Solving Korteweg-de Vries equations with Discontinuous Galerkin methods

Department of Mathematics, Linköping University

Markus Henrik Hellberg Mediaa

LITH-MAT-EX–2023/02–SE

Credits:	**30 hp**
Level:	**A**
Supervisor:	**Andrew Ross Winters**, Department of Mathematics, Linköping University
Examiner:	**Fredrik Berntsson**, Department of Mathematics, Linköping University
Linköping:	**March 2023**

Abstract

In this thesis the Discontinuous Galerkin approximation performance applied to the Korteweg–de Vries equation is investigated. This equation is nonlinear with a third spatial derivative and can be used for shallow water movement. The thesis includes a background in numerical methods on conservation laws, Discontinuous Galerkin methods and the Korteweg-de Vries equation. To approximate the third order derivative, the thesis reformulates Korteweg-de Vries equation as a system of first order equations in order to apply Discontinuous Galerkin effectively. The thesis presents two choices of numerical fluxes, central- and alternating flux which both show promising convergence results, similar to that of the first order problem. Stability of the numerical approximation is proved analytically while convergence is shown numerically. The central flux appear to have spectral convergence, $\mathcal{O}(h^m)$ for even approximation order m and sub-optimal convergence for m odd while alternating flux shows spectral convergence for all approximation orders, h being the discretization mesh. However, the central flux is found, in practice, to be only half as stiff and thus one should choose the numerical flux by the problem at hand.

Keywords:

Discontinuous Galerkin, Korteweg-de Vries, Conservation Laws, Partial Differential Equations, Numerical methods, Higher order spatial derivative, Heat Equation, Burgers' Equation, Collocation, Lobatto Gauss Legendre, Stiffness, Nodal Approximation, Finite Volume, Finite Element.

Acknowledgements

This thesis was written largely on the road and the homes of friends and family. The completion lies on the support of these people, thank you Jonatan, Saewon, Ellen, Kalle, Linda, Martin, Linnea, Mathias, Hanna, Arvid, Niclas, Erik, Gerson, Christoffer, Andreas and family Melvin, Maja, Tone and Ulf. I also want to thank my co-authors and dogs Nala and Zuri who has been very good girls.

Lastly I wish to thank my supervisor Andrew for his expertise, availability and guidance.

Nomenclature

- $\mathbf{R}^{\mathcal{N}}$ denotes the Euclidian $\mathcal{N}$-space, $||x|| = (\sum_j x_j^2)^{\frac{1}{2}}$
- $\mathcal{L}^p$ denotes the space of $\mathcal{L}^p(\Omega)$ integrable functions, $||x(t)||_{\mathcal{L}^p(\Omega)} = (\int_\Omega |x(t)|^2\, dt)^{\frac{1}{2}}$. Specifically the $\mathcal{L}^2$ space has a inner norm defined by $(x, y)_\Omega = \int_\Omega x(t)y(t)\, dt$
- $\mathcal{H}^q$ denotes the space of functions for which the Sobolev norm is bounded, i.e. $||x||_q^2 = \sum_{l=0}^q ||x^{(l)}||_\Omega^2 < \infty$.
- KdV - Korteweg-de Vries
- PDE - Partial Differential Equation
- FE - Finite Element
- FD - Finite Difference
- FV - Finite Volume
- DG - Discontinuous Galerkin

Contents

Chapter 1

Introduction

Consider a physical quantity u which cannot be compressed and do not emerge or disappear from anywhere but the boundary flux. Such a quantity abides by a conservation law, meaning that the total quantity do not change as time progresses without transportation to or from the space in which it resides. The thesis will be limited to one spatial dimensional $u = u(x,t)$ where $(x,t) \in \{\Omega_x \otimes \Omega_t\}$. Mathematically, we describe this conservation law as

$$\int_{\Omega_x} u(x,t)\,dx = C, \quad \forall t \in \Omega_t, \forall x \in \Omega_x. \tag{1.1}$$

Conservation laws are present in fluid dynamics as conservation of momentum or energy systems with energy conservation. Through the language of mathematics we can interpret physical models, like those regarding conservation laws, and express them as a Partial Differential Equation (PDE). The conservation laws of interest for this thesis describe conservation of mass or momentum and are associated with hyperbolic PDEs [8]. This class of PDEs differ from elliptic and parabolic PDEs in that perturbations in the initial data does not immediately affect all points of the domain and is thus capable of describing waves of information. The simplest hyperbolic PDE is the linear advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial au}{\partial x} = 0, a \in \mathbb{R} \tag{1.2}$$

where $f(u) = au$ is the flux and the wavespeed is the result of differentiating the flux with respect to the spatial term. For the simple PDE (1.2) the value of the wave speed is the constant a. To learn more about the method of characteristics, see Hesthaven's Numerical methods to conservation laws [9]. The

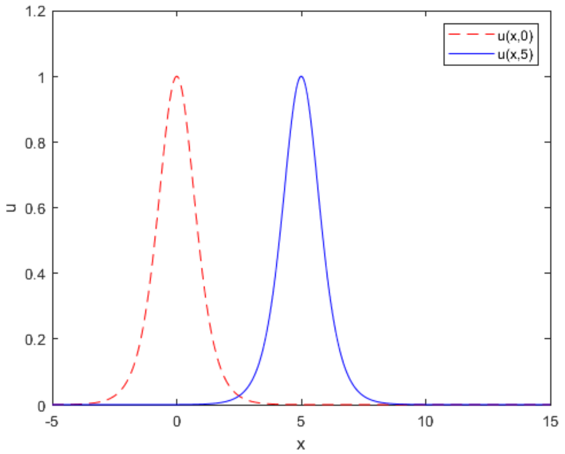

Figure 1.1: Dashed line is the initial condition and solid is the result after 5 seconds of advection transportation. Linearity of the PDE preserves the initial condition's shape.

sign of the wave speed, $a > 0$ or $a < 0$ corresponds to a right and left moving wave respectively. An example solution to the PDE (1.2) can be found by the method of characteristics with initial conditions $u(x, 0)$ =sech$^2(x)$ yields the exact solution $u(x, t)$ =sech$^2(x - at)$ [12]. This sample solution is plotted in Figure 1.1 with speed speed $a = 1$. The initial condition of with hyperbolic secant function, sech, was chosen arbitrarily but it is informative because the resulting solution resembles a propagating wave. The example illustrates a phenomenom due to the flux linearity where the solution is the initial condition, but transported with speed a. This only holds for linear advection equations but highlights what can be expected from linearity.

With some physical context for PDEs, it is now time to describe the equation at the core of this thesis, Korteweg-de Vries (KdV) equation. It is a nonlinear PDE with a third spatial derivate. Solutions to the PDE belongs to the classification solitons as it is a remarkably stable localized wave solution (1.3).

$$u_t + u_{xxx} + 6uu_x = 0, \quad x \in \mathbf{R}. \tag{1.3}$$

The soliton was first observed in 1834 by J.S Russel who studied waves of water in a narrow channel which preserves their shape and appearence as they move over large distances and away from its heap of water, there is no elevation [12].

The same equation (1.3) has also emerged in theory of plasmas and quantum mechanics. An exact solution is $u(x,t) = \frac{1}{2}a\text{sech}^2(x - at)$, where sech is the hyperbolic secant which decays exponentially as $x \to \pm\infty$. The existence of this well behaved exact solution will later allow precise convergence tests of the approximate solution u_h. This behaviour is further examined in the article by Walter Strauss [12]. The wave speed is a and amplitude is $a/2$. There is a soliton for every $a > 0$ and its appearence is decided by a. If a is large the soliton will be tall, fast and thin while a small a will yield a short, slow and wide soliton [14]. A solution of particular interest for the KdV equation has linear behaviour as two waves interacting. If a taller wave overtakes a smaller wave, they will have a very nonlinear interaction but emerge identical as they started. Apart from a small time delay they behave as if they were linear. Walter Strauss [12] mentions that while this is expected for linear problems, finding nonlinear equations with this behaviour was a complete suprise at the time.

Although the method of characteristics worked nicely on the earlier constant, linear advection problem, it falls short on more complex equations such as the KdV equation. Therefore, in general, a numerical approximation which retains the original properties of the KdV equation is the best for which we can hope for. Desirable properties of such a numerical approximation are that it is efficient, stable and converges to the exact solution with low computational cost. In many problems posed with general geometries the Discontinuous Galerkin (DG) methods have been succesfully applied. In these cases, the typical Finite Element (FE) method fall short as it assumes continuity, often unvalid for hyperbolic PDEs (section 2.2). Although Finite Volume (FV) methods can capture the discontinuities, or shocks, it is only low order accurate. Hence, a combination of these methods, the DG methods has been successfully applied to a wide range of applications, especially when the problem has a dominant first-order term or when it's advection dominated [11]. High-order methods like the DG can resolve waves on fewer degress of freedom and maintain accuracy over large time scales. But how well does it behave when it is introduced to a third spatial derivative, such as in the Korteweg-de Vries (KdV) equation. To tackle this derivative, the DG implementation requires some modification which in turn might yield instability.

1.1 Questions at the centre

To formulate the challenges above, the questions at centre of this thesis summarizes to:

- How would DG methods work on higher spatial derivative terms?

- How does the spatial approximation involving third order derivative terms influence the time integration?
- Can DG methods properly resolve physically relevant solutions to the KdV equation, e.g., solitons?
- How can one construct a DG spatial discretization to the KdV equation that is stable? That is, the discretization properly captured energy estimates from the continuous PDE analysis.

1.2 Method

The thesis will begin with a background chapter on numerical methods for conservation laws. It will highlight what we wish to retain as we move the continuous analysis to the discretized solution space. The behaviour of the solution might provide several challenges due to lack of smoothness, development of shocks and approximation errors that needs to be handled with caution. Following this, the paper will briefly look into the FE and FV method to properly motivate the advantages of the DG method. The bigger challenges such as flux will be underlined and relevant work in the area will be investigated. A DG construction which satisfies the criterias for stability and convergence will be presented and motivated analytically. Lastly numerical results will verify the analytical work. Some conclusive thoughts on DG methods applied to the KdV equation will round of the thesis.

Chapter 2

Background

Before we apply Discontinuous Galerkin methods on the KdV equation, we build up the necessary toolbox to deal with numerical methods. As DG methods are composed of the Finite Element and Finite Volume methods, we will look into their respective strengths and flaws, the mathematics behind and how to implement the methods numerically.

2.1 Approximation space

We begin by assuming that the exact solution $u \in \mathcal{H}^1$ to bound the integral of u and allow derivatives of $\mathcal{L}^2$ functions, more about Sobolov spaces in [2, ch 3]. u can either be projected as a series of orthogonal $\mathcal{L}^2$ integrable functions

$$u = \sum_{i=0}^{\infty} \tilde{u}_i \varphi_i \tag{2.1}$$

where φ_i spans u or interpolated

$$u = \sum_{i=0}^{\infty} u(x_i, t) P_i(x) \tag{2.2}$$

by some interpolating polynomial P. Mathematically there is little difference but the code-wise implementation differs. The basis expansion is called a modal approximation while the interpolation approach is called a nodal approximation. More about nodal and modal approximations in [1]. As numerical approximations cannot use infinite series expansions, a truncation is needed, leaving an approximate solution $u_h \in \mathbf{P}^m$. Higher order m minimizes the residual between

u and u_h such that higher frequency functions can be properly represented.

We begin the discussion of modal approximations by underlining the importance of choosing suitable basis functions for a given problem. Using basis functions that are largely non-zero globally would open the door to Spectral Methods which is widely used due to their exponential convergence rate on smooth problems [9]. Fourier representation utilizes such basis functions but as they need to be periodic, the usefulness is limited. Additionally, in formation of shocks they fall short due to the large number of nodes needed to encapsule the localized shockwave. Hence spectral methods are not suitable for solving conservation laws where shocks are prone to emerge in non-linear solutions. Therefore we exclude basis functions that are non-zero globally and instead consider basis functions being non-zero only locally. An example of such a basis could be the Legendre polynomials. The unknowns in the modal approximation are the expansion coefficients $\tilde{u}_i$. To find these, we first define each basis function $\varphi_j \in D_j$. The coefficients are then calculated by $\mathcal{L}^2$ projection via

$$\tilde{u}_i = (u(x,t), \varphi_i(x))_{D_i} = \int_{D_j} u(x,t)\varphi_i(x)\,dx. \tag{2.3}$$

If we instead consider a nodal approximation the solution is interpolatory [9], meaning that for some chosen $m+1$ gridpoints $x_i \in \{x_0, x_1, ..x_m\}$ on our spatial axis the exact solution is calculated and in turn used as coefficients. We are free to choose an interpolating polynomial $P_i(x)$ with a particularly useful basis being the Lagrange polynomials [9]

$$P_i(x) = l_i(x) = \prod_{\substack{k=0\\k\neq i}}^{m} \frac{x - x_k}{x_i - x_k} \tag{2.4}$$

such that

$$u_h(x,t) = \sum_{i=0}^{m} u(x_i,t) l_i(x), \quad x_i \in \{x_0, x_1, ..x_m\}, \tag{2.5}$$

while emphasizing that $u(x_i, t)$ is still time dependent. For this thesis we will use a nodal approach with Lagrange polynomials. The nodal approach will provide some numerical advantages as the global solution is patched together which will be discussed in detail in later sections.

2.2 Shocks and weak solutions

Consider Burgers' equation

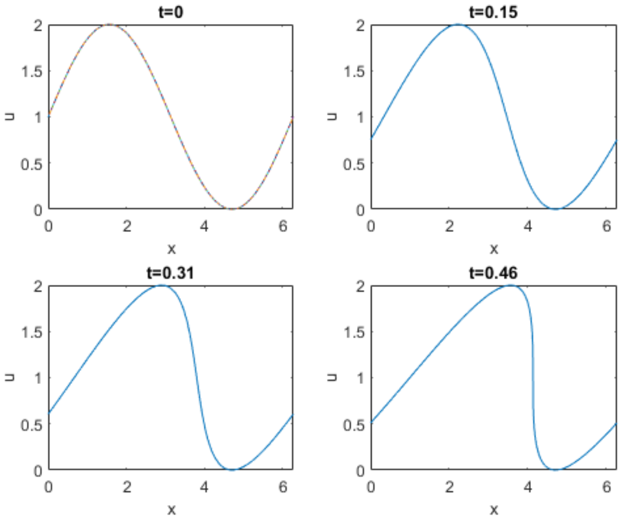

Figure 2.1: Burgers' equation shows an emergin shock wave.

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = 0 \tag{2.6}$$

where the flux is u by the chain rule applied to the spatial term. If this is computed with the initial conditions $u(x, 0) = 1+\sin(x), x \in [0, 2\pi]$ some interesting consequences of nonlinearity makes itself apparent, see Figure 2.1. As the figure shows, a discontinuity forms as time progresses from our initial conditions. The resulting discontinuity or shock wave means that a classical solution to the PDE does not exist as smoothness disappear. The development of numerical tools must consider these properties. As $\frac{\partial u}{\partial x}$ grows to infinity, the numerical computation will be put at a halt. Hence a new way to model the equation is needed. Distribution theory handles with this discontinuity problem by using test functions. Recall the generic hyperbolic PDE

$$u_t + f(u)_x = 0 \tag{2.7}$$

which may contain a nonlinear flux. By applying u on a test function $\phi \in \mathcal{C}^\infty$ we define the weak form of equation (2.7) as

$$(u_t + f(u)_x, \phi)_{\Omega_x} = \int_{\Omega_x} (u_t + f(u)_x)\, \phi dx = 0. \tag{2.8}$$

Partial integration gives

$$\int_{\Omega_x} u_t \phi \, dx + [f(u)\phi]_{\partial\Omega_x} - \int_{\Omega_x} f(u)\phi_x \, dx = 0 \tag{2.9}$$

and thus the solution u to above, called a weak solution, solved the issue of classical derivative due to the shock. However, the weak solution removes uniqueness, making infinitely many solutions possible [12]. To address this issue, we will work from the criterias to recover the true solution, namely entropy conditions [9, eq 2.7]. The entropy criterias stems from the characteristic lines and how they must run into the shock since no true solution can have characteristics emerging from a shock, i.e., once information enters a shock, there is no way to recover it. From Hesthaven [9, eq 2.10], we state the Lax entropy condition

$$f'(u_l) > s > f'(u_r), \quad f''(u) > 0 \tag{2.10}$$

where u_l and u_r are the left and right values of the shockwave and s is the shockspeed. The first derivatie of the flux is related to the wavespeed while the second derivative introduces convexity. We also borrow corollary 2.4 from Hesthaven [9]; *If u is a weak solution with a convex flux and it satisfies an entropy condition, the solution is unique.* With confidence that we can recover the unique solution from the weak form, we assemble what we have so far in the Finite Element (FE) method.

2.3 Finite Element method

Finite Element methods approximate the solution by piecewise polynomials evaluated at arbitrary nonoverlapping elements D_j which together fill the physical domain [10]

$$\Omega_x = \bigcup_{j=1}^{N} D_j. \tag{2.11}$$

These elements can be of different shape and size. In 1D they are simply nonoverlapping line segments. The individual elements are connected together as a mesh. The advantages of FE methods are its efficiency on complex boundaries with little difficulty; however, the methods are not free from drawbacks, read more at the introduction by Hesthaven [10]. As the FE method relies on a finite number of polynomial basis functions to represent the solution it lacks accuracy when the solution is non-smooth. Additionally, if the solution is smooth but, complex the FE method becomes computionally expensive as it requires more

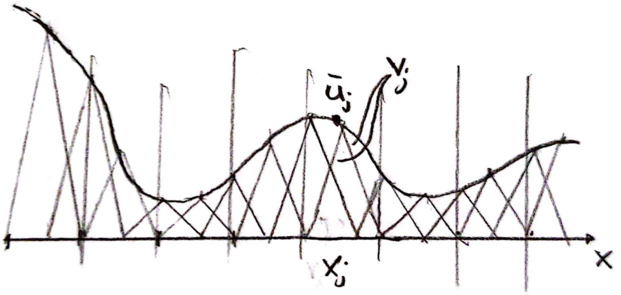

Figure 2.2: Finite Element method illustration with triangular basis functions.

degrees of freedom to capture the complex wave behaviour [3]. So, the solution and flux is approximated by some finite sum of basis functions

$$u_h(x) = \sum_{i=0}^{m} \hat{u}_i \varphi_i(x), \quad \varphi_i \in \mathcal{H}^1(\mathbb{R}) \tag{2.12}$$

and

$$f_h(x) = \sum_{i=0}^{m} \hat{f}_i \varphi_i(x), \quad \varphi_i \in \mathcal{H}^1(\mathbb{R}) \tag{2.13}$$

where φ_i are the basis functions for the entire spatial domain Ω_x but non-zero for a few elements only [5, ch 2.4], see Figure 2.2. By using basis functions non-zero over a small part of the domain only, the discrete problem becomes more localized and in turn faster to compute [5]. To allow discontinuities in the scalar hyperbolic PDE

$$\frac{\partial u_h}{\partial t} + \frac{\partial f_h(u)}{\partial x} = 0 \tag{2.14}$$

without blowing up the solution due to the derivates, we consider the weak form of the governing equation for some test function $\phi \in \mathcal{L}^2$ and require the test functions are orthogonal to the residual in $\mathcal{L}^2$ norm such that

$$\int_{\Omega_x} R_h(x,t)\, dx = \int_{\Omega_x} \left(\frac{\partial u_h}{\partial t} + \frac{\partial f(u_h)}{\partial x} \right) \phi(x)\, dx = 0 \tag{2.15}$$

where the residual is derived from

$$\begin{aligned} &\left|\left(\frac{\partial u}{\partial t}+\frac{\partial f(u)}{\partial x}\right)-\left(\frac{\partial u_h}{\partial t}+\frac{\partial f(u_h)}{\partial x}\right)\right| \\ =&\left|0-\left(\frac{\partial u_h}{\partial t}+\frac{\partial f(u_h)}{\partial x}\right)\right|. \end{aligned} \tag{2.16}$$

From equations (2.12), (2.13) and (2.15) we find that the approximate solution is defined by the weak formulation

$$\begin{aligned} &\int_{\Omega_x} \frac{\partial}{\partial t}\left(\sum_{i=0}^{m} \hat{u}_i \varphi_i(x)\right) \phi(x)\, dx + \\ &\int_{\Omega_x} \frac{\partial}{\partial x}\left(\sum_{i=0}^{m} \hat{f}_i \varphi_i(x)\right) \phi(x)\, dx = 0. \end{aligned} \tag{2.17}$$

By selecting test functions which span the same space as the basis functions, i.e.

$$(\cdot,\phi)=0 \iff (\cdot,\varphi_i)=0 \quad \forall i \in 0,..,m \tag{2.18}$$

and define a mass matrix M and stiffness matrix S by the elements

$$\begin{aligned} M_{i,j} &= \int_{D_j} \varphi_i(x)\varphi_j(x)\, dx \\ S_{i,j} &= \int_{D_j} \varphi_i(x)\frac{\mathrm{d}\varphi_j(x)}{\mathrm{d}x}\, dx \end{aligned} \tag{2.19}$$

we retrieve the Galerkin approximation method [9, pg 379] and rewrite equation 2.48 to

$$M\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t}+S\mathbf{f}_h=0 \tag{2.20}$$

where $\mathbf{u}_h=[u_1,u_2,...u_N]^T$ and $\mathbf{f}_h=[f_1,f_2,...f_N]^T$. The integrals in equation (2.19) needs evaluation which is not always trivial and therefore quadrature will be used, further discussed in section 2.6. The elements are coupled together by requiring continuity at their interfaces.

High-order accuracy is achieved by increasing the polynomial degree of the basis functions, m. However the globally defined basis functions and residual being orthogonal to the same set of globally defined test functions implies that the FE method is implicit and M must be inverted, at high computational cost and limitation to our time solver. It is also typically restricted to $\mathcal{H}^1$ problems to avoid dealing with discontinuities in $\mathcal{L}^2$. Next, we will be looking into an alternative approach, the Finite Volume (FV) method [10].

2.4 Finite Volume method

If the test functions are assumed to be a constant 1 and we apply partial integration to the weak formulation, a method is obtained which discretize the integral form instead of the differential strong form. This leads to piecewise constants that are assembled to reproduce the global solution, see Figure 2.3. The key prospect of this possibility is its encapsuling of shocks. Introducing the grid $(x_j, t^n) = (jh, kn)$ where k and n are the spatial and temporal uniform grid size, respectively. $x_{j+1/2}, x_{j-1/2}$ marks the cell boundaries.

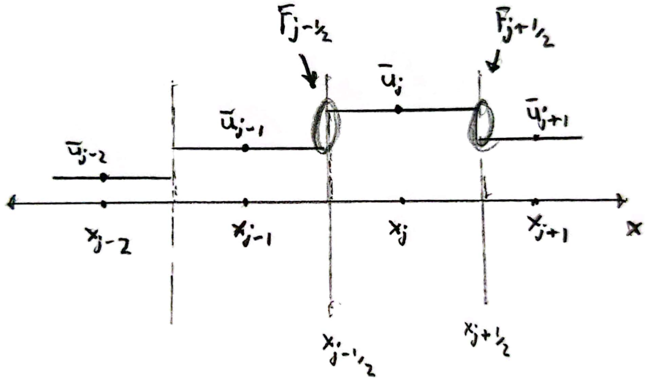

Figure 2.3: Illustration of the Finite Volume method

$$\frac{\partial}{\partial t}\int_{D_x} u\,dx + f(u(x_{j+1/2})) - f(u(x_{j-1/2})) = 0. \tag{2.21}$$

or the formulation on a single finite volume cell

$$\begin{aligned}\int_{x-1/2}^{x+1/2} [u(x, t^{n+1}) - u(x, t^n)]\,dx &= \\ -\int_{t^n}^{t^{n+1}} [f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t))]\,dt.\end{aligned} \tag{2.22}$$

The cell averages are

$$\overline{u}^n = \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t^n)\,dx \tag{2.23}$$

and the flux

$$F^n_{j+1/2} = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(u(x_{j+1/2},t))\,dx. \tag{2.24}$$

Hence we have the numerical scheme known as the Finite Volume (FV) method

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \frac{k}{h}\left[F^n_{j+1/2} - F^n_{j-1/2}\right]. \tag{2.25}$$

This is shock-preserving [9] but changes our unknown to be the cell average. A key stage at this point for the thesis is the introduction of the numerical flux. Each interface, $x_{\pm 1/2}$ is shared between two cells. The flux at these interfaces therefore needs to be computed by some flux function. The function takes the values of the solution from the left and right and return a unique value of the flux up to the sign of an outward pointing normal vector, to ensure global conservation. To construct these fluxes we back up a bit.

2.5 Monotonicity and conservation

Conservation in the continuous sense means that

$$\frac{d}{dt}||u|||_{\mathcal{L}^2} = 0. \tag{2.26}$$

The numerical method must carry this property. To ensure this we use [9, eq 4.1], if a scheme can be written as

$$U_j^{n+1} = U_j^n - \frac{k}{h}\left[F^n_{j+1/2} - F^n_{j-1/2}\right] \tag{2.27}$$

then it is in conservation form. Here the numerical flux is

$$F_{j+1/2}n = F(U^n_{j-p},...,U^n_{j+q}), F_{j-1/2}n = F(U^n_{j-p-1},...,U^n_{j+q-1}) \tag{2.28}$$

and p, q is the number of left and right cells that the scheme depends on as time integrates. Building upon this we use theorem 4.3 from Hesthaven [9]; *If conservation form is established and the flux fulfills consistency, Lipschitz continuity and the scheme converges to total variation bounded solution in $\mathcal{L}^1$,* *then the solution to the discrete form is a weak solution to the conservation law.* Hence, if the constructed numerical flux fulfills consistency and Lipschitz continuity, we can expect that the computed solution is conservative. Now, to better connect the results above and the implemented schemes, we return to the general conservative form

$$U_j^{n+1} = U_j^n - \frac{k}{h}\left[F_{j+1/2}^n - F_{j-1/2}^n\right] = G(U_{j-p-1}^n, ..., U_{j+q}^n) = G(U^n). \quad (2.29)$$

The scheme is called monotone if the operator $G(\cdot)$ is non-decreasing in all arguments. With monotonicity follows stability and guarantees that no oscillations can be created by the numerical approximation. It also implies that the scheme holds the desired properties of the continuous case and a solution obtained by a monotone flux satisfies all entropy conditions [9].

There are many monotone fluxes to consider but in this thesis we will consider upwind, downwind and central flux. While Lax-Friedrichs flux is regularly used in the FV community it relies on a limiter depending on wavespeed. As we formulate our method for the KdV solution it shall be clear how this approach does not always work. Listed below are several choices of the numerical flux function, including the central, upwind and downwind flux where u_L, u_R are the left and right values of u at an interface $x_{j\pm1/2}$ [9]. Upwind and downwind flux originates from the characteristics of the problem and wave propagation, where information travels upwind with the wave if $a > 0$ and hence the left hand value is solely used. Vice-versa for downwind. Upwind and downwind often gives excellent results for strictly advection problems [9]. The fluxes are constructed as follows;

- Lax-Friedrichs flux

$$F_{LF}(u_L, u_R) = \frac{f(u_L) + f(u_R)}{2} - \frac{C}{2}(u_R - u_L), \;\; C \geq \max|f'(u)|. \quad (2.30)$$

- Central flux

$$\hat{u}(u_L, u_R) = \frac{u_L + u_R}{2} \quad (2.31)$$

- Upwind flux

$$\hat{u}(u_L, u_R) = u_L \quad (2.32)$$

- Downwind flux

$$\hat{u}(u_L, u_R) = u_R \quad (2.33)$$

and in turn $F(u_L, u_R) = f(\hat{u})$ for the three latter fluxes. The last term in Lax-Friedrichs flux is the speed limiter which also adds some dissipation, often helpful to remove numerical oscillations. However, the strength of monotonicity comes at cost. In fact, a monotone scheme is, at most $\mathcal{O}(h)$ accurate and for discontinuous solutions we cannot expect better than $\mathcal{O}(\sqrt{h})$ for the FV method, this is the classical result of Godunov [9, sec 5.2]. This is highly discouraging as we seek high order approximations at low computational cost.

In essence, the FV scheme left is of a explicit semi-discrete form which offers high flexibility in the choice of time-solver. It is also highly localized thus allowing parallel computation of each cell average and it imposes no condition on the grid structure, allowing high geometric flexibility [10]. The main drawback being $\mathcal{O}(h)$ convergence at best.

2.6 Quadrature

Before proceeding with the Discontinuous Galerkin methods, we equip ourselves with tools to numerically solve integrals, like those present in the weak formulation of the PDE (2.19). Explicitly solving integrals is not a cheap operation and not always an easy task to do analytically, especially in higher spatial dimensions or on general element geometries. Hence, we approximate the integral with quadrature [1]

$$\int_a^b u(x)\,dx = \sum_{k=0}^{N} u(x_k)w_j + E \tag{2.34}$$

where x_k, w_k are the abscissas and weights respectively and E is the approximation error. The choice of abscissas and weights are called quadrature rules. The quadrature rules that provides maximum precision, i.e. being exact for the highest order polynomials are Gauss quadrature rules. Further reading on quadrature is referred to Kopriva [1]. There are many types of Gauss quadrature rules to consider but for this thesis Legendre Gauss Lobatto rules will be utilized. The associated abscissas and weights are

$$x_k = +1, -1 \text{ and zeros of } L'_m(x) \tag{2.35}$$

$$w_k = \frac{2}{m(m+1)} \frac{1}{[L_m(x_k)]^2} \tag{2.36}$$

for $k = 0, .., m$ which are exact for all polynomials of order $2m - 1$ or less. The Legendre polynomials $L_m(x)$ can be constructed and evaluated from the recursion formula [1]

$$L_{m+1}(x) = \frac{2m+1}{m+1} x L_m(x) - \frac{m}{m+1} L_{m-1}(x) \tag{2.37}$$

with $L_0(x) = 1, L_1(x) = x$. There are alternatives such as Jacobi-Gauss quadrature and Legendre-Gauss which relaxes our abscissas to exclude the endpoints -1,+1, increasing the exactness to $2m + 1$ or less. However, the inclusion of endpoints is of great importance as they allow precise evaluation of boundary contributions, prevalent in the DG methods.

2.7 Discontinuous Galerkin methods

Combining the best aspects of the FE method and FV method, Discontinuous Galerkin methods embeds the idea of numerical flux and slope limiters from the FV method with the geometric flexibility and basis functions of the FE method. It is a method where the FE method is applied locally and glued together globally using the FV method. The resulting combination gives DG methods the following strong advantages [7]:

- The order of the approximation depends on the exact solution, its smoothness, and by choosing an approximation of sufficient degree .
- The discontinuous elements results in a block diagonal mass matrices and the size of these matrices depend on the degree of freedom within each loacl element. Compare this to the FE method where the mass matrix is defined globally [9, pg 382]. Hence the invertion of these blocks are parallelizable.
- As the underlying elements are arbitraryily choosen, DG methods are well suited for complicated geometries.
- DG methods does not require a complicated treatment of the boundary conditions to acquire high-order accuracy.
- DG methods allows flexibility in the choice of flux which in turn incorporates desired physical properties.

We shall now investigate how this is achieved.

2.8 Basis functions

To relax the requirement on uniform spatial grid Ω_x, just as the FE, DG methods split the space to arbitrary nonoverlapping, elements D_j

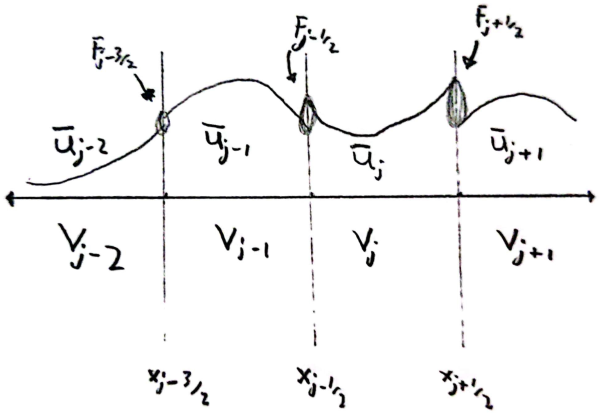

Figure 2.4: Illustration of the Discontinuous Galerkin method with a polynomial approximation on each element.

$$\Omega_x = \bigcup_{j=1}^{N} D_j, \quad D_j = [x_{j-1/2}, x_{j+1/2}]. \tag{2.38}$$

In 1D each element is a line segment but as the dimension increases, this approach allows the spatial elements of general polygons. The key difference to the FE method is that the basis functions in DG are defined locally on each element, instead of the entire computational domain. From here on u_h will drop the subscript h to improve readability and unless stated otherwise u_j is the approximate solution. Hence for each element j we get an approximation u_j with the basis expansion on a given element,

$$u_j(x,t) = \sum_{i=0}^{m} \hat{u}_{j,i}(t)\varphi_{j,i}(x) \tag{2.39}$$

where $\varphi_{j,i} \in \mathcal{S}^m := \left\{\varphi_{j,i} \in \mathcal{L}^2(\mathbb{R}) : \varphi_{j,i} \mid_{D_j} \in \mathcal{P}_m\right\}$, the function space of integrable polynomials order m restricted on D_j. Similarly the collocated flux can be expressed as

$$f_j(x,t) = \sum_{i=0}^{m} \hat{f}_{j,i}(t)\varphi_{j,i}(x) \tag{2.40}$$

We notice that this representation does not impose the need of continuity between each element as in FE method and also allows different basis functions for each element, although the latter is rarely a used property in practice [10]. Now, for each element D_j the local solution, built on locally defined basis functions $\varphi_{j,i}$ will be evaluated at $x_{j-1/2}$ and $x_{j+1/2}$, resulting in duplicate unknowns as neighbouring interfaces overlap at the interface, yielding a vector of unknowns

$$\mathbf{u} = [\underbrace{u_{1/2}, u_{3/2}}_{D_1}, \underbrace{u_{3/2}, u_{5/2}}_{D_2} \ldots \underbrace{u_{N-1/2}, u_{N+1/2}}_{D_N}]^T \tag{2.41}$$

which is $2N$ long instead of $N+1$ as in the FE method. With the local solution represented by u_j and the residual required to be orthogonal to the polynomial function space $\mathcal{S}^m$ we have found a way to recover the N local statements

$$\int_{D_j} \left(\frac{\partial u_j}{\partial t} + \frac{\partial f(u_j)}{\partial x} \right) \varphi_{j,i}(x)\, dx = 0 \quad \forall i \in 0..m, \forall j \in 1...N. \tag{2.42}$$

At this stage we have several local solutions and seek a method to find a suitable global solution which promises uniqueness despite the endpoints being multiply defined across element interfaces. If we select the constant basis function $\varphi_{j,0} = 1/h_j$ we recover the FV method for which this exact issue has been solved [10]. By partially integrating $f_x\varphi$ of equation (2.42), we obtain

$$\int_{D_j} \frac{\partial u_j}{\partial t} \varphi_{j,i}\, dx - \int_{D_j} f_j \frac{\mathrm{d}\varphi_{j,i}}{\mathrm{d}x}\, dx = -\left[f_i \varphi_{j,i}\right]_{x_{j-1/2}}^{x_{j+1/2}}. \tag{2.43}$$

We recall the nodal expression (2.5) and express u with a Lagrange interpolation basis

$$u(x,t) = \sum_{i=0}^{m} u(x_i,t) l_i(x), \quad x_i \in \{x_0, x_1, ..x_m\} \in D_j \tag{2.44}$$

Approximating u and f in their basis expansion form on the left hand side we get

$$\begin{aligned} &\int_{D_j} \frac{\partial}{\partial t} \left(\sum_{i=0}^{m} u_j(x_i,t) l_i(x) \right) \varphi_{j,i}(x)\, dx - \\ &\int_{D_j} \left(\sum_{k=0}^{m} f(u_j(x_k,t) l_k(x) \right) \frac{\partial \varphi_{j,i}}{\partial x}(x)\, dx = \\ &- \left[f_i \varphi_{j,i}\right]_{x_{j-1/2}}^{x_{j+1/2}} \end{aligned} \tag{2.45}$$

We now apply the Galerkin approximation ansatz as in the FE method, i.e. $\varphi_i(x) = l_i(x)$. At this stage we notice the need to modify our equations for quadrature application. The gauss-type quadratures lives on the interval [-1,1] where as the elements are of length h_j. To solve this we use a variable mapping in order to apply the quadrature rules. The mapping looks as follows

$$x(r) = x_{j-1/2} + \frac{1+r}{2}(x_{j+1/2} - x_{j-1/2}), \quad \forall x \in D_j, r \in [-1,1] \tag{2.46}$$

and with this we get, focusing on a specific element D_j (dropping index j),

$$\begin{aligned} &\frac{h}{2}\int_{-1}^{1} \frac{\partial}{\partial t}\sum_{k=0}^{m} u(r_k,t) l_k(r) l_i(r)\, dr - \\ &\int_{-1}^{1} \sum_{k=0}^{m} f(u(r_k,t) l_k(r) \frac{\partial l_i}{\partial r}(r)\, dr = \\ &-\left[f_i l_i\right]_{-1}^{1}. \end{aligned} \tag{2.47}$$

In (2.47), the factor $\frac{h}{2}$ dropped out of the first term as the Jacobian from the variable swap, h being the element size. This cancels itself out due to the derivative present on the second term. We have assumed collocation on the flux values to make the factorization possible. We move the time dependent u, f out of the integrals as they are indepentent of r. (2.47) now looks as follows,

$$\begin{aligned} &\frac{h_j}{2}\sum_{k=0}^{m} \frac{\partial}{\partial t} u(r_k,t) \int_{-1}^{1} l_k(r) l_i(r)\, dr - \\ &\sum_{k=0}^{m} f(u(r_k,t) \int_{-1}^{1} l_k(r) \frac{\partial l_i}{\partial r}(r)\, dr = \\ &-\left[f_i l_i\right]_{-1}^{1}. \end{aligned} \tag{2.48}$$

We approximate the integrals with quadrature, discussed in section 2.6. The integrand is the product of the polynomial basis functions, namely $l_k(r) l_i(r)$ and is thus of order $2m$. This means that LGL inexactly evaluates the integrand and we lose 1 order of exactness. By deciding the interpolation nodes l_i, l_k to collocate with the LGL abscissas $\{r_l\}_{l=0}^{m}$ we can utilize the Kronecker delta property of the Lagrange polynomial. While collocation makes mass- and stiffness matrices very computationally efficient, it introduces an aliasing error. The error introduced can lead to a reduction of one in the rate of convergence, but not more than that.

Theorem 2.8.1 *If $u \in \mathcal{L}^2$ and $I_m u \in \mathbf{P}^m$, the aliasing error introduced by the interpolation operator I_m is*

$$||u - I_m u||_{\mathcal{L}^2(D_j)} \leq N^{(-1/2)} |u|_{D_j,1} \tag{2.49}$$

where

$$|u|^2_{D_j,q} = ||u^{(q)}||^2_{\mathcal{L}^2(D_j)}, \tag{2.50}$$

is the Sobolev seminorm and $u^{(q)}$ is the qth derivative of u.

Proof can be found in theorem 12.6 in Hesthaven's Numerical Methods for Conservation Laws [9] with $q = 0, p = 1$ applied. With the numerical aliasing error of collocation stated, we continue by looking at the gain by using the collocation. We first state the quadrature approximated version of (2.48)

$$\begin{aligned} &\frac{h}{2} \sum_{k=0}^{m} \frac{\partial}{\partial t} u(r_k, t) \sum_{l=0}^{m} l_k(r_l) l_i(r_l) w_l - \\ &\sum_{k=0}^{m} f(u(r_k, t) \sum_{l=0}^{m} l_k(r_l) \frac{\partial l_i}{\partial r} |_{r_l} w_l \end{aligned} \tag{2.51}$$

and the corresponding mass- and stiffness matrices' entries

$$M_{ki} = \frac{h}{2} \sum_{l=0}^{m} l_k(r_l) l_i(r_l) w_l \tag{2.52}$$

and

$$S_{ki} = \sum_{l=0}^{m} l_k(r_l) \frac{\partial l_i}{\partial r} |_{r_l} w_l. \tag{2.53}$$

With the kronecker delta property of the collocated quadrature points and Lagrange polynomials

$$l_i(r_l) = \delta_{il} = \begin{cases} 1, & i = l \\ 0, & i \neq l \end{cases} \tag{2.54}$$

the mass matrix will become diagonal and is trivially inverted. The differentiation of the Lagrange polynomial, required in the stiffness matrix, is done accordingly

$$
\begin{aligned}
\frac{\partial l_i}{\partial r}\,|_r = \frac{\partial}{\partial r}\prod_{\substack{k=0\\k\neq 1}}^{m}\frac{r-r_k}{r_j-r_k} &= \frac{1}{r_i-r_0}\left(\frac{\prod_{\substack{k=1\\k\neq i}}^{m}(r-r_k)}{\prod_{\substack{k=1\\k\neq i}}^{m}(r_i-r_k)}\right)+...\\
...+\frac{1}{r_i-r_m}\left(\frac{\prod_{\substack{k=0\\k\neq i}}^{m-1}(r-r_k)}{\prod_{\substack{k=0\\k\neq i}}^{m-1}(r_i-r_k)}\right) &= \sum_{\substack{n=0\\n\neq i}}^{m}\left(\frac{\prod_{\substack{k=0\\k\neq i\\k\neq n}}^{m}(r-r_k)}{\prod_{\substack{k=0\\k\neq i}}^{m}(r_i-r_k)}\right)
\end{aligned}
\tag{2.55}
$$

With the time dependent unknown solution values $\mathbf{u} = [u(x_0,t),.,u(x_m,t)]^T$, collocated flux values $\mathbf{f} = [f(u(x_0,t)),..,f(u(x_m,t))]^T$ and basis functions $\mathbf{l} = [l_0(r),...,l_m(r)]^T$, we arrive at the semi-discrete weak form of the local DG scheme,

$$
\begin{aligned}
&\mathrm{M}\mathbf{u}_t - \mathrm{S}\mathbf{f} = -\left[f\mathbf{l}\right]_{-1}^{1} \iff\\
&\mathbf{u}_t = \mathrm{M}^{-1}\left(\mathrm{S}\mathbf{f} - \left[f\mathbf{l}\right]_{-1}^{1}\right).
\end{aligned}
\tag{2.56}
$$

From these N expressions, which each denotes the solution on each of the elements, we now assemble the global solution much like in the FV method with the help of numerical fluxes, $\hat{f}$ at each interface.

The performance of the DG scheme highlights its composition of the best from both FV and FE schemes, Hesthaven shows that DG with a monotone flux is stable and satisfies all cell entropy conditions [9, thm 12.8], i.e. $\frac{d}{dt}||u_h||_{\mathcal{L}^2} \leq 0$ for a scalar nonlinear conservation law. He also shows that it converges $\mathcal{O}(h^{m+1})$ for linear convection equations with upwind flux [9, pg 384]. This is the DG method for a conservative law containing a single spatial derivative. While this is well used and tested the question now is how it should be modified to work on the KdV equation which contain a spatial derivative of third order. But before that we notice that we only have a method which discretizes in space explicitly. We have so far not mentioned how to move forward in time with the semi-discrete expression.

2.9 Time stepping

The explicit semi-discrete form allows flexibility in the choice of time stepper. A third order Runge Kutta seems sufficient for most DG problems and will thus be used in this thesis [9, 11]. The size of the explicit time step itself is however of great importance as the time step dictates how quickly an approximate solution can be obtained numerically. If the highest wave-speed is a and $ak/h < 1/2$ the physical wave will travel faster than our numerical wave and information disappears [9], thus violating the physics of the problem and leading to a numerical instability. The Courant–Friedrichs–Lewy (CFL) condition is derived from this physical perspecive and ensures stability. It means that the discrete wave information travels at a rate that respects the continuous domain of dependence [13]. By defining CFL$=\max|f'|\frac{\Delta t}{\Delta x}$ to be the Courant number, we can expect stable time integration by bounding the CFL number [9]. For the conservation law the time step will scale as, presuming a uniform gridsize h_j,

$$k \leq \mathrm{CFL}_1 \frac{h}{\max|f'|} \tag{2.57}$$

where CFL_1 scales on the polynomial order for the basis functions of the scheme by $\mathrm{CFL}_1 \propto m^{-2}$ for DG methods [9]. For second-order time-dependent problems, i.e.

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = \frac{\partial}{\partial x} \nu_j(u) \frac{\partial u}{\partial x} \tag{2.58}$$

where ν_j is the viscosity coefficient we expect [9, eq 12.74]

$$k \leq \mathrm{CFL}_2 \frac{h^2}{\max|\nu_j|}, \quad \mathrm{CFL}_2 \propto m^{-4}. \tag{2.59}$$

2.10 Approximating higher-order derivative with DG methods

Applying DG methods to the KdV equation requires some modification due to the nonlinearity and third spatial derivative. Recall the KdV equation

$$u_t + u_{xxx} + 6uu_x = 0 \tag{2.60}$$

and imagine that a direct application of the DG methodology to the nonlinear flux function $f(u) = u_{xx} + 3u^2$. But as Hesthaven shows in [10, pg 244] this methodology leads to an unstable numerical scheme that lacks convergence. Instead, other discretization techniques are sought. To motivate such an alternative discretization strategy, we take inspiration from the heat equation

$$u_t - u_{xx} = 0, \tag{2.61}$$

discussed in Hesthaven's books [10, 9] as well as the article of Shu and Yan [11]. The general idea is to introduce an auxiliary variable q such that the heat equation can be written as a system of first order equations

$$u_t - u_{xx} = 0 \iff \begin{cases} u_t = q_x \\ q = u_x. \end{cases} \tag{2.62}$$

Designing the fluxes for the auxilary variable q requires a more care as the idea of a wave speed, such as with the Lax-Friedrichs flux (section 2.5), does not translate to these artificial variables. While u represent waves the auxiliary variables holds no such property and hence the Jacobian and characteristics strategy breaks. As q is introduced for the design of high-order derivative terms such as with the heat equation, where information flows both ways across the interfaces, it makes sense to use a Bassi-Rebay 1 coupling, or a central numerical flux [6, 4]

$$\hat{q} = \frac{q^+ + q^-}{2}. \tag{2.63}$$

The other option is to use upwind alternated with downwind as based on Shu's and Yan's work [11]. The alternating upwind, downwind flux is motivated by the general good convergence rates that we have in the case of one spatial derivative and is constructed as

$$\hat{u}_{j\pm1/2} = u^-_{j\pm1/2}, \ \ \hat{q}_{j\pm1/2} = q^+_{j\pm1/2} \text{ or vice versa,} \tag{2.64}$$

$$\hat{u}_{j\pm1/2} = u^+_{j\pm1/2}, \ \ \hat{q}_{j\pm1/2} = q^-_{j\pm1/2}. \tag{2.65}$$

The choice between alternating fluxes, known as Local Discontinuous Galerkin (LDG), or central fluxes is problem-dependent as we shall see in in the important results below which highlights what happens when a system of first order equations is applied to the heat equation.

- Theorem 7.2 and 7.3 from [10]; the heat equation with central flux is stable and converges as $\mathcal{O}(h^m)$ if m is odd and $\mathcal{O}(h^{m+1})$ if m is even. Meaning that there is some sub-optimal order of accuracy for odd m.

- Equation 1.11 and theorem 12.27 from [11, 9] respecively; Local Discontinuous Galerkin (LDG) shows $\mathcal{O}(h^m)$ applied to the heat equation for both odd and even m.

- The LDG tends to be stiff, i.e. requiring small time steps to fulfill the CFL condition [10].

With this in mind for a diffusion type problem, we turn our heads to the KdV equation, a convection-diffusion type problem. Hence one could expect some of these properties to be prevalent as we move forward.

Chapter 3

Results

As the KdV equation is a convection-diffusion type problem, i.e. quantity both travels in our spatial domain and diffuse, a brief look at the strictly diffusion problem, the heat equation, can inspire the numerical flux design. We formulate KdV as a system of first order equations, just as with the heat equation, in the following way,

$$u_t + u_{xxx} + 6uu_x = 0 \iff \begin{cases} u_t = -(3u^2 + q)_x \\ q = p_x \\ p = u_x. \end{cases} \tag{3.1}$$

This allows us to use the same DG method for the convection equation to solve 3.1. That is, we seek approximations $u, q, p \in V_h$ such that for all test functions $\varphi, \phi, \theta \in V_h$

$$\int_{D_j} u_t \varphi \, dx - \int_{D_j} (3u^2 + q)\varphi_x \, dx = -\left[(\hat{3u^2} + \hat{q})\varphi\right]_{-1}^{1} \tag{3.2}$$

$$\int_{D_j} q\phi \, dx + \int_{D_j} p\phi_x \, dx = [\hat{p}\phi]_{-1}^{1} \tag{3.3}$$

$$\int_{D_j} p\theta \, dx + \int_{D_j} u\theta_x \, dx = [\hat{u}\theta]_{-1}^{1} \tag{3.4}$$

where $\hat{3u^2}$, $\hat{u}$, $\hat{q}$ and $\hat{p}$ is the numerical flux functions that we are yet to construct. Following a similar discretization strategy of high-order LGL quadratures and collocations we had in DG Section 2.7 we determine the semidiscretization on each element

$$\mathrm{M}\mathbf{u}_x - \mathrm{S}(3\mathbf{u}^2 + \mathbf{q}) = -\left[(\widehat{3u^2} + \hat{q})\mathbf{l}\right]_{-1}^{1} \tag{3.5}$$

$$\mathrm{M}\mathbf{q} + \mathrm{S}\mathbf{p} = [\hat{p}\mathbf{l}]_{-1}^{1} \tag{3.6}$$

$$\mathrm{M}\mathbf{p} + \mathrm{S}\mathbf{u} = [\hat{u}\mathbf{l}]_{-1}^{1} \tag{3.7}$$

The numerical flux for the nonlinear flux term $3u^2$ grants some flexibility in the choice as long as it is entropy conservative [9]. We use the classical Lax-Friedrichs flux defined in Section 2.5 but other monotone fluxes which yield entropy conservation could also be used. Recall the Lax Friedrichs flux,

$$F_{LF}(u_L, u_R) = \frac{f(u_L) + f(u_R)}{2} - \frac{C}{2}(u_R - u_L), \;\; C \geq \max|f'(u)|. \tag{3.8}$$

C can be a global estimate over the entire domain or a local estimate only involving the values of u_L and u_R. Both holds but in this thesis C will be based on the global domain. For u, q, p we will use either central flux or alternating upwind and downwind flux. The key conditions for alternating fluxes to work is that $\hat{q} = q^+$ and not $\hat{q} = q^-$ otherwise it will not numerically converge [11].

3.1 Choice of time step

Likely, the DG semidiscretization will be very stiff as it works on the KdV equation based on the discussion in Section 2.9. We loosely follow the pattern induced by the order of derivative and should expect a stable stepsize in time by

$$k \leq \mathrm{CFL}\frac{h^3}{\max|f'|}, \quad \mathrm{CFL} \propto m^{-6}. \tag{3.9}$$

This time step restriction will be numerically tested by first running the central and LDG fluxes with a time step neglecting temporal errors to produce $\mathcal{L}^2$ errors for each m, N, denoted $e_{m,N,\text{exact}}$. From these results CFL tests will be conducted where the $\mathcal{L}^2$ error for different CFL numbers, $e_{m,N,\text{CFL}}$ is compared to $e_{m,N,\text{exact}}$. The tests will begin with CFL$= 0.9$ and will be multiplied with 0.9 until relative error

$$\frac{|e_{m,N,\text{CFL}} - e_{m,N,\text{exact}}|}{e_{m,N,\text{CFL}}} = e_{m,N,\text{relative}} \tag{3.10}$$

is less than some ϵ.

3.2 Convergence and stability for the KdV equation

For LDG methods, stability for the non-linear case and convergence for the linear case is proved by Shu and Yun [11]. Convergence, although only proved in the linear case

$$u_t + u_x + u_{xxx} = 0 \tag{3.11}$$

follow the inequality

$$||u - u_h||_{\mathcal{L}^2}^2 = \int_{\Omega_x} (u(x,t) - u_h(x,t))^2 \, dx \leq Ch^{2m+1} \tag{3.12}$$

where u, u_h is the exact and approximate solution and C depends on the derivatives of u and time t. For the central flux, stability and convergence is demonstrated accordingly.

Theorem 3.1 *The Discontinuous Galerkin method applied on the KdV equation with central flux for u and the auxilary variables as well as the entropy conservative Burgers' flux for the nonlinear f is stable*

Proof of thm 3.1 We seek to prove

$$\frac{d}{dt}||u_h||_{\mathcal{L}^2(\Omega_x)}^2 \leq 0, \tag{3.13}$$

i.e. discrete stability. The proof is inspired by [10] for the heat equation. Recall the KdV equation in first order form (3.1), where we replace the nonlinear flux $3u^2$ with the value f

$$u_t + u_{xxx} + 6uu_x = 0 \iff \begin{cases} u_t = -(f+q)_x \\ q = p_x \\ p = u_x. \end{cases} \tag{3.14}$$

The corresponding weak form of this first order system is

$$\int_{D_j} u_t \varphi \, dx - \int_{D_j} (f+q)\varphi_x \, dx = -\left[(\hat{f} + \hat{q})\varphi\right]_{\partial D_j} \tag{3.15}$$

$$\int_{D_j} q\phi \, dx + \int_{D_j} p\phi_x \, dx = [\hat{p}\phi]_{\partial D_j} \tag{3.16}$$

$$\int_{D_j} p\theta \, dx + \int_{D_j} u\theta_x \, dx = [\hat{u}\theta]_{\partial D_j} \, . \tag{3.17}$$

We now apply partial integration on the second integral in each row one more time to achieve the strong form

$$\int_{D_j} u_t \varphi \, dx + \int_{D_j} (f+q)_x \varphi \, dx = \Big[(f + q - (\hat{f} + \hat{q}))\varphi \Big]_{\partial D_j} \tag{3.18}$$

$$\int_{D_j} q\phi \, dx - \int_{D_j} p_x \phi \, dx = - \left[(p - \hat{p})\phi \right]_{\partial D_j} \tag{3.19}$$

$$\int_{D_j} p\theta \, dx - \int_{D_j} u\theta_x \, dx = - \left[(u - \hat{u})\theta \right]_{\partial D_j} . \tag{3.20}$$

We define the bilinear operator, $\mathbf{B}_h$ over an element D_j to be

$$\begin{aligned} &\mathbf{B}_h(u, q, p; \varphi, \phi, \theta) = \\ &\int_{D_j} u_t \varphi \, dx + \int_{D_j} (f+q)_x \varphi \, dx - \Big[(f + q - (\hat{f} + \hat{q}))\varphi \Big]_{\partial D_j} + \\ &\int_{D_j} q\phi \, dx - \int_{D_j} p_x \phi \, dx + \left[(p - \hat{p})\phi \right]_{\partial D_j} + \\ &\int_{D_j} p\theta \, dx - \int_{D_j} u\theta_x \, dx + \left[(u - \hat{u})\theta \right]_{\partial D_j} . \end{aligned} \tag{3.21}$$

It therefore follows that $\mathbf{B}_h = 0$ from the strong form definition if φ, ϕ, θ are choosen from $\mathcal{P}^N$, when u, q, p is a weak solution. Selecting the test functions to be $\varphi, \phi, \theta = u, -p, q$ yields

$$\begin{aligned} &\mathbf{B}_h(u, q, p; u, -p, q) = \\ &\int_{D_j} u \cdot u_t \, dx + \int_{D_j} (f+q)_x u \, dx - \Big[(f + q - (\hat{f} + \hat{q}))u \Big]_{\partial D_j} + \\ &\int_{D_j} q(-p) \, dx - \int_{D_j} p_x(-p) \, dx + \left[(p - \hat{p})(-p) \right]_{\partial D_j} + \\ &\int_{D_j} pq \, dx - \int_{D_j} uq_x \, dx + \left[(u - \hat{u})q \right]_{\partial D_j} . \end{aligned} \tag{3.22}$$

Note that

$$\int_{D_j} u \cdot u_t \, dx = \int_{D_j} \frac{\partial}{\partial t} \frac{u^2}{2} \, dx = \frac{1}{2} \frac{\partial}{\partial t} \int_{D_j} u^2(x,t) \, dx = \frac{1}{2} \frac{\partial}{\partial t} ||u||^2_{\mathcal{L}^2(D_j)} \tag{3.23}$$

and since we require

$$\mathbf{B}_h = 0 \tag{3.24}$$

and

$$\frac{1}{2}\frac{\partial}{\partial t}||u||^2_{\mathcal{L}^2(D_j)} \leq 0 \tag{3.25}$$

it follows that we must show, after rearranging terms,

$$\begin{aligned}&\int_{D_j} uf_x \, dx + \int_{D_j} uq_x \, dx - \Big[(f + q - (\hat{f} + \hat{q}))u\Big]_{\partial D_j} + \\ &\int_{D_j} pdp - [p(p - \hat{p})]_{\partial D_j} - \int_{D_j} qdu + \\ &[(u - \hat{u})q]_{\partial D_j} \geq 0\end{aligned} \tag{3.26}$$

for the stability proof to hold. Summarizing, what we have done so far is integrate all the first order equations over the domain and applied partial integration to achieve the weak form. Afterwards the boundary terms has been replaced by the numerical fluxes and the entire system was then partially integrated again to achieve the strong form of the DG approximation. Now,

$$\begin{aligned}&\int_{D_j} uf_x \, dx = [uf]_{\partial D_j} - \int_{D_j} u_x f \, dx = [uf]_{\partial D_j} - \int_{D_j} u_x f \frac{\partial x}{\partial u} \, du = \\ &[uf]_{\partial D_j} - \int_{D_j} f \, du = [uf]_{\partial D_j} - F(u)\end{aligned} \tag{3.27}$$

where

$$F(u) = \int_{D_j} f \, du$$

with similar definitions for $\int uq_x \, dx$ and $Q(u)$. Without loss of generality, we consider the interface between two elements and defining the contribution from the left and right element as ξ_L and ξ_R respectively to have

$$\xi = -F(u) + u\hat{f} - \frac{p^2}{2} + p\hat{p} - 2Q(u) + uq - \hat{u}q. \tag{3.28}$$

According to (3.26) we seek to prove that $\xi_L - \xi_R \geq 0$. Thus it is sufficient to examine

$$\begin{aligned}\xi_L - \xi_R = &-F(u^-) + u^-\hat{f} - \frac{(p^-)^2}{2} + p^-\hat{p} - 2Q(u^-) + u^-q^- - \hat{u}q^- \\ &\left(-F(u^+) + u^+\hat{f} - \frac{(p^+)^2}{2} + p^+\hat{p} - 2Q(u^+) + u^+q^+ - \hat{u}q^+\right)\end{aligned} \tag{3.29}$$

where we define the jump between two quantities to be $[\cdot] = (\cdot)^+ - (\cdot)^-$. With some algebraic manipulations we rewrite (3.29) to be

$$\xi_L - \xi_R = [F(u)] - [u]\,\hat{f} + 2\,[Q(u)] - 2\,[u]\,\hat{q}. \tag{3.30}$$

Now, with central flux for q, $\hat{q} = \frac{q^+ + q^-}{2}$, we get with the integral mean value theorem applied to the quantity $Q(u)$ with $\alpha \in [u^-, u^+]$

$$[Q(u)] = Q'(\alpha)[u] = q(\alpha)[u] \implies (q(\alpha) - \hat{q})\,[u] \geq 0. \tag{3.31}$$

The inequality in (3.31) is a result from the standard condition for monotone E-flux [9]. With the entropy conservative Burgers' flux $\hat{f} = (u^+)^2 + u^+u^- + (u^-)^2$ we see that

$$[F(u)] - [u]\hat{f} = (u^+)^3 - (u^-)^3 - (u^+ - u^-)((u^+)^2 + u^+u^- + (u^-)^2) = 0. \tag{3.32}$$

We highlight that Lax Friedrichs flux also yields an acceptable inequality as follows

$$\begin{aligned}[F(u)] - [u]\hat{f} &= (u^+)^3 - (u^-)^3 - (u^+ - u^-)((u^+)^2 + u^+u^- + (u^-)^2) \\ &+ \frac{1}{2}C[u]^2 = \frac{1}{2}C[u]^2 \geq 0\end{aligned} \tag{3.33}$$

as $[u]^2$ and C is always positive and thus also fulfills the desired inequality but with added dissipation. Thus, assuming periodic boundary conditions all elementwise inequalities above will cancel itself out as we sum over all elements. This leaves us with $\frac{\partial}{\partial t}||u||^2_{\mathcal{L}^2(\Omega_x)} \leq 0$. ■

3.3 Numerical tests

The setup for all numerical simulations are as follows. The DG method is applied on the KdV first order form (3.1). The parameters for our simulations will be

- $\Omega_x = [-10, 12]$, runtime $T = 0.5$
- True solution: $u(x,t) = 2\text{sech}^2(x - 4t)$
- Initial conditions: $u(x,0) = 2\text{sech}^2(x)$
- Boundary conditions: Periodic
- Mesh: Non-uniform, LGL
- $N = 32, 64, 128, 256$
- $m = 1, 2, 3, 4$.

With these global parameters and functions implemented in *Matlab* we will investigate the numerical performance of central flux and LDG, alternating upwind and downwind fluxes. The CFL number is taken small enough to neglect temporal errors while investigating m, N convergence for LDG and central fluxes. To evaluate the stiffness of the LDG and central approximations, i.e.

how small the CFL number has to be for stability, different CFL numbers will be compared to a very small CFL number of 0.00005. The largest CFL numbers which yield a relative error ≤ 0.05 will indicate which method is the stiffest, LDG or central flux. For further details on the implementation and numerical testing routines see the code included in the appendix.

3.4 Central flux results

The DG methods with central flux (CF) for u, q, p and Lax-Friedrichs flux (3.8) for f produced $\mathcal{L}^2$ errors in table 3.1. The CFL value is 0.00005, small enough to rule out temporal errors. N being the number of elements and m indicating the overall order of the method.

$\mathcal{L}^2$ error, KdV equation				
$m\backslash N$	N=32	N=64	N=128	N=256
1	8.0907e-01	2.5353e-01	8.2056e-02	2.7785e-02
2	1.7439e-01	1.8033e-02	2.0180e-03	3.1716e-04
3	9.5370e-03	7.6100e-04	1.1627e-04	1.6374e-05
4	1.4098e-03	6.3536e-05	1.5385e-06	7.9098e-08

Table 3.1: KdV eqution $u_t + u_{xxx} + 6uu_x = 0$. $u(x, 0) = 2\text{sech}^2(x)$. Periodic boundary condition. Non uniform meshes with N elements. $\mathcal{L}^2$ error for central flux

To investigate stable CFL numbers, i.e., CFL numbers for which the relative error is 0.05 or less, numerical results are displayed in table 3.2.

Stable CFL numbers for central flux				
$m\backslash N$	N=32	N=64	N=128	N=256
1	5.9049e-01	2.0589e-01	4.2391e-02	9.6977e-03
2	9.2651e-02	1.9076e-02	4.3640e-03	1.1093e-03
3	2.6167e-02	5.9863e-03	1.3695e-03	3.4810e-04
4	1.0138e-02	2.3192e-03	5.8951e-04	1.4985e-04

Table 3.2: Stable CFL numbers. DG on KdV. Central flux. Relative error $\leq 5\%$.

To better illustrate the converging results we plot the exact solution and the central flux approximation for the example problem stated in Section 3.3 for $m = 1, 2$ and $N = 32, 64$ in Figure 3.1,3.2,3.3 and 3.4. Here the exact solution is black dashed and the approximate solution is colored differently for each element.

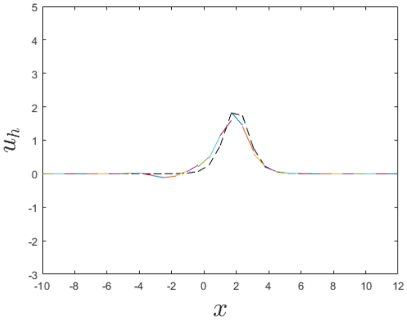

Figure 3.1: CF, $m = 1, N = 32$

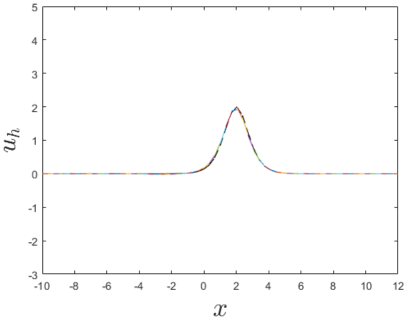

Figure 3.2: CF, $m = 1, N = 64$

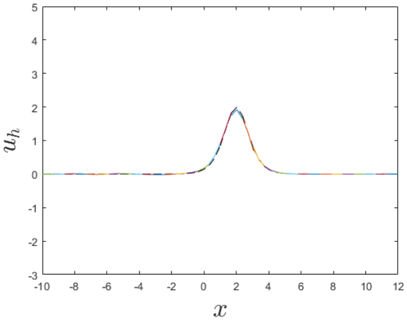

Figure 3.3: CF, $m = 2, N = 32$

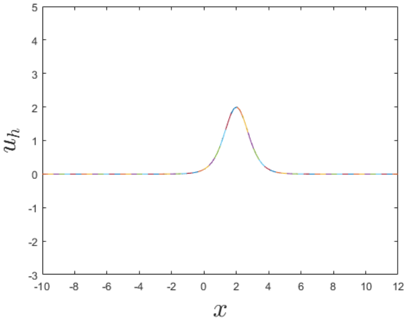

Figure 3.4: CF, $m = 2, N = 64$

3.5 Alternating flux (LGD) results

The $\mathcal{L}^2$ errors for alternating flux, i.e. $\hat{u} = u^-$, $\hat{q} = q^+$ and $\hat{p} = p^+$ are shown in table 3.3. Similarly to the central flux, Lax-Friedrichs flux is chosen for f.

$\mathcal{L}^2$ error, KdV equation, LDG				
$m\backslash N$	N=32	N=64	N=128	N=256
1	1.0374e+00	5.3071e-01	1.9033e-01	7.0300e-02
2	9.9424e-02	1.5605e-02	2.7361e-03	4.8217e-04
3	1.0690e-02	8.1607e-04	7.4421e-05	6.6309e-06
4	4.9140e-04	4.7418e-05	2.1528e-06	1.0864e-07

Table 3.3: KdV eqution $u_t + u_{xxx} + 6uu_x = 0$. $u(x, 0) = 2\text{sech}^2(x)$. Periodic boundary condition. Non uniform meshes with N elements. $\mathcal{L}^2$ error for LDG. CFL = 0.00005.

And stable CFL numbers are displayed in table 3.4.

Stable CFL numbers, LDG				
$m\backslash N$	N=32	N=64	N=128	N=256
1	2.8243e-01	9.8477e-02	2.7813e-02	7.0697e-03
2	2.7813e-02	1.0775e-02	2.7389e-03	6.5501e-04
3	2.7389e-03	2.4650e-03	6.2658e-04	1.4985e-04
4	6.2658e-04	6.2658e-04	2.1847e-04	5.8053e-05

Table 3.4: Stable CFL numbers. DG on KdV. Central flux. Relative error $\leq 5\%$

Similarly as we did in the central flux, we plot the exact solution and the LDG approximation for the example problem stated in Section 3.3 for $m = 1, 2$ and $N = 32, 64$ in Figure 3.5,3.6,3.7 and 3.8. Here the exact solution is black dashed and the approximate solution is colored differently for each element.

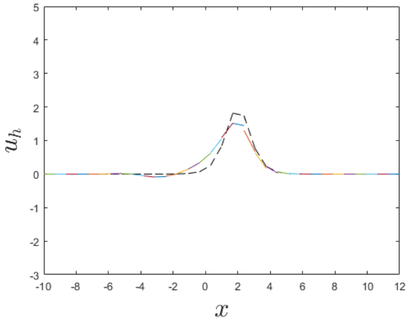

Figure 3.5: LDG, $m = 1, N = 32$

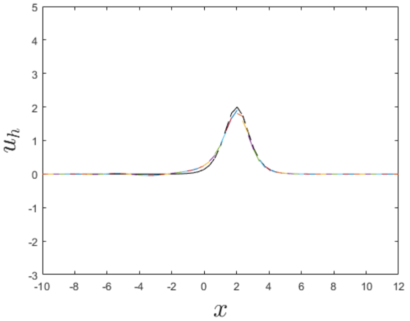

Figure 3.6: LDG, $m = 1, N = 64$

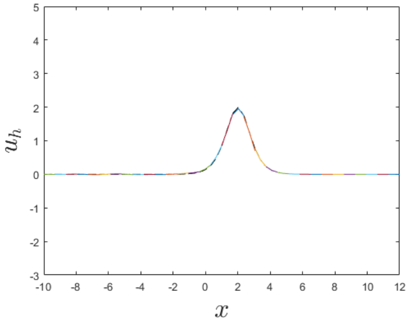

Figure 3.7: LDG, $m = 2, N = 32$

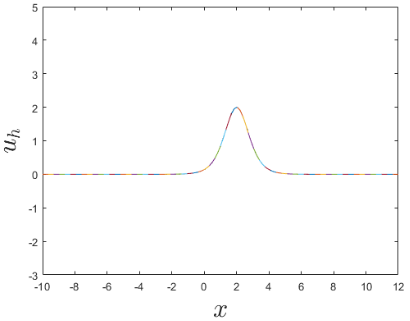

Figure 3.8: LDG, $m = 2, N = 64$

3.6 Comparing central flux to LDG

To better understand the accuracy and applicability of the results, we compare convergence between central flux and LDG for m and N in figure 3.9 and 3.10. As can be seen, the convergence rates for both methods are alike and of high order. The errors drop almost spectrally for m and h-convergence is shown for N. Comparing the two methods, no strong strengths or flaws can be concluded except the alternating m convergence for central flux. Much like discussed when it is applied to the heat equation, it shows lower convergence for m being even.

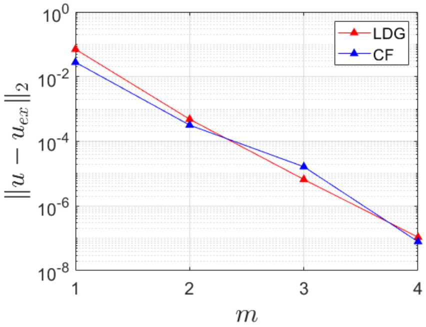

Figure 3.9: m convergence

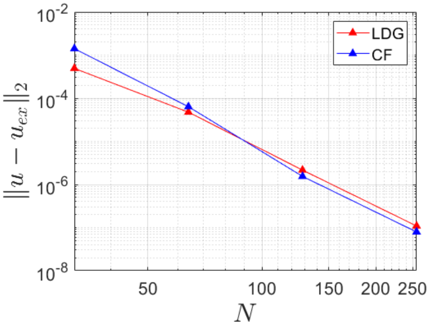

Figure 3.10: N convergence

The CFL numbers required for converging computation, i.e. CFL numbers which yield relative error < 0.05, are plotted for both central flux and LDG in Figures 3.11 and 3.12 together with the expected stability requirements, h^3 and m^{-6}, we see results in favour of central flux which can approximately double its timestep and still provide the same stability as LDG [9]. The stability requirements, although not broken, appear to be very relaxed which could be attributed to the smoothness of the KdV solution. Note that the stability requirements are proportional and can therefore be shifted, it is the slope that is of interest.

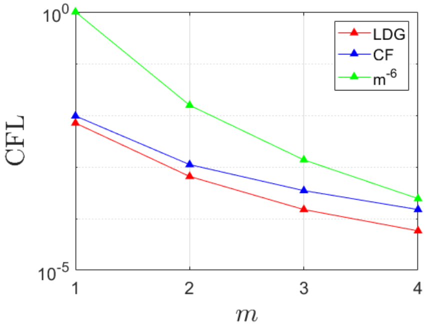

Figure 3.11: Converging CFL numbers for m

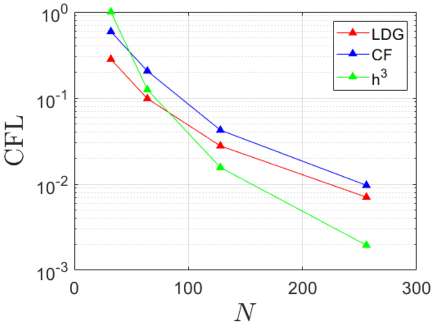

Figure 3.12: Converging CFL numbers for N

Chapter 4

Conclusion

The Discontinuous Galerkin method appears well suited for capturing Korteweg-de Vries equation, see Figure 4.1. The high order approximation shows promising convergence, $\mathcal{O}(h^m)$, despite the high spatial differentiation for alternating fluxes. For even m, central flux also holds spectral convergence but shows some sub-optimality when m is odd. This is likely due to the dissipation which is to be expected as we discussed the heat equation in section 2.10. The solver

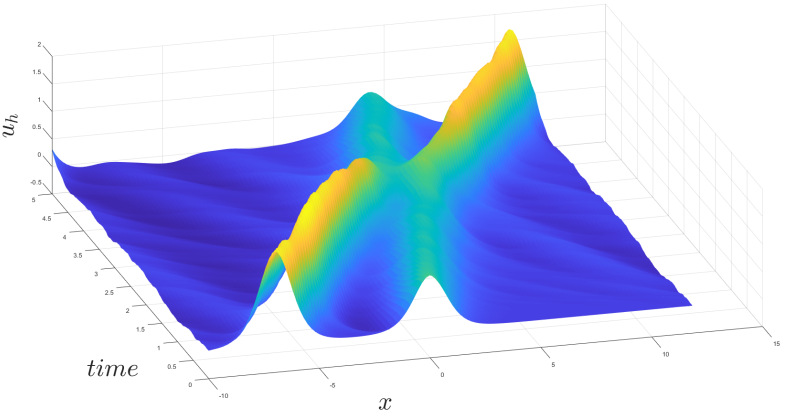

Figure 4.1: Korteweg-de Vries equation solved with Discontinuous Galerkin, central flux, $m = 2$,$N = 64$. 5 seconds of wave propagation.

is clearly stiff compared to a first order spatial derivative with the alternating flux appearing twice as stiff as the central flux and thus the flux should be decided depending on the problem at hand. Although the numerical results are promising there is a lot left for future research. Questions could revolve around stability and convergence in the multi-dimensional case. Additionally, analytical proof of nonlinear convergence for the central and alternating fluxes would greatly strengthen the applicability of DG methods applied to the KdV. We also highlight that the stiffness results relies on the semidiscrete formulation of the DG. Perhaps implicit time integration method yields other stiffness results.

Bibliography

[1] David A. Kopriva. *Implementing Spectral Methods for Partial Differential Equations.* Springer, 2009.

[2] Robert A.Adams and John J.F Fournier. *Sobolov Spaces 2E.* Elsevier Ltd, 2003.

[3] Khoei Amir R. *Extended finite element method : theory and applications.* Wiley, 2015.

[4] Douglas N Arnold, Franco Brezzi, Bernardo Cockburn, and L Donatella Marini. Unified analysis of discontinuous galerkin methods for elliptic problems. *SIAM journal on numerical analysis*, 39(5):1749–1779, 2002.

[5] Szabo Barna Aladar. *Introduction to finite element analysis: formulation, verification and validation.* Wiley, 2011.

[6] Francesco Bassi and Stefano Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible navier–stokes equations. *Journal of computational physics*, 131(2):267–279, 1997.

[7] George Cockburn, Bernardo & E. Karniadakis and Chi-Wang Shu(Eds.). *Discontinuous Galerkin Methods: Theory, Computation and Applications.* Springer, 2000.

[8] Peter D. Lax. *Hyperbolic Partial Differential Equations.* AMS, 2017.

[9] Jan S. Hesthaven. *Numerical Methods for Conservation Laws From Analysis to Algorithms.* SIAM - Society for Industrial and Applied Mathematics, 2018.

[10] Jan S. Hesthaven and Tim Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis and Applications.* Springer, 2008.

[11] Chi-Wang Shu and Jue Yan. A local discontinuous galerkin method for kdv type equations. *SIAM J. Numer. Anal.*, 40(2):769–791, 2002.

[12] Walter A Strauss. *Partial Differential Equations- An Introduction 2E.* John Wiley & Sons Inc, United States, 2008.

[13] John C Strikwerda. *Finite difference schemes and partial differential equations.* SIAM, 2004.

[14] Gerald Beresford Whitham. *Linear and nonlinear waves.* John Wiley & Sons, 2011.

Linköping University Electronic Press