u}|^2, |\boldsymbol{u}|^2 \right\rangle \right) = 0.$$

Proof. Since the mixed system is equivalent to the vmKdV system the proof is clear through Proposition 8.1.1, however, for illustrative purposes we present it in full as it will become the basis for the design of our numerical scheme. To begin note

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F_4(\boldsymbol{u}) &= \langle \boldsymbol{u}_x, \boldsymbol{u}_{xt} \rangle - \left\langle \frac{1}{2} |\boldsymbol{u}|^2 \boldsymbol{u}, \boldsymbol{u}_t \right\rangle \\ &= - \langle \boldsymbol{u}_{xx}, \boldsymbol{u}_t \rangle - \frac{1}{2} \left\langle |\boldsymbol{u}|^2 \boldsymbol{u}, \boldsymbol{u}_t \right\rangle \\ &= - \langle \boldsymbol{v}, \boldsymbol{u}_t \rangle . \end{aligned}$$

Now making use of (8.10)

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F_4(\boldsymbol{u}) &= \langle \boldsymbol{v}, \boldsymbol{v}_x + \boldsymbol{w} \rangle \\ &= \langle \boldsymbol{v}, \boldsymbol{w} \rangle \\ &= \left\langle \frac{1}{2} |\boldsymbol{u}|^2 \boldsymbol{u} + \boldsymbol{u}_{xx}, \boldsymbol{w} \right\rangle . \end{aligned}$$

Note that from (8.10) we can see that $\boldsymbol{w} \cdot \boldsymbol{u} = 0$ and hence

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F_4(\boldsymbol{u}) &= \langle \boldsymbol{u}_{xx}, \boldsymbol{w} \rangle \\ &= \left\langle \boldsymbol{u}_{xx}, |\boldsymbol{u}|^2 \boldsymbol{u}_x - (\boldsymbol{u}_x \cdot \boldsymbol{u}) \boldsymbol{u} \right\rangle . \end{aligned} \tag{8.12}$$

Now, through an integration by parts we have

$$\begin{aligned}\left\langle |\boldsymbol{u}|^2 \boldsymbol{u}_x, \boldsymbol{u}_{xx} \right\rangle &= -\left\langle \left(|\boldsymbol{u}|^2 \boldsymbol{u}_x \right)_x, \boldsymbol{u}_x \right\rangle \\ &= -2 \left\langle \boldsymbol{u} \cdot \boldsymbol{u}_x, \boldsymbol{u}_x \cdot \boldsymbol{u}_x \right\rangle + \left\langle |\boldsymbol{u}|^2 \boldsymbol{u}_{xx}, \boldsymbol{u}_x \right\rangle\end{aligned}$$

and hence

$$\left\langle |\boldsymbol{u}|^2 \boldsymbol{u}_x, \boldsymbol{u}_{xx} \right\rangle = -\left\langle \boldsymbol{u} \cdot \boldsymbol{u}_x, \boldsymbol{u}_x \cdot \boldsymbol{u}_x \right\rangle. \tag{8.13}$$

In addition,

$$\begin{aligned}\left\langle \boldsymbol{u}_x \cdot \boldsymbol{u}, \boldsymbol{u} \cdot \boldsymbol{u}_{xx} \right\rangle &= -\left\langle \left(\boldsymbol{u}_x \cdot \boldsymbol{u}\boldsymbol{u} \right)_x, \boldsymbol{u}_x \right\rangle \\ &= -\left\langle \boldsymbol{u}_{xx} \cdot \boldsymbol{u}, \boldsymbol{u} \cdot \boldsymbol{u}_x \right\rangle - 2 \left\langle \boldsymbol{u}_x \cdot \boldsymbol{u}_x, \boldsymbol{u} \cdot \boldsymbol{u}_x \right\rangle\end{aligned}$$

and hence

$$\left\langle \boldsymbol{u}_x \cdot \boldsymbol{u}, \boldsymbol{u} \cdot \boldsymbol{u}_{xx} \right\rangle = -\left\langle \boldsymbol{u}_x \cdot \boldsymbol{u}_x, \boldsymbol{u} \cdot \boldsymbol{u}_x \right\rangle. \tag{8.14}$$

Substituting (8.13) and (8.14) into (8.12) concludes the proof. □

8.2 Temporal discretisation

For the reader's convenience we will present an argument for designing the temporally discrete scheme in the spatially continuous setting. As in the prequel, we consider a time interval $[0, T]$ subdivided into a partition of N consecutive adjacent subintervals whose endpoints are denoted $t_0 = 0 < t_1 < \ldots < t_N = T$. The n-th timestep is defined as $\tau_n := t_{n+1} - t_n$. We will consistently refer to temporally discrete functions through superscripts, i.e., $y^n(x)$ is the temporally discrete approximation of $y(t, x)$ at $t = t_n$. We also denote $y^{n+\frac{1}{2}} := \frac{1}{2}(y^n + y^{n+1})$.

We consider the temporal discretisation of the mixed system (8.10) as follows: Given $\boldsymbol{u}^0$, for $n \in [0, N]$ find $\boldsymbol{u}^{n+1}$ such that

$$\begin{aligned}\frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^n}{\tau_n} + \boldsymbol{v}_x^{n+1} + \boldsymbol{w}^{n+1} &= \boldsymbol{0} \\ \boldsymbol{v}^{n+1} - \frac{1}{2}\left(|\boldsymbol{u}^n|^2 + \left|\boldsymbol{u}^{n+1}\right|^2 \right) \boldsymbol{u}^{n+\frac{1}{2}} - \boldsymbol{u}_{xx}^{n+\frac{1}{2}} &= \boldsymbol{0} \\ \boldsymbol{w}^{n+1} - \left|\boldsymbol{u}^{n+\frac{1}{2}}\right|^2 \boldsymbol{u}_x^{n+\frac{1}{2}} + \left(\boldsymbol{u}_x^{n+\frac{1}{2}} \cdot \boldsymbol{u}^{n+\frac{1}{2}} \right) \boldsymbol{u}^{n+\frac{1}{2}} &= \boldsymbol{0}.\end{aligned} \tag{8.15}$$

Remark 8.2.1 (Structure of the temporal discretisation)**.** *The temporal discretisation*

given in (8.15) *is* not *a Runge-Kutta method. It resembles a Crank-Nicolson discretisation, however the treatment of the nonlinearity is different. This treatment is similar to that conducted in our temporal discretisations in Chapter 5. It is formally of second order and is constructed such that it satisfies the next theorem. Although construction of higher order methods is possible they become very complicated to write down so we will not press this point here.*
Notice that the diagnostic variables $\boldsymbol{v}^{n+1}$ *and* $\boldsymbol{w}^{n+1}$ *are* not *evaluated at the midpoint, suggesting our temporal discretisation does not resemble Crank-Nicolson. This is not the case. In fact, this does not effect our temporal discretisation due to the variables' diagnostic nature. Through eliminating the diagnostic variables we observe a temporal discretisation which is a* second order *perturbation of Crank-Nicolson.*

Theorem 8.2.2 (Conservativity of the temporal discretisation)**.** *Let* $\{\boldsymbol{u}^n\}_{n=0}^N$ *be a temporally discrete solution of* (8.15) *then we have*

$$F_4(\boldsymbol{u}^n) = F_4(\boldsymbol{u}^0) \quad \forall n \in [0, N].$$

Proof. It suffices to show that

$$F_4(\boldsymbol{u}^{n+1}) - F_4(\boldsymbol{u}^n) = 0$$

and then the result follows inductively. So

$$\begin{aligned}
2\left(F_4(\boldsymbol{u}^{n+1}) - F_4(\boldsymbol{u}^n)\right) &= \left\langle \boldsymbol{u}_x^{n+1}, \boldsymbol{u}_x^{n+1} \right\rangle - \langle \boldsymbol{u}_x^n, \boldsymbol{u}_x^n \rangle \\
&\quad - \frac{1}{4}\left\langle \left|\boldsymbol{u}^{n+1}\right|^2, \left|\boldsymbol{u}^{n+1}\right|^2 \right\rangle + \frac{1}{4}\left\langle |\boldsymbol{u}^n|^2, |\boldsymbol{u}^n|^2 \right\rangle \\
&= \left\langle \boldsymbol{u}_x^{n+1} - \boldsymbol{u}_x^n, \boldsymbol{u}_x^{n+1} + \boldsymbol{u}_x^n \right\rangle \\
&\quad - \frac{1}{4}\left\langle \boldsymbol{u}^{n+1} - \boldsymbol{u}^n, \left|\boldsymbol{u}^{n+1}\right|^2 \boldsymbol{u}^{n+1} + \left|\boldsymbol{u}^{n+1}\right|^2 \boldsymbol{u}^n \right\rangle \\
&\quad - \frac{1}{4}\left\langle \boldsymbol{u}^{n+1} - \boldsymbol{u}^n, |\boldsymbol{u}^n|^2 \boldsymbol{u}^{n+1} + |\boldsymbol{u}^n|^2 \boldsymbol{u}^n \right\rangle \\
&= -\left\langle \left(\boldsymbol{u}^{n+1} - \boldsymbol{u}^n\right), \boldsymbol{u}_{xx}^{n+1} + \boldsymbol{u}_{xx}^n \right\rangle \\
&\quad - \frac{1}{2}\left\langle \boldsymbol{u}^{n+1} - \boldsymbol{u}^n, \left|\boldsymbol{u}^{n+1}\right|^2 \boldsymbol{u}^{n+\frac{1}{2}} + |\boldsymbol{u}^n|^2 \boldsymbol{u}^{n+\frac{1}{2}} \right\rangle \\
&= -\left\langle \boldsymbol{u}^{n+1} - \boldsymbol{u}^n, \boldsymbol{v}^{n+1} \right\rangle,
\end{aligned}$$

through expanding differences, integrating by parts and using the scheme (8.15). Now,

again using the scheme

$$\begin{aligned} 2\left(F_4(\boldsymbol{u}^{n+1}) - F_4(\boldsymbol{u}^n)\right) &= \tau_n \left\langle \boldsymbol{v}_x^{n+1} + \boldsymbol{w}^{n+1}, \boldsymbol{v}^{n+1} \right\rangle \\ &= \tau_n \left\langle \boldsymbol{w}^{n+1}, \boldsymbol{v}^{n+1} \right\rangle \\ &= \tau_n \left\langle \boldsymbol{w}^{n+1}, \frac{1}{2}\left(\left|\boldsymbol{u}^{n+1}\right|^2 \boldsymbol{u}^{n+\frac{1}{2}} + |\boldsymbol{u}^n|^2 \boldsymbol{u}^{n+\frac{1}{2}} \right) + \boldsymbol{u}_{xx}^{n+\frac{1}{2}} \right\rangle \\ &= \tau_n \left\langle \boldsymbol{w}^{n+1}, \boldsymbol{u}_{xx}^{n+\frac{1}{2}} \right\rangle, \end{aligned}$$

in view of the orthogonality condition $\boldsymbol{w}^{n+1} \cdot \boldsymbol{u}^{n+\frac{1}{2}} = 0$ following from the third equation of (8.15). Now we may use the definition of $\boldsymbol{w}^{n+1}$ and the identities (8.13) and (8.14) to conclude.

□

Remark 8.2.3 (Conservation of other invariants)**.** *This discretisation does not lend itself to conservation of other invariants, for example even the quadratic invariant F_2 is not conserved under this scheme. A class of Runge-Kutta methods that are able to exactly conserve all quadratic invariants are the Gauss-Radau family, this is because they are* symplectic*, see Chapter 2. When one considers higher order invariants, it seems that schemes must be designed individually and there is no class that can exactly conserve all.*

8.3 Spatial and full discretisation

In this section we describe the discretisation which we analyse for the approximation of (8.1). We show that the scheme has a temporally constant energy functional consistent with that of the original PDE system.
Recall we define our spatial partition and the appropriate finite element spaces as follows.

Definition 8.3.1 (Finite element space)**.** *We discretise* (8.1) *spatially using a piecewise polynomial continuous finite element method. To that end we let $S^1 := [0,1]$ be the unit interval with matching endpoints and choose*

$$0 = x_0 < x_1 < \cdots < x_M = 1.$$

Note that in the numerical experiments we take a larger periodic interval, however for clarity of presentation we restrict our attention in this section to S^1. We denote $\mathcal{J}_m =$

$[x_m, x_{m+1}]$ *to be the* m*–th subinterval and let* $h_m := x_{m+1} - x_m$ *be its length. We impose that the ratio* h_m/h_{m+1} *is bounded from above and below for* $m = 0, \ldots, M-1$*. Let* $\mathbb{P}_q(\mathcal{J}_m)$ *denote the space of polynomials of degree* q *on the element* $\mathcal{J}_m$*, then the* discontinuous *finite element space is given by*

$$\mathbb{V}_q = \{U : S^1 \to \mathbb{R} : U|_{\mathcal{J}_m} \in \mathbb{P}_q(\mathcal{J}_m) \text{ for } m = 0, ..., M-1\}.$$

Further to this we define the continuous *finite element space as*

$$\mathbb{V}_q^C = \mathbb{V}_q \cap C^0\left(S^1\right),$$

where C^0 *denotes the space of continuous functions.*

Throughout this section we will use capital Latin letters to denote spatially discrete trial functions and Greek letters to denote discrete test functions.

Remark 8.3.2 (Use of the discontinuous finite element space)**.** *For clarity of exposition in this chapter we primarily focus on a numerical scheme utilising the* continuous *finite element space. There is no reason we have to design our scheme with such a restriction, in fact our scheme and all analytical results herein also apply to an appropriately defined spatially discontinuous scheme. We shall discuss this discontinuous scheme briefly in Remark 8.3.8.*

8.3.1 Spatial discretisation

Before we give the discretisation let us first consider a direct semi-discretisation of the mixed system (8.10), to find $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W} \in \mathbb{V}_q^C$ such that

$$\begin{aligned}
\langle (\boldsymbol{U}_t + \boldsymbol{V}_x + \boldsymbol{W}), \boldsymbol{\phi} \rangle = 0 \qquad & \forall \boldsymbol{\phi} \in \mathbb{V}_q^C \\
\left\langle \boldsymbol{V} - \frac{1}{2}|\boldsymbol{U}|^2 \boldsymbol{U}, \boldsymbol{\psi} \right\rangle + \langle \boldsymbol{U}_x, \boldsymbol{\psi}_x \rangle = 0 \qquad & \forall \boldsymbol{\psi} \in \mathbb{V}_q^C \\
\left\langle \left(\boldsymbol{W} - |\boldsymbol{U}|^2 \boldsymbol{U}_x + (\boldsymbol{U}_x \cdot \boldsymbol{U}) \boldsymbol{U} \right), \boldsymbol{\chi} \right\rangle = 0 \qquad & \forall \boldsymbol{\chi} \in \mathbb{V}_q^C.
\end{aligned}$$

One may run through the calculation in the proof of Proposition 8.1.7 analogously to see that

$$\frac{\mathrm{d}}{\mathrm{d}t} F_4(\boldsymbol{U}) = \langle \boldsymbol{V}, \boldsymbol{W} \rangle, \tag{8.16}$$

whereby in the continuous case one uses the fact that $\boldsymbol{v}$ and $\boldsymbol{w}$ are orthogonal. In the discrete setting there is no reason why this should be the case and, indeed, except in very special cases, it is not. This necessitates a formulation that forces $\langle \boldsymbol{V}, \boldsymbol{W} \rangle = 0$ thus ensuring conservation of $F_4(\boldsymbol{U})$. We achieve this through a Lagrange multiplier approach encapsulated by the following spatially discrete scheme, to seek $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W} \in \mathbb{V}_q^C$ and $P \in \mathbb{R}/\{0\}$ such that

$$\begin{aligned}
\langle \boldsymbol{U}_t + \boldsymbol{V}_x + \boldsymbol{W}, \boldsymbol{\phi} \rangle = 0 \qquad & \forall \boldsymbol{\phi} \in \mathbb{V}_q^C \\
\left\langle \boldsymbol{V} - \frac{1}{2} |\boldsymbol{U}|^2 \boldsymbol{U}, \boldsymbol{\psi} \right\rangle + \langle \boldsymbol{U}_x, \boldsymbol{\psi}_x \rangle = 0 \qquad & \forall \boldsymbol{\psi} \in \mathbb{V}_q^C \\
\left\langle \boldsymbol{W} - |\boldsymbol{U}|^2 \boldsymbol{U}_x + (\boldsymbol{U}_x \cdot \boldsymbol{U}) \boldsymbol{U}, \boldsymbol{\chi} \right\rangle = 0 \qquad & \forall \boldsymbol{\chi} \in \mathbb{V}_q^C \\
P \langle \boldsymbol{V}, \boldsymbol{\chi} \rangle + \zeta \langle \boldsymbol{V}, \boldsymbol{W} \rangle = 0 \qquad & \forall \zeta \in \mathbb{R}/\{0\} \\
\boldsymbol{U}^0 = \Pi \boldsymbol{u}_0, &
\end{aligned} \tag{8.17}$$

where Π represents the L_2 projection into the finite element space.

Theorem 8.3.3 (Conservativity of the spatially discrete scheme)**.** *Let $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, P$ solve the spatially discrete formulation* (8.17) *then*

$$\frac{\mathrm{d}}{\mathrm{d}t} F_4(\boldsymbol{U}) = \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{2} \langle \boldsymbol{U}_x, \boldsymbol{U}_x \rangle - \frac{1}{8} \left\langle |\boldsymbol{U}|^2, |\boldsymbol{U}|^2 \right\rangle \right) = 0.$$

Proof. An analogous argument to the proof of Proposition 8.1.7 yields (8.16). To conclude pick $\zeta = P$ and $\boldsymbol{\chi} = \boldsymbol{W}$ to see that

$$2P \langle \boldsymbol{V}, \boldsymbol{W} \rangle = 0,$$

as required. □

Remark 8.3.4 (Compatibility of the scheme with the Poisson bracket)**.** *Notice that in view of Theorem 8.3.3 the spatial discretisation is compatible with the Poisson structure of the vmKdV system, indeed, using the same formulation as* (8.11) *we have that the numerical scheme can be written as*

$$\boldsymbol{U}_t = \{\boldsymbol{U}, \boldsymbol{F}_4(\boldsymbol{U})\} = -\left(\Pi\left(\boldsymbol{V}_x\right) + \boldsymbol{W}\right),$$

where Π denotes the L_2 orthogonal projector onto $\mathbb{V}_q^C$. In addition, the evolution of the

Hamiltonian can be described consistently

$$\frac{\mathrm{d}}{\mathrm{d}t}F_4(\boldsymbol{U}) = \{F_4(\boldsymbol{U}), F_4(\boldsymbol{U})\} = 0.$$

It is important to note that the evolution of other quantities, such as further invariants are not compatible with this structure, for example

$$\frac{\mathrm{d}}{\mathrm{d}t}F_2(\boldsymbol{U}) = \{F_2(\boldsymbol{U}), F_4(\boldsymbol{U})\} \neq 0.$$

Remark 8.3.5 (Issues with preserving unsigned invariants)**.** *Typically, preserving an invariant of a continuous problem numerically leads to many desirable properties, such as boundedness of the numerical scheme and a methodology of obtaining a priori error bounds, see [176]. Unfortunately this depends on the energy being signed as this often leads to invariant* inducing a norm*. Consider, for example, the vmKdV-type equation described by*

$$\boldsymbol{u}_t - \frac{3}{2}\boldsymbol{u}\cdot\boldsymbol{u}\boldsymbol{u}_x + \boldsymbol{u}_{xxx} = \boldsymbol{0}. \tag{8.18}$$

We can show, through a similar argument to Proposition 8.1.1, that (8.18) *possesses the sign definite Hamiltonian*

$$\widetilde{F}_4(\boldsymbol{u}) := \frac{1}{2}\left\langle \boldsymbol{u}_x, \boldsymbol{u}_x\right\rangle + \frac{1}{8}\left\langle |\boldsymbol{u}|^2, |\boldsymbol{u}|^2\right\rangle = \frac{1}{2}\left\|\boldsymbol{u}_x\right\|^2_{L_2(S^1)} + \frac{1}{8}\left\|\boldsymbol{u}\right\|^4_{L_4(S^1)}\,\mathrm{d}x.$$

The conservation of this invariant immediately implies that, for any $t > 0$,

$$\frac{1}{2}\left\|\boldsymbol{u}_x(t)\right\|^2_{L_2(S^1)} + \frac{1}{8}\left\|\boldsymbol{u}(t)\right\|^4_{L_4(S^1)} = \frac{1}{2}\left\|\boldsymbol{u}_x(0)\right\|^2_{L_2(S^1)} + \frac{1}{8}\left\|\boldsymbol{u}(0)\right\|^4_{L_4(S^1)},$$

guaranteeing stability of solutions without the necessity of the interpolation arguments of Proposition 8.1.3. Similarly to our spatially discrete scheme for vmKdV, we can define the spatially discrete scheme for this vmKdV-type problem by seeking $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W} \in \mathbb{V}_q^C$ and

$P \in \mathbb{R}/\{0\}$ *such that*

$$\begin{aligned}
\langle \boldsymbol{U}_t + \boldsymbol{V}_x + \boldsymbol{W}, \boldsymbol{\phi} \rangle &= 0 \qquad \forall \boldsymbol{\phi} \in \mathbb{V}_q^C \\
\left\langle \boldsymbol{V} + \frac{1}{2} |\boldsymbol{U}|^2 \boldsymbol{U}, \boldsymbol{\psi} \right\rangle + \langle \boldsymbol{U}_x, \boldsymbol{\psi}_x \rangle &= 0 \qquad \forall \boldsymbol{\psi} \in \mathbb{V}_q^C \\
\left\langle \boldsymbol{W} + |\boldsymbol{U}|^2 \boldsymbol{U}_x - (\boldsymbol{U}_x \cdot \boldsymbol{U}) \boldsymbol{U}, \boldsymbol{\chi} \right\rangle &= 0 \qquad \forall \boldsymbol{\chi} \in \mathbb{V}_q^C \\
P \langle \boldsymbol{V}, \boldsymbol{\chi} \rangle + \zeta \langle \boldsymbol{V}, \boldsymbol{W} \rangle &= 0 \qquad \forall \zeta \in \mathbb{R}/\{0\} \\
\boldsymbol{U}^0 &= \Pi \boldsymbol{u}_0.
\end{aligned}$$

Through an almost identical argument to that in the proof of Theorem 8.3.3 we find that this spatially discrete scheme conserves the Hamiltonian functional in the sense that

$$\frac{\mathrm{d}}{\mathrm{d}t} \widetilde{F}_4(\boldsymbol{U}) = \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{2} \|\boldsymbol{U}_x\|_{L_2(S^1)}^2 + \frac{1}{8} \|\boldsymbol{U}\|_{L_4(S^1)}^4 \right) = 0,$$

so our scheme is numerically stable in the sense that

$$\frac{1}{2} \|\boldsymbol{U}_x(t)\|_{L_2(S^1)}^2 + \frac{1}{8} \|\boldsymbol{U}(t)\|_{L_4(S^1)}^4 = \frac{1}{2} \|\boldsymbol{U}_x(0)\|_{L_2(S^1)}^2 + \|\boldsymbol{U}(0)\|_{L_4(S^1)}^4,$$

so we have boundedness of the numerical approximation in not only L_4 but also in a discrete H^1 norm. We expect that the space of exact solutions for (8.18) *and* (8.1) *are quite different and leave the quantification of solutions of this problem for future work. For our spatially discrete scheme for vmKdV such elegant stability bounds cannot be shown, and conservation of $F_4(\boldsymbol{U})$ ultimately tells us that*

$$\frac{1}{2} \|\boldsymbol{U}_x(t)\|_{L_2(S^1)}^2 - \frac{1}{8} \|\boldsymbol{U}(t)\|_{L_4(S^1)}^4 = \frac{1}{2} \|\boldsymbol{U}_x(0)\|_{L_2(S^1)}^2 - \|\boldsymbol{U}(0)\|_{L_4(S^1)}^4,$$

which, while suggesting that the numerical scheme is likely bounded, does not conclusively prove it. The stability analysis and a priori analysis for this scheme are left as future work.

8.3.2 Fully discrete scheme

Making use of the semi discretisations developed in §8.2 and §8.3.1 we consider a fully discrete approximation that consists of finding a sequence of functions $\boldsymbol{U}^{n+1}, \boldsymbol{V}^{n+1}, \boldsymbol{W}^{n+1} \in$

$\mathbb{V}_q^C$ and $P^{n+1} \in \mathbb{R}/\{0\}$ such that for each $n \in [0, N-1]$ we have

$$\begin{aligned}
\left\langle \frac{\boldsymbol{U}^{n+1} - \boldsymbol{U}^n}{\tau_n} + \boldsymbol{V}_x^{n+1} + \boldsymbol{W}^{n+1}, \boldsymbol{\phi} \right\rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q^C \\
\left\langle \boldsymbol{V}^{n+1} - \frac{1}{2}\left(|\boldsymbol{U}^n|^2 + \left|\boldsymbol{U}^{n+1}\right|^2 \right) \boldsymbol{U}^{n+1/2}, \boldsymbol{\psi} \right\rangle + \left\langle \boldsymbol{U}_x^{n+1/2}, \boldsymbol{\psi}_x \right\rangle &= 0 \qquad \forall \psi \in \mathbb{V}_q^C \\
\left\langle \boldsymbol{W}^{n+1} - \left|\boldsymbol{U}^{n+1/2}\right|^2 \boldsymbol{U}_x^{n+1/2} + \left(\boldsymbol{U}_x^{n+1/2} \cdot \boldsymbol{U}^{n+1/2} \right) \boldsymbol{U}^{n+1/2}, \boldsymbol{\chi} \right\rangle &= 0 \qquad \forall \chi \in \mathbb{V}_q^C \\
P^{n+1} \left\langle \boldsymbol{V}^{n+1}, \boldsymbol{\chi} \right\rangle + \zeta \left\langle \boldsymbol{V}^{n+1}, \boldsymbol{W}^{n+1} \right\rangle &= 0 \qquad \forall \zeta \in \mathbb{R}/\{0\} \\
\boldsymbol{U}^0 &= \Pi \boldsymbol{u}_0,
\end{aligned} \tag{8.19}$$

where Π denotes the L_2 orthogonal projector into the finite element space $\mathbb{V}_q^C$ at the initial point in time. This is the direct discretisation of the mixed system (8.10) with the temporal discretisation as that proposed in §8.2 with an additional equation for a unknown real number that represents a Lagrange multiplier ensuring $\langle \boldsymbol{V}^n, \boldsymbol{W}^n \rangle = 0$ for all n.

Remark 8.3.6 (Adaptivity)**.** *Note that our method permits spatial adaptivity over time, that is to say that our spatial mesh, and therefore our finite element spaces, can change from one time step to the next. However, we shall assume that our spatial mesh is fixed over all time here for simplicity. Note that multiple complications arise from allowing spatial adaptivity as discussed in Chapter 6 and §7.4, and for the arguments in the sequel to hold without the introduction of nonstandard interpolation operators we require* nested *finite element spaces.*

Theorem 8.3.7 (Conservativity of the fully discrete scheme)**.** *Let $\{\boldsymbol{U}^n\}_{n=0}^N$ be the fully discrete scheme generated by* (8.19)*, then we have that*

$$F_4(\boldsymbol{U}^n) = F_4(\boldsymbol{U}^0) \quad \forall n \in [0, N].$$

Proof. It suffices to show that

$$F_4(\boldsymbol{U}^{n+1}) - F_4(\boldsymbol{U}^n) = 0$$

and then the result follows inductively. To this end

$$\begin{aligned}
2\left(F_4(\boldsymbol{U}^{n+1}) - F_4(\boldsymbol{U}^n)\right) &= \left\langle \boldsymbol{U}_x^{n+1}, \boldsymbol{U}_x^{n+1} \right\rangle - \langle \boldsymbol{U}_x^n, \boldsymbol{U}_x^n \rangle \\
&\quad - \frac{1}{4}\left\langle \left|\boldsymbol{U}^{n+1}\right|^2, \left|\boldsymbol{U}^{n+1}\right|^2 \right\rangle + \frac{1}{4}\left\langle |\boldsymbol{U}^n|^2, |\boldsymbol{U}^n|^2 \right\rangle \\
&= \left\langle \boldsymbol{U}_x^{n+1} - \boldsymbol{U}_x^n, \boldsymbol{U}_x^{n+1} + \boldsymbol{U}_x^n \right\rangle \\
&\quad - \frac{1}{4}\left\langle \boldsymbol{U}^{n+1} - \boldsymbol{U}^n, \left|\boldsymbol{U}^{n+1}\right|^2 \boldsymbol{U}^{n+1} + \left|\boldsymbol{U}^{n+1}\right|^2 \boldsymbol{U}^n \right\rangle \\
&\quad - \frac{1}{4}\left\langle \boldsymbol{U}^{n+1} - \boldsymbol{U}^n, |\boldsymbol{U}^n|^2 \boldsymbol{U}^{n+1} + |\boldsymbol{U}^n|^2 \boldsymbol{U}^n \right\rangle \\
&= -\left\langle \boldsymbol{U}^{n+1} - \boldsymbol{U}^n, \boldsymbol{V}^{n+1} \right\rangle,
\end{aligned}$$

through expanding differences and using the second equation of (8.19). Now, using the first equation of (8.19)

$$\begin{aligned}
2\left(F_4(\boldsymbol{U}^{n+1}) - F_4(\boldsymbol{U}^n)\right) &= \tau_n \left\langle \boldsymbol{V}_x^{n+1} + \boldsymbol{W}^{n+1}, \boldsymbol{V}^{n+1} \right\rangle \\
&= \tau_n \left\langle \boldsymbol{W}^{n+1}, \boldsymbol{V}^{n+1} \right\rangle \\
&= 0
\end{aligned}$$

using the fourth equation of (8.19) with $\zeta = P^{n+1}$ and $\boldsymbol{\chi} = \boldsymbol{W}^{n+1}$, concluding the proof. □

Remark 8.3.8 (A *discontinuous* fully discrete scheme)**.** *As mentioned in Remark 8.3.2, we do not necessarily need to assume that our finite element space is continuous to yield a conservative scheme. In fact, as we discussed in Chapter 4 there are significant benefits to a discontinuous approximation. If the scheme is discontinuous it is possible to reduce the numerical scheme from a system to primal form which would significantly improve the performance.*

To introduce a discontinuous finite element scheme we borrow operators representing discrete first and second spatial derivatives from Chapter 5. When referring to jumps $[\![\cdot]\!]$ *and averages* $\{\cdot\}$ *in this chapter it is important to note that these definitions are all* componentwise*, i.e., they act on each vector component independently and do not change the vectorial dimension of a given function. Let* $\boldsymbol{Z} \in \mathbb{V}_q$*, then recall we define the first discrete*

spatial derivative through $\mathcal{G} : \mathbb{V}_q \to \mathbb{V}_q$ *such that*

$$\langle \mathcal{G}(\boldsymbol{Z}), \boldsymbol{\phi} \rangle = \langle \boldsymbol{Z}_x, \boldsymbol{\phi} \rangle - \sum_{m=0}^{M-1} [\![\boldsymbol{Z}_m]\!] \cdot \{ \boldsymbol{\phi}_m \} \qquad \forall \phi \in \mathbb{V}_q.$$

Additionally, recall the symmetric interior penalty bilinear form, given in Definition 5.1.6,

$$\begin{aligned} \mathcal{A}_h(\boldsymbol{Z}, \boldsymbol{\phi}) = \langle \boldsymbol{Z}_x, \boldsymbol{\phi}_x \rangle + \sum_{m=0}^{M-1} & - [\![\boldsymbol{Z}_m]\!] \cdot \{ \boldsymbol{\phi}_{xmm} \} \\ & - [\![\boldsymbol{\phi}_m]\!] \cdot \{ \boldsymbol{Z}_{xm} \} + \frac{\sigma}{h_m} [\![\boldsymbol{Z}_{xm}]\!] \cdot [\![\boldsymbol{\phi}_{xm}]\!] \qquad \forall \boldsymbol{\phi} \in \mathbb{V}_q, \end{aligned}$$

where σ *is a sufficiently large constant to guarantee stability. Recall that* $\mathcal{G}$ *satisfies the* discrete integration by parts *identity*

$$\langle \mathcal{G}(\boldsymbol{Z}), \boldsymbol{\phi} \rangle = - \langle \boldsymbol{Z}, \mathcal{G}(\boldsymbol{\phi}) \rangle, \tag{8.20}$$

and the symmetric interior penalty form is bilinear. *With these definitions in mind we can introduce the discontinuous fully discrete scheme as follows. Seek the sequence of functions* $\boldsymbol{U}^{n+1}, \boldsymbol{V}^{n+1}, \boldsymbol{W}^{n+1} \in \mathbb{V}_q$ *and* $P^{n+1} \in \mathbb{R}/\{0\}$ *such that for each* $n \in [0, N-1]$ *we have*

$$\begin{aligned} \left\langle \frac{\boldsymbol{U}^{n+1} - \boldsymbol{U}^n}{\tau_n} + \mathcal{G}\left(\boldsymbol{V}^{n+1}\right) + \boldsymbol{W}^{n+1}, \boldsymbol{\phi} \right\rangle &= 0 \qquad \forall \phi \in \mathbb{V}_q \\ \left\langle \boldsymbol{V}^{n+1} - \frac{1}{2}\left(|\boldsymbol{U}^n|^2 + \left|\boldsymbol{U}^{n+1}\right|^2 \right) \boldsymbol{U}^{n+1/2}, \boldsymbol{\psi} \right\rangle + \mathcal{A}_h\left(\boldsymbol{U}^{n+1/2}, \boldsymbol{\psi}\right) &= 0 \qquad \forall \psi \in \mathbb{V}_q \\ \left\langle \boldsymbol{W}^{n+1} - \left|\boldsymbol{U}^{n+1/2}\right|^2 \mathcal{G}\left(\boldsymbol{U}^{n+1/2}\right) + \left(\mathcal{G}\left(\boldsymbol{U}^{n+1/2}\right) \cdot \boldsymbol{U}^{n+1/2}\right) \boldsymbol{U}^{n+1/2}, \boldsymbol{\chi} \right\rangle &= 0 \qquad \forall \chi \in \mathbb{V}_q \\ P^{n+1} \left\langle \boldsymbol{V}^{n+1}, \boldsymbol{\chi} \right\rangle + \zeta \left\langle \boldsymbol{V}^{n+1}, \boldsymbol{W}^{n+1} \right\rangle &= 0 \qquad \forall \zeta \in \mathbb{R}/\{0\} \\ \boldsymbol{U}^0 &= \Pi \boldsymbol{u}_0, \end{aligned}$$

where Π *denotes the* L_2 *orthogonal projector into the finite element space* $\mathbb{V}_q$ *at the initial point in time. Following the methodology outlined in the proof of Theorem 8.3.7, along with the bilinearity of* $\mathcal{A}_h(\cdot, \cdot)$ *and* (8.20)*, we have that the discrete energy*

$$\hat{F}_4(\boldsymbol{U}^n) = \frac{1}{2} \mathcal{A}_h(\boldsymbol{U}, \boldsymbol{U}) - \frac{1}{8} \left\langle |\boldsymbol{U}|^2, |\boldsymbol{U}|^2 \right\rangle,$$

is preserved over time, i.e.,

$$\hat{F}_4(\boldsymbol{U}^n) = \hat{F}_4(\boldsymbol{U}^0) \quad \forall n \in [0, N].$$

Note that as the finite element spaces are discontinuous we can employ local discontinuous Galerkin techniques, similarly to Remark 4.2.6, to rewrite our scheme in primal form speeding up the implementation by orders of magnitude.

8.4 Numerical experiments

In this section we illustrate the performance of the method proposed through a series of numerical experiments. Similarly to the prequel the brunt of the computational work has been carried out using Firedrake [153]. We employ a Gauss quadrature of order $4q$, where q is the degree of the finite element space, to minimise quadrature error introduced into the implementation. Indeed, at this degree all integrals are performed exactly with the exception of the projection of the initial condition. When computing errors we shall utilise a $4q + 4$ degree Gauss quadrature. The nonlinear system of equations are then approximated using the PETSc [20, 21] Newton line search method with a tolerance of 10^{-12} on each time step. A combination of Paraview and Matplotlib have been used as a visualisation tool. The code written for this purpose is freely available at [106]. For each benchmark test we fix the polynomial degree q and compute a sequence of solutions with $h = h(i) = 2^{-i}$ and τ chosen either so $\tau \ll h$, to make the temporal discretisation error negligible, or so $\tau = h$ so temporal discretisation error dominates. This is done for a sequence of refinement levels, $i = l, \dots, L$. We have previously used S^1 as the unitary periodic domain. For our numerical experiments, we have scaled the domain to $[0, 40]$ for computational convenience.

Remark 8.4.1 (Numerical deviation in F_4)**.** *While the analysis shows that our scheme* exactly *preserves the energy over arbitrarily long time, the implementation relies on linear and nonlinear solvers that inherently require further approximation. The result of this is that the energy may deviate* locally *up to the tolerance of the linear and nonlinear solvers which introduces the possibility of these errors propagating over time. In our numerical tests we focus on studying the global deviation in time, $F_4(U^n) - F_4(U^0)$, which* includes *any propagation arising from solver or precision errors.*

8.4.1 Test 1 - Asymptotic benchmarking of a 1-soliton solution

We take $d = 2$ and

$$\boldsymbol{u}_0 = \frac{2\mu}{\cosh\left(\left(\mu\left(x - c_\mu\right)\right)\right)}\boldsymbol{E}, \tag{8.21}$$

over the periodic domain $S^1([0,40])$ with $\boldsymbol{E} = (0.8, (1-0.8^2)^{0.5})^{\mathrm{T}}$, $\mu = 1$ and $c_\mu = 20$. The exact solution is then given by (8.5). We take a uniform timestep and uniform meshes that are fixed with respect to time. Convergence results are shown in Figure 8.2 and conservativity over long time is given in Figure 8.1. Note that for the 1-soliton solution we have $\boldsymbol{W}^n \equiv \boldsymbol{0}$ for all n in which case the Lagrange multiplier is not required as $\langle \boldsymbol{V}^n, \boldsymbol{W}^n \rangle = 0$ trivially for all n.

For this test case we also investigate how well the qualitative structure of the solution is captured, similarly to §5.4 and [32] through the amplitude error, phase error and shape error of a single soliton over time. Recall that we define $\boldsymbol{X}$ as the set of Lagrange degrees of freedom of our finite element functions.

Further recall the following definitions. The amplitude error for U_i is then given by

$$\max_{\boldsymbol{X}} U_i - \max_{\boldsymbol{X}} u_i.$$

If the amplitude error is positive then the numerical soliton is larger than the exact solution, and vice versa. Similarly we define the phase error as

$$e_{p_i} = \operatorname*{\texttt{argmax}}_{\boldsymbol{X}} U_i - \operatorname*{\texttt{argmax}}_{\boldsymbol{X}} u_i,$$

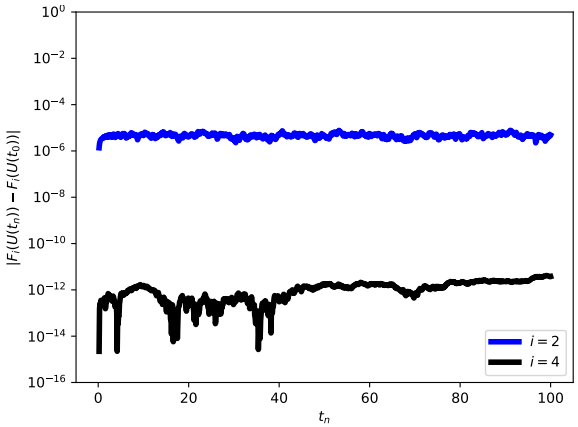
where `argmax` represents the spatial coordinate associated to the maximum over $\boldsymbol{X}$. If the phase error is positive then the numerical approximation is moving faster than the exact solution, and vice versa. Note that this discrete measure of the error cannot detect shifts in phase which are smaller than the distance between degrees of freedom. The amplitude and phase errors for this test case can be seen in Figure 8.3 and Figure 8.4 respectively.

Additionally recall that we define the shape error by shifting the exact solution by the distance of the phase error and computing the L_2 error at fixed times, i.e., the discrete shape error is

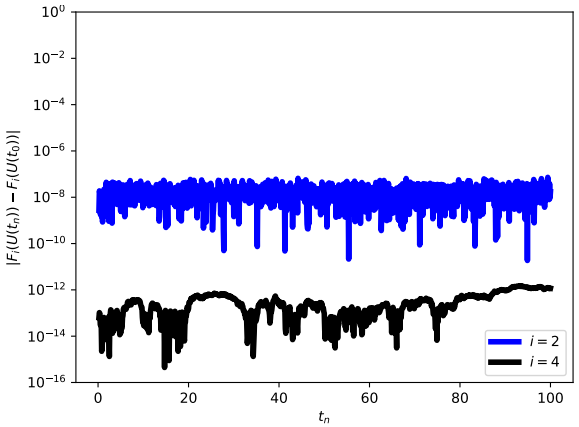
$$\left\| u_i(x + e_{p_i}, t_n) - U_i(x, t_n) \right\|_{L_2(S^1([0,40]))}.$$

The shape error for this test can be seen in Figure 8.5.

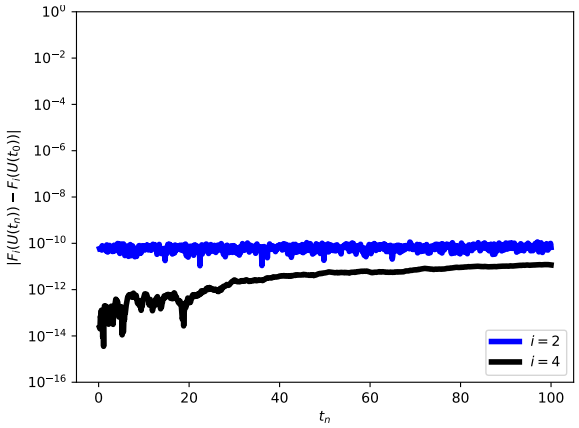
Figure 8.1: Here we examine the conservative discretisation scheme with various polynomial degrees, q, approximating the exact solution (8.5) with initial conditions given by (8.21). We show the deviation in the two invariants F_i, $i = 2, 4$, corresponding to momentum and energy. In each test we take a fixed spatial discretisation parameter of $h = 0.25$ and fixed time step of $\tau = 0.001$. Notice that in each case the deviation in energy is smaller than the solver tolerance of 10^{-12} and the deviation in momentum is bounded. In addition, as the degree of approximation is increased the deviation in momentum becomes smaller, in this case by around two orders of magnitude per polynomial order. The simulations are simulated for long time to test conservativity with $T = 100$ in each case.



(a) Here $q = 1$.

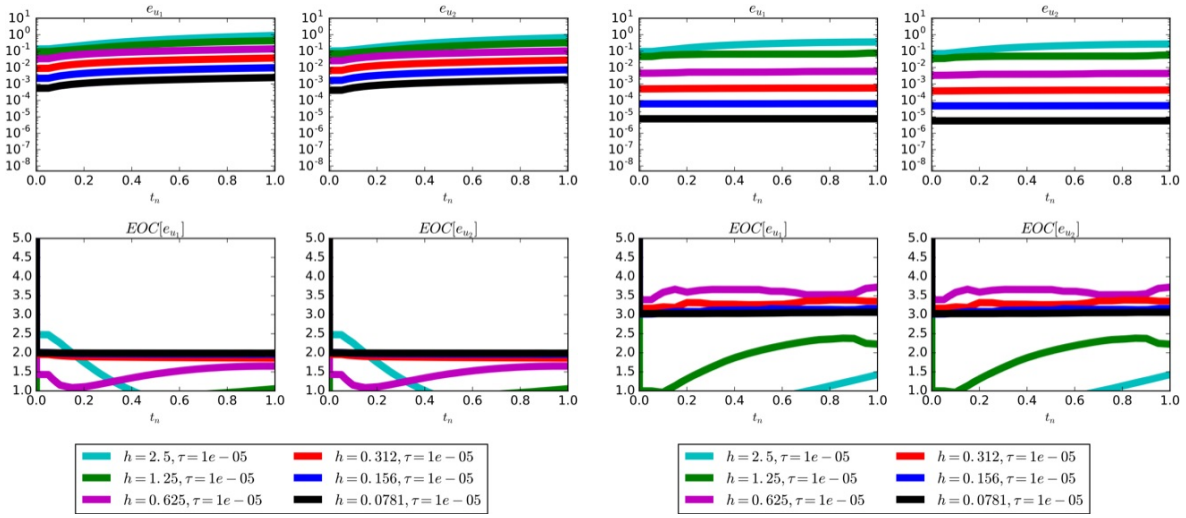


(b) Here $q = 2$.



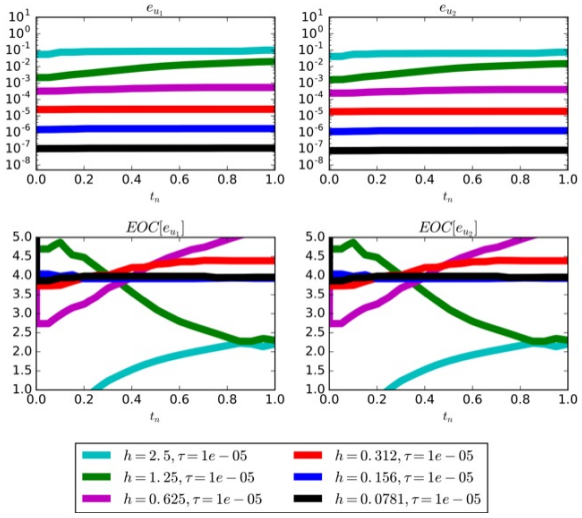
(c) Here $q = 3$.

Figure 8.2: Here we examine the conservative discretisation scheme with various polynomial degrees, q, approximating the exact solution (8.5) with initial conditions given by (8.21). We show the errors measured in the $L_\infty(0, t_n; L_2(S^1([0, 40])))$ norm for each component of the system and the EOC for test runs that benchmark both the spatial and temporal discretisation and show that the scheme is of optimal order. We use $e_{u_i} := \|u_i - U_i\|_{L_\infty(0,t_n;L_2(S^1([0,40])))}$ for $i = 1, 2$, the components of the solution $\boldsymbol{u} = (u_1, u_2)^{\mathrm{T}}$ and numerical approximation $\boldsymbol{U} = (U_1, U_2)^{\mathrm{T}}$.

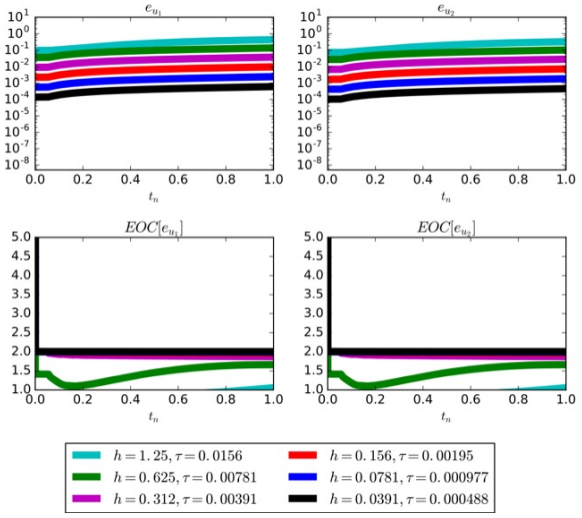


(a) Here $q = 1$ and we fix $\tau = 0.00001$. This is sufficiently small that the spatial discretisation error dominates.

(b) Here $q = 2$ and we fix $\tau = 0.00001$. This is sufficiently small that the spatial discretisation error dominates.

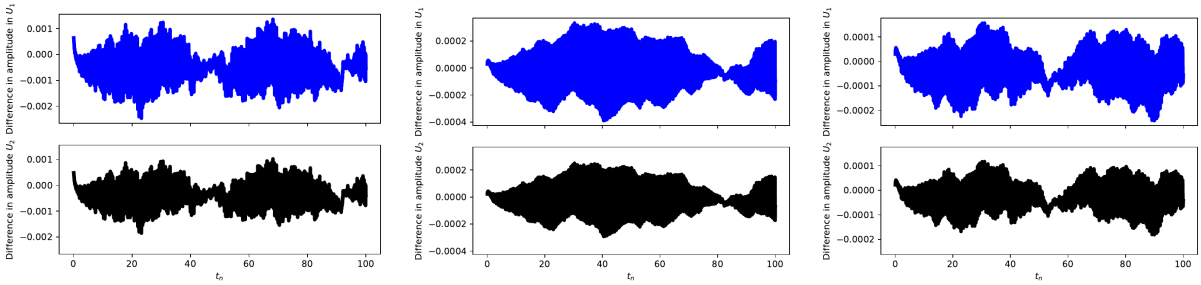


(c) Here $q = 3$ and we fix $\tau = 0.00001$. This is sufficiently small that the spatial discretisation error dominates.



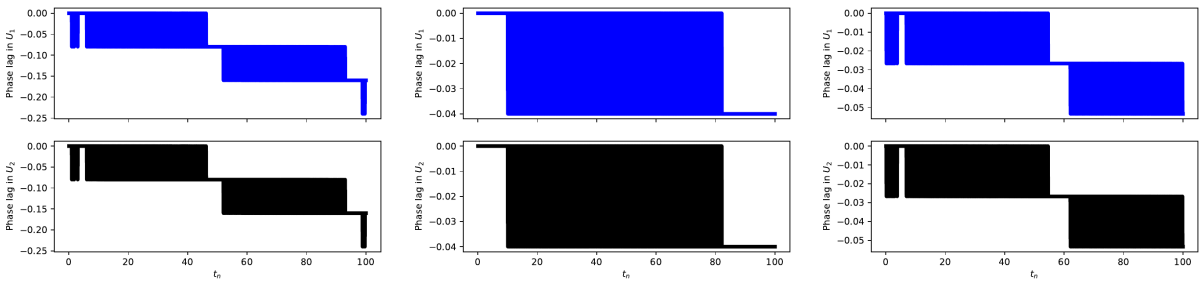
(d) Here $q = 2$ and on every refinement level we choose a coupling $\tau = Ch$. Note that the time discretisation error here dominates.

Figure 8.3: We examine the difference in maximal amplitude of the solutions for various polynomial degrees, q, compared to the exact solution (8.5) with initial conditions given by (8.21). When the difference in amplitude is positive the numerical approximation is larger than the exact solution, and when it is negative the numerical approximation is smaller. We plot these difference for both vector components of the soliton independently. In these simulations we take a fixed spatial discretisation parameter of $h = 0.08$ and a fixed time step of $\tau = 0.05$. We notice that the deviation in amplitude is bounded over time and decreases with polynomial degree.



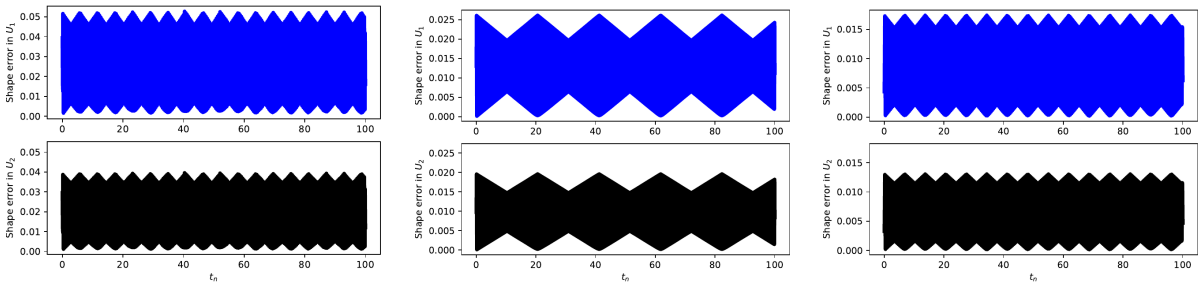
(a) Here $q = 1$. (b) Here $q = 2$. (c) Here $q = 3$.

Figure 8.4: We examine the difference in speed between numerical solutions for various polynomial degrees, q, and the exact solution (8.5) with initial conditions given by (8.21). We track this difference by looking at the spatial coordinate of the maximal amplitude for the numerical and exact solutions, and take the difference. We plot these differences for both vector components of the soliton independently. In these simulations we take a fixed spatial discretisation parameter of $h = 0.08$ and a fixed time step of $\tau = 0.05$. We notice that the phase error decreases over time, i.e., the numerical solution travels slower than its exact counterpart.



(a) Here $q = 1$. (b) Here $q = 2$. (c) Here $q = 3$.

Figure 8.5: We examine the difference in the shape between numerical solutions for various polynomial degrees, q, and the exact solution (8.5) with initial conditions given by (8.21). Mathematically we can write the shape error as $\min_{y\in S^1([0,40])} \|u_i(x+y,t_n) - U_i(x,t_n)\|_{L_2(S^1([0,40]))}$ for each component of the solution. In these simulations we take a fixed spatial discretisation parameter of $h = 0.08$ and a fixed time step of $\tau = 0.05$. We notice that the shape error does not propagate over long time and decreases as the polynomial degree increases.



(a) Here $q = 1$. (b) Here $q = 2$. (c) Here $q = 3$.

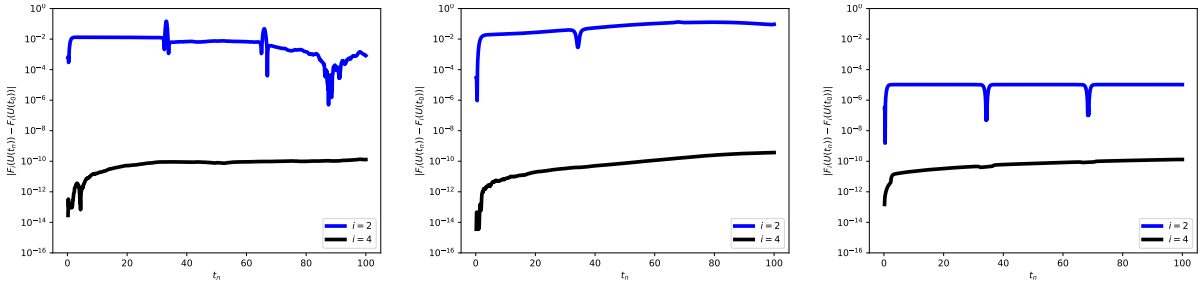
8.4.2 Test 2 - Asymptotic benchmarking of a 2-soliton solution

We take $d = 2$ and

$$\boldsymbol{u}_0 = \frac{F_{\mu,\nu}}{G}\boldsymbol{E}_1 + \frac{F_{\nu,\mu}}{G}\boldsymbol{E}_2, \tag{8.22}$$

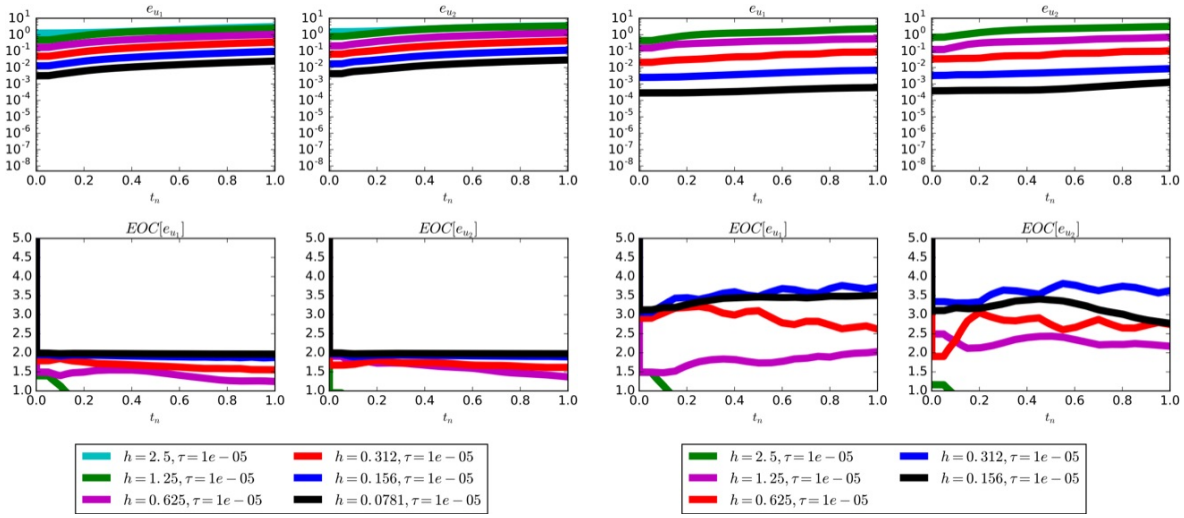
with $F_{\mu,\nu}$ given in (8.7) G given in (8.8). The parameters are $\boldsymbol{E_1} = (1,0)^{\mathrm{T}}$, $\boldsymbol{E_2} = (0,1)^{\mathrm{T}}$, $\mu = \sqrt{2}$, $\nu = \sqrt{3}$, $c_\nu = 24.9, c_\mu = 25.1$. The exact solution is then given by (8.1.1). We take a uniform timestep and uniform meshes that are fixed with respect to time. Convergence results are shown in Figure 8.7 and conservativity over long time is given in Figure 8.6. Note that for 2-soliton solution we have $\boldsymbol{W}^n \neq \boldsymbol{0}$ in general in which case the Lagrange multiplier is required to ensure $\langle \boldsymbol{V}^n, \boldsymbol{W}^n \rangle = 0$ for all n and that the results of Theorem 8.3.7 hold.

Figure 8.6: We examine the conservative discretisation scheme with various polynomial degrees, q, approximating the exact solution (8.6) with initial conditions given by (8.22). We show the deviation in the two invariants F_i, $i = 2, 4$, corresponding to momentum and energy respectively. In each test we take a fixed spatial discretisation parameter of $h = 0.25$ and fixed time step of $\tau = 0.001$. Notice that in each case the deviation in energy is smaller than the solver tolerance of 10^{-12} and the deviation in momentum is bounded. The simulations are simulated for long time to test conservativity with $T = 100$ in each case.



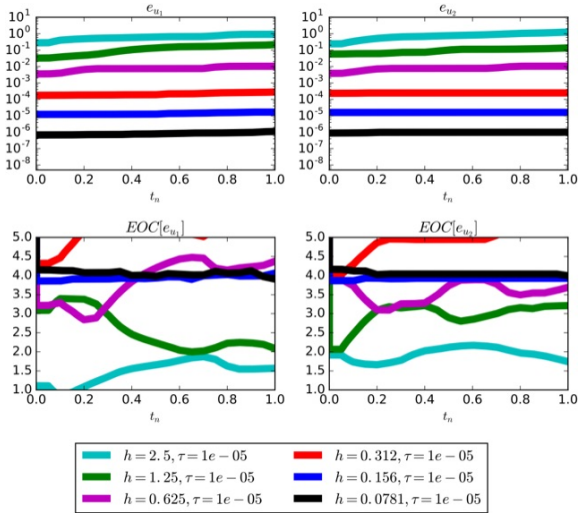
(a) Here $q = 1$. (b) Here $q = 2$. (c) Here $q = 3$.

Figure 8.7: We examine the conservative discretisation scheme with various polynomial degrees, q, approximating the exact solution (8.6) with initial conditions given by (8.22). We show the errors measured in the $L_\infty(0, t_n; L_2(S^1([0, 40])))$ norm for each component of the system and the EOC for test runs that benchmark both the spatial and temporal discretisation and show that the scheme is of the correct order. We use $e_{u_i} := \|u_i - U_i\|_{L_\infty(0,t_n;L_2(S^1([0,40])))}$ for $i = 1, 2$, the components of the solution $\boldsymbol{u} = (u_1, u_2)^{\mathrm{T}}$.

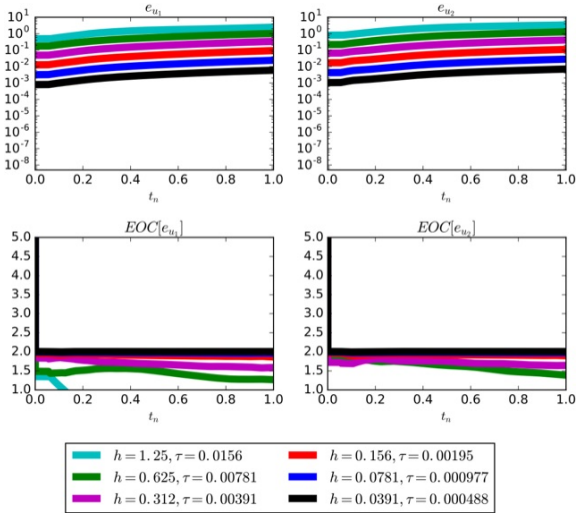


(a) Here $q = 1$ and we fix $\tau = 0.00001$. This is sufficiently small that the spatial discretisation error dominates.

(b) Here $q = 2$ and we fix $\tau = 0.00001$. This is sufficiently small that the spatial discretisation error dominates.



(c) Here $q = 3$ and we fix $\tau = 0.00001$. This is sufficiently small that the spatial discretisation error dominates.



(d) Here $q = 2$ and on every refinement level we choose a coupling $\tau = Ch$. Note that the time discretisation error here dominates.

8.4.3 Test 3 - Dynamics of 2 and 3-soliton solutions

8.4.3.1 Subtest 1

We take $d = 2$ and

$$\boldsymbol{u}_0 = \frac{F_{\mu,\nu}}{G}\boldsymbol{E}_1 + \frac{F_{\nu,\mu}}{G}\boldsymbol{E}_2, \tag{8.23}$$

with $F_{\mu,\nu}$ given in (8.7) G given in (8.8). The parameters are $\boldsymbol{E_1} = (\frac{9}{10}, \frac{\sqrt{19}}{10})^{\mathrm{T}}$, $\boldsymbol{E_2} = (\frac{1}{10}, \frac{3\sqrt{11}}{10})^{\mathrm{T}}$, $\mu = \sqrt{2}$, $\nu = \sqrt{3}$, $c_\nu = 10, c_\mu = 13$. Figure 8.8 shows some plots of the dynamics of the numerical approximation.

8.4.3.2 Subtest 2

We take $d = 2$ and

$$\boldsymbol{u}_0 = \frac{F_{\mu,\nu}}{G}\boldsymbol{E}_1 + \frac{F_{\nu,\mu}}{G}\boldsymbol{E}_2, \tag{8.24}$$

with $F_{\mu,\nu}$ given in (8.7) G given in (8.8). The parameters are $\boldsymbol{E_1} = (1, 0)^{\mathrm{T}}$, $\boldsymbol{E_2} = (0, 1)^{\mathrm{T}}$, $\mu = -\sqrt{2}$, $\nu = \sqrt{3}$, $c_\mu = 9, c_\nu = 13$. Figure 8.9 shows some plots of the dynamics of the numerical approximation. We also examine the difference in the amplitude and phase between the numerical and exact solitons in each vector component before and after the soliton interaction in Table 8.1. The amplitude and phase errors are calculated as described in the one soliton case.

8.4.3.3 Subtest 3

In addition to the 2-soliton interactions we also take the opportunity to examine the dynamics of a 3-soliton interaction. We take $d = 2$ and

$$\boldsymbol{u}_0 = \sum_{i=1}^{3} \frac{2\mu_i}{\cosh\left(\left(\mu_i\left(x - c_{\mu_i}\right)\right)\right)}\boldsymbol{E_i} \tag{8.25}$$

with $\boldsymbol{E}_1 = \boldsymbol{E}_3 = (1, 0)^{\mathrm{T}}, \boldsymbol{E}_2 = (0, 1)^{\mathrm{T}}$, $\mu_1 = \frac{19}{10}, \mu_2 = -\frac{40}{25}, \mu_3 = \frac{13}{10}$ and $c_{\mu_1} = 4, c_{\mu_2} = 12, c_{\mu_3} = 21$. Figure 8.10 shows the dynamics of the numerical approximation.

Figure 8.8: The dynamics of the approximation generated by the conservative discretisation scheme with polynomial degree $q = 1$ approximating a smooth solution with initial conditions given by (8.23).

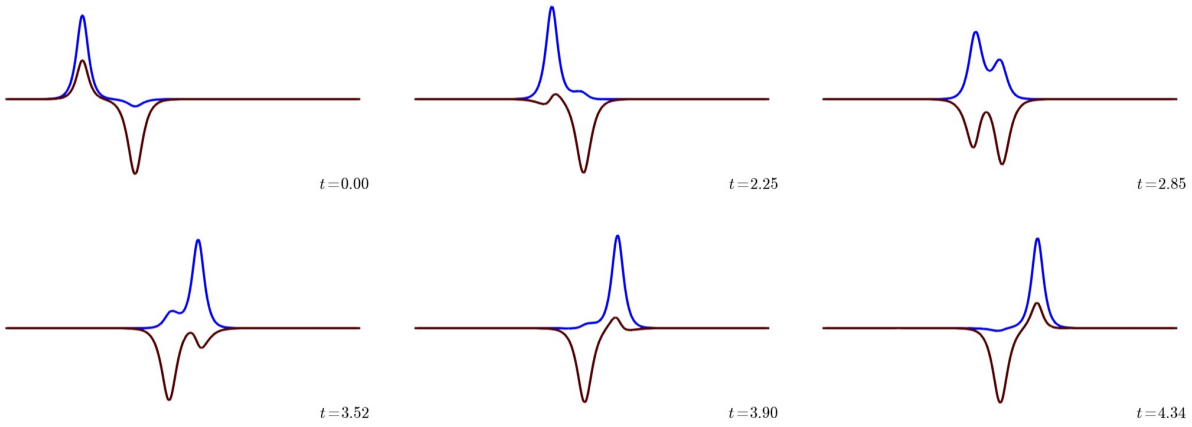


Figure 8.9: The dynamics of the approximation generated by the conservative discretisation scheme with polynomial degree $q = 1$ approximating a smooth solution with initial conditions given by (8.24).

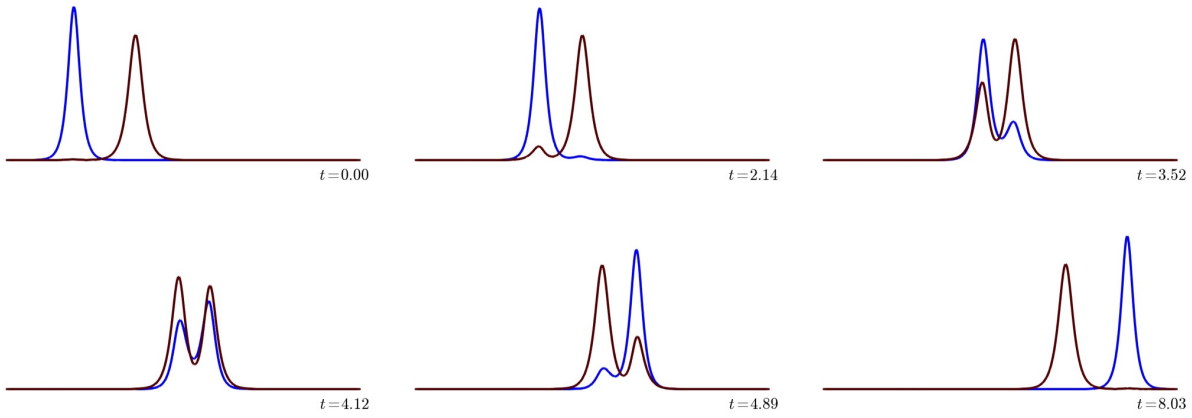


Table 8.1: The phase and amplitude errors committed by the conservative discretisation scheme approximating the smooth solution (8.24). We display the minimal and maximal errors both before and after the soliton interactions for various coupled temporal and spatial discretisations and various polynomial degrees. We notice, as in the one soliton case, our numerical solitons travel slower than their exact counterparts both before and after soliton interactions, with the phase error generally increasing after the soliton interactions. Note that when the phase error is measured to be zero this *does not* mean that the phase of the numerical scheme is exact, only that the phase error is smaller than the distance between the degrees of freedom, $\frac{h}{q}$, of our numerical approximation. The error in amplitude also increases after the soliton interaction, but remains reasonably small and decreases with the polynomial degree.

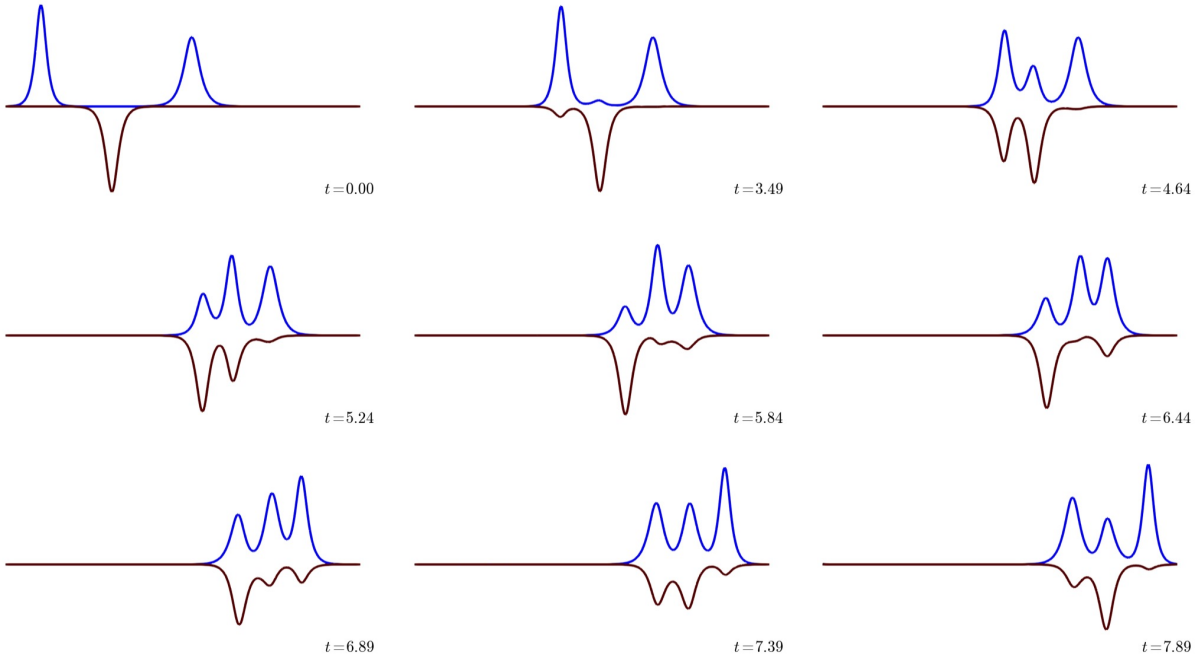
τ	h	Degree	Phase error for $t \in [0, 2.0]$		Phase error for $t \in [6.0, 8.0]$		Amplitude error for $t \in [0, 2.0]$		Amplitude error for $t \in [6.0, 8.0]$	
			min	max	min	max	min	max	min	max
2.0e-02	3.2e-01	1	-3.2e-01	0.0e+00	-1.3e+00	-6.4e-01	-2.5e-01	3.2e-03	-2.6e-01	2.1e-02
		2	-1.6e-01	0.0e+00	-1.6e-01	0.0e+00	-2.0e-02	8.8e-03	-3.9e-02	2.3e-02
		3	-1.1e-01	0.0e+00	-1.1e-01	0.0e+00	-6.2e-03	4.3e-03	-1.4e-02	1.2e-02
1.0e-02	1.6e-01	1	-1.6e-01	0.0e+00	-4.8e-01	-1.6e-01	-7.0e-02	-3.8e-03	-7.6e-02	-6.1e-03
		2	0.0e+00	0.0e+00	-8.0e-02	0.0e+00	-2.1e-03	1.2e-03	-4.4e-03	2.8e-03
		3	-5.3e-02	0.0e+00	-5.3e-02	0.0e+00	-7.9e-04	7.7e-04	-2.3e-03	2.2e-03
5.0e-03	8.0e-02	1	-8.0e-02	0.0e+00	-8.0e-02	0.0e+00	-1.6e-02	-5.7e-04	-1.8e-02	-4.8e-03
		2	-4.0e-02	0.0e+00	0.0e+00	0.0e+00	-2.2e-04	1.1e-04	-5.3e-04	3.7e-04
		3	-2.7e-02	0.0e+00	-2.7e-02	0.0e+00	-1.2e-04	7.7e-05	-3.5e-04	2.7e-04
2.5e-03	4.0e-02	1	-4.0e-02	0.0e+00	-4.0e-02	0.0e+00	-3.5e-03	-5.9e-05	-4.9e-03	4.0e-05
		2	-2.0e-02	0.0e+00	0.0e+00	0.0e+00	-2.6e-05	1.4e-05	-7.6e-05	3.6e-05
		3	-1.3e-02	0.0e+00	-1.3e-02	0.0e+00	-1.9e-05	1.1e-05	-5.6e-05	3.5e-05

(a) Phase and amplitude errors for the first vector component of the solution.

τ	h	Degree	Phase error for $t \in [0, 2.0]$		Phase error for $t \in [6.0, 8.0]$		Amplitude error for $t \in [0, 2.0]$		Amplitude error for $t \in [6.0, 8.0]$	
			min	max	min	max	min	max	min	max
2.0e-02	3.2e-01	1	-3.2e-01	0.0e+00	-9.6e-01	-3.2e-01	-1.4e-01	-1.2e-02	-1.0e-01	-9.3e-03
		2	0.0e+00	0.0e+00	0.0e+00	0.0e+00	-5.0e-03	1.2e-03	-7.5e-03	3.9e-03
		3	0.0e+00	0.0e+00	-1.1e-01	0.0e+00	-8.2e-04	7.2e-04	-2.9e-03	2.8e-03
1.0e-02	1.6e-01	1	-1.6e-01	0.0e+00	-3.2e-01	0.0e+00	-3.2e-02	-1.1e-03	-2.8e-02	-4.4e-03
		2	-8.0e-02	0.0e+00	0.0e+00	0.0e+00	-4.0e-04	1.9e-04	-3.3e-04	2.9e-04
		3	0.0e+00	0.0e+00	0.0e+00	0.0e+00	-1.2e-04	7.0e-05	-3.5e-04	4.9e-04
5.0e-03	8.0e-02	1	-8.0e-02	0.0e+00	-8.0e-02	0.0e+00	-7.0e-03	-1.9e-04	-9.1e-03	1.6e-03
		2	0.0e+00	0.0e+00	0.0e+00	0.0e+00	-3.9e-05	1.8e-05	-1.4e-05	9.7e-05
		3	0.0e+00	0.0e+00	0.0e+00	0.0e+00	-1.9e-05	1.0e-05	-4.4e-05	8.0e-05
2.5e-03	4.0e-02	1	-4.0e-02	0.0e+00	-4.0e-02	0.0e+00	-1.5e-03	-2.0e-05	-2.1e-03	7.0e-05
		2	0.0e+00	0.0e+00	0.0e+00	0.0e+00	-7.6e-06	7.7e-06	-5.7e-07	2.7e-05
		3	0.0e+00	0.0e+00	0.0e+00	0.0e+00	-3.4e-06	2.1e-06	-5.6e-06	1.5e-05

(b) Phase and amplitude errors for the second vector component of the solution.

Figure 8.10: The dynamics of the approximation generated by the conservative discretisation scheme with polynomial degree $q = 1$ approximating a smooth solution with initial conditions given by (8.25).



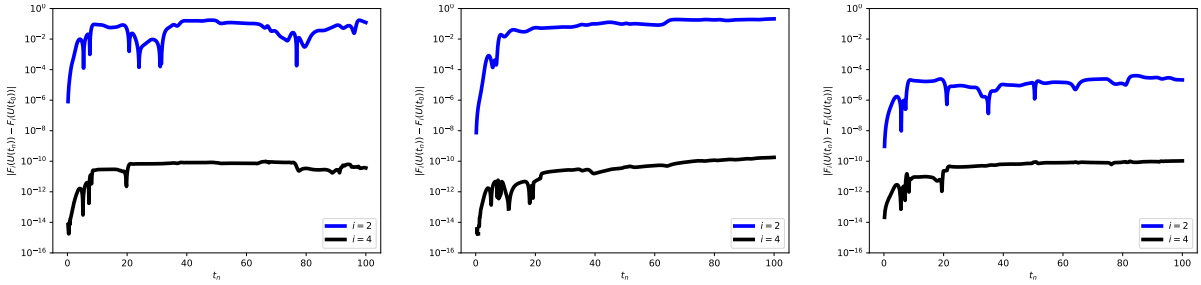
8.4.4 Test 4 - Propagation of solitary waves from smooth initial data

We take $d = 2$ and $\boldsymbol{u}_0 = (u_{0,1}, u_{0,2})$ with

$$\begin{aligned} u_{0,1} &= \sin\left(\frac{\pi}{20}x\right) \\ u_{0,2} &= \cos\left(\frac{\pi}{10}x\right). \end{aligned} \tag{8.26}$$

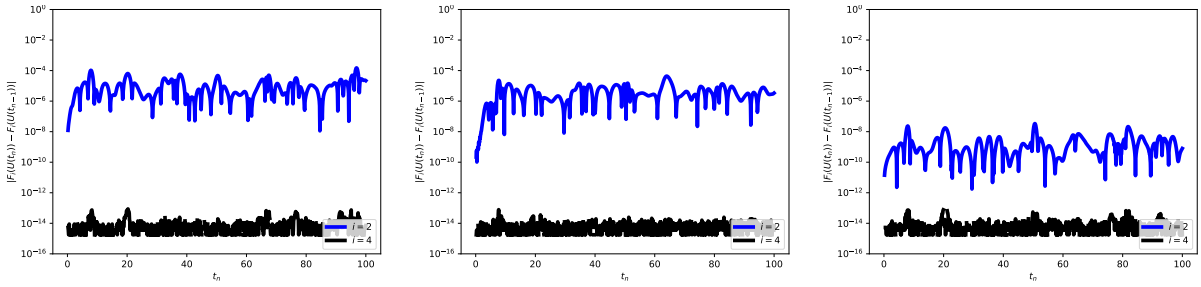
The solution here is smooth and solitary waves begin to form quickly into the simulation. Plots of the solutions are given in Figure 8.13 as well as conservativity plots in Figure 8.11.

Figure 8.11: The conservative discretisation scheme with various polynomial degrees, q, approximating the solution to (8.1) with initial conditions given by (8.26). We show the deviation in the two invariants F_i, $i = 2, 4$, corresponding to momentum and energy respectively. In each test we take a fixed spatial discretisation parameter of $h = 0.25$ and fixed time step of $\tau = 0.001$. Notice that in each case the deviation in energy is smaller than the solver tolerance of 10^{-12} and the deviation in momentum is bounded. The simulations are simulated for long time to test conservativity with $T = 100$ in each case.



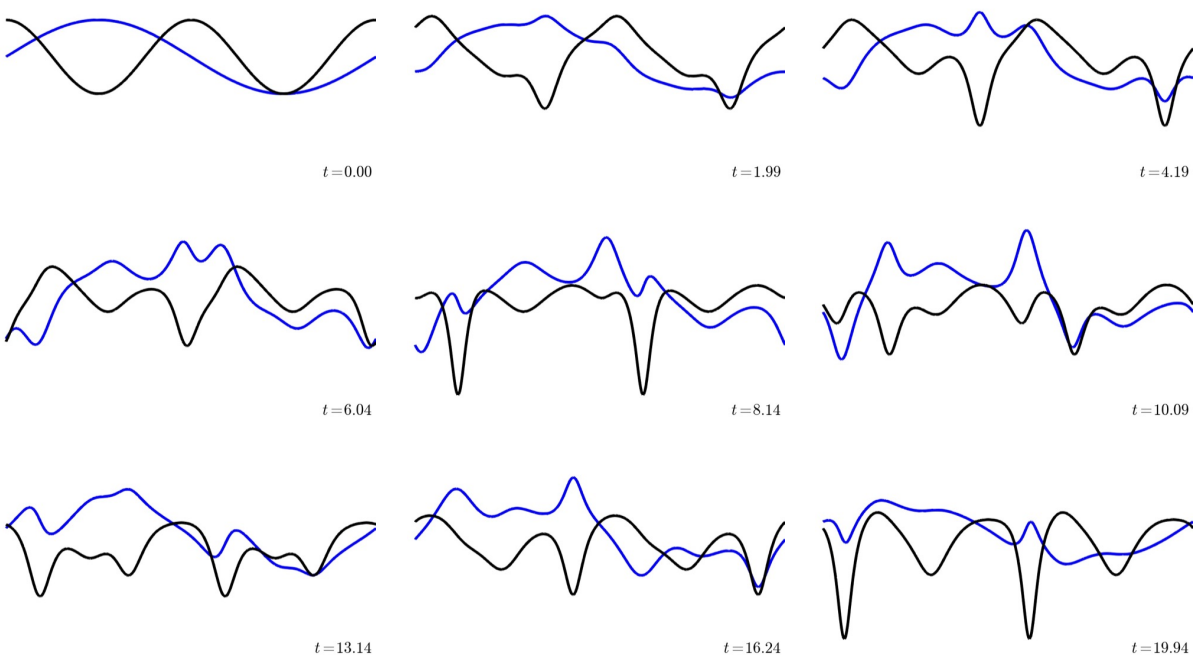
(a) $q = 1$ (b) $q = 2$ (c) $q = 3$

Figure 8.12: The conservative scheme with various polynomial degrees, q, approximating the solution to (8.1) with initial conditions given by (8.26). We now show the deviation in the two invariants F_i, $i = 2, 4$, *locally*, that is, for each n, we measure $|F_4(U^n) - F_4(U^{n-1})|$. In each test we take a fixed spatial discretisation parameter of $h = 0.25$ and fixed time step of $\tau = 0.001$. Notice that in each case the deviation in energy is smaller than the solver tolerance of 10^{-12} and the deviation in momentum is bounded. The simulations run over long time to test conservativity with $T = 100$ in each case. This result should be compared to the study of the global deviation $|F_4(U^n) - F_4(U^0)|$ given in Figure 8.11 which accumulates in time through propagation of precision errors over time.



(a) $q = 1$ (b) $q = 2$ (c) $q = 3$

Figure 8.13: Here we show the dynamics of the approximation generated by the conservative discretisation scheme with polynomial degree $q = 1$ approximating the solution to (8.1) with initial conditions given by (8.26). Notice that initially, dispersive waves emanate from the discontinuity.



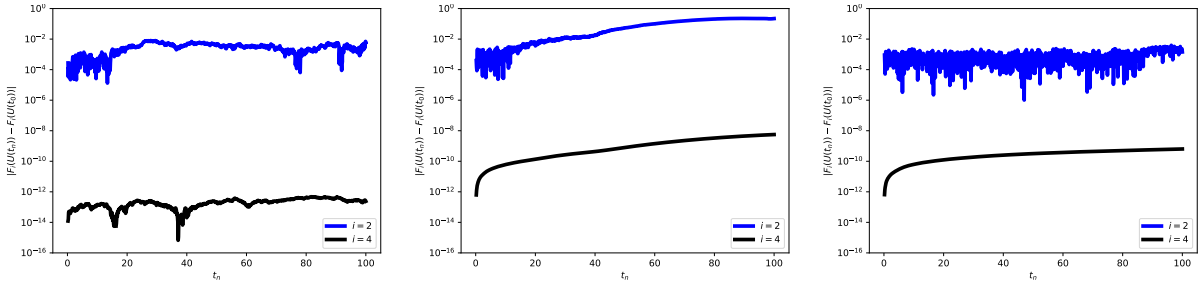
8.4.5 Test 5 - Solution with discontinuous initial data

We take $d = 2$ and $\boldsymbol{u}_0 = (u_{0,1}, u_{0,2})$ with

$$u_{0,1} = \begin{cases} 1 \text{ for } x \in [10, 20] \\ 0 \text{ otherwise.} \end{cases}$$
$$u_{0,2} = \begin{cases} 0 \text{ for } x \in [20, 30] \\ 1 \text{ otherwise.} \end{cases} \tag{8.27}$$

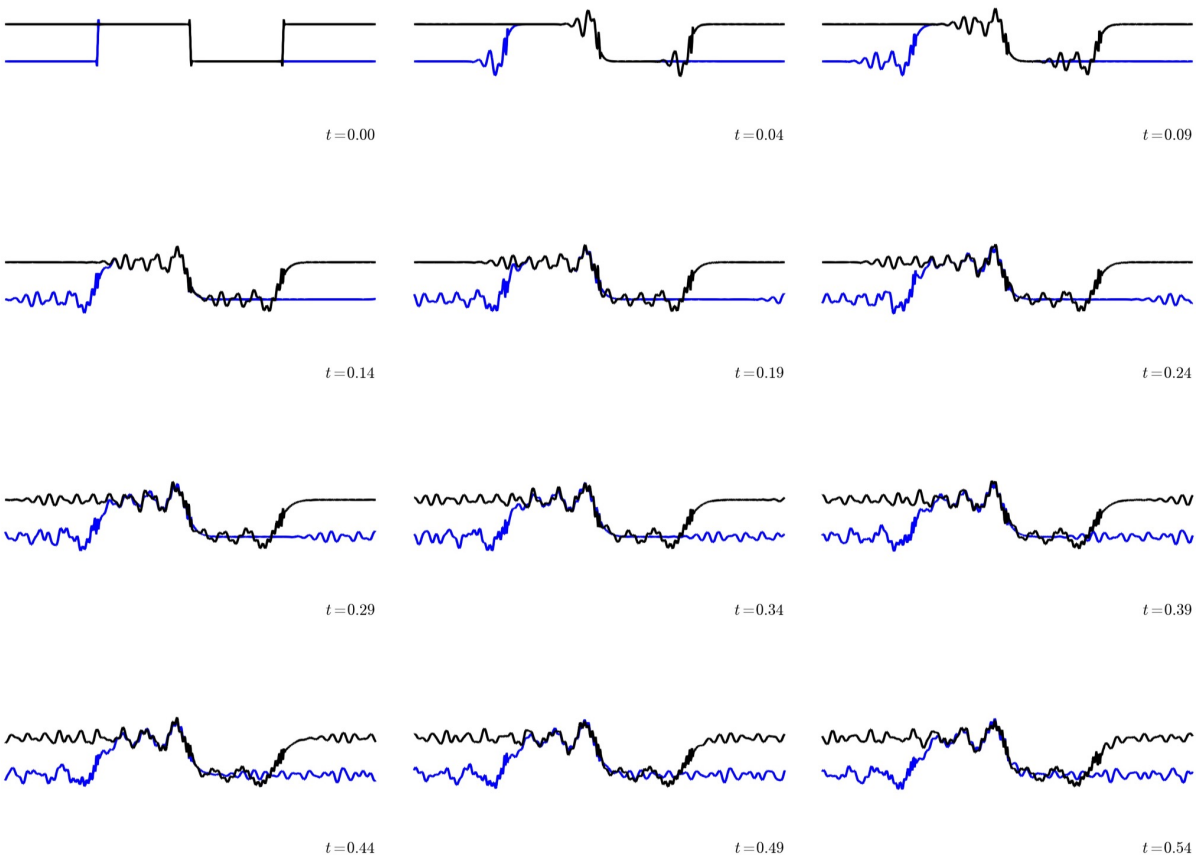
The solution here is discontinuous in both components. This is a particularly tough scenario to simulate as there is no guarantee of classical solutions. We align the mesh to the discontinuities so that the discrete energy at the initial condition makes sense. Plots of the solutions are given in Figure 8.15 as well as conservativity plots in Figure 8.14. The phenomena demonstrated in this experiment are related to the observations in [82] where, for the scalar KdV equation, arguments based on Whitham's modulation theory are presented.

Figure 8.14: Here we examine the conservative discretisation scheme with various polynomial degrees, q, approximating the solution to (8.1) with initial conditions given by (8.27). We show the deviation in the two invariants F_i, $i = 2, 4$, corresponding to momentum and energy respectively. In each test we take a fixed spatial discretisation parameter of $h = 0.25$ and fixed time step of $\tau = 0.001$. Notice that in each case the deviation in energy is smaller than the solver tolerance of 10^{-12} and the deviation in momentum is bounded. The simulations are simulated for long time to test conservativity with $T = 100$ in each case.



(a) $q = 1$ (b) $q = 2$ (c) $q = 3$

Figure 8.15: Here we show the dynamics of the approximation generated by the conservative discretisation scheme with polynomial degree $q = 1$ approximating the solution to (8.1) with initial conditions given by (8.27). Notice that initially, dispersive waves emanate from the discontinuity.



8.5 Conclusion

In this chapter we have constructed a Galerkin approximation for the vmKdV equation that is consistent with the underlying algebraic structure of the PDE. We have proven that both the semi-discretisations as well as the fully discrete problems are conservative and numerically shown that this is true in practice. In addition, we have given numerical evidence to suggest that the method is of optimal order, that is,

$$\|\boldsymbol{u} - \boldsymbol{U}\|_{L_\infty((0,T);L_2(S^1))} = \mathcal{O}(\tau^2 + h^{q+1}).$$

We expect methods designed in this fashion, which is quite generic, to be successful in the simulation of geophysical fluid flows.
Extensions to work in this chapter will be carried out by exploring applications to further systems of Hamiltonian PDEs, a thorough error analysis, and exploiting the possibility of constructing schemes with multiple conserved quantities.

Index

Bibliography

[1] K. Abe and O. Inoue, *Fourier expansion solution of the Korteweg-de Vries equation*, J. Comput. Phys., 34 (1980), pp. 202–210.

[2] M. Ablowitz and P. Clarkson, *Solitons, nonlinear evolution equations and inverse scattering*, vol. 149, Cambridge university press, 1991.

[3] M. J. Ablowitz and H. Segur, *Solitons and the inverse scattering transform*, vol. 4, SIAM, 1981.

[4] P. Adamapoulou and G. Papamikos, *On the hierarchy of the vector modified KdV equation*, In preparation, (2017).

[5] M. Ainsworth, *A posteriori error estimation for discontinuous Galerkin finite element approximation*, SIAM J. Numer. Anal., 45 (2007), pp. 1777–1798.

[6] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics, John Wiley & Sons, New York, 2000.

[7] M. E. Alexander and J. L. Morris, *Galerkin methods applied to some model equations for non-linear dispersive waves*, J. Comput. Phys., 30 (1979), pp. 428–451.

[8] M. S. Alnæs, A. Logg, K. B. Ølgaard, M. E. Rognes, and G. N. Wells, *Unified form language: a domain-specific language for weak formulations and partial differential equations*, ACM Trans. Math. Software, 40 (2014), pp. Art. 9, 37.

[9] S. C. Anco, *Hamiltonian flows of curves in $G/\mathrm{SO}(N)$ and vector soliton equations of mKdV and sine-Gordon type*, SIGMA Symmetry Integrability Geom. Methods Appl., 2 (2006), pp. Paper 044, 17.

[10] S. C. Anco, N. T. Ngatat, and M. Willoughby, *Interaction properties of complex modified korteweg–de vries (mkdv) solitons*, Physica D: Nonlinear Phenomena, 240 (2011), pp. 1378–1394.

[11] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[12] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2001/02), pp. 1749–1779.

[13] D. N. Arnold, R. S. Falk, and R. Winther, *Finite element exterior calculus, homological techniques, and applications*, Acta Numer., 15 (2006), pp. 1–155.

[14] H. Askes and A. Rodríguez-Ferran, *A combined rh-adaptive scheme based on domain subdivision. formulation and linear examples*, International Journal for numerical methods in engineering, 51 (2001), pp. 253–273.

[15] I. Babuška, *Error-bounds for finite element method*, Numerische Mathematik, 16 (1971), pp. 322–333.

[16] ——, *The finite element method with Lagrangian multipliers*, Numer. Math., 20 (1972/73), pp. 179–192.

[17] D. Bai and L. Zhang, *The finite element method for the coupled Schrödinger–KdV equations*, Physics Letters A, 373 (2009), pp. 2237–2244.

[18] M. J. Baines, M. E. Hubbard, and P. K. Jimack, *A moving mesh finite element algorithm for the adaptive solution of time-dependent partial differential equations with moving boundaries*, Appl. Numer. Math., 54 (2005), pp. 450–469.

[19] C. T. H. Baker, *The numerical treatment of integral equations*, Clarendon Press, Oxford, 1977. Monographs on Numerical Analysis.

[20] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, D. A. May, L. C. McInnes, R. T. Mills, T. Munson, K. Rupp, P. Sanan, B. F. Smith, S. Zampini, H. Zhang, and H. Zhang, *PETSc users manual*, Tech. Rep. ANL-95/11 - Revision 3.9, Argonne National Laboratory, 2018.

[21] S. Balay, W. D. Gropp, L. C. McInnes, and B. F. Smith, *Efficient management of parallelism in object oriented numerical software libraries*, in Modern Software Tools in Scientific Computing, E. Arge, A. M. Bruaset, and H. P. Langtangen, eds., Birkhäuser Press, 1997, pp. 163–202.

[22] W. Bangerth, R. Hartmann, and G. Kanschat, *deal.II—a general-purpose object-oriented finite element library*, ACM Trans. Math. Software, 33 (2007), pp. Art. 24, 27.

[23] R. E. Bank and J. Xu, *Asymptotically exact a posteriori error estimators. II. General unstructured grids*, SIAM J. Numer. Anal., 41 (2003), pp. 2313–2332 (electronic).

[24] E. Bänsch, F. Karakatsani, and C. Makridakis, *The effect of mesh modification in time on the error control of fully discrete approximations for parabolic equations*, Appl. Numer. Math., 67 (2013), pp. 35–63.

[25] F. Barros, S. Proença, and C. de Barcellos, *On error estimator and p-adaptivity in the generalized finite element method*, International Journal for Numerical Methods in Engineering, 60 (2004), pp. 2373–2398.

[26] A. Beléndez, C. Pascual, D. Méndez, T. Beléndez, and C. Neipp, *Exact solution for the nonlinear pendulum*, Revista brasileira de ensino de física, 29 (2007), pp. 645–648.

[27] R. Bellman, *The stability of solutions of linear differential equations*, Duke Math. J., 10 (1943), pp. 643–647.

[28] G. Benettin and A. Giorgilli, *On the Hamiltonian interpolation of near-to-the-identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys., 74 (1994), pp. 1117–1143.

[29] M. Berzins, P. J. Capon, and P. K. Jimack, *On spatial adaptivity and interpolation when using the method of lines*, Applied Numerical Mathematics, 26 (1998), pp. 117 – 133.

[30] J. Bona, V. Dougalis, and D. Mitsotakis, *Numerical solution of KdV–KdV systems of Boussinesq equations: I. The numerical scheme and generalized solitary waves*, Mathematics and Computers in Simulation, 74 (2007), pp. 214–228.

[31] J. L. Bona, H. Chen, O. Karakashian, and Y. Xing, *Conservative, discontinuous Galerkin-methods for the generalized Korteweg-de Vries equation*, Math. Comp., 82 (2013), pp. 1401–1432.

[32] J. L. Bona, V. A. Dougalis, O. A. Karakashian, and W. R. McKinney, *Conservative, high-order numerical schemes for the generalized Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London Ser. A, 351 (1995), pp. 107–164.

[33] S. Brenner and R. Scott, *The mathematical theory of finite element methods*, vol. 15, Springer Science & Business Media, 2007.

[34] T. J. Bridges and S. Reich, *Multi-symplectic integrators: numerical schemes for Hamiltonian PDEs that conserve symplecticity*, Phys. Lett. A, 284 (2001), pp. 184–193.

[35] C. J. Budd, W. Huang, and R. D. Russell, *Adaptivity with moving grids*, Acta Numer., 18 (2009), pp. 111–241.

[36] C. J. Budd and J. F. Williams, *Moving mesh generation using the parabolic Monge-Ampère equation*, SIAM J. Sci. Comput., 31 (2009), pp. 3438–3465.

[37] J. C. Butcher, *A stability property of implicit Runge-Kutta methods*, BIT Numerical Mathematics, 15 (1975), pp. 358–361.

[38] ——, *A history of Runge-Kutta methods*, Appl. Numer. Math., 20 (1996), pp. 247–260. Selected keynote papers presented at 14th IMACS World Congress (Atlanta, GA, 1994).

[39] ——, *Numerical methods for ordinary differential equations*, John Wiley & Sons, Ltd., Chichester, third ed., 2016. With a foreword by J. M. Sanz-Serna.

[40] M. P. Calvo, A. Iserles, and A. Zanna, *Numerical solution of isospectral flows*, Math. Comp., 66 (1997), pp. 1461–1486.

[41] W. Cao, W. Huang, and R. D. Russell, *An r-adaptive finite element method based upon moving mesh PDEs*, J. Comput. Phys., 149 (1999), pp. 221–244.

[42] E. Celledoni, R. I. McLachlan, D. I. McLaren, B. Owren, G. R. W. Quispel, and W. M. Wright, *Energy-preserving Runge-Kutta methods*, M2AN Math. Model. Numer. Anal., 43 (2009), pp. 645–649.

[43] Y. Chen, B. Cockburn, and B. Dong, *Superconvergent HDG methods for linear, stationary, third-order equations in one-space dimension*, Math. Comp. AMS, (2016). Published electronically.

[44] R. W. Clough, *The finite element method in plane stress analysis*, in Proceedings of 2nd ASCE Conference on Electronic Computation, Pittsburgh Pa., 1960.

[45] B. Cockburn, *Continuous dependence and error estimation for viscosity methods*, Acta Numer., 12 (2003), pp. 127–180.

[46] B. Cockburn, M. Luskin, C.-W. Shu, and E. Süli, *Enhanced accuracy by post-processing for finite element methods for hyperbolic equations*, Math. Comp., 72 (2003), pp. 577–606.

[47] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework*, Math. Comp., 52 (1989), pp. 411–435.

[48] D. Cohen, T. Matsuo, and X. Raynaud, *A multi-symplectic numerical integrator for the two-component Camassa-Holm equation*, J. Nonlinear Math. Phys., 21 (2014), pp. 442–453.

[49] D. Cohen, B. Owren, and X. Raynaud, *Multi-symplectic integration of the Camassa-Holm equation*, J. Comput. Phys., 227 (2008), pp. 5492–5512.

[50] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.

[51] K. Deckelnick, G. Dziuk, and C. M. Elliott, *Computation of geometric partial differential equations and mean curvature flow*, Acta Numer., 14 (2005), pp. 139–232.

[52] A. Dedner, B. Kane, R. Klöfkorn, and M. Nolte, *Python framework for hp adaptive discontinuous Galerkin method for two phase flow in porous media*, arXiv preprint arXiv:1805.00290, (2018).

[53] A. Dedner and M. Ohlberger, *A new hp-adaptive DG scheme for conservation laws based on error control*, in Hyperbolic problems: theory, numerics, applications, Springer, Berlin, 2008, pp. 187–198.

[54] L. Demkowicz, J. Oden, and W. Rachowicz, *A new finite element method for solving compressible navier-stokes equations based on an operator splitting method and hp adaptivity*, Computer methods in applied mechanics and engineering, 84 (1990), pp. 275–326.

[55] L. Demkowicz, W. Rachowicz, and P. Devloo, *A fully automatic hp-adaptivity*, in Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala), vol. 17, 2002, pp. 117–142.

[56] P. Deuflhard, P. Leinen, and H. Yserentant, *Concepts of an adaptive hierarchical finite element code*, IMPACT of Computing in Science and Engineering, 1 (1989), pp. 3–35.

[57] P. Díez and A. Huerta, *A unified approach to remeshing strategies for finite element h-adaptivity*, Computer Methods in Applied Mechanics and Engineering, 176 (1999), pp. 215–229.

[58] Z. Dong, E. H. Georgoulis, and T. Pryer, *Recovered finite element methods on polygonal and polyhedral meshes*, arXiv preprint arXiv:1804.08259, (2018).

[59] W. Dörfler, *A time and space adaptive algorithm for the linear time-dependent schrödinger equation*, Numerische Mathematik, 73 (1996), pp. 419–448.

[60] J. Douglas, Jr. and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[61] M. Dumbser, M. Kaser, and E. F. Toro, *An arbitrary high-order discontinuous galerkin method for elastic waves on unstructured meshes-v. local time stepping and p-adaptivity*, Geophysical Journal International, 171 (2007), pp. 695–717.

[62] T. Dupont, *Mesh modification for evolution equations*, Math. Comp., 39 (1982), pp. 85–107.

[63] S. l. Eidnes, B. Owren, and T. r. Ringholm, *Adaptive energy preserving methods for partial differential equations*, Adv. Comput. Math., 44 (2018), pp. 815–839.

[64] M. Elliotis, G. Georgiou, and C. Xenophontos, *Solving Laplacian problems with boundary singularities: a comparison of a singular function boundary integral method with the p/hp version of the finite element method*, Appl. Math. Comput., 169 (2005), pp. 485–499.

[65] A. Ern and J.-L. Guermond, *Theory and practice of finite elements*, vol. 159 of Applied Mathematical Sciences, Springer-Verlag, New York, 2004.

[66] D. Estep, *A posteriori error bounds and global error control for approximation of ordinary differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1–48.

[67] D. Estep and D. French, *Global error control for the continuous Galerkin finite element method for ordinary differential equations*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 815–852.

[68] D. J. Estep and A. M. Stuart, *The dynamical behavior of the discontinuous Galerkin method and related difference schemes*, Math. Comp., 71 (2002), pp. 1075–1103.

[69] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[70] L. Faddeev and L. Takhtajan, *Hamiltonian methods in the theory of solitons*, Springer Science & Business Media, 2007.

[71] J. E. Flaherty and P. K. Moore, *Integrated space-time adaptive hp-refinement methods for parabolic systems*, Appl. Numer. Math., 16 (1995), pp. 317–341.

[72] H. Flanders, *Differential forms with applications to the physical sciences*, Dover Books on Advanced Mathematics, Dover Publications, Inc., New York, second ed., 1989.

[73] D. A. French and J. W. Schaeffer, *Continuous finite element methods which preserve energy properties for nonlinear problems*, Appl. Math. Comput., 39 (1990), pp. 271–295.

[74] E. H. Georgoulis and T. Pryer, *Recovered finite element methods*, Comput. Methods Appl. Mech. Engrg., 332 (2018), pp. 303–324.

[75] J. Giesselmann, C. Makridakis, and T. Pryer, *Energy consistent discontinuous Galerkin methods for the Navier-Stokes-Korteweg system*, Math. Comp., 83 (2014), pp. 2071–2099.

[76] ——, *A posteriori analysis of discontinuous Galerkin schemes for systems of hyperbolic conservation laws*, SIAM J. Numer. Anal., 53 (2015), pp. 1280–1303.

[77] J. Giesselmann and T. Pryer, *Energy consistent discontinuous Galerkin methods for a quasi-incompressible diffuse two phase flow model*, ESAIM Math. Model. Numer. Anal., 49 (2015), pp. 275–301.

[78] ——, *Reduced relative entropy techniques for a priori analysis of multiphase problems in elastodynamics*, BIT, 56 (2016), pp. 99–127.

[79] ——, *A posteriori analysis for dynamic model adaptation in convection-dominated problems*, Mathematical Models and Methods in Applied Sciences, 27 (2017), pp. 2381–2423.

[80] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, Classics in Mathematics, Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[81] O. Gonzalez, *Time integration and discrete Hamiltonian systems*, J. Nonlinear Sci., 6 (1996), pp. 449–467.

[82] T. Grava, *Whitham modulation equations and application to small dispersion asymptotics and long time asymptotics of nonlinear dispersive equations*, in Rogue and Shock Waves in Nonlinear Dispersive Media, Springer, 2016, pp. 309–335.

[83] T. H. Gronwall, *Note on the derivatives with respect to a parameter of the solutions of a system of differential equations*, Ann. of Math. (2), 20 (1919), pp. 292–296.

[84] A. Guillou and J. L. Soulé, *La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation*, Rev. Française Informat. Recherche Opérationnelle, 3 (1969), pp. 17–44.

[85] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration*, vol. 31 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006. Structure-preserving algorithms for ordinary differential equations.

[86] P. C. Hammer and J. W. Hollingsworth, *Trapezoidal methods of approximating solutions of differential equations*, Math. Tables Aids Comput., 9 (1955), pp. 92–96.

[87] P. Hansbo, *A note on energy conservation for hamiltonian systems using continuous time finite elements*, Commun. Numer. Meth. Engng., 17 (2001), pp. 863–869.

[88] A. Harten, *High resolution schemes for hyperbolic conservation laws [MR0701178 (84g:65115)]*, J. Comput. Phys., 135 (1997), pp. 259–278. With an introduction by Peter Lax, Commemoration of the 30th anniversary {of J. Comput. Phys.}.

[89] A. Harten, P. D. Lax, and B. van Leer, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Rev., 25 (1983), pp. 35–61.

[90] M. Hénon and C. Heiles, *The applicability of the third integral of motion: Some numerical experiments*, Astronom. J., 69 (1964), pp. 73–79.

[91] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier-Stokes problem. I. Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.

[92] R. Hirota, *Exact envelope-soliton solutions of a nonlinear wave equation*, Journal of Mathematical Physics, 14 (1973), pp. 805–809.

[93] R. Hirota, *The direct method in soliton theory*, vol. 155, Cambridge University Press, 2004.

[94] H. Hofer and E. Zehnder, *Symplectic invariants and Hamiltonian dynamics*, Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks], Birkhäuser Verlag, Basel, 1994.

[95] D. D. Holm, T. Schmah, and C. Stoica, *Geometric mechanics and symmetry*, vol. 12 of Oxford Texts in Applied and Engineering Mathematics, Oxford University Press, Oxford, 2009. From finite to infinite dimensions, With solutions to selected exercises by David C. P. Ellis.

[96] A. Hrennikoff, *Solution of problems of elasticity by the framework method*, J. Appl. Mech., 8 (1941), pp. A–169–A–175.

[97] A. Huerta, A. Rodríguez-Ferran, P. Díez, and J. Sarrate, *Adaptive finite element strategies based on error assessment*, International Journal for Numerical Methods in Engineering, 46 (1999), pp. 1803–1818.

[98] C. Hufford and Y. Xing, *Superconvergence of the local discontinuous Galerkin method for the linearized Korteweg-de Vries equation*, J. Comput. Appl. Math., 255 (2014), pp. 441–455.

[99] T. J. R. Hughes and G. M. Hulbert, *Space-time finite element methods for elastodynamics: formulations and error estimates*, Comput. Methods Appl. Mech. Engrg., 66 (1988), pp. 339–363.

[100] T. J. R. Hughes, W. K. Liu, and T. K. Zimmermann, *Lagrangian-Eulerian finite element formulation for incompressible viscous flows*, Comput. Methods Appl. Mech. Engrg., 29 (1981), pp. 329–349.

[101] A. Iserles, *Composite methods for numerical solution of stiff systems of ODEs*, SIAM J. Numer. Anal., 21 (1984), pp. 340–351.

[102] ——, *A first course in the numerical analysis of differential equations*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, second ed., 2009.

[103] T. Itoh and K. Abe, *Hamiltonian-conserving discrete canonical equations based on variational difference quotients*, J. Comput. Phys., 76 (1988), pp. 85–102.

[104] J. Jackaman, *Geometric integration: A brief overview*, Master's thesis, University of Kent, 2014.

[105] ——, *Finite element methods as geometric integrators for Hamiltonian initial value problems*, Master's thesis, University of Reading and Imperial College London, 2015.

[106] J. Jackaman, *A conservative Galerkin method for the vectorial modified Korteweg deVries equation*, in `http://dx.doi.org/10.5281/zenodo.600668`, 2017.

[107] J. Jackaman, G. Papamikos, and T. Pryer, *The design of conservative finite element discretisations for the vectorial modified kdv equation*, Applied Numerical Mathematics, (2018).

[108] J. Jackaman and T. Pryer, *Invariant preserving schemes for the kdv equation*, In preparation.

[109] ——, *Recovered finite element methods in time*, In preparation.

[110] ——, *Conservative galerkin methods for dispersive hamiltonian problems*, arXiv preprint arXiv:1811.09999, (2018).

[111] A. Jameson, W. Schmidt, and E. Turkel, *Numerical solution of the Euler equations by finite volume methods using Runge Kutta time stepping schemes*, in 14th fluid and plasma dynamics conference, 1981, p. 1259.

[112] B. Jeremić and C. Xenophontos, *Application of the p-version of the finite element method to elastoplasticity with localization of deformation*, Comm. Numer. Methods Engrg., 15 (1999), pp. 867–876.

[113] C. Johnson, *Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 908–926.

[114] O. Karakashian and C. Makridakis, *A posteriori error estimates for discontinuous Galerkin methods for the generalized Korteweg–de Vries equation*, Math. Comp., 84 (2015), pp. 1145–1167.

[115] O. Karakashian and W. McKinney, *On optimal high-order in time approximations for the Korteweg-de Vries equation*, Math. Comp., 55 (1990), pp. 473–496.

[116] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.

[117] P. Kersten, I. S. Krasil'shchik, A. M. Verbovetsky, and R. Vitolo, *Hamiltonian Structures for General PDEs*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 187–198.

[118] E. Kesici, B. Pelloni, T. Pryer, and D. Smith, *A numerical implementation of the unified Fokas transform for evolution problems on a finite interval*, European J. Appl. Math., 29 (2018), pp. 543–567.

[119] D. Kordeweg and G. de Vries, *On the change of form of long waves advancing in a rectangular channel, and a new type of long stationary wave*, Philos. Mag, 39 (1895), pp. 422–443.

[120] F. Kretzschmar, A. Moiola, I. Perugia, and S. M. Schnepp, *A priori error analysis of space-time Trefftz discontinuous Galerkin methods for wave problems*, IMA J. Numer. Anal., 36 (2016), pp. 1599–1635.

[121] O. Lakkis and C. Makridakis, *Elliptic reconstruction and a posteriori error estimates for fully discrete linear parabolic problems*, Math. Comp., 75 (2006), pp. 1627–1658 (electronic).

[122] O. Lakkis, C. Makridakis, and T. Pryer, *A comparison of duality and energy a posteriori estimates for $l_\infty(0,t;l_2(\omega))$ in parabolic problems*, Mathematics of Computation, 84 (2015), pp. 1537–1569.

[123] O. Lakkis and T. Pryer, *Gradient recovery in adaptive finite-element methods for parabolic problems*, IMA Journal of Numerical Analysis, 32 (2011), pp. 246–278.

[124] P. D. Lax, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467–490.

[125] ——, *Almost periodic solutions of the KdV equation*, SIAM Rev., 18 (1976), pp. 351–375.

[126] T. E. Lee, M. J. Baines, and S. Langdon, *A finite difference moving mesh method based on conservation for moving boundary problems*, J. Comput. Appl. Math., 288 (2015), pp. 1–17.

[127] B. Leimkuhler and S. Reich, *Simulating Hamiltonian dynamics*, vol. 14 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2004.

[128] H. Liu and J. Yan, *A local discontinuous Galerkin method for the Korteweg-de Vries equation with boundary effect*, J. Comput. Phys., 215 (2006), pp. 197–218.

[129] M. Luskin and R. Rannacher, *On the smoothing property of the Crank-Nicolson scheme*, Applicable Anal., 14 (1982/83), pp. 117–135.

[130] R. S. MacKay and J. D. Meiss, *Hamiltonian Dynamical Systems*, CRC Press, 1987.

[131] F. Magri, *A simple model of the integrable hamiltonian equation*, Journal of Mathematical Physics, 19 (1978), pp. 1156–1162.

[132] C. Makridakis and R. H. Nochetto, *Elliptic reconstruction and a posteriori error estimates for parabolic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 1585–1594 (electronic).

[133] ——, *A posteriori error analysis for higher order dissipative methods for evolution problems*, Numer. Math., 104 (2006), pp. 489–514.

[134] G. Mari Beffa, J. A. Sanders, and J. P. Wang, *Integrable systems in three-dimensional riemannian geometry*, Journal of nonlinear science, 12 (2002), pp. 143–167.

[135] J. E. Marsden, G. W. Patrick, and S. Shkoller, *Multisymplectic geometry, variational integrators, and nonlinear PDEs*, Comm. Math. Phys., 199 (1998), pp. 351–395.

[136] R. I. McLachlan, *Composition methods in the presence of small parameters*, BIT, 35 (1995), pp. 258–268.

[137] R. I. McLachlan, G. R. W. Quispel, and N. Robidoux, *Geometric integration using discrete gradients*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 1021–1045.

[138] R. I. McLachlan and A. Stern, *Multisymplecticity of hybridizable discontinuous galerkin methods*, arXiv preprint arXiv:1705.08609, (2017).

[139] D. I. McLaren and G. R. W. Quispel, *Integral-preserving integrators*, J. Phys. A, 37 (2004), pp. L489–L495.

[140] D. Meidner and B. Vexler, *Adaptive space-time finite element methods for parabolic optimization problems*, SIAM J. Control Optim., 46 (2007), pp. 116–142.

[141] H. Mirzaee, L. Ji, J. K. Ryan, and R. M. Kirby, *Smoothness-increasing accuracy-conserving (SIAC) postprocessing for discontinuous Galerkin solutions over structured triangular meshes*, SIAM J. Numer. Anal., 49 (2011), pp. 1899–1920.

[142] R. M. Miura, *Korteweg-de Vries equation and generalizations. I. a remarkable explicit nonlinear transformation*, Journal of Mathematical Physics, 9 (1968), pp. 1202–1204.

[143] Y. Miyatake and T. Matsuo, *A note on the adaptive conservative/dissipative discretization for evolutionary partial differential equations*, J. Comput. Appl. Math., 274 (2015), pp. 79–87.

[144] P. Müller, C. Garrett, and A. Osborne, *Rogue waves*, Oceanography, 18 (2005), pp. 66–75.

[145] E. Noether, *Invariant variation problems*, Transport Theory Statist. Phys., 1 (1971), pp. 186–207. Translated from the German (Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II 1918, 235–257).

[146] S. Novikov, S. Manakov, L. Pitaevskii, and V. Zakharov, *Theory of solitons: the inverse scattering method*, Springer Science & Business Media, 1984.

[147] P. J. Olver, *Applications of Lie groups to differential equations*, vol. 107 of Graduate Texts in Mathematics, Springer-Verlag, New York, second ed., 1993.

[148] ——, *Dispersive quantization*, Amer. Math. Monthly, 117 (2010), pp. 599–610.

[149] J. A. Pava, J. L. Bona, M. Scialom, et al., *Stability of cnoidal waves*, Advances in Differential Equations, 11 (2006), pp. 1321–1374.

[150] T. Pryer, *Applications of nonvariational finite element methods to Monge-Ampère type equations*, in Numerical mathematics and advanced applications 2011, Springer, Heidelberg, 2013, pp. 441–448.

[151] T. Pryer, *Discontinuous Galerkin methods for the p-biharmonic equation from a discrete variational perspective*, Electron. Trans. Numer. Anal., 41 (2014), pp. 328–349.

[152] G. R. W. Quispel and G. S. Turner, *Discrete gradient methods for solving ODEs numerically while preserving a first integral*, J. Phys. A, 29 (1996), pp. L341–L349.

[153] F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. T. McRae, G.-T. Bercea, G. R. Markall, and P. H. J. Kelly, *Firedrake: automating the finite element method by composing abstractions*, ACM Trans. Math. Software, 43 (2017), pp. Art. 24, 27.

[154] C. Rogers and W. Schief, *Bäcklund and Darboux transformations: geometry and modern applications in soliton theory*, vol. 30, Cambridge University Press, 2002.

[155] I. Roulstone and J. Norbury, *A Hamiltonian structure with contact geometry for the semi-geostrophic equations*, J. Fluid Mech., 272 (1994), pp. 211–233.

[156] M. Salle and V. Matveev, *Darboux transformations and solitons.*, 1991.

[157] P. H. Sammon, *Convergence estimates for semidiscrete parabolic equation approximations*, SIAM J. Numer. Anal., 19 (1982), pp. 68–92.

[158] J. A. Sanders and J. P. Wang, *Integrable systems in n-dimensional riemannian geometry*, Moscow Mathematical Journal, 3 (2003), pp. 1369–1393.

[159] J. M. Sanz-Serna, *An explicit finite-difference scheme with exact conservation properties*, J. Comput. Phys., 47 (1982), pp. 199–210.

[160] ——, *Runge-Kutta schemes for Hamiltonian systems*, BIT, 28 (1988), pp. 877–883.

[161] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian problems*, vol. 7 of Applied Mathematics and Mathematical Computation, Chapman & Hall, London, 1994.

[162] J.-M. Sanz-Serna and M.-P. Calvo, *Numerical hamiltonian problems*, Courier Dover Publications, 2018.

[163] H. Schamel and K. Elsässer, *The application of the spectral method to nonlinear wave propagation*, J. Computational Phys., 22 (1976), pp. 501–516.

[164] M. Schmich and B. Vexler, *Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations*, SIAM J. Sci. Comput., 30 (2007/08), pp. 369–393.

[165] A. Schmidt and K. G. Siebert, *Design of adaptive finite element software*, vol. 42 of Lecture Notes in Computational Science and Engineering, Springer-Verlag, Berlin, 2005. The finite element toolbox ALBERTA, With 1 CD-ROM (Unix/Linux).

[166] T. Shepherd, *Symmetries, Conservation Laws, and Hamiltonian Structure in Geophysical Fluid Dynamics*, Advances in Geophysics, 32 (1990), pp. 287 – 338.

[167] J. Simo, N. Tarnow, and K. Wong, *Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics*, Computer Methods in Applied Mechanics and Engineering, 100 (1992), pp. 63 – 116.

[168] P. Šolín, J. Červený, and I. Doležel, *Arbitrary-level hanging nodes and automatic adaptivity in the hp-FEM*, Math. Comput. Simulation, 77 (2008), pp. 117–132.

[169] P. Solín and L. Demkowicz, *Goal-oriented hp-adaptivity for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 449–468.

[170] M. Spivak, *A comprehensive introduction to differential geometry. Vol. I*, Publish or Perish, Inc., Wilmington, Del., second ed., 1979.

[171] G. Strang and G. J. Fix, *An analysis of the finite element method*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973. Prentice-Hall Series in Automatic Computation.

[172] E. Tadmor, *The numerical viscosity of entropy stable schemes for systems of conservation laws. I*, Math. Comp., 49 (1987), pp. 91–103.

[173] T. E. Tezduyar, *Stabilized finite element formulations for incompressible flow computations*, in Advances in applied mechanics, Vol. 28, vol. 28 of Adv. Appl. Mech., Academic Press, Boston, MA, 1992, pp. 1–44.

[174] T. E. Tezduyar, S. Sathe, R. Keedy, and K. Stein, *Space-time finite element techniques for computation of fluid-structure interactions*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2002–2027.

[175] V. Thomée, *Negative norm estimates and superconvergence in Galerkin methods for parabolic problems*, Math. Comp., 34 (1980), pp. 93–113.

[176] V. Thomée, *Galerkin finite element methods for parabolic problems*, vol. 25 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[177] G. Vallis, *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation*, Cambridge Univ. Press, Cambridge, 2006.

[178] R. Verfürth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Math. Comp., 62 (1994), pp. 445–475.

[179] G. Wanner, *Runge-Kutta-methods with expansion in even powers of h*, Computing (Arch. Elektron. Rechnen), 11 (1973), pp. 81–85.

[180] R. Winther, *A conservative finite element method for the Korteweg-de Vries equation*, Math. Comp., 34 (1980), pp. 23–43.

[181] K. Wright, *Some relationships between implicit Runge-Kutta, collocation Lanczos τ methods, and their stability properties*, Nordisk Tidskr. Informationsbehandling (BIT), 10 (1970), pp. 217–227.

[182] Y. Xu and C.-W. Shu, *Error estimates of the semi-discrete local discontinuous Galerkin method for nonlinear convection-diffusion and KdV equations*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3805–3822.

[183] J. Yan and C.-W. Shu, *A local discontinuous Galerkin method for KdV type equations*, SIAM J. Numer. Anal., 40 (2002), pp. 769–791 (electronic).

[184] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A, 150 (1990), pp. 262–268.

[185] G. Zhong and J. E. Marsden, *Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators*, Phys. Lett. A, 133 (1988), pp. 134–139.

[186] J. Zhu and O. Zienkiewicz, *Adaptive techniques in the finite element method*, Communications in applied numerical methods, 4 (1988), pp. 197–204.

[187] O. C. Zienkiewicz and J. Z. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, Internat. J. Numer. Methods Engrg., 24 (1987), pp. 337–357.