NTNU

Norwegian University of
Science and Technology

On Operator Splitting for the Viscous Burgers' and the Korteweg-de Vries Equations

Espen Birger Nilsen

Master of Science in Physics and Mathematics
Submission date: June 2011
Supervisor: Helge Holden, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem Description

Study the paper *Operator Splitting for Partial Differential Equations with Burgers Nonlinearity*, by Holden, Lubich and Risebro. Extend the approach in the article to also yield for Godunov splitting for the viscous Burgers' equation, the Korteweg–de Vries equation and other equations. Perform numerical experiments for the viscous Burgers' equation and the Korteweg–de Vries equation, to test the operator splitting method in practice.

Assignment given: 17 January 2011,
Supervisor: Helge Holden.

On Operator Splitting for the Viscous Burgers' and the Korteweg–de Vries Equations

Espen Birger Nilsen

Department of Mathematical Sciences

NTNU 2011

Mathematics is a game played according to certain simple rules
with meaningless marks on paper
David Hilbert

Preface

This paper constitutes my master's thesis, written at the Norwegian University of Science and Technology (NTNU) the last semester of a five years Master's degree. The thesis was written at the Department of Mathematical Sciences under the supervision of Professor Helge Holden.

When I first heard about the operator splitting method, I was fascinated by the somewhat naive idea. It seemed to good to be true that it was possible to split a partial differential equation into several subequations, solve each of the subequations for small time steps, and concatenate the solutions from the subequations to yield the solution of the original equation.

When I started working on the thesis, Helge and I agreed that I should go into the depth of the article *Operator splitting for partial differential equations with Burgers nonlinearity*, cf. [11], and extend the framework in the article to also yield for the operator splitting of Godunov type. The road to achieve this goal has been long and interesting, and I have had the privilege to study several different topics, which all was combined to achieve the main goal of this thesis. The study started with the Peano kernel theorem, which I had never heard about and which I today think gives a pretty elegant result. I continued with studying the abstract differential calculus in Banach spaces. I spent a lot of time studying the fractional Sobolev spaces, and I struggled trying to prove the Banach algebra property of this space, and it was with great satisfaction that I finally managed to find a proof. With all the necessary results in my tool box, I could take on the operator splitting method. At last, when I had understood the details in the proof of the Strang splitting in [11] and in addition found a proof for the Godunov splitting, I started with a numerical study of the operator splitting method. Throughout the numerical study I had to study different numerical methods for the subequations from the splitting approach, and implement them to work in an operator splitting framework. It was fascinating to actually see that the operator splitting method worked in practice.

I am satisfied with the final result. Trough the semester I have been a lot on my own, trying to get all the details fitted together, and I have been fortunate enough to get to work on a subject I find interesting. Therefore, I really feel that this is *my* thesis, and that it is a worthy end to a five year long study of physics and mathematics.

A few thanks are in order. First, I want to thank my supervisor Helge Holden for fruitful discussions and excellent guidance throughout the semester. My brother Håvard Johannes Nilsen has earned himself a big thanks for proof reading this text. Last but not least, I thank my girlfriend Elisabeth Raknes Brekke for all the support and encouragement she has given me.

Trondheim, June 5, 2011.

Espen Birger Nilsen

Abstract

We discuss numerical quadratures in one and two dimensions, which is followed by a discussion regarding the differentiation of general operators in Banach spaces. In addition, we discuss the standard and fractional Sobolev spaces $H^s(\mathbb{R})$, and prove several properties of these spaces.

We show that the operator splitting methods of the Godunov type and Strang type applied to the viscous Burgers' equation, $u_t = u_{xx} + uu_x$, and the Korteweg–de Vries (KdV) equation, $u_t = u_{xxx} + uu_x$, (and other equations), have the correct convergence rates in $H^s(\mathbb{R})$, for arbitrary integer $s \geq 1$. In the proofs we use the new framework originally introduced in [11].

We investigate the Godunov method and Strang method numerically for the viscous Burgers' equation and the KdV equation, and present different numerical methods for the subequations from the splitting. We numerically check the convergence rates for the split step size Δt, in addition with other aspects for the numerical methods. We find that the operator splitting methods work well numerically for the two equations. For the viscous Burgers' equation, we find that several combination of numerical solvers for the subequations work well on the test problems, while we for the KdV equation find only one combination of numerical solvers which works well on all test problems.

Contents

1 Introduction

Throughout the last century, several new physical phenomena has been discovered, and to describe them mathematically it is possible to use partial differential equations. The theory of quantum mechanics and the motion of fluid substances are examples of such phenomena which has given arise to new equations; the Schrödinger[1] and Navier[2]–Stokes[3] equations, respectively. The complexity of the equations have increased due to an increase in the complexity of the phenomena which the equations are a model of. To be able to solve these new equations, new solving methods have been developed. The popularity and usability of these new solving methods have varied. The operator splitting method is a successful solving method, which have been widely studied for several years. The strategy behind the operator splitting approach is to "divide and conquer" the problem.

The idea behind the operator splitting approach is to split the partial differential equation into subequations, each which hopefully models different physical aspects, and solve each of these subequations for small time steps. To form a full solution, the solutions at each time step are concatenated in a special order, which hopefully yield the correct solution to the original equation. A beauty of the method is that the splitting of the full equation into subequations is easy to understand, even for non-mathematician.

From a modelling point of view, one wishes that a model of a physical phenomenon should involve some physical aspects like for instance convection or diffusion. These aspects give partial differential equations which contain terms that mathematically are different, and which often result in major challenges in the analysis of the model. By applying the operator splitting method to such problems, the different physical characteristics can be separated, so that each of the subequations describe these aspects mathematically. From a mathematical point of view, this is brilliant since the different terms are separated in different subequations, such that the mathematical behaviour is separated. In addition, it is (hopefully) possible to solve these subequations more easily than the full equation. For an introduction to the operator splitting technique from a mathematical view point, we recommend [12].

[1]Erwin Schrödinger, 12 August 1887 – 4 January 1961, Austrian physicist. Received his Ph.D. in 1910 in Vienna. Professor among other places at Stuttgart, Zurich and Oxford. Known as one of the fathers to quantam mechanics, formulated in 1926, which in few words are summarized by the famous Scrödinger equation, for which he recieved the Nobel prize in Physics in 1933, in collaboration with P. A. M. Dirac.

[2]Claude-Louis Navier, 10 February 1785 – 21 August 1836, French engineer and physicist. He was admitted to the French Academy of Science in 1824, and took up a professorship at the École Nationale des Ponts et Chausses in 1830. Formulated the general theory of elasticity in a mathematicaly usable form. Made several contributions to the theory of structural analysis, and is recognized as one of the founders of this theory. Today, he is most famous for the development of the Navier–Stokes equation in fluid mechanics.

[3]Sir George Gabriel Stokes, 13 August 1819 – 1 February 1903, British mathematician and physicist. Made major contributions to fluid dynamics, optics and mathematical physics. Several important laws from these fields are named after him. In 1849, Stokes was appointed to the Lucasian professorship of mathematics at Cambridge, a position which he held until his death in 1903. Given an honorary doctorate at the University of Oslo in 1902.

The main focus in this text is to apply the operator splitting method on two famous equations; *the viscous Burgers' equation* and *the Korteweg–de Vries equation*, and we will study the operator splitting method from an analytical and numerical point of view. Before we start with the mathematical treatment, we give a brief historical background for the two equations.

The viscous Burgers' equation is a nonlinear partial differential equation of second order, which has its origin in the famous Navier–Stokes equation. In 1939 Johannes Martinus Burgers[4] simplified the Navier–Stokes equation, which resulted in what today is called the viscous Burgers' equation. The equation is built up by a nonlinear convection term and a linear diffusion term, which often is referred to as the viscosity effect.

The equation appears in fluid dynamics and in general engineering as a simplified model for turbulence, boundary layer behaviour, mass transport and wave propagation in acoustics. In the last decades the equation has also been used to model traffic flows. The equation is described in the following way,

$$u_t + uu_x = \kappa\, u_{xx}$$

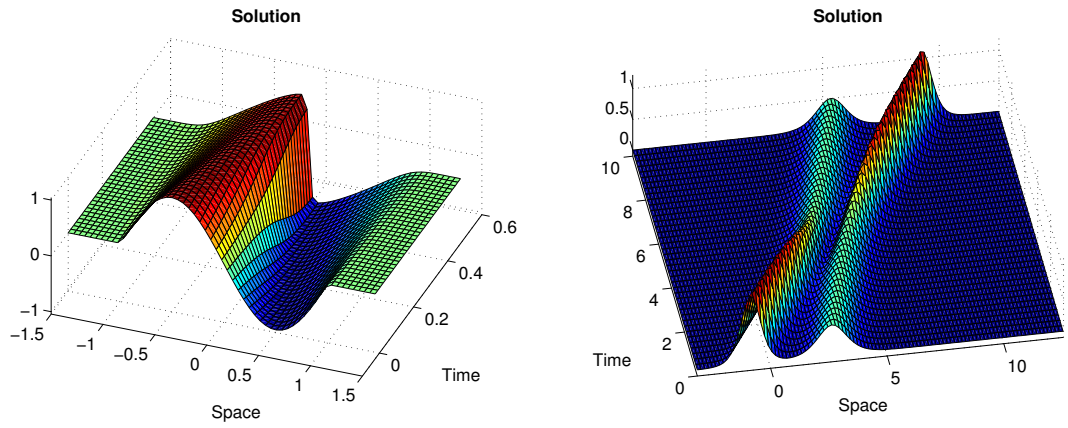
where $\kappa \geq 0$ is the viscosity factor determing the amount of viscosity added to the solutions. The nonlinear term uu_x transports the information, while the second order term u_{xx} smooths the information. The sign in front of the nonlinear term identify the direction of the transport. The sign in front of u_{xx} is important, since a negative sign results in an ill-posed problem.

The viscous Burgers' equation has been studied and applied for many decades, resulting in the existence of many analytical solutions, series approximations and numerical solutions for a wide range of boundary conditions. In that sense, the study of the equation is finished. For an example of a solution of the equation see Figure 1.1.

The Korteweg–de Vries (KdV) equation has a more interesting historical background. The history starts back in 1834, when the Scottish naval engineer John Scott Russel[5] observed a boat which was drawn by two horses in a narrow channel. When the boat stopped he observed a sharp, but smoothed heap of water propagating along the water surface. He jumped on the horseback and followed the wave for several kilometers. This wave shape was new and so interesting that he started practical and theoretical investigations of the phenomenon he just had observed. Through his studies he found several properties of these waves. The wave is stable and can travel over large distances, compared to normal waves which tends to be smoothed out if they travel over long distances. Another interesting property is that two waves never will merge; one small wave is overtaken by a large one. Russel called the phenomenon "Wave of translation" — today such waves are called *solitons* or *solitory waves*.

[4] Johannes Martinus Burgers, January 13 1895 – June 7 1981, Dutch physicist. Received his Ph.D. in 1918, under supervision of Paul Ehrenfest. The father of Burgers' equation, the Burgers vector in dislocation theory and the Burgers material in viscoelasticity.

[5] John Scott Russell, 9 May 1808 – 8 June 1882, Scottish naval engineer. Famous for the building of the iron sailing steam ship Great Eastern, in collaboration with Isambard Kingdom Brunel, and the discovery of the soliton phenomenon in fluid dynamics.



(a) The viscous Burgers' equation

(b) The Korteweg–de Vries equation

Figure 1.1: Solutions to the viscous Burgers' equation and the Korteweg–de Vries equation. In (a) observe how the initial wave is getting steeper and steeper as it travels to the right, but the wave never breaks. In (b) observe how the two solitons intersect without changing velocities or shape.

The first theoretical investigation of the soliton phenomen was done by Lord Rayleigh[6] and Joseph Boussinesq[7] in the 1870s. In a paper from 1895, Diederik Korteweg[8] and Gustav de Vries[9] investigated the solution phenomenon, and took the theory of solutions a step further. They found the partial differential equation, which the soliton phenomenon satisfies. The equation is given as

$$u_t + uu_x + \kappa u_{xxx} = 0,$$

and is named after the two authors. The coefficient κ denote the sharpness of the solitons. A small κ results in sharp solitons, while a large κ result in smoother solitons. Since a steep soliton moves faster than a smoother soliton, κ also in some sense identifies the velocity of the soliton. The paper by Korteweg and de Vries has become a milestone

[6] John William Strutt Rayleigh, 12 November 1842 – 30 June 1919, English physicist. Famous for major contributions in many fields in physics (physical chemistry, optics, acoustics, wave motion, and more), and published more than 400 papers. In 1894 he discovered, in collaboration with N. F. Ramsay, the element argon in the atmosphere, for which he received the Nobel prize in Physics in 1904. Received an honorary doctorate at the University of Oslo in 1902.

[7] Joseph Valentin Boussinesq, 13 March 1842 – 19 February 1929, French mathematician and physicist. Made significant contributions to the theory of hydrodynamics, vibration, light and heat. Appointed professor at Faculty of Sciences of Lille from 1872 to 1886. From 1896 to his retirement in 1918 he was professor of mechanics at Faculty of Sciences of Paris.

[8] Diederik Johannes Korteweg, 31 March 1848 – 10 May 1941, Dutch mathematician. Received his Ph.D. from the University of Amsterdam in 1878. Remembered as the joint discoverer of the Korteweg–de Vries equation.

[9] Gustav de Vries, 22 January 1866 – 16 December 1934, Dutch mathematician. Finished his Ph.D. in 1894 under supervision of Diedrik Korteweg. Remembered for the discovery of the Korteweg–de Vries equation.

in the theory of solitons, which was developed further in the 1960s. Nowadays, the KdV equaton is well-studied and several regularity results regarding the solutions are proven. It is used to model several different physical phenomena like waves in plasma and ion-acoustic waves. We refer to [17] for an discussion of the KdV eqaution for different physical phenomena. An example of a solution is given in Figure 1.1.

The operator splitting method is an approximation method, and when the method is applied it yields some error which is dependent on the split step size and how the solutions of the subequations are concatenated. Therefore, to use this method in practice one needs to prove that the operator splitting solution converges to the exact solution of the full equation when the split step size tends to zero. The litterature contains several convergence proofs and techniques for different partial differential equations. In [13], a new general analytical approach which involves introducing a two dimensional approximation is used to prove convergence for the KdV equation. Recently, in [11] another analytical framework is introduced, which is used to prove convergence for a wide class of partial differential equations which has a so-called Burgers' nonlinearity. The class includes the viscous Burgers' and the KdV equations.

In this text we use the analytical approach presented in [11], to prove convergence rates for the splitting of *Godunov*[10] and *Strang*[11] types, which formally is introduced in Section 5. To be ahead of the text, the Godunov splitting formally converges as $\mathcal{O}(\Delta t)$, while the Strang splitting converges as $\mathcal{O}\left((\Delta t)^2\right)$. In [11] the correct convergence rate for the Strang splitting is proven. We adopt this idea, and use the same technique to prove the correct convergence rate for the Godunov splitting as well.

The approach in [11] relies heavily on the differential theory of operators in Banach[12] spaces and the error terms of one and two dimensional numerical quadratures, together with Sobolev[13] space properties. Therefore, we start with a discussion of nu-

[10]Sergei Konstantinovich Godunov, 17 July 1929 – , Russian mathematician. Famous for contributions in applied and numerical mathematics, for instance the Godunov's methods for conservation laws. His contributions have had major impact on science and engineering. Became a member of the Russian Academy of Sciences in 1994, and a honorary professor of the University of Michigan in 1997. Currently, Professor at the Sobolev Institute of Mathematics of the Russian Academy of Sciences in Novosibirsk, Russia.

[11]William Gilbert Strang, 27 November 1934 – , American mathematician. Recieved his Ph.D. in 1959 at University of California, Los Angeles. Known for contributions to the finite element theory, the calculus of variations, wavelet analysis and linear algebra. Has written several textbooks, which have become classics in the teaching of mathematics. Currently, Professor of Mathematics at the Massachusetts Institute of Technology (MIT).

[12]Stefan Banach, 30 March 1892 – 31 August 1945, Polish mathematican. His doctoral thesis (1922) introduced the basic ideas of functional analysis, which became a totally new branch of mathematics. In 1922 he became a professor at Lwów Polytechnic. During the following years he made several publications on functional analysis and linear metric spaces. Was accepted as a member of the Polish Academy of Learning in 1924. One of the founders of the "Lwów School of Mathematics", which had meetings in the Scottish Café. Banach published the first monograph on the general theory of linear-metric space in 1931. Diagnosed with lung cancer during World War Two, and died in August 1945.

[13]Sergei Lvovich Sobolev, 6 October 1908 – 3 January 1989, Soviet mathematician. Famous for introducing the distribution theory and the Sobolev spaces. He was a Moscow State University professor from 1935 – 1957, and participated in the nuclear weapon program in the USSR. Played an important role in the establishment and development of Novosibirsk State University.

merical quadrature formulas, before we continue with a discussion including proofs of the necessary results from the differential theory. The Sobolev spaces are introduced and presented, and we give proofs of for instance the Banach-algebra property for the standard and fractional Sobolev spaces. Then, the operator splitting method is discussed in details, and we give proofs of the abovementioned convergence rates for the viscous Burgers' and KdV equations (and other equations). At the very end of this text, we present numerical experiments with the use of the operator splitting method.

2 Differentiation in Banach Spaces

In this section we consider the differentiability of general operators in Banach spaces, and we start by defining the Fréchet[14] differential of an operator, which is followed up by a discussion of some results regarding the definition. At the end of this section we prove a variational of parameters formula.

Remark: The notation for differential theory in vector spaces gets quickly rather messy, and is not standardized in the literature. We adopt the notation in [3], and try to be consistent with this notation in the remainder of the text.

2.1 The Fréchet Differential

The Fréchet differential is the generalization in Banach spaces of the usual differential of a function in Euclidean[15] spaces. The definition is as follows.

Definition 2.1. *Let X and Y be Banach spaces, and let V be an open subset in X, $V \subset X$. A mapping $T : V \to Y$ is Fréchet differentiable at a point v in V if there exists A in $L(X, Y)$ such that*

$$\lim_{\|h\|_X \to 0} \frac{\|T(v+h) - T(v) - A(h)\|_Y}{\|h\|_X} = 0. \tag{2.1}$$

A is called the Fréchet differential of T at a point v, and is denoted by

$$A = dT(u).$$

T is said to be Fréchet differentiable in V if it is differentiable for all v in V .

Introducing the "little-oh" notation, requirement (2.1) is the same as requiring

$$\|T(v+h) - T(v) - A(h)\|_Y = o(\|h\|_X).$$

From the definition, we see that the differential of an operator is unique and linear. Moreover, if A satisfies (2.1), then A is continuous at v if and only if T is continuous at v. The generalization to Banach spaces of the chain rule in Euclidean spaces, is given in the following lemma.

Proposition 2.2. *Let X, Y and Z be Banach spaces and let $T : V \to Y$ and $S : U \to Z$, where $T(V) \subset U$, V and U are open subsets of X and Y, respectively. The composite map is defined as*

$$S \circ T : V \to Z, \quad (S \circ T)(v) = S(T(v)).$$

[14] Maurice René Fréchet, 2 September 1878 – 4 June 1973, French mathematician. Made major contribution to topology and introduced the concept of metric spaces, as a part of his doctoral thesis in 1906. Also made contributions to the the field of statistics, probability and calculus. Independently of Riesz, he discovered the representation theorem in the space of Lebesgue square integrable functions.

[15] Euclid, lived about 300 BC, Greek mathematician. His *Elements* is one of the most influential works in the history of mathematics, introducing what we today call Eucledian geometry. *Elements* has been one of the main textbooks in the teaching of mathematics up to the 20th Century.

If T is differentiable at v in V and S is differentiable at $u = T(v)$ in U, then $S \circ T$ is differentiable at v and

$$d(S \circ T)(v)h = dS(v)[dT(u)h].$$

From the above proposition we see that the differential of $S \circ T$ at v is the composition of the linear maps $dF(v)$ and $dG(F(u))$. We define the Fréchet derivative, as a mapping between Banach spaces.

Definition 2.3. *Let $T : V \to Y$ be Fréchet differentiable in V. The mapping*

$$T' : V \to L(X, Y), \quad T' : v \to dT(v),$$

is called the Fréchet derivative of T.

The higher order differentials will be defined in the same manner as the first order differential. Let T be in $C^1(V, Y)$ and consider the derivative $T' : V \to L(X, Y)$.

Definition 2.4. *Let X and Y be Banach spaces and $V \subset X$. Let $T : V \to Y$ be continuous and consider v in V. T is twice Fréchet-differentiable at v if T' is differentiable at v. The second differential of T is defined as*

$$d^2T(v) = dF'(v).$$

If T is twice differentiable for all v in V, T is said to be twice differentiable in V.

From this definition, it is clear that $d^2T(v)$ is a continuous linear map from X to $L(X, Y)$, more formally $d^2T(v)$ is in $L(X, L(X, Y))$. This mapping is a symmetric bilinear mapping on X, which is proven in [3, Ch. 1.]. The space $L(X, L(X, Y))$ is isometric isomorphic with $L_2(X, Y)$. In the following a value of $d^2T(v)$ at a pair (g, h) will be denoted as

$$d^2T(v)[g, h].$$

If T is twice differentiable in V, then the second derivative, T'', is a mapping from $V \to L_2(X, Y)$.

The definition of the $(n + 1)$-th derivate is done by induction. Let $T : V \to Y$ be n times differentiable in V. Let the n-th differential at a point v in V be represented with a continuous n-linear mapping from $X \times X \times \ldots \times X \to Y$. As was the case for the second derivative, there is an isometric isomorphism between $L(X, \ldots, L(X, Y))$ and $L_n(X, Y)$. Let $T^{(n)} : V \to L_n(X, Y)$ denote the mapping

$$F^{(n)} : v \to d^nT(v).$$

The $(n + 1)$-th differential at v is defined as the differential of $T^{(n)}$, namely

$$d^{(n+1)}T(v) = dT^{(n)}(v) \in L_{n+1}(X, Y).$$

If T is n times differentiable in V and the nth derivative is continuous from V to $L_n(X,Y)$, then T is in $C^n(V,Y)$. The value of the differential at a point $(h_1,\ldots,h_n)$ is denoted by

$$d^nT(v)[h_1,\ldots,h_n].$$

If $h = h_1 = \ldots = h_n$, $d^nT(v)[h]^n$ will be written for short. The mapping $(h_1,\ldots,h_n) \to d^nT(v)[h_1,\ldots,h_n]$ is symmetric, if T is n times differentiable in V.

We illustrate the Fréchet derivative using two operators for which the derivative will be used later on. Let A be an general operator in $L(X,Y)$ and define $B : X \to Y$ by $B(v) = vv_x$. Using Definition 2.1, the first derivative of A is found as,

$$\lim_{\|h\|_X\to 0} \frac{\|A(v+h) - A(v) - dA(v)[h]\|_Y}{\|h\|_X} = \lim_{\|h\|_X\to 0} \frac{\|A(h) - dA(v)[h]\|_Y}{\|h\|_X} = 0,$$

where we have used the linearity of A. This yield

$$dA(v)[h] = A(h). \tag{2.2}$$

Furthermore, for B we get

$$\begin{aligned}\lim_{\|h\|_X\to 0} &\frac{\|(v+h)(v+h)_x - vv_x - dB(v)[h]\|_Y}{\|h\|_X} \\ &= \lim_{\|h\|_X\to 0} \frac{\|vh_x + hv_x + hh_x - dB(v)[h]\|_Y}{\|h\|_X} = 0,\end{aligned}$$

from which

$$dB(v)[h] = (vh)_x = vh_x + v_xh. \tag{2.3}$$

The higher order derivatives are found similarly. For A we get

$$\begin{aligned}\lim_{\|k\|_X\to 0} &\frac{\|dA(v+k) - dA(v) - d^2A(v)\|_{L(X,Y)}}{\|k\|_X} \\ &= \lim_{\|k\|_X\to 0} \sup_{\|h\|_X=1} \frac{\|dA(v+k)[h] - dA(v)[h] - d^2A(v)[h,k]\|_Y}{\|k\|_X} \\ &= \lim_{\|k\|_X\to 0} \sup_{\|h\|_X=1} \frac{\|A(h) - A(h) - d^2A(v)[h,k]\|_Y}{\|k\|_X} = 0,\end{aligned}$$

which gives

$$d^2A(v)[h,k] = 0, \tag{2.4}$$

and all the higher derivatives is 0. Furthermore, for B we obtain

$$\begin{aligned}\lim_{\|k\|_X\to 0} \sup_{\|h\|_X=1} &\frac{\|dB(v+k)[h] - dB(v)[h] - d^2B(v)[h,k]\|_Y}{\|k\|_X} \\ &= \lim_{\|k\|_X\to 0} \sup_{\|h\|_X=1} \frac{\|(vh)_x + (kh)_x - (vh)_x - d^2B(v)[h,k]\|_Y}{\|k\|_X} = 0,\end{aligned}$$

which gives

$$d^2B(v)[h,k] = (hk)_x = h_xk + hk_x. \tag{2.5}$$

Since $d^2B(v)$ do not depend on v,

$$d^nB(f)[g,h,k] = 0,$$

for $n \geq 3$.

As a final remark, we note that all the vector spaces was assumed to be Banach spaces, but only the properties of a normed space was used in the definitions. Thus, the completeness assumptions on the vector spaces could have been leaved out. This would have given a slightly more general theory. However, the spaces we will look at in the remains of this text are Banach spaces.

2.2 A Variation of Parameters Formula

In the theory of linear partial differential equations, the Duhamel's principle[16] is a solution method for nonhomogeneous initial value problems, which involves first solving a homogeneous version of the initial value problem, from which the solution is integrated to yield a solution of the nonhomogeneous initial value problem. In [5, Ch. 2.3.] the Duhamel's principle is introduced and used to solve the nonhomogeneous heat equation over $\mathbb{R}^n \times (0,\infty)$. The Duhamel's principle has wide applicability to both linear ordinary differential equations and other partial differential equations. The idea behind the *variation of parameters* formulas, is to create a framework where formulas similar to the Duhamel's principle, can be utilized in a more general setting.

To emphasize the idea consider the two initial value problems

$$v_t = A(v), \quad v|_{t=t_0} = v_0, \tag{2.6}$$

and

$$u_t = A(u) + B(u), \quad u|_{t=t_0} = u_0, \tag{2.7}$$

where A and B are some differential operators. Let $v(t;t_0,v_0)$ and $u(t;t_0,u_0)$ denote the solutions of (2.6) and (2.7) with initial data (t_0,v_0) and (t_0,u_0), respectively. The question is if we can write the solution of (2.7) using the solution of (2.6) in the following way,

$$u(t;t_0,u_0) = v(t;t_0,u_0) + \int_{t_0}^{t} v\big(t;s,B(u(s;t_0,u_0))\big)\,ds.$$

The above equation is an example of a variation of parameters formula.

[16] Jean-Marie Duhamel, 5 February 1797 – 29 April 1872, French physicist and mathematician. Proposed a theory dealing with transmission of heat in crystal structures, and made contributions to infinitesimal calculus.

To be more precise regarding A and B, let X and Y be Hilbert[17] spaces where X is continuously imbedded in Y, and let $A : X \to Y$, and $B : X \to Y$. Furthermore, assume A is in $L(X, Y)$ and B is in $C^1(X, Y)$ and require $\|v\|_Y \leq \|v\|_X$ for all v in X. At last, assume $v(t; t_0, v_0)$ is $C^1([t_0, T], Y(\text{or } X))$ for a fixed $T > t_0$ and for all v_0 in $X(\text{or } Y)$.

To prove the variation of parameters formula, we need the following lemma.

Lemma 2.5. *The Fréchet derivatives of $v(t; , t_0, v_0)$ satisfy: For v_0 and w in Y*

$$\frac{\partial}{\partial v_0} v(t; t_0, v_0)[w] = v(t; t_0, w), \tag{2.8}$$

and for v_0 in Y,

$$\frac{\partial}{\partial t_0} v(t; t_0, v_0) = -v(t; t_0, A(v_0)). \tag{2.9}$$

Proof. From Definition 2.1 we see that to prove (2.8) we have to show

$$\lim_{\|w\|_Y \to 0} \frac{\|v(t; t_0, v_0 + w) - v(t; t_0, v_0) - \frac{\partial}{\partial v_0} v(t; t_0, v_0)[w]\|_Y}{\|w\|_Y} = 0.$$

From the linearity with respect to the initial data, we have

$$v(t; t_0, v_0 + w) = v(t; t_0, v_0) + v(t; t_0, w),$$

which gives, by inserting into the above equation,

$$\frac{\partial}{\partial v_0} v(t; t_0, v_0)[w] = v(t; t_0, w),$$

and (2.8) is proven.

To prove (2.9), we define

$$z(\tau) = v(t; t_0 + \tau, v_0) - v(t; t_0, v_0) + \tau v(t; t_0, A(v_0)),$$

and we have to show $\|z\|_Y = o(\tau)$. The solution of (2.6) at $t = t_0 + \alpha$ and $t = t_0$ is the same as long as the initial data is given at the same time, thus we have

$$v(t + \alpha; s + \alpha, v_0) = v(t; s, v_0).$$

Using this result gives

$$z(\tau) = v(t - \tau; t_0, v_0) - v(t; t_0, v_0) + \tau v(t; t_0, A(v_0)).$$

[17] David Hilbert, 23 January 1862 – 14 February 1943, German mathematician. Recognized as one of the most influental and universial mathematicians of the 19th and 20th Centuries. He contributed to the theory of geometry, integral equations and functional analysis. He also contributed in physics in topics like kinectic gass theory and the theory of relativity. Received his Ph.D. in 1885. Professor at the University of Königsberg from 1886 to 1895, and at the University of Göttingen from 1895 to 1936. He is also known for his statement of the *23 unsolved problems*, at a congress in Paris in 1900. Some of them are still unsolved today.

Differentiation wrt. τ gives

$$z_\tau(\tau) = -v_t(t-\tau; t_0, v_0) + v(t; t_0, A(v_0)) = -A\big(v(t-\tau; t_0, v_0)\big) + v(t; t_0, A(v_0)),$$

where we have used (2.6). Thus,

$$\begin{aligned}\|z_\tau\|_Y &\le \|A(v(t-\tau; t_0, v_0))\|_Y + \|v(t; t_0, A(v_0))\|_Y \\ &\le \|A\|_{L(X,Y)}\|v(t-\tau; t_0, v_0)\|_X + \|v(t; t_0, A(v_0))\|_Y \le C_1 + C_2 \le C,\end{aligned}$$

where the second inequality follows from that A is in $L(X,Y)$, and the third inequality follows from that $v(t; t_0, v_0)$ is in $C^1([t_0, T], Y)$. Since $z(0) = 0$ we have

$$z(\tau) = \int_0^\tau z_\tau(s)\, ds,$$

which results in

$$\|z(\tau)\|_Y \le \int_0^\tau \|z_\tau(s)\|_Y\, ds \le C\tau,$$

from the estimate for $\|z_\tau\|_Y$. Hence, $\|z\|_Y = o(\tau)$ and (2.9) is proven. □

Using this lemma, we can prove the variation of parameters formula.

Theorem 2.6 (Variation of parameters formula)**.** *Let A be in $L(X,Y)$ and B in $C^1(X,Y)$ and assume that* (2.6) *has a unique bounded solution $v(t; t_0, v_0)$ in $C^1([t_0, T]; X)$ or $v(t; t_0, v_0)$ in $C^1([t_0, T]; Y)$. Furthermore, assume* (2.7) *has a unique solution $u(t; t_0, v_0)$ in $C^1([t_0, T]; X)$ for all v_0 in X. Then*

$$u(t; t_0, u_0) = v(t; t_0, u_0) + \int_{t_0}^t v\big(t; s, B(u(s; t_0, u_0))\big)\, ds.$$

Proof. Differentiation of v gives

$$\begin{aligned}\frac{d}{ds}v\big(t; s, u(s; t_0, u_0)\big) = &\frac{\partial}{\partial s}v\big(t; s, u(s; t_0, u_0)\big) \\ &+ \frac{\partial}{\partial v_0}v\big(t; s, u(s; t_0, u_0)\big)\Big[\frac{\partial}{\partial t}u(s; t_0, u_0)\Big],\end{aligned}$$

which by (2.8) and (2.9) are transformed to,

$$\frac{d}{ds}v\big(t; s, u(s; t_0, u_0)\big) = -v\big(t; s, A(u(s; t_0, u_0))\big) + v\big(t; s, \frac{\partial}{\partial t}u(s; t_0, u_0)\big).$$

By using (2.7) and the linearity of the initial data, we obtain

$$\begin{aligned}\frac{d}{ds}v\big(t; s, u(s; t_0, u_0)\big) &= -v\big(t; s, A(u(s; t_0, u_0))\big) \\ &\quad + v\big(t; s, A(u(s; t_0, u_0)) + B(u(s; t_0, u_0))\big) \\ &= v\big(t; s, B(u(s; t_0, u_0))\big).\end{aligned}$$

From (2.7) we have that $v(t; s, u(s; t_0, u_0))|_{s=t_0} = v(t; t_0, u_0)$ and from (2.6) we get $v(t; s, u(s; t_0, u_0))|_{s=t} = u(t; t_0, u_0)$. Hence,

$$u(t; t_0, u_0) - v(t; t_0, u_0) = \int_{t_0}^{t} v(t; s, B(u(s; t_0, u_0))) \, ds,$$

which proves the theorem. □

The assumptions for both Lemma 2.5 and Theorem 2.6 are restrictive and leads to a result which is not as general as possible. In our defense, we do this simplification because the assumptions satisfy the operator splitting problem which is studied later on. However, in [19, Ch. 4.2] it is shown that Lemma 2.5 also is valid if A is nonlinear, and Theorem 2.6 is extended to yield for Banach spaces. The proofs presented in [19, Ch. 4.2.] are much longer and more complicated as linearity of A and uniqueness of the solution of (2.6) and (2.7) are not assumed to hold. This leads to an interesting discussion regarding the uniqueness problem, and the continuity and differentiability with respect to the initial data of the solutions of (2.7).

3 Numerical Integration

In numerics the evaluations of the definite integrals

$$\int_a^b f(x)\,dx, \quad \text{and} \quad \int_a^b \int_c^d f(x,y)\,dy\,dx,$$

leads to the so-called quadrature formulas. The main focus in this section is the error term for one and two dimensional quadrature formulas, and we start with a treatment of the former. Due to Giuseppe Peano[18] the error in the one dimensional case, is given in a very compact and elegant form, which we prove in the fascinating *Peano kernel theorem*. We use this theorem to find error formulas for two one dimensional quadrature formulas, and in addition we find a error formula for a two dimensional quadrature formula.

3.1 One Dimensional Quadratures

The quadrature formula yield an approximation to the integral, and can in general be given as

$$I(f) = \sum_{k=0}^{m_0} w_{k0} f(x_{k0}) + \sum_{k=0}^{m_1} w_{k1} f'(x_{k1}) + \ldots + \sum_{k=0}^{m_n} w_{kn} f^{(n)}(x_{kn}), \tag{3.1}$$

where w_{ki} are the weights, $\{m_i\}_i \subset \mathbb{N}$, $\{x_{ki}\}_{k,i} \subset [a,b]$ and f is in $C^n[a,b]$. The error is given as

$$E(f) = I(f) - \int_a^b f(x)\,dx, \tag{3.2}$$

and tells us of how good (3.1) approximates the integral. Several different forms for (3.2) exist, and we focus on the elegant Peano kernel form, which we present below.

3.1.1 The Peano Kernel Theorem

The Peano kernel theorem relies on *a priori* knowledge of the involved quadrature rule. To be precise, we must know the degree of the polynomials which (3.1) integrates exactly. For (3.1), it is possible to find such degree n using for instance a Taylor[19] series expansion.

Assume f is in $C^{n+1}([a,b])$ and that (3.1) integrates all p in $\mathbb{P}_n$ exactly. Expanding f in a Taylor series yield

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)(x-a)^k}{k!} + \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n\,dt = p_n(x) + r_n(x),$$

[18] Giuseppe Peano, 27 August 1858 – 20 April 1932, Italian mathematician. Founder of the set theory, and the standard axiomatization of the natural numbers is named the Peano axioms in his honour. He made contributions to the method of mathematical induction, and wrote over 200 books and papers. In addition, he created an international language, *Latino sine flexione*, which is without grammar but has a rich vocabular.

[19] Brook Taylor, 18 August 1685 – 30 November 1731, English mathematician. Entered St John's College, Cambridge in 1701, and received the degrees of LL.B. and LL.D., in 1709 and 1714 respectively. Known for the Taylor's theorem and the Taylor series, and for contributions to the theory of differentials.

which is valid since f is in $C^{n+1}([a,b])$. The error E is a linear operator, thus $E(f) = E(p_n + r_n) = E(p_n) + E(r_n) = E(r_n)$, where $E(p_n) = 0$ because (3.1) integrates all p in $\mathbb{P}_n$ exactly. Thus, by using (3.2),

$$\begin{aligned} E(f) = E(r_n) &= I(r_n) - \int_a^b r_n(x)\,dx \\ &= I(r_n) - \frac{1}{n!}\int_a^b \int_a^x f^{(n+1)}(t)(x-t)^n\,dt\,dx \\ &= I(r_n) - \frac{1}{n!}\int_a^b \int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\,dx, \end{aligned}$$

where we have introduced

$$(x-t)_+^n = \begin{cases} (x-t)^n & \text{if } x \geq t, \\ 0 & \text{if } x < t. \end{cases}$$

Using (3.1) for r_n gives

$$\begin{aligned} E(r_n) &= I\left(\frac{1}{n!}\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\right) - \frac{1}{n!}\int_a^b \int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\,dx \\ &= \frac{1}{n!}\sum_{k=0}^{m_0} w_{k0}\int_a^b f^{(n+1)}(t)(x_{k0}-t)_+^n\,dt \\ &\quad + \frac{1}{n!}\sum_{k=0}^{m_1} w_{k1}\frac{d}{dx}\left[\int_a^b f^{(n+1)}(t)(x_{k1}-t)_+^n\,dt\right] \\ &\quad + \ldots + \frac{1}{n!}\sum_{k=0}^{m_n} w_{kn}\frac{d^n}{dx^n}\left[\int_a^b f^{(n+1)}(t)(x_{kn}-t)_+^n\,dt\right] \\ &\quad - \frac{1}{n!}\int_a^b \int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\,dx. \end{aligned} \tag{3.3}$$

The integration and differentiation in the sums have to be interchanged, to yield what we want. Thus, the following needs a proof,

$$\frac{d^k}{dx^k}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\right] = \int_a^b f^{(n+1)}(t)\,\frac{d^k}{dx^k}\left[(x-t)_+^n\right]\,dt, \tag{3.4}$$

for $1 \leq k \leq n$. For $k < n$ this follows immediately since $(x-t)_+^n$ is $n-1$ times continuously differentiable. In particular, for $k = n-1$ we get

$$\frac{d^{n-1}}{dx^{n-1}}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\right] = \int_a^b f^{(n+1)}(t)\,\frac{d^{n-1}}{dx^{n-1}}\left[(x-t)_+^n\right]\,dt,$$

which by evaluating the differentiation on the right hand side leads to

$$\begin{aligned} \frac{d^{n-1}}{dx^{n-1}}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\right] &= n!\int_a^b f^{(n+1)}(t)\,(x-t)_+\,dt \\ &= n!\int_a^x f^{(n+1)}(t)\,(x-t)\,dt. \end{aligned}$$

The integral on the right hand side is differentiable as a function of x, because the integrand is jointly continuous in x and t. Thus, by the fundamental theorem of calculus,

$$\begin{aligned}\frac{d}{dx}\left[\frac{d^{n-1}}{dx^{n-1}}\left[\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\right]\right] &= n!\,f^{(n+1)}(x)(x-x) + n!\int_a^x f^{(n+1)}(t)\,dt \\ &= 0 + \int_a^x f^{(n+1)}(t)\,\frac{d^n}{dx^n}[(x-t)^n]\,dt \\ &= \int_a^b f^{(n+1)}(t)\,\frac{d^n}{dx^n}[(x-t)_+^n]\,dt.\end{aligned}$$

This proves that (3.4) holds for $k = n$. We return to (3.3), and interchange the two operators,

$$\begin{aligned}E(r_n) = &\frac{1}{n!}\sum_{k=0}^{m_0} w_{k0}\int_a^b f^{(n+1)}(t)(x_{k0}-t)_+^n\,dt \\ &+ \frac{1}{n!}\sum_{k=0}^{m_1} w_{k1}\int_a^b f^{(n+1)}(t)\,\frac{d}{dx}\left[(x_{k1}-t)_+^n\right]\,dt \\ &+ \ldots + \frac{1}{n!}\sum_{k=0}^{m_n} w_{kn}\int_a^b f^{(n+1)}(t)\,\frac{d^n}{dx^n}\left[(x_{kn}-t)_+^n\right]\,dt \\ &- \frac{1}{n!}\int_a^b\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\,dx.\end{aligned}$$

From the continuity of the integrand $f^{(n+1)}(t)(x-t)_+^n$, we interchange the integration order of the double integral,

$$\frac{1}{n!}\int_a^b\int_a^b f^{(n+1)}(t)(x-t)_+^n\,dt\,dx = \frac{1}{n!}\int_a^b f^{(n+1)}(t)\int_a^b (x-t)_+^n\,dx\,dt.$$

Thus,

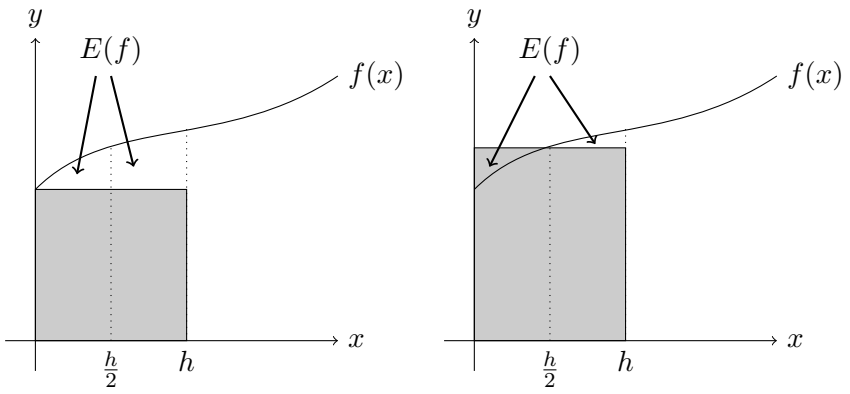
$$\begin{aligned}E(r_n) = &\frac{1}{n!}\int_a^b f^{(n+1)}(t)\left(\sum_{k=0}^{m_0} w_{k0}(x_{k0}-t)_+^n + \sum_{k=0}^{m_1} w_{k1}\,\frac{d}{dx}\left[(x_{k1}-t)_+^n\right]\right. \\ &\left. + \ldots + \sum_{k=0}^{m_n} w_{kn}\,\frac{d^n}{dx^n}\left[(x_{kn}-t)_+^n\right]\right)dt - \frac{1}{n!}\int_a^b f^{(n+1)}(t)\int_a^b (x-t)_+^n\,dx\,dt \\ = &\frac{1}{n!}\int_a^b f^{(n+1)}(t)\,E_x\left((x-t)_+^n\right)\,dt.\end{aligned}$$

Hence, the error of the one dimensional quadrature rule has a compact and elegant form. This is the essential of the classic Peano kernel theorem, which we have proven above.

Theorem 3.1 (Peano kernel theorem)**.** *If f is in $C^{n+1}([a,b])$ and I is a quadrature rule given in* (3.1) *that integrates all p in $\mathbb{P}_n$ exactly, then*

$$E(f) = I(f) - \int_a^b f(x)\,dx = \frac{1}{n!}\int_a^b f^{(n+1)}(t)\,K(t)\,dt,$$

where $K(t) = E_x\left((x-t)_+^n\right)$ is the Peano kernel.



(a) The rectangle rule (b) The midpoint rule

Figure 3.1: The rectangle and midpoint rule for evaluating the definite integral $\int_0^h f(x)\,dx$. The rectangle rule (a) approximates the integral with a rectangle with height at one of the endpoints, while the midpoint rule (b) uses a rectangle with height in the middle of the interval.

3.1.2 The Rectangle Rule

The simplest quadrature rule in one dimension is the rectangle rule, which is given as

$$\int_0^h f(x)dx \approx f(0) \cdot h, \tag{3.5}$$

and the rule integrates all f in $\mathbb{P}_0$ exactly. This rule is illustrated in Figure 3.1. The error is

$$E(f) = hf(0) - \int_0^h f(x)\,dx,$$

and from Theorem 3.1 we get the Peano kernel for the rectangle rule, which is given as

$$K_R(t) = E_x\left((x-t)_x^0\right) = h(0-t)_+^0 - \int_0^h (x-t)_+^0\,dx = -\int_t^h dx = t - h. \tag{3.6}$$

Thus the error for the rectangle rule can be written as

$$E_R(f) = \int_0^h K_R(t) f'(t)\,dt.$$

3.1.3 The Midpoint Rule

Another simple quadrature rule is the midpoint rule, which integrates all f in $\mathbb{P}_1$ exactly. It is given as

$$\int_0^h f(x)dx \approx f(h/2) \cdot h, \tag{3.7}$$

and is illustrated in Figure 3.1. Since (3.7) is exact for all p in $\mathbb{P}_1$, it is possible to obtain two Peano kernels. Similarly as above, the first order Peano kernel is found as

$$K_{M_1}(t) = E_x\left((x-t)_+^0\right) = h\left(\frac{h}{2}-t\right)_+^0 - \int_0^h (x-t)_+^0\,dx = h - \int_t^h dx = t,$$

while the second order Peano kernel becomes

$$\begin{aligned} K_{M_2}(t) &= E_x\left((x-t)_+^1\right) = h\left(\frac{h}{2}-t\right)_+ - \int_0^h (x-t)_+\,dx \\ &= h\left(\frac{h}{2}-t\right) - \int_t^h (x-t)\,dx = h\left(\frac{h}{2}-t\right) - \frac{(h-t)^2}{2}. \end{aligned} \tag{3.8}$$

Hence, there exist two error formulas for the midpoint rule,

$$\begin{aligned} E_{M_1}(f) &= \int_0^h K_{M_1}(t) f'(t)\,dt, \\ E_{M_2}(f) &= \int_0^h K_{M_2}(t) f''(t)\,dt. \end{aligned}$$

3.2 Two Dimensional Quadratures

We will now derive a two dimensional midpoint rule for the double integral

$$\int_0^h \int_0^x f(x,y)\,dy\,dx,$$

where the integration domain is shown in Figure 3.2. The starting point is the two dimensional Taylor series expansion, which we briefly discuss. In one dimension, the Taylor series expansion for $F(x)$ in $C^{n+1}([0,1])$ is given as

$$F(1) = F(0) + F'(0) + \frac{F''(0)}{2!} + \cdots + \frac{F^{(n)}(0)}{n!} + \frac{F^{(n+1)}(\xi)}{(n+1)!}, \tag{3.9}$$

for ξ in $[0,1]$. Define the parametrization of F as

$$F(t) = f(a+th, b+tk), \tag{3.10}$$

for some $f(x,y)$ and t in $[0,1]$. We assume that $f(x,y)$ has continuous partial derivatives up to order $n+1$ at all points in an open set containing the line segment joining the points (a,b) and $(a+h,b+k)$ in its domain. For simplicity, we will only derive the Taylor's formula for $n=1$ (see [1, Ch. 12.9.] for a discussion for general n). The derivatives of $F(t)$ is given as

$$\begin{aligned} F'(t) &= hf_x(x+th, y+tk) + kf_y(x+th, y+tk), \\ F''(t) &= h^2 f_{xx}(x+th, y+tk) + 2khf_{xy}(x+th, y+tk) + k^2 f_{yy}(x+th, y+tk). \end{aligned}$$

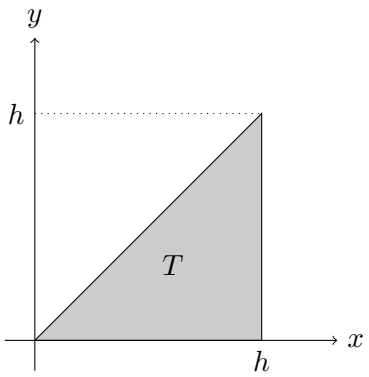


Figure 3.2: The domain $T = \{(x, y) : 0 \le y \le x \le h\}$ for the integral $\int_0^h \int_0^x f(x, y)\, dy\, dx$.

Thus, by using (3.9) and (3.10), we obtain

$$\begin{aligned} F(1) = f(a+h, b+k) = f(a,b) + hf_x(a+h, b+k) + kf_y(a+h, b+k) \\ + \frac{1}{2}\Big(h^2 f_{xx}(a+\xi h, b+\xi k) + 2khf_{xy}(a+\xi h, b+\xi k) \\ + k^2 f_{yy}(a+\xi h, b+\xi k)\Big) + \ldots, \end{aligned}$$

where the dots involves higher order derivatives on f which not are of interests. Letting $h = x - a$ and $k = y - b$, we obtain the second order Taylor formula for $f(x, y)$,

$$\begin{aligned} f(x,y) = f(a,b) + (x-a)f_x(a,b) + (y-b)f_y(a,b) \\ + \frac{1}{2}\Big(h^2 f_{xx}(a+\xi(x-a), b+\xi(y-b)) \\ + 2khf_{xy}(a+\xi(x-a), b+\xi(y-b)) \\ + k^2 f_{yy}(a+\xi(x-a), b+\xi(y-b))\Big). \end{aligned} \tag{3.11}$$

Returning to the double integral, we obtain using (3.11),

$$\int_0^h \int_0^x f(x,y)\, dy\, dx = \int_0^h \int_0^x f(a,b)\, dy\, dx + R(f), \tag{3.12}$$

where

$$\begin{aligned} R(f) = \int_0^h \int_0^x \Big((x-a)f_x(a,b) + (y-b)f_y(a,b)\Big)\, dy\, dx \\ + \frac{1}{2}\int_0^h \int_0^x \Big(h^2 f_{xx}(a+\xi(x-a), b+\xi(y-b)) \\ + 2khf_{xy}(a+\xi(x-a), b+\xi(y-b)) \\ + k^2 f_{yy}(a+\xi(x-a), b+\xi(y-b))\Big)\, dy\, dx. \end{aligned} \tag{3.13}$$

The integral on the right-hand-side in (3.12) is just the area of integration domain times the function itself. Thus, an approximation for the double integral is given as

$$\int_0^h \int_0^x f(x,y)\,dy\,dx = \frac{h^2}{2} f(a,b) + R(f), \tag{3.14}$$

and the error is given as

$$E(f) = \int_0^h \int_0^x f(x,y)\,dy\,dx - \frac{h^2}{2} f(a,b) = R(f),$$

which by (3.13) is bounded as

$$\begin{aligned}
|E(f)| \le{}& \max_T |f_x| \Big| \int_0^h \int_0^x (x-a)\,dy\,dx \Big| + \max_T |f_y| \Big| \int_0^h \int_0^x (y-b)\,dy\,dx \Big| \\
&+ \frac{h^2}{2} \max_T |f_{xx}| \int_0^h \int_0^x (x-a)^2\,dy\,dx \\
&+ 2kh \max_T |f_{xy}| \int_0^h \int_0^x (x-a)(y-b)\,dy\,dx \\
&+ \frac{k^2}{2} \max_T |f_{yy}| \int_0^h \int_0^x (y-b)^2\,dy\,dx,
\end{aligned}$$

where $T = \{(x,y) : 0 \le y \le x \le h\}$ and is illustrated in Figure 3.2. By evaluating all of the above integrals, we get

$$\begin{aligned}
|E(f)| \le{}& \max_T \Big|\frac{\partial f}{\partial x}\Big| \Big|\frac{1}{3}h^3 - \frac{a}{2}h^2\Big| + \max_T \Big|\frac{\partial f}{\partial y}\Big| \Big|\frac{1}{6}h^3 - \frac{b}{2}h^2\Big| \\
&+ \frac{h^2}{2} \max_T \Big|\frac{\partial^2 f}{\partial x^2}\Big| \Big|\frac{1}{4}h^4 - \frac{2a}{3}h^3 + \frac{a^2}{2}h^2\Big| \\
&+ kh \max_T \Big|\frac{\partial^2 f}{\partial x \partial y}\Big| \Big|\frac{1}{4}h^4 - \frac{2b}{3}h^3 - \frac{a}{3}h^3 + abh^2\Big| \\
&+ \frac{k^2}{2} \max_T \Big|\frac{\partial^2 f}{\partial y^2}\Big| \Big|\frac{1}{12}h^4 - \frac{b}{3}h^3 + \frac{b^2}{2}x^2\Big|,
\end{aligned} \tag{3.15}$$

which is the error bound for the two dimensional midpoint rule in (3.14). We let this formula end the discussion regarding numerical quadratures in two dimensions, and return to it in the analysis of the operator splitting method.

4 Sobolev Spaces

The Sobolev spaces possess many interesting properties and have wide applications in analysis and the theory of partial differential equations. These spaces has been a subject for extensive studies through the last century, which has resulted in several ways of defining these spaces. The most used way of defining the Sobolev spaces is with the use of distribution theory and the introduction of weak derivatives. A second way of defining the Sobolev spaces is by a development of monotone, absolutely continuous functions and functions of bounded variations in one variable. In [5, Ch. 5.] the author gives an introduction in the former case, while in [20] the author gives an introduction in the latter case.

We are interested in the Sobolev spaces which forms a Hilbert space. These spaces are denoted as $H^s(\mathbb{R}) = W^{s,2}(\mathbb{R})$, where s is an integer. The inner product and norm are defined as

$$(u, v)_{H^s} = \sum_{j=0}^{s} \int_{\mathbb{R}} \partial_x^j u(x)\, \partial_x^j v(x)\, dx \qquad \text{and} \qquad \|u\|_{H^s} = \sqrt{(u, u)_{H^s}}. \tag{4.1}$$

We see that $H^s(\mathbb{R})$ contains all functions which has weak derivatives up to order s in $L^2(\mathbb{R})$, and we remark that $H^0(\mathbb{R}) = L^2(\mathbb{R})$. In the context of Fourier[20] transforms, this is equivalent to require that

$$\xi^\alpha \hat{u}(\xi) \in L^2(\mathbb{R}),$$

for all non-negative integer $\alpha \leq s$. From this requirement we see that it is possible to define $H^s(\mathbb{R})$ imposing suitable condition on the Fourier transform of u, instead of using condition on the weak derivatives of u. Doing so leads to the definition of the *fractional Sobolev spaces*, which are defined for all s in $\mathbb{R}$.

In the estimation which follows the convergence analysis of the operator splitting method, we use heavily two important results regarding $H^s(\mathbb{R})$. The first result is the imbedding of $H^s(\mathbb{R})$ in $L^\infty(\mathbb{R})$ and a natural relation between the norms in the two spaces. The second result is the *Banach algebra* property of $H^s(\mathbb{R})$, which states that $H^s(\mathbb{R})$ forms an algebra over $\mathbb{R}$, and that the product inequality $\|uv\|_{H^s} \leq C_s \|u\|_{H^s} \|v\|_{H^s}$ for u and v in $H^s(\mathbb{R})$ is valid.

This section is divided as follows: First we prove the imbedding of $H^s(\mathbb{R})$ in $L^\infty(\mathbb{R})$ and the Banach algebra property using the definition based on the weak derivatives of the Sobolev spaces. We then introduce the fractional Sobolev spaces and discuss some results regarding these spaces. This discussion is followed up by proofs of the two abovementioned results.

[20] Jean Baptiste Joseph Fourier, 21 March 1768 – 16 May 1830, French mathematician and physicist. Made contributions to the theory of heat transfer, and developed the theory of harmonic analysis and Fourier series. The Fourier transform and Fourier's Law are named in his honour. Took a prominent part in his own district in promoting the French Revolution.

4.1 Standard Sobolev Spaces

Consider $H^s(\mathbb{R})$ defined when s is a positive integer, with inner product and norm as in (4.1). From the definition, we observe that $H^r(\mathbb{R})$ is continously imbedded in $H^s(\mathbb{R})$ for $r > s$, which results in that the respective norms are comparable in the following way

$$\|u\|_{H^s} \leq C\|u\|_{H^r},$$

for u in $H^r(\mathbb{R})$. We first show that $H^s(\mathbb{R})$ is imbedded in $L^\infty(\mathbb{R})$ for $s \geq 1$.

Lemma 4.1. *If u is in $H^s(\mathbb{R})$ for $s \geq 1$, then u is in $L^\infty(\mathbb{R})$. Moreover,*

$$\|u\|_{L^\infty} \leq \frac{1}{\sqrt{2}}\|u\|_{H^1} \leq C_s\|u\|_{H^s},$$

where C_s depends only on s.

Proof. The proof uses the Cauchy[21]–Schwarz[22] inequality, $\|uv\|_{L^1} \leq \|u\|_{L^2}\,\|v\|_{L^2}$, and the Young's[23] inequality, $ab \leq a^2/2 + b^2/2$.

Let u be in $H^1(\mathbb{R})$. Then we get, using the above inequalities and the triangle inequality,

$$\begin{aligned}
|u(y)^2| &= \Big| \int_{-\infty}^{y} \frac{1}{2}\partial_x(u(x)^2)\,dx - \int_{y}^{\infty} \frac{1}{2}\partial_x(u(x)^2)\,dx \Big| \\
&= \Big| \int_{-\infty}^{y} u(x)u'(x)\,dx - \int_{y}^{\infty} u(x)u'(x)\,dx \Big| \\
&\leq \int_{-\infty}^{y} |u(x)u'(x)|\,dx + \int_{y}^{\infty} |u(x)u'(x)|\,dx \\
&= \int_{-\infty}^{\infty} |u(x)u'(x)|\,dx \leq \|u\|_{L^2}\|u'\|_{L^2} \leq \frac{1}{2}\left(\|u\|_{L^2}^2 + \|u'\|_{L^2}^2\right) = \frac{1}{2}\|u\|_{H^1}^2.
\end{aligned}$$

By taking the supremum and the square root, we get

$$\|u\|_{L^\infty} \leq \frac{1}{\sqrt{2}}\|u\|_{H^1}.$$

The last inequality follows from the definition of the Sobolev norm, by adding the L^2-norm of the (weak) derivatives up to order s. □

[21] Augustin Louis Cauchy, 21 August 1789 – 23 May 1857, French mathematician. One of the pioneers of analysis, especially the theory of convergence and limits. Started the project of formulating and proving the theorems of infinitesimal calculus in a rigorous manner, and defined continuity in terms of infinitesimals. In addition, he is the author of several important theorems in complex analysis.

[22] Karl Hermann Amandus Schwarz, 25 January 1843 – 30 November 1921, German mathematician. Originally studied chemistry, but on advice from Ernst Kummer, he changed to mathematics. Became a member of the Berlin Academy of Science and a professor at the University of Berlin in 1892. Today known for his work in complex analysis.

[23] William Henry Young, 20 October 1863 – 7 July 1942, English mathematician. Worked on measure theory, Fourier series, and differential calculus, and made contributions to the study of functions of several complex variables.

A Banach algebra is a Banach space X which is an algebra, and which satisfy

$$\|xy\| \leq C\|x\|\|y\| \quad \text{for all } x, y \in X,$$

where C is a constant not dependent on the involved norms. To check that $H^s(\mathbb{R})$ satisfy the algebra properties is straightforward, and we omit these details. The proof of the inequality relies on the imbedding of $H^s(\mathbb{R})$ in $L^\infty(\mathbb{R})$.

Lemma 4.2. *The space $H^s(\mathbb{R})$ is a Banach algebra for $s \geq 1$. In particular, if u, v are in $H^s(\mathbb{R})$ for $s \geq 1$, then*

$$\|uv\|_{H^s} \leq C_s \|u\|_{H^s}\|v\|_{H^s},$$

where C_s depends only on s.

Proof. Since the Sobolev norm is a sum of (weak) derivatives of u and v, it is sufficient to show that for all $r \leq s$

$$\|\partial_x^r(uv)\|_{L^2} \leq C_s \|u\|_{H^s}\|v\|_{H^s}.$$

Consider $\partial_x^r(uv)$ and expand it using Leibniz'[24] rule

$$\partial_x^r(uv) = \sum_{j=0}^{r} \binom{r}{j} \partial_x^j u \, \partial_x^{r-j} v.$$

By the triangle inequality it is sufficient to look at one term in the above sum. Moreover, we need to be careful in the estimation of the term, since when we vary j and s we get different orders of the derivatives on u and v, which is not necessarily bounded in $H^s(\mathbb{R})$. However, we get for $r < s$ and $0 \leq j \leq r$

$$\begin{aligned}\|\partial_x^j u \, \partial_x^{r-j} v\|_{L^2}^2 &= \int_{-\infty}^{\infty} \left(\partial_x^j u\right)^2 \left(\partial_x^{r-j} v\right)^2 dx \leq \|\partial_x^j u\|_{L^\infty}^2 \int_{-\infty}^{\infty} \left(\partial_x^{r-j} v\right)^2 dx \\ &\leq C_s \|u\|_{H^{j+1}}^2 \|\partial_x^{r-j} v\|_{L^2}^2 \leq C_s \|u\|_{H^s}^2 \|v\|_{H^{r-j}}^2 \leq C_s \|u\|_{H^s}^2 \|v\|_{H^s}^2,\end{aligned}$$

since $j + 1 \leq r + 1 \leq s$ and $r - j \leq s$. For $r = s$ and $0 \leq j < r$ we get, using same technique as above,

$$\|\partial_x^j u \, \partial_x^{s-j} v\|_{L^2}^2 \leq C_s \|u\|_{H^s}^2 \|v\|_{H^s}^2.$$

We are left with one case; when $r = s = j$,

$$\|\partial_x^s u \, v\|_{L^2}^2 = \int_{-\infty}^{\infty} (\partial_x^s u)^2 \, (v)^2 \, dx \leq \|v\|_{L^\infty}^2 \|\partial_x^s u\|_{L^2}^2 \leq C_s \|v\|_{H^s}^2 \|u\|_{H^s}^2.$$

[24] Gottfried Wilhelm Leibniz, 1 July 1646 – 14 November 1716, German philosopher and mathematician. One of the founders of the infinitesimal calculus (the other was I. Newton), and introduced the symbols "dx, dy,dy/dx". He proved standard differential rules, and introduced the principles of integration. In addition, he introduced the symbol "$\int$" for the integral, the "$\cdot$" for multiplication and the terms "function" and "coordinate". Most of Leibniz' terminology is still used today. In philosophy, Leibniz is most known for his optimism, and was was one of the 17th century advocates of rationalism. See [25, pp. 273–275.] for a funny introduction to Leibniz' reasoning.

By taking the square root of the above estimates, and summing up all the derivatives, we get

$$\|uv\|_{H^s} \leq C_s \, \|u\|_{H^s} \|v\|_{H^s},$$

and the lemma is proven. □

The Banach algebra property can be extended to yield for all Sobolev spaces $W^{s,p}(\Omega)$, where Ω is a domain in $\mathbb{R}^n$, using the density of $C^\infty(\Omega)$ and the Sobolev imbedding theorems. See [2, Ch.4.] for a proof for this general case.

4.2 Fractional Sobolev Spaces

Before we introduce the fractional Sobolev spaces, we have to discuss some results regarding the Fourier transform and the convolution on $L^1(\mathbb{R})$ and $L^2(\mathbb{R})$. The Fourier transform is defined on $L^1(\mathbb{R})$ as

$$\hat{u}(\xi) = (\mathscr{F}u)(\xi) = \int_{\mathbb{R}} e^{-2\pi i \xi x} u(x) \, dx,$$

and the inverse Fourier transform is defined as

$$\check{u}(\xi) = (\mathscr{F}^{-1}u)(\xi) = \int_{\mathbb{R}} e^{2\pi i \xi x} u(x) \, dx.$$

Furthermore, the convolution of u and v is defined as

$$(u * v)(x) = \int_{\mathbb{R}} u(x-t) \, v(t) \, dt = \int_{\mathbb{R}} u(\tau) \, v(x-\tau) \, d\tau.$$

To extend the Fourier transform to yield for a function v in $L^2(\mathbb{R})$ we consider the space of Schwartz[25] functions $\mathscr{S}(\mathbb{R})$. The Schwartz space is defined as

$$\mathscr{S}(\mathbb{R}) = \{ f \in C^\infty(\mathbb{R}) \mid \lim_{|x| \to \infty} x^\alpha \, \partial_x^\beta f(x) \to 0 \quad \forall \, \alpha, \beta \in \mathbb{N} \}.$$

From the definition we see that for a function in $\mathscr{S}(\mathbb{R})$, the function and all its derivatives decrease faster than an arbitrary polynomial when $|x|$ tends to infinity. In addition, $\mathscr{S}(\mathbb{R})$ is a vector space over the complex numbers and is invariant under the Fourier transform, and the convolution is a continuous operator from $\mathscr{S}(\mathbb{R}) \times \mathscr{S}(\mathbb{R})$ to $\mathscr{S}(\mathbb{R})$. A standard example of a function which is in $\mathscr{S}(\mathbb{R})$ is $f(x) = e^{-|x|^2}$.

The Schwartz space is dense in $L^2(\mathbb{R})$ (in fact in $L^p(\mathbb{R})$ for $1 \leq p < \infty$, see [26, Thm. 5.2.5.]), and the extension of the Fourier transform to $L^2(\mathbb{R})$ is done by a density argument for the transform on $\mathscr{S}(\mathbb{R})$. A useful result is that the Fourier transform interchange the convolution and multiplication on $L^2(\mathbb{R})$. To be more formal, we have

$$\widehat{(u * v)}(\xi) = \hat{u}(\xi) \cdot \hat{v}(\xi) \quad \text{and} \quad \widehat{(u \cdot v)}(\xi) = (\hat{u} * \hat{v})(\xi).$$

[25] Laurent Schwartz, 5 March 1915 – 4 July 2002, French mathematician. Founded the theory of distributions, which he achieved the Fields medal for in 1950, as the first French mathematician. The theory of distributions clarifies the mysteries of the Dirac's delta function and the Heaviside function, and is now of capital importance to the theory of partial differential equations.

For a more detailed treatment of the Fourier transform and convolution see [7].

The fractional Sobolev space $\mathscr{H}^s(\mathbb{R})$ is defined as

$$\mathscr{H}^s(\mathbb{R}) = \Big\{ f \in \mathscr{S}'(\mathbb{R}) \,|\, (1+|\xi|^2)^{s/2} \hat{f}(\xi) \in L^2(\mathbb{R}) \Big\},$$

for all s in $\mathbb{R}$. We equip $\mathscr{H}^s(\mathbb{R})$ with the inner product

$$(u,v)_{\mathscr{H}^s} = \int_{\mathbb{R}} (1+|\xi|^2)^s \, \hat{u}(\xi) \, \overline{\hat{v}(\xi)} \, d\xi, \tag{4.2}$$

which in turn induce the norm

$$\|u\|_{\mathscr{H}^s} = \left(\int_{\mathbb{R}} (1+|\xi|^2)^s \, |\hat{u}(\xi)|^2 \, d\xi \right)^{\frac{1}{2}}. \tag{4.3}$$

The Schwartz space have a close connection with the fractional Sobolev space, which is given in the next theorem.

Theorem 4.3. *The Schwartz space $\mathscr{S}(\mathbb{R})$ is dense in $\mathscr{H}^s(\mathbb{R})$.*

Proof. To prove the theorem, we have to prove that $\mathscr{S}(\mathbb{R})$ is a subset in $\mathscr{H}^s(\mathbb{R})$, and that an arbitrary sequence in $\mathscr{S}(\mathbb{R})$ converges to a limit in $\mathscr{H}^s(\mathbb{R})$.

From the definition of $\mathscr{S}(\mathbb{R})$, it follows that $\mathscr{S}(\mathbb{R}) \subset \mathscr{H}^s(\mathbb{R})$ for all s in $\mathbb{R}$. Let u be in $\mathscr{H}^s(\mathbb{R})$, which gives that $(1+|\xi|^2)^{s/2}\hat{u}(\xi)$ in $L^2(\mathbb{R})$ by definition. Since $\mathscr{S}(\mathbb{R})$ is dense in $L^2(\mathbb{R})$, there exists a sequence $\{\varphi_j\}_{j=0}^{\infty}$ in $\mathscr{S}(\mathbb{R})$ such that

$$\varphi_j(\xi) \to (1+|\xi|^2)^{s/2}\hat{u}(\xi), \quad \text{as } j \to \infty.$$

Since $\mathscr{S}(\mathbb{R})$ is invariant under multiplication by polynomials, is $(1+|\xi|^2)^{-s/2}\varphi_j(\xi)$ in $\mathscr{S}(\mathbb{R})$ for all j. Thus, the functions

$$\psi(x)_j = \Big(\mathscr{F}^{-1}\big((1+|\xi|^2)^{-s/2}\varphi_j(\xi)\big) \Big)(x) \in \mathscr{S}(\mathbb{R}) \text{ for all } j,$$

since $\mathscr{S}(\mathbb{R})$ is invariant under the Fourier transform. Furthermore, we get, using that $\mathscr{F}(\psi_j(x))(\xi) = (1+|\xi|^2)^{-s/2}\varphi_j(\xi)$,

$$\begin{aligned} \|u-\psi_j\|_{\mathscr{H}^s}^2 &= \int_{\mathbb{R}} (1+|\xi|^2)^s |\hat{u}(\xi) - (1+|\xi|^2)^{-s/2}\varphi_j(\xi)|^2 \, d\xi \\ &= \int_{\mathbb{R}} \left| (1+|\xi|^2)^{s/2}\hat{u}(\xi) - \varphi_j(\xi) \right|^2 d\xi \to 0 \quad \text{as } j \to \infty. \end{aligned}$$

Hence, since the sequence $\{\varphi_j\}_{j=0}^{\infty}$ was chosen arbitrarily in $\mathscr{S}(\mathbb{R})$, and the limit u is in $\mathscr{H}^s(\mathbb{R})$, the Schwartz space $\mathscr{S}(\mathbb{R})$ is dense in $\mathscr{H}^s(\mathbb{R})$. □

From the theorem an important property of $\mathscr{H}^s(\mathbb{R})$ follows. Since $\mathscr{S}(\mathbb{R})$ is dense, every function in $\mathscr{H}^s(\mathbb{R})$ can be approximated by a sequence in $\mathscr{S}(\mathbb{R})$. From the definition of $\mathscr{S}(\mathbb{R})$, every Schwartz function (and all its derivatives) decays to zero at infinity. Hence, every function in $\mathscr{H}^s(\mathbb{R})$ decays to zero at infinity. This observation is very important.

We want $\mathscr{H}^s(\mathbb{R})$ to be an extension of $H^s(\mathbb{R})$ to a larger class of spaces. We see immediately that $\mathscr{H}^s(\mathbb{R})$ includes a larger class of spaces since s is allowed to be on the whole real line. If $\mathscr{H}^s(\mathbb{R})$ is an valid extension, $\mathscr{H}^s(\mathbb{R})$ have to coincide with $H^s(\mathbb{R})$ when s is a non-negative integer. To prove this, we show that $\mathscr{H}^s(\mathbb{R})$ is complete, and thus forms a Hilbert space. Then we show that $\mathscr{H}^s(\mathbb{R})$ and $H^s(\mathbb{R})$ coincides when s is a non-negative integer.

Theorem 4.4. *$\mathscr{H}^s(\mathbb{R})$ is a Hilbert space with respect to the inner product in* (4.2).

Proof. We must prove that every Cauchy sequence in $\mathscr{H}^s(\mathbb{R})$ converge to a limit in $\mathscr{H}^s(\mathbb{R})$. Let $\{u_j\}_{j=1}^{\infty}$ be a Cauchy sequence in $\mathscr{H}^s(\mathbb{R})$. By definition, we then have

$$(1+|\xi|^2)^{s/2}\,\hat{u}_j(\xi) \in L^2(\mathbb{R}) \quad \text{for all } j,$$

and we have to find a limit in $\mathscr{H}^s(\mathbb{R})$ for the sequence. Since $L^2(\mathbb{R})$ is complete, the Cauchy sequence converges to a function v in $L^2(\mathbb{R})$. We define

$$w(\xi) = (1+|\xi|^2)^{-s/2}\,v(\xi),$$

and put $f(x) = \mathscr{F}^{-1}(w(\xi))(x)$, which implies $\hat{f}(\xi) = (1+|\xi|^2)^{-s/2}\,v(\xi)$. Then, we obtain

$$\|f\|_{\mathscr{H}^s}^2 = \int_{\mathbb{R}} (1+|\xi|^2)^s\,|\hat{f}(\xi)|^2\,d\xi = \int_{\mathbb{R}} |v(\xi)|^2\,d\xi = \|v\|_{L^2}^2 < \infty,$$

which proves that f is in $\mathscr{H}^s(\mathbb{R})$. Thus,

$$\begin{aligned}\lim_{j\to\infty} &\left(\int_{\mathbb{R}} \left|(1+|\xi|^2)^{s/2}(\hat{u}_j(\xi)-\hat{f}(\xi))\right|^2 d\xi\right)^{\frac{1}{2}} \\ &= \lim_{j\to\infty}\left(\int_{\mathbb{R}} \left|(1+|\xi|^2)^{s/2}\hat{u}_j(\xi)-v(\xi)\right|^2 d\xi\right)^{\frac{1}{2}} = 0.\end{aligned}$$

Hence, $\mathscr{H}^s(\mathbb{R})$ is a complete inner product space. □

The next lemma proves that the two spaces coincides when s is non-negative integer, thus $\mathscr{H}^s(\mathbb{R})$ is an extension of $H^s(\mathbb{R})$.

Lemma 4.5. *If s is a non-negative integer, then the two spaces $\mathscr{H}^s(\mathbb{R})$ and $H^s(\mathbb{R})$ coincides.*

Proof. We have to check that if u is in $\mathscr{H}^s(\mathbb{R})$ then u also is in $H^s(\mathbb{R})$, and vice versa.

Let u be in $\mathscr{H}^s(\mathbb{R})$, which implies u is in $\mathscr{S}'(\mathbb{R})$ by definition. Thus its Fourier transform exists and $\mathscr{F}(\partial_x^\alpha u(x)) = (i\xi)^\alpha \hat{u}(\xi)$ for all non-negative integers α and $\alpha \le s$. Using the Plancherel[26]–Parseval[27] equality, we get

$$\int_{\mathbb{R}} |\partial_x^\alpha u(x)|^2\,dx = \int_{\mathbb{R}} |\xi^\alpha \hat{u}(\xi)|^2\,d\xi \le \int_{\mathbb{R}} (1+|\xi|^2)^s\,|\hat{u}(\xi)|^2\,d\xi < \infty,$$

[26]Michel Plancherel, 16 January 1885 – 4 March 1967, Swiss mathematician. Known for the Plancherel–Parseval theorem in harmonic analysis.

[27]Marc-Antoine Parseval, 27 April 1755 – 16 August 1836, French mathematician. Known for the Plancherel–Parseval theorem in harmonic analysis.

where we have used the inequality $|x|^\alpha \leq (1+|x|^2)^{\alpha/2} \leq (1+|x|^2)^{s/2}$ for $\alpha \leq s$. This proves that all the weak derivatives of u up to order s are in $L^2(\mathbb{R})$. Hence, $\mathscr{H}^s(\mathbb{R}) \subseteq H^s(\mathbb{R})$.

Now, assume u is in $H^s(\mathbb{R})$, which implies that $\partial_x^\alpha u$ is in $L^2(\mathbb{R})$ for all $\alpha \leq s$. Then we get,

$$\begin{aligned}\int_{\mathbb{R}} (1+|\xi|^2)^s \, |\hat{u}(\xi)|^2 \, d\xi &= \int_{\mathbb{R}} \sum_{k=0}^{s} \binom{s}{k} |\xi|^{2k} \, |\hat{u}(\xi)|^2 \, d\xi \leq C_s \int_{\mathbb{R}} |\xi^\alpha \hat{u}(\xi)|^2 \, d\xi \\ &\leq C_s \int_{\mathbb{R}} |\partial_x^\alpha u(x)|^2 \, dx < \infty,\end{aligned}$$

where we have used the binomial sum,

$$\sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k = (a+b)^n.$$

This proves that $H^s(\mathbb{R}) \subseteq \mathscr{H}^s(\mathbb{R})$, and we have $H^s(\mathbb{R}) = \mathscr{H}^s(\mathbb{R})$ when s is a non-negative integer. □

To simplify the notation, we denote both $\mathscr{H}^s(\mathbb{R})$ and $H^s(\mathbb{R})$ with $H^s(\mathbb{R})$ for all s. It is possible to obtain the Sobolev imbedding theorems for $H^s(\mathbb{R})$ using the fractional definition of the spaces, but we leave these details out.

We now prove equal results for the fractional Sobolev spaces as the results in Lemmas 4.1 and 4.2. The lemma below states that if $s > 1/2$ then $H^s(\mathbb{R})$ is imbedded in $L^\infty(\mathbb{R})$.

Lemma 4.6. *If u is in $H^s(\mathbb{R})$ for $s > 1/2$, then u is in $L^\infty(\mathbb{R})$. Moreover,*

$$\|u\|_{L^\infty} \leq C_s \, \|u\|_{H^s},$$

where C_s only depends on s.

Proof. Let u be in $H^s(\mathbb{R})$. By the Fourier inversion formula, see [7], we get

$$|u(x)| \leq \int_{\mathbb{R}} |e^{2\pi i y x} \hat{u}(y)| \, dy = \|\hat{u}\|_{L^1}.$$

If we can prove that $\hat{u}$ is in $L^1(\mathbb{R})$ we are done. Using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\|\hat{u}\|_{L^1} &= \int_{\mathbb{R}} |\hat{u}(y)| \, dy = \int_{\mathbb{R}} (1+|y|^2)^{s/2} \, (1+|y|^2)^{-s/2} \, |\hat{u}(y)| \, dy \\ &\leq \left(\int_{\mathbb{R}} (1+|y|^2)^{-s} \, dy \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}} (1+|y|^2)^s \, |\hat{u}(y)|^2 \, dy \right)^{\frac{1}{2}} \\ &= \left(\int_{\mathbb{R}} (1+|y|^2)^{-s} \, dy \right)^{\frac{1}{2}} \|u\|_{H^s}.\end{aligned}$$

The integral on the right-hand-side above is finite for $s > 1/2$. Hence $|u(x)| \leq C_s \|u\|_{H^s}$ where

$$C_s = \left(\int_{\mathbb{R}} (1 + |y|^2)^{-s} \, dy \right)^{\frac{1}{2}}.$$

By taking the supremum, we get $\|u\|_{L^\infty} \leq C_s \|u\|_{H^s}$. □

We now prove the Banach algebra property when s is in $\mathbb{R}$, and we use the ideas introduced in [6]. For u and v in $H^s(\mathbb{R})$, we get using (4.3),

$$\|uv\|_{H^s}^2 = \int_{\mathbb{R}} (1 + |x|^2)^s \, |\widehat{uv}(x)|^2 \, dx = \int_{\mathbb{R}} (1 + |x|^2)^s \, |(\hat{u} * \hat{v})(x)|^2 \, dx,$$

where we have used that the Fourier transform on $L^2(\mathbb{R})$ interchanges convolution and multiplication. For notational easiness we define the weight function $w_s(x) = (1 + |x|^2)^{s/2}$, such that

$$\|uv\|_{H^s}^2 = \|(\hat{u} * \hat{v}) w_s\|_{L^2}^2 = \int_{\mathbb{R}} w_s^2(x) \, |(\hat{u} * \hat{v})(x)|^2 \, dx. \tag{4.4}$$

To obtain the desirable results, we need to investigate the integrand in the above norm carefully. We start with the following lemma, which shows that w_s is subadditive.

Lemma 4.7. *For $s \geq 0$ the weight function $w_s(x)$ satisfies*

$$w_s(x + y) \leq C_s (w_s(x) + w_s(y)),$$

where C_s is only dependent on s.

Proof. The weight function w_s is convex for $s \geq 0$, which implies that w_s satisfies

$$w_s(tx + (1 - t)y) \leq t w_s(x) + (1 - t) w_s(y),$$

for any t in $[0, 1]$ and x, y in $\mathbb{R}$. Using this property we get

$$\begin{aligned} w_s(x + y) = (1 + |x + y|^2)^{s/2} &\leq \left(4 + 4 \Big| \frac{x}{2} + \frac{y}{2} \Big|^2 \right)^{s/2} \\ &= 2^s \left(1 + \Big| \frac{x}{2} + \frac{y}{2} \Big|^2 \right)^{s/2} = 2^s \, w_s \left(\frac{x}{2} + \frac{y}{2} \right) \\ &\leq 2^{s-1} (w_s(x) + w_s(y)) = C_s \, (w_s(x) + w_s(y)). \end{aligned}$$

□

Consider the integrand in (4.4), and observe that

$$|w_s(x) \, (\hat{u} * \hat{v})(x)| = \Big| w_s(x) \int_{\mathbb{R}} \hat{u}(y) \, \hat{v}(x - y) \, dy \Big| \leq \int_{\mathbb{R}} w_s(x) \, |\hat{u}(y)| \, |\hat{v}(x - y)| \, dy.$$

With the substitution $z = x - y$ and Lemma 4.7, we get

$$\begin{aligned}
|w_s(x)\,(\hat{u} * \hat{v})(x)| &\le C_s\Big(\int_{\mathbb{R}} w_s(z)\,|\hat{u}(x-z)|\,|\hat{v}(z)|\,dz \\
&\quad + \int_{\mathbb{R}} w_s(y)\,|\hat{u}(x-z)|\,|\hat{v}(z)|\,dz\Big) \\
&= C_s\Big(\int_{\mathbb{R}} w_s(x-y)\,|\hat{u}(y)|\,|\hat{v}(x-y)|\,dy \\
&\quad + \int_{\mathbb{R}} w_s(x-z)\,|\hat{u}(x-z)|\,|\hat{v}(z)|\,dz\Big) \\
&= C_s\,\big(|\hat{u}| * |\hat{v}|w_s + |\hat{u}|w * |\hat{v}|\big)(x).
\end{aligned} \tag{4.5}$$

Thus, by using the convolution inequality $\|u * v\|_{L^2} \le \|u\|_{L^1}\|v\|_{L^2}$ and (4.5), we obtain

$$\begin{aligned}
\|uv\|_{H^s}^2 &= \|(\hat{u} * \hat{v})w_s\|_{L^2} \le C_s(\|\hat{u}\|_{L^1}\|\hat{v}w_s\|_{L^2} + \|\hat{u}w_s\|_{L^2}\|\hat{v}\|_{L^1}) \\
&= C_s(\|\hat{u}\|_{L^1}\|v\|_{H^s} + \|u\|_{H^s}\|\hat{v}\|_{L^1}).
\end{aligned}$$

From the proof of Lemma 4.6 we recall that the inequality $\|\hat{u}\|_{L^1} \le C_s\,\|u\|_{H^s}$ is valid for $s > 1/2$. Hence

$$\|uv\|_{H^s} \le C_s\,\|u\|_{H^s}\|v\|_{H^s} \quad \text{for } s > 1/2,$$

and we have proven the Banach algebra property for the fractional Sobolev space. We summarize this results in the following lemma.

Lemma 4.8. *$H^s(\mathbb{R})$ is a Banach algebra for $s > 1/2$. In particular, for u, v in $H^s(\mathbb{R})$ we have*

$$\|uv\|_{H^s} \le C_s\|u\|_{H^s}\|v\|_{H^s},$$

where C_s depends only on s.

As a final remark, we mention that the results in Lemmas 4.1, 4.2, 4.6 and 4.8 also are valid for the multidimensional case, but the results becomes dependent on the space dimension n as well. The proof are somewhat similar, but one needs to be a bit more tedious and delicate in getting the desirable inequalites.

5 Operator Splitting

The operator splitting method is an approximation method, which involves splitting the different terms in a partial differential equation into subequations, solving the subequations for small time steps Δt, and concatenate the solutions at the end of each time step. Dependent on how the solutions of the subequations are concatenated give different splitting methods. In this section we consider two operator splitting methods: the Godunov and Strang splitting methods. When forming the operator splitting solution, we make in general some error, which is dependent on Δt. What is of great interest is how fast the error converges. Formally, the Godunov splitting converges as $\mathcal{O}(\Delta t)$, while the Strang splitting converges as $\mathcal{O}\left((\Delta t)^2\right)$.

This section as a whole is devoted to analytical prove the above converge rates for the two splitting methods, using a new framework recently introduced in [11]. In [11], the correct convergence rate for the Strang splitting in Sobolev spaces is proven, for a large class of partial differential equations. We follow this outline, and in addition we adopt the ideas from the framework to prove the correct convergence rates for the Godunov splitting, as well.

This section is divided as follows: We begin with a general formulation of the operator splitting method for an abstract differential equation, before we give a sketch of the new framework. A statement of the class of partial differential equations which we discuss then follows. This is followed up by two sections which discuss and prove some results for the corresponding subequations from the splitting approach. Then we prove the correct convergence rate for the Godunov splitting, before we prove the correct convergence rate for the Strang splitting. At the end, a few comments are given.

5.1 General Formulation

Assume the time $T > 0$ is fixed and consider a general partial differential equation

$$u_t = C(u), \quad t \in [0, T], \quad u|_{t=0} = u_0, \tag{5.1}$$

where $C(u)$ is a differential operator (typically in the spatial variable) between some normed spaces, say X, and assume u_0 and the solution $u(t)$ are in X. We assume that the Taylor series expansion is valid for $u(t)$, which results in

$$u(t) = u(0) + t\, u_t(0) + \mathcal{O}(t^2).$$

If we replace the second term in the above series with (5.1) we get

$$u(t) = u_0 + t\, C(u_0) + \mathcal{O}(t^2).$$

Furthermore, assume $C(u)$ can be written as a sum of more elementary operators, say

$$C(u) = A(u) + B(u),$$

which yield

$$u(t) = u_0 + t\big(A(u_0) + B(u_0)\big) + \mathcal{O}(t^2).$$

The operator splitting method is built up as follows: Fix a positive and small time step Δt, and discretize the time with n steps such that $t_n \leq n\Delta t$. Instead of solving equation (5.1) directly, we solve the two subequations

$$v_t = A(v) \qquad \text{and} \qquad w_t = B(w),$$

for each time step, and concatenate the solutions. The simplest form for an operator splitting solution of (5.1) is formed solving the first subequation using the solution from the second subequation as initial data when solving at each time step. Writing out this procedure gives,

$$u_{n+1} = \Pi^{\Delta t}(u_n) = \Phi_A^{\Delta t}\left(\Phi_B^{\Delta t}(u_n)\right) = \Phi_A^{\Delta t} \circ \Phi_B^{\Delta t}(u_n) = \left[\Phi_A^{\Delta t} \circ \Phi_B^{\Delta t}\right]^n (u_0), \tag{5.2}$$

where u_n is the operator splitting solution at time t_n, and $\Phi_A^t(v_0)$ and $\Phi_B^t(w_0)$ are the exact solution operators of the above subequations at time t with initial data v_0 and w_0, respectively. This is the well-known *Godunov splitting method*.

Other and more sophisticated methods for forming an operator splitting solution of (5.1) are created by solving the two subequations for different split step sizes, and compose the solution operators in a more complicated way. The composition of the solution operators can potentially be done in several clever ways, each naturally resulting in different operator splitting formulas. However, by solving one of the subequations for half the step size composed with the solution of the other subequation for a full time step, we obtain the famous *Strang splitting method*, which is given as

$$\begin{aligned} u_{n+1} &= \Psi^{\Delta t}(u_n) = \Phi_A^{\Delta t/2}\left(\Phi_B^{\Delta t}\left(\Phi_A^{\Delta t/2}(u_n)\right)\right) \\ &= \Phi_A^{\Delta t/2} \circ \Phi_B^{\Delta t} \circ \Phi_A^{\Delta t/2}(u_n) = \left[\Phi_A^{\Delta t/2} \circ \Phi_B^{\Delta t} \circ \Phi_A^{\Delta t/2}\right]^n (u_0). \end{aligned} \tag{5.3}$$

We hope that both (5.2) and (5.3) converge towards the correct solution of (5.1), when the time step Δt tends to 0, that is

$$u(t) = \lim_{\Delta t \to 0} \left[\Phi_A^{\Delta t} \circ \Phi_B^{\Delta t}\right]^n (u_0) = \lim_{\Delta t \to 0} \left[\Phi_A^{\Delta t/2} \circ \Phi_B^{\Delta t} \circ \Phi_A^{\Delta t/2}\right]^n (u_0).$$

Taking this one step further, we hope that by forming the operator splitting solution with (5.3) instead of (5.2), we gain something. This "something" is the convergence rate for the error between $u(t)$ and u_n. Formally, the Godunov splitting (5.2), converges as

$$\|u_n - u(t_n)\|_X \leq \mathcal{O}(\Delta t),$$

while the Strang splitting (5.3), converges as

$$\|u_n - u(t_n)\|_X \leq \mathcal{O}\left((\Delta t)^2\right).$$

The major task in what follows is to prove the convergence rates for the two splitting methods in Sobolev spaces.

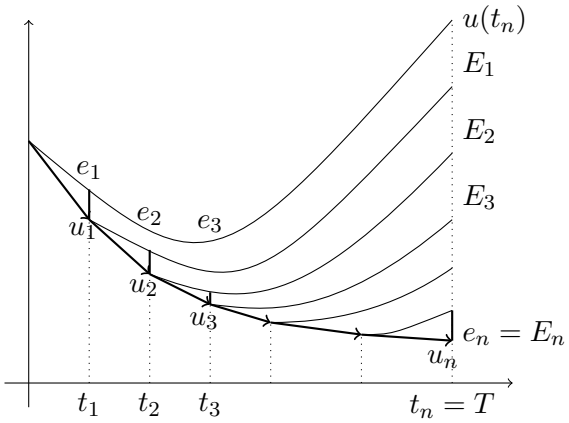


Figure 5.1: Schematic view of the global error estimation. e_i is the local error at step i and are symbolized with bold lines , u_i is the approximate solution at t_i and are symbolized with bold arrows, E_i is the error at the end point. The solid lines are the exact solution. Observe how the local errors at each step is transported to the end point $t = T$. To obtain the global error all the "transported" local errors are added up.

5.2 Sketch of the Framework

The main idea of the framework in [11] is to use a standard argument from error estimation of numerical methods. We find an estimate of the *local error*, which is the error after performing *one* step with the operator splitting method, before we add up all the local errors from *each* step. This yield the *global error*, which is what we are after. The procedure is illustrated in Figure 5.1.

The keypoint in the new approach in [11] is to use error terms for numerical quadratures, in Peano kernel form using Theorem 3.1, to estimate the local errors in $H^s(\mathbb{R})$, where $H^s(\mathbb{R})$ is the Sobolev space introduced in Section 4 and s is an arbitrary non-negative integer. The local and global estimates are grounded on a well-posedness theory of the involved partial differential equations, which is presented below. In addition, a Taylor series expansion and a variation of parameters formula are used to obtain the local estimates. These foundations yield an estimation of the local error which is delicate and elegant, and which involve the abovementioned error forms in combination with differential calculus and estimation tools in $H^s(\mathbb{R})$.

The summation of the local errors is dependent on the regularity results for the two subequations (5.5) and (5.6). The framework as a whole will become clear in the sections which follows.

5.3 Statement of the Problem

With the general formulation in Section 5.1 in mind, we formulate the problem which we shall delve into. Consider the initial value problem

$$u_t = P(\partial_x)u + uu_x, \quad u|_{t=0} = u_0, \tag{5.4}$$

where x is in $\mathbb{R}$ and t is in the intervall $[0, T]$ for a fixed time $T > 0$, and P is a polynomial of degree l. We require that u_0 and $u(t)$ are in $H^s(\mathbb{R})$, where the order s is specified in details later. Applying the operator splitting method to (5.4), and splitting it into two subequations gives

$$v_t = A(v) = P(\partial_x)v \tag{5.5}$$

and

$$w_t = B(w) = ww_x, \tag{5.6}$$

where the latter is the inviscid Burgers' equation, and B is the *Burgers' operator*.

The initial value problem (5.4) includes a wide class of equations, but our focus is on the viscous Burgers' equation,

$$u_t = u_{xx} + uu_x, , \tag{5.7}$$

and the Korteweg–de Vries (KdV) equation,

$$u_t = u_{xxx} + uu_x. \tag{5.8}$$

Two other equations, which falls into the class, are for instance the Benney–Lin equation, $u_t = -u_{xxx} - \beta(u_{xx} + u_{xxxx}) - u_{xxxxx} + uu_x$, and the Kawahara equation, $u_t = u_{xxxxx} - u_{xxx} + uu_x$. These two equations are not discussed further in this text.

When we apply the operator splitting method to (5.4), several questions arises immediately. The success of the operator splitting method, in the sense that it produces correct solutions, is dependent on the two terms on the right-hand-side in (5.4). It is a well-known fact that (5.6) potentially can produce discontinuous solutions, independent of the smoothness of the initial condition. The solutions of (5.5) is naturally dependent on the form of the polynomial. Thus, it is way too much to hope for that the operator splitting method works well on all equations in the class of (5.4)!

The full solutions of (5.7) and (5.8) are smooth, and we can in worst case be put in a situation where (5.6) produces a shock when a Burgers' step is performed in the forming of the operator splitting solution. In combination with steps with (5.5) this can potentially go very wrong. To conquer the problem with the potentially shock creation in (5.6), the idea is to choose small enough time steps Δt such that (5.6) never has time to produce a discontinuity in the solution. The problem is that there is no automechanism in that such an Δt exists. However, in Section 5.4 we formally prove that such Δt exists, in addition with other results for (5.6). On the other hand, we are not able to say anything about the solutions of (5.5) without some kind of constraints on A. It turns out that, in combination with the results for (5.6), the polynomial in (5.5) needs to be

linear with degree $l \geq 2$ such that the corresponding Sobolev norm of the solution do not increase. The formal requirement and proofs are given in Section 5.5.

As mentioned above, the upcoming analysis relies heavily on a well-posedness theory for (5.4) in $H^s(\mathbb{R})$. For simplicity in the refering, we list the well-posedness requirements for (5.4) in addition with the assumptions for u_0 and $u(t)$, as hypotheses for arbitrary order $k \geq 0$ of $H^k(\mathbb{R})$, and specify for which k they should hold in details for the Godunov and Strang splittings below.

The first hypothesis is about the local well-posedness of the solutions to (5.4).

Hypothesis 5.1 (Local well-posedness)**.** *For a fixed time T, there exists $R > 0$ such that for all u_0 in $H^k(\mathbb{R})$ with $\|u_0\|_{H^k} \leq R$, there exists a unique strong solution u in $C([0,T], H^k)$ of (5.4). In addition, for the initial data u_0 there exists a constant $K(R,T) < \infty$, such that*

$$\|\tilde{u}(t) - u(t)\|_{H^k} \leq K(R,T)\, \|\tilde{u}_0 - u_0\|_{H^k}, \tag{5.9}$$

for two arbitrary solutions u and $\tilde{u}$, corresponding to two different initial data u_0 and $\tilde{u}_0$.

The requirement in (5.9) is the same as requering that u_0 is local *Lipschitz continuous*. The last hypothesis requires that the solution and the initial data are bounded in the Sobolev spaces.

Hypothesis 5.2 (Boundedness)**.** *The solution $u(t)$ and the initial data u_0 of (5.4) are both in $H^k(\mathbb{R})$, and are bounded as*

$$\|u(t)\|_{H^k} \leq R < \rho \quad \text{and} \quad \|u_0\|_{H^k} \leq C < \infty,$$

for $0 \leq t \leq T$.

We define the following set of integers, which we keep fixed throughout this entire section,

$$s \geq 1, \quad p = s + 2l - 1, \quad q = s + l - 1 = p - l. \tag{5.10}$$

where $l \geq 2$ is the degree of the polynomial in (5.5). We specify for which integers the hypotheses should hold in the lemmas and theorems for the two splitting methods.

5.4 Results for the Inviscid Burgers' Equation

As mentioned in the previous subsection, one crucial point with the forming of the operator splitting solution of (5.4), is that the inviscid Burgers' equation (5.6) can produce a shock in the operator splitting solution while the full solution of (5.4) remains smooth. The workaround for this problem is to choose a uniformly and small enough split step Δt, such that (5.6) never has time to create a shock in the operator splitting solution. Thus, we have to estimate the solutions of (5.6) carefully.

Showing that there exist a small time step Δt which prevents (5.6) from producing a shock for the solution $\Phi_B^t(w_0)$ in a Sobolev space, is a rather delicate calculation, and

this type of estimation was first introduced in [13]. A discontinuity in $\Phi_B^t(w_0)$ results in the L^2-norm of the corresponding derivatives blowing up, which results in that the $H^s(\mathbb{R})$-norm also blowing up. Thus, to show that $\Phi_B^t(w_0)$ is smooth, it is enough to show that the corresponging norm is finite. In the following lemma, we prove that there exist a small Δt such that $\Phi_B^t(w_0)$ remains smooth in Sobolev spaces.

Lemma 5.3. *For p and q in (5.10) assume the solution $\Phi_B^t(w_0) = w(t)$ of (5.6) with initial data w_0 in $H^p(\mathbb{R})$, satisfies $\|\Phi_B^t(w_0)\|_{H^q} \leq \alpha$ for $0 \leq t \leq \Delta t$. Then $\Phi_B^t(w_0)$ is in $H^p(\mathbb{R})$ and in particular*

$$\|\Phi_B^t(w_0)\|_{H^p} \leq e^{c\alpha t}\|w_0\|_{H^p},$$

where c is independent of w_0 and Δt.

Proof. In this proof we use C as a general constant, which can take many different values. From the definition of norm in $H^p(\mathbb{R})$, we find that $w(t)$ satisfies

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\Phi_B^t(w_0)\|_{H^p}^2 = \frac{1}{2}\frac{d}{dt}\|w\|_{H^p}^2 = \frac{1}{2}\frac{d}{dt}\sum_{j=0}^{p}\int_{\mathbb{R}}(\partial_x^j w)^2\,dx = \sum_{j=0}^{p}\int_{\mathbb{R}}\partial_x^j w\,\partial_x^j w_t\,dx \\ = (w, w_t)_{H^p} = (w, ww_x)_{H^p} = \sum_{j=0}^{p}\int_{\mathbb{R}}\partial_x^j w\,\partial_x^j(ww_x)\,dx \\ = \sum_{j=0}^{p}\sum_{k=0}^{j}\binom{j}{k}\int_{\mathbb{R}}\partial_x^j w\,\partial_x^{k+1}w\,\partial_x^{j-k}w\,dx,\end{aligned}$$

where the last equality comes from the Leibniz' rule. If we prove the bound for the norm for $w(t)$ in $H^p(\mathbb{R})$, then we also have proved that $w(t)$ is in $H^p(\mathbb{R})$. To obtain this bound, we need to estimate each term in the above sum carefully. The crucial point is the order on the derivatives of w, which varies for the different terms in the sum. To estimate each term, as a standard technique we use the imbedding of $H^s(\mathbb{R})$ in $L^\infty(\mathbb{R})$ and move one term out of the integral, and bound it in $H^p(\mathbb{R})$. The remaining of the integral is estimated using the Cauchy–Schwarz inequality. The problem with this technique is that we have to vary which term we move out of the integral, to ensure that the estimate still is valid in $H^p(\mathbb{R})$. For the cases $j < p$, this is not a problem and all terms can be estimated using same argument. On the other side, when $j = p$ we have to change which term we move out of the integral. In addition, order of at most $p + 1$ derivatives on w appears in one term. This is a more critical problem, which we have to treat specially. To this end, we divide the sum into different parts, and estimate each part such that everything becomes clear.

For $j < p$, we obtain for each term in the sum

$$\begin{aligned}\Big|\int_{\mathbb{R}}\partial_x^j w\,\partial_x^{k+1}w\,\partial_x^{j-k}w\,dx\Big| &\leq \int_{\mathbb{R}}|\partial_x^j w\,\partial_x^{k+1}w\,\partial_x^{j-k}w|\,dx \leq \|\partial_x^j w\|_{L^\infty}\int_{\mathbb{R}}|\partial_x^{k+1}w\,\partial_x^{j-k}w|\,dx \\ &\leq \left\|\partial_x^j w\right\|_{L^\infty}\left\|\partial_x^{\max\{k+1,j-k\}}w\right\|_{L^2}\left\|\partial_x^{\min\{k+1,j-k\}}w\right\|_{L^2} \\ &\leq C\,\|w\|_{H^p}\,\|w\|_{H^p}\,\|w\|_{H^q} \leq C\,\|w\|_{H^p}^2\,\|w\|_{H^q} \\ &\leq C\alpha\,\|w\|_{H^p}^2,\end{aligned}$$

where we have used Lemma 4.1, and the fact that

$$\min\{k+1, j-k\} \leq \frac{j+1}{2} \leq \frac{p}{2} = \frac{s-1}{2} + l \leq s - 1 + l = p - l = q$$
$$\max\{k+1, j-k\} \leq j + 1 \leq p,$$

since $p \geq 2l$.

For $j = p$ we distinguish the different terms in the sum. For $k \leq q$ and $k \neq q-1$, we estimate

$$\begin{aligned}\Big| \int_{\mathbb{R}} \partial_x^p w \, \partial_x^{k+1} w \, \partial_x^{p-k} w \, dx \Big| &\leq \|\partial_x^{k+1} w\|_{L^\infty} \int_{\mathbb{R}} |\partial_x^p w \, \partial_x^{j-k} w| \, dx \\ &\leq \|\partial_x^{k+1} w\|_{L^\infty} \|\partial_x^p w\|_{L^2} \|\partial_x^{p-k} w\|_{L^2} \\ &\leq C \|\partial_x w\|_{H^{k+1}} \|w\|_{H^p} \|w\|_{H^{p-k}} \\ &\leq C \|w\|_{H^{k+2}} \|w\|_{H^p} \|w\|_{H^{p-k}}.\end{aligned}$$

To get a bound, the above inequality is divided in two cases; when $k+2 \leq q$ and when $k = q$. For the first case we obtain

$$\Big| \int_{\mathbb{R}} \partial_x^p w \, \partial_x^{k+1} w \, \partial_x^{p-k} w \, dx \Big| \leq C \|w\|_{H^q} \|w\|_{H^p} \|w\|_{H^p} \leq C\alpha \|w\|_{H^p}^2,$$

from $\|w\|_{H^{k+2}} \leq \|w\|_{H^q}$ since $k+2 \leq q$ and $\|w\|_{H^{p-k}} \leq \|w\|_{H^p}$. For the second case, we get

$$\Big| \int_{\mathbb{R}} \partial_x^p w \, \partial_x^{k+1} w \, \partial_x^{p-k} w \, dx \Big| \leq C \|w\|_{H^p} \|w\|_{H^p} \|w\|_{H^q} \leq C\alpha \|w\|_{H^p}^2,$$

where we have used that $p - q = l \leq l + r - 1 = q$, and $q + 2 \leq q + l = p$.

We are left with three cases; $k+1 = q$, $q+2 \leq k+1 \leq p$ and $k = p = j$. For the first case we get

$$\begin{aligned}\Big| \int_{\mathbb{R}} \partial_x^p w \, \partial_x^{k+1} w \, \partial_x^{p-k} w \, dx \Big| &\leq \|\partial_x^p w\|_{L^2} \|\partial_x^{k+1} w\|_{L^2} \|\partial_x^{p-k} w\|_{L^\infty} \\ &\leq C \|w\|_{H^p} \|w\|_{H^{k+1}} \|w\|_{H^{p-k+1}} \leq C\alpha \|w\|_{H^p}^2,\end{aligned}$$

because $k + 1 = q \leq p$ and $p - k + 1 = l + 2 \leq 2l \leq p$.

For the second case we get same result as above, but now we use that for $k+1 \geq q+2$ we have $p - k + 1 \leq p - q \leq l \leq q$, in the estimation.

For the third case, when $k = p = j$, we need to be a bit more careful. To get rid of the derivative of order $p+1$ on w, we first perform a partial integration, followed up by similar arguments as above. Thus,

$$\begin{aligned}\Big| \int_{\mathbb{R}} \partial_x^p w \, \partial_x^{p+1} w \, w \, dx \Big| &= \frac{1}{2} \Big| \int_{\mathbb{R}} \partial_x \left(\partial_x^p w\right)^2 \, w \, dx \Big| \\ &= \frac{1}{2} \Big| \Big[\left(\partial_x^p w\right)^2 \, w \Big]_{\infty}^{\infty} - \int_{\mathbb{R}} \left(\partial_x^p w\right)^2 \, \partial_x w \, dx \Big| \\ &= \frac{1}{2} \int_{\mathbb{R}} \Big| \left(\partial_x^p w\right)^2 \, \partial_x w \Big| \, dx \leq \|\partial_x w\|_{L^\infty} \|\partial_x^p w\|_{L^2}^2 \\ &\leq C \|w\|_{H^2} \|w\|_{H^p}^2 \leq C \|w\|_{H^q} \|w\|_{H^p}^2 \leq C\alpha \|w\|_{H^p}^2,\end{aligned}$$

since $2 \leq s + l - 1 = q$ when $l \geq 2$.

All in all we get, by summing up the above estimates, the following inequality

$$\frac{d}{dt}\|w(t)\|_{H^p}^2 = \|w(t)\|_{H^p}\frac{d}{dt}\|w(t)\|_{H^p} \leq c\alpha\|w(t)\|_{H^p}^2,$$

which leads to

$$\frac{d}{dt}\|w(t)\|_{H^p} \leq c\alpha\|w(t)\|_{H^p}.$$

By integrating this inequality, and using that $w(0) = w_0$, the result follows. □

The next lemma shows that if the initial condition for (5.6) are bounded in $H^k(\mathbb{R})$, then the solution itself is also bounded for an t which is dependent on the bound for the intitial condition.

Lemma 5.4. *Assume $\|w_0\|_{H^k} \leq K$ for some $k \geq 1$. Then there exists $\bar{t}(K) > 0$ such that $\|\Phi_B^t(w_0)\|_{H^k} \leq 2K$ for $0 \leq t \leq \bar{t}(K)$.*

Proof. By doing the same calculations as in the proof of Lemma 5.3 with k instead of m and using the bound for u_0 in $H^k(\mathbb{R})$, we arrive with the following inequality

$$\|w(t)\|_{H^k}\frac{d}{dt}\|w(t)\|_{H^k} \leq c\,\|w(t)\|_{H^k}^3,$$

which simplifies to

$$\frac{d}{dt}\|w(t)\|_{H^k} \leq c\,\|w(t)\|_{H^k}^2.$$

By comparing with the solution of the differential equation $y' = cy^2$, we see that if we want $\|\Phi_B^t(w_0)\|_{H^k} \leq 2K$, we must integrate the above inequality a time $\bar{t}$ which is dependent on the bound K. □

In the proofs of the convergence rates for (5.2) and (5.3), we need to expand $\Phi_B^t(w_0)$ using Taylor series expansions of first and second order. Thus, $\Phi_B^t(w_0)$ needs to be continuous, such that the expansions are valid. The following lemma proves the sufficient continuity.

Lemma 5.5. *If $\|w_0\|_{H^{s+l}} \leq C_0$ for $s \geq 1$ and $l \geq 2$, then there exists $\bar{t}$ depending only on C_0, such that the solution $w(t)$ of (5.6) is $C^3([0,\bar{t}], H^s)$.*

Proof. Let t be in $[0,\bar{t}]$, with $\bar{t}$ from Lemma 5.4, and define

$$\widetilde{w}(t) = w_0 + tB(w_0) + \int_0^t (t-s)dB\big(w(s)\big)\big[B(w(s))\big]ds,$$

where $dB(\cdot)[\cdot]$ is the Fréchet derivative. Calculating the second derivative of $\widetilde{w}$, gives using (2.3),

$$\begin{aligned}\widetilde{w}_{tt} &= dB\big(w(s)\big)\big[B(w(s))\big] \\ &= w_x B(w) + w(B(w))_x = ww_x^2 + w(ww_x)_x = 2ww_x^2 + w^2w_{xx},\end{aligned}$$

from which we have that $\widetilde{w}$ is in $C^2\left([0,\bar{t}],H^s\right)$. To prove that $w=\widetilde{w}$, we must show that the two functions satisfies the same differential equation and the same initial conditions. By differentiation (5.6) with respect to t, we get

$$\begin{aligned} w_{tt} = B(w)_t = (ww_x)_t = w_t w_x + ww_{xt} = ww_x^2 + w(ww_x)_x \\ = ww_x^2 + ww_x^2 + w^2 w_{xx} = 2ww_x^2 + w^2 w_{xx} = \widetilde{w}_{tt}, \end{aligned}$$

which shows that w and $\widetilde{w}$ satisfies the same equation. From the definition of $\widetilde{w}$, we see that $\widetilde{w}(0)=u_0$ and $\widetilde{w}_t(0)=B(u_0)=w_t$. Thus, we have shown that $w=\widetilde{w}$.

To prove that w is $C^3([0,\bar{t}],H^s)$, we find for $\widetilde{w}$, using (5.6),

$$\begin{aligned} \widetilde{w}_{ttt} &= 2(ww_x^2)_t + (w^2 w_{xx})_t = 2(w_t w_x^2 + 2ww_x w_{xt}) + (2ww_t w_{xx} + w^2 w_{xxt}) \\ &= 2\left((ww_x)w_x^2 + 2ww_x(ww_x)_x\right) + 2w^2 w_x w_{xx} + w^2(ww_x)_{xx} \\ &= 2\left(ww_x^3 + 2ww_x w_x^2 + 2w^2 w_x w_{xx}\right) + 2w^2 w_x w_{xx} + 3w^2 w_x w_{xx} + w^2 w w_{xxx} \\ &= 2\left(3ww_x^3 + 2w^2 w_x w_{xx}\right) + 5w^2 w_x w_{xx} + w^3 w_{xxx} \\ &= 6ww_x^3 + 9w^2 w_x w_{xx} + w^3 w_{xxx}, \end{aligned}$$

from which it follows that $\widetilde{w}$ is $C^3([0,\bar{t}],H^s)$. The lemma is proven. □

5.5 Results for the Polynomial Equation

We give constraints for the polynomial equation in (5.5), which in combination with the results for (5.6) in the previous subsection, yield the sufficient results for the operator splitting analysis which follows. As mentioned, the critical point for (5.5) (in combination with those for (5.6)), is that the Sobolev norm do not increase. We state which properties A must satisfy to yield this property in the first lemma below. In addition we prove that A is a continuous operator.

Lemma 5.6. *Let P be a linear polynomial of degree $l \geq 2$ with constant coefficients, which satisfies*

$$\operatorname{Re} P(i\xi) \leq 0 \text{ for all } \xi \in \mathbb{R}. \tag{5.11}$$

In addition, let m be a integer such that $m \geq l$, and assume v_0 is in $H^{m+l}(\mathbb{R})$ and the solution $\Phi_A^t(v_0)=v(t)$ of (5.5) is in $H^m(\mathbb{R})$ and satisfies

$$\int_{\mathbb{R}} \left(\partial_x^{j+l/2} v\right)^2 dx < \infty,$$

for all $j \leq m$ and l even. Then $\Phi_A^t(v_0)$ has a non-increasing norm in $H^m(\mathbb{R})$, in particular

$$\|\Phi_P^t(v_0)\|_{H^m} \leq \|v_0\|_{H^{m+l}}.$$

Proof. Using the assumptions, P is given as $P(x) = \sum_{\alpha=2}^{l} a_\alpha x^\alpha$, where a_α is in $\mathbb{R}$ for all α. Thus, (5.5) becomes

$$v_t = a_l\, \partial_x^l v + a_{l-1}\, \partial_x^{l-1} v + \cdots + a_2\, \partial_x^2 v.$$

The time evolution of $\Phi_A^t(u_0)$ is given as

$$\frac{1}{2}\frac{d}{dt}\|\Phi_A^t(v_0)\|_{H^m}^2 = (v, v_t)_{H^m} = \sum_{j=0}^{m} \int_{\mathbb{R}} \partial_x^j v\, (a_l \partial_x^{j+l} + \cdots + a_2 \partial_x^{j+2}) v\, dx. \tag{5.12}$$

It is sufficient to estimate one general term in the above sum, say

$$\sum_{j=0}^{m} \int_{\mathbb{R}} \partial_x^j v\, a_l \partial_x^{j+l} v\, dx = a_l \sum_{j=0}^{m} \int_{\mathbb{R}} \partial_x^j v\, \partial_x^{j+l} v\, dx.$$

By partial integration the above equation turns into

$$\begin{aligned} a_l \sum_{j=0}^{m} \int_{\mathbb{R}} \partial_x^j v\, \partial_x^{j+l} v\, dx &= a_l \sum_{j=0}^{m} \Big([\partial_x^j v\, \partial_x^{j+l-1} v]_{-\infty}^{\infty} - \int_{\mathbb{R}} \partial_x^{j+1} v\, \partial_x^{j+l-1} v\, dx \Big) \\ &= -a_l \sum_{j=0}^{m} \int_{\mathbb{R}} \partial_x^{j+1} v\, \partial_x^{j+l-1} v\, dx, \end{aligned}$$

where we have used that the derivatives on v of order up to m decay to zero when $x \to \pm\infty$. Performing partial integration together with the decay property for the derivatives of v subsequently, we get if l even

$$a_l \sum_{j=0}^{m} \int_{\mathbb{R}} \partial_x^j v\, \partial_x^{j+l} v\, dx = a_l \sum_{j=0}^{m} (-1)^{l-1} \int_{\mathbb{R}} \left(\partial_x^{j+l/2} v\right)^2 dx = -a_l \sum_{j=0}^{m} \int_{\mathbb{R}} \left(\partial_x^{j+l/2} v\right)^2 dx.$$

By the property given in (5.11), the coefficient a_l is such that the right-hand-side of the above equation is negative, that is $a_l > 0$. We write this for simplicity as

$$a_l \sum_{j=0}^{m} \int_{\mathbb{R}} \partial_x^j v\, \partial_x^{j+l} v\, dx = -\sum_{j=0}^{m} \int_{\mathbb{R}} \left(\partial_x^{j+l/2} v\right)^2 dx = -\|\partial_x^{l/2} v\|_{H^m}. \tag{5.13}$$

If l is odd, we obtain by partial integration

$$\begin{aligned} a_l \sum_{j=0}^{m} \int_{\mathbb{R}} \partial_x^j v\, \partial_x^{j+l} v\, dx &= a_l \sum_{j=0}^{m} (-1)^l \int_{\mathbb{R}} \partial_x \left(\partial_x^{j+(l-1)/2} v\right)^2 dx \\ &= -a_l \sum_{j=0}^{m} \left[(\partial_x^{j+(l-1)/2} v)^2 \right]_{\infty}^{\infty} = 0. \end{aligned} \tag{5.14}$$

By using the estimates in (5.13) and (5.14), we get for (5.12),

$$\frac{1}{2}\frac{d}{dt}\|\Phi_A^t(v_0)\|_{H^m}^2 = -C\|\partial_x^{l/2} v\|_{H^m} \leq 0,$$

where C is a constant. Solving the differential equation gives

$$\|\Phi_A^t(v_0)\|_{H^m} \leq \|v_0\|_{H^{m+l}}.$$

□

The next lemma proves the continuity of P.

Lemma 5.7. *Let l and m be integers such that $l \geq 1$ and $m \geq 1$. Assume P is a linear polynomial with constant coefficients of degree l. If v is in $H^m(\mathbb{R})$, then $P(\partial_x)v$ is in $H^{m-l}(\mathbb{R})$ and the mapping $P : H^m \to H^{m-l}$ is continuous.*

Proof. We prove this using the fractional definition of the Sobolev spaces given in Section 4.2. Let $P(x) = \sum_{\alpha=0}^{l} a_\alpha x^\alpha$ where a_α is in $\mathbb{R}$ for all α. Recall, the Fourier transform has the property $\mathscr{F}(\partial_x^\alpha v) = (ix)^\alpha \hat{v}$. Hence

$$\begin{aligned}
\|P(\partial_x)v\|_{H^{m-l}} &= \big\|(1+|\cdot|^2)^{m-l}\hat{v}(\cdot)\big\|_{L^2} = \Big(\int_{\mathbb{R}} (1+|x|^2)^{m-l} |\mathscr{F}\left(P(\partial_x)v\right)(x)|^2\,dx\Big)^{1/2} \\
&= \Big(\int_{\mathbb{R}} (1+|x|^2)^{m-l} |P(ix)\hat{v}(x))|^2\,dx\Big)^{1/2} \\
&= \Big(\int_{\mathbb{R}} (1+|x|^2)^{m-l} \Big|\sum_{\alpha=0}^{l} a_\alpha (ix)^\alpha\Big|^2 |\hat{v}(x)|^2\,dx\Big)^{1/2}.
\end{aligned}$$

We need the following inequality

$$|x^\alpha| = |x|^\alpha = (|x|^2)^{\alpha/2} \leq (1+|x|^2)^{\alpha/2} \leq (1+|x|^2)^{l/2},$$

for all $\alpha \leq l$. Using this inequality and the triangle inequality, we get

$$\begin{aligned}
\|P(\partial_x)v\|_{H^{m-l}} &\leq \Big(\int_{\mathbb{R}} (1+|x|^2)^{m-l} \Big|\sum_{\alpha=0}^{l} a_\alpha (1+|x|^2)^{l/2}\Big|^2 |\hat{v}(x)|^2\,dx\Big)^{1/2} \\
&= \Big|\sum_{\alpha=0}^{l} a_\alpha\Big| \Big(\int_{\mathbb{R}} (1+|x|^2)^{m-l} (1+|x|^2)^{l} |\hat{v}(x)|^2\,dx\Big)^{1/2} \\
&\leq \sum_{\alpha=0}^{l} |a_\alpha| \Big(\int_{\mathbb{R}} (1+|x|^2)^{m} |\hat{v}(x)|^2\,dx\Big)^{1/2} \leq \sum_{\alpha=0}^{l} |a_\alpha| \|v\|_{H^m}.
\end{aligned}$$

This proves that $P(\partial_x)v$ is in $H^{m-l}(\mathbb{R})$. To prove the continuity, consider a sequence $\{v_j\}_{j=0}^{\infty} \subset H^m(\mathbb{R})$ which converges towards v in $H^m(\mathbb{R})$, that is

$$\lim_{j\to\infty} \|v_j - v\|_{H^m} = 0.$$

Then, using the above inequality, we get

$$\lim_{j\to\infty} \|P(\partial_x)(v_j - v)\|_{H^{m-l}} \leq \lim_{j\to\infty} \sum_{\alpha=0}^{l} |a_\alpha| \|v_j - v\|_{H^m} = 0,$$

which proves the continuity of the operator P. □

5.6 Godunov Splitting

In the previous subsections, we have presented and proven several results which now will prove useful. From the discussion about the new framework, it should not come as a surprise that we first estimate the local error for the Godunov splitting (5.2), before we use this estimate to find a bound for the global error.

5.6.1 Local Error

Lemma 5.8. *Let $s \geq 1$ be an integer and assume Hypothesis 5.2 holds for $k = s + l$ for the solution $u(t) = \Phi^t_{A+B}(u_0)$ of (5.4). If the initial data u_0 is in $H^{s+l}(\mathbb{R})$, then the local error of the Godunov splitting (5.2) is bounded in $H^s(\mathbb{R})$ by*

$$\left\|\Pi^{\Delta t}(u_0) - \Phi^{\Delta t}_{A+B}(u_0)\right\|_{H^s} \leq c_1 \, (\Delta t)^2,$$

where c_1 depends on $\|u_0\|_{H^{s+l}}$ and where Δt is a small time step.

Proof. In this proof C is a general constants which can take several values, and R is as given in Hypothesis 5.2. Recall the definition of A and B in (5.5) and (5.6), respectively. We start with

$$B(\varphi(s)) - B(\varphi(0)) = \int_0^s dB(\varphi(\sigma))\left[\dot\varphi(\sigma)\right] d\sigma,$$

and define $\varphi(\sigma) = \Phi_A^{(s-\sigma)}(u(\sigma))$, and get for the integrand in the above equation

$$\begin{aligned} dB(\varphi(\sigma))\left[\dot\varphi(\sigma)\right] &= -A\Phi_A^{(s-\sigma)}(u(\sigma)) + \Phi_A^{(s-\sigma)}(\dot u(\sigma)) \\ &= \Phi_A^{(s-\sigma)}(-A(u(\sigma))) + \Phi_A^{(s-\sigma)}((A+B)(u(\sigma))) = \Phi_A^{(s-\sigma)}(B(u(\sigma))), \end{aligned}$$

which yield

$$B(u(s)) = B(\Phi_A^s(u_0)) + \int_0^s dB\left(\Phi_A^{(s-\sigma)}(u(\sigma))\right)\left[\Phi_A^{(s-\sigma)}(B(u(\sigma)))\right] d\sigma, \tag{5.15}$$

where we have used that $\varphi(0) = \Phi_A^s(u(0)) = \Phi_A^s(u_0)$. Using the variation of parameters formula in Theorem 2.6, the exact solution of (5.4) is given as

$$\Phi^t_{A+B}(u_0) = \Phi^t_A(u_0) + \int_0^t \Phi_A^{(t-s)}\Big(B(u(s))\Big)\, ds. \tag{5.16}$$

To find the exact solution after one split step, we insert (5.15) into (5.16) and evaluate at $t = \Delta t$,

$$u(\Delta t) = \Phi_A^{\Delta t}(u_0) + \int_0^{\Delta t} \Phi_A^{(\Delta t - s)}\Big(B(\Phi_A^s(u(s)))\Big)\, ds + e_{G,1}, \tag{5.17}$$

where

$$e_{G,1} = \int_0^{\Delta t}\int_0^s \Phi_A^{(\Delta t - s)}\left(dB\left(\Phi_A^{(s-\sigma)}(u(\sigma))\right)\left[\Phi_A^{(s-\sigma)}(B(u(\sigma)))\right]\right) d\sigma\, ds. \tag{5.18}$$

One step with the Godunov splitting (5.2), is given as

$$u_1 = \Pi^{\Delta t}(u_0) = \Phi_A^{\Delta t}\left(\Phi_B^{\Delta t}(u_0)\right). \tag{5.19}$$

From Lemma 5.5, the exact solution $\Phi_B^{\Delta t}(u_0)$ of (5.6) can be expanded using Taylor series expansion, for Δt sufficiently small. Thus

$$\Phi_B^{\Delta t}(v) = v + \Delta t B(v) + (\Delta t)^2 \int_0^1 (1-\theta) dB(\Phi_B^{\theta\Delta t}(v)) \left[B(\Phi_B^{\theta\Delta t}(v))\right] d\theta,$$

for a general vector v. By inserting the expansion into (5.19) we get for $v = u_0$

$$u_1 = \Phi_A^{\Delta t}(u_0) + \Delta t\, \Phi_A^{\Delta t}(B(u_0)) + e_{G,2},$$

where

$$e_{G,2} = (\Delta t)^2 \int_0^1 (1-\theta) \Phi_A^{\Delta t}\Big(dB(\Phi_B^{\theta\Delta t}(u_0)) \left[B(\Phi_B^{\theta\Delta t}(u_0))\right]\Big) d\theta.$$

The error between the exact and the operator splitting solution, after one step becomes

$$u_1 - u(\Delta t) = \Delta t\, \Phi_A^{\Delta t}(B(u_0)) - \int_0^{\Delta t} \Phi_A^{(\Delta t - s)}\Big(B(\Phi_A^s(u(s)))\Big)\, ds + (e_{G,2} - e_{G,1}).$$

For simplicity, we define

$$f(s) = \Phi_A^{(\Delta t - s)}\Big(B(\Phi_A^s(u_0))\Big), \tag{5.20}$$

from which the above equation can be rewritten as

$$u_1 - u(\Delta t) = \Delta t\, f(0) - \int_0^{\Delta t} f(s)\, ds + (e_{G,2} - e_{G,1}).$$

We are at the keypoint of the new framework. By looking carefully at the terms in the equation above, we recognize the two first terms as the error of the rectangle rule, given in (3.5), for the integral of $f(s)$ over the interval $[0, \Delta t]$. By using the Peano kernel for the rectangle rule, the equation above turns into

$$u_1 - u(\Delta t) = \int_0^{\Delta t} K_R(t) f'(t)\, dt + (e_{G,2} - e_{G,1}),$$

where $K_R(t)$ is given in (3.6) and $f'(t)$ is the Fréchet derivative. If we use the substitution $\theta = t/\Delta t$ and (3.6), the integral is transformed to

$$\int_0^{\Delta t} K_R(t) f'(t)\, dt = (\Delta t)^2 \int_0^1 (\theta - 1) f'(\theta\, \Delta t)\, d\theta = (\Delta t)^2 \int_0^1 K_R(\theta) f'(\theta\, \Delta t)\, d\theta.$$

Thus, the error after one step between the operator splitting solution and the exact solution, is given as

$$u_1 - u(\Delta t) = (\Delta t)^2 \int_0^1 K_R(\theta) f'(\theta\, \Delta t)\, d\theta + (e_{G,2} - e_{G,1}),$$

and to obtain the bound for the error, we apply the H^s-norm and use the triangle inequality,

$$\begin{aligned}\|u_1 - u(\Delta t)\|_{H^s} &\le (\Delta t)^2 \int_0^1 \|K_R(\theta) f'(\theta\,\Delta t)\|_{H^s}\, d\theta + \|(e_{G,2} - e_{G,1})\|_{H^s} \\ &\le (\Delta t)^2 \int_0^1 \|f'(\theta\,\Delta t)\|_{H^s}\, d\theta + \|e_{G,2}\|_{H^s} + \|e_{G,1}\|_{H^s}. \end{aligned} \tag{5.21}$$

The remaining of the proof contains the estimation in $H^s(\mathbb{R})$ of the three terms on the right hand side.

We start with the integrand in (5.21). Using the differential rules introduced in Section 2, we find the Fréchet derivative of f as

$$\begin{aligned} f'(s) &= -\Phi_A^{(\Delta t - s)}\Big(dA\big(\Phi_A^s(u_0)\big)\big[B(\Phi_A^s(u_0))\big]\Big) + \Phi_A^{(\Delta t - s)}\Big(dB\big(\Phi_A^s(u_0)\big)\big[A(\Phi_A^s(u_0))\big]\Big) \\ &= -\Phi_A^{(\Delta t - s)}\Big(dA\big(\Phi_A^s(u_0)\big)\big[B(\Phi_A^s(u_0))\big] + dB\big(\Phi_A^s(u_0)\big)\big[A(\Phi_A^s(u_0))\big]\Big) \\ &= -\Phi_A^{(\Delta t - s)}\left[A, B\right]\left(\Phi_A^s(u_0)\right), \end{aligned}$$

where $[A,B](v) = dA(v)[B(v)] - dB(v)[A(v)]$ is the so-called Lie[28] commutator. Lemma 5.6 gives that $\Phi_A^t(u_0)$ do not increase the Sobolev norm, and therefore it is sufficient to consider the commutator for a general vector v. Using A and B given in (5.5) and (5.6), respectively, and their respective derivatives in (2.2) and (2.3), we get

$$\begin{aligned}[A,B](v) &= P(\partial_x)(vv_x) - v\left(P(\partial_x)v\right)_x - \left(P(\partial_x)v\right)v_x = \partial_x^l(vv_x) - v\partial_x^{l+1}v - (\partial_x^l v)v_x \\ &= \left(\sum_{k=0}^{l}\binom{l}{k}\partial_x^k v\,\partial_x^{l+1-k}v\right) - v\partial_x^{l+1}v - (\partial_x^l v)v_x = \sum_{k=1}^{l-1}\binom{l}{k}\partial_x^k v\,\partial_x^{l+1-k}v. \end{aligned}$$

Thus, using Lemma 4.2,

$$\begin{aligned}\|f'(s)\|_{H^s} &= \Big\|\sum_{k=1}^{l-1}\binom{l}{k}\partial_x^k v\,\partial_x^{l+1-k}v\Big\|_{H^s} \le \sum_{k=1}^{l-1}\binom{l}{k}\|\partial_x^k v\,\partial_x^{l+1-k}v\|_{H^s} \\ &\le C\sum_{k=1}^{l-1}\|\partial_x^k v\|_{H^s}\|\partial_x^{l+1-k}v\|_{H^s} \le C\|v\|_{H^{s+l}}^2, \end{aligned}$$

[28] Marius Sophus Lie, 17 December 1842 – 18 February 1899, Norwegian mathematician. Beside N. H. Abel, the most famous Norwegian mathematician in history. Founder of the theory of continuous transformation groups, which now is called Lie groups (and algebras). He applied the theory to the study of geometry and differential equations. Received his Ph.D. at the University of Oslo in 1871 with a thesis entitled *On a class of geometric transformations*. The Norwegian Parliament established an extraordinary professorship for him the year after. In 1886 Lie became a professor in Leipzig, and was in addition the administrator of the Mathematical Institute. Student from all over Europe was sent to Leipzig to follow his lectures. He returned to the University of Oslo in September 1898, where he stayed until his death half a year later. He was made Honorary Member of the London Mathematical Society in 1878, Member of the French Academy of Sciences in 1892, Foreign Member of the Royal Society of London in 1895 and foreign associate of the National Academy of Sciences of the United States of America in 1895.

Thus, by using $v = \Phi_A^s(u_0)$ in combination with Lemma 5.6, we get

$$\|f'(s)\|_{H^s} \leq C\|\Phi_A^s(u_0)\|_{H^{s+l}}^2 \leq C\|u_0\|_{H^{s+l}}^2.$$

Hence, the integral in (5.21) is bounded as

$$(\Delta t)^2 \int_0^1 \|f'(\theta\,\Delta t)\|_{H^s}\, d\theta \leq C\|u_0\|_{H^{s+l}}^2 (\Delta t)^2. \tag{5.22}$$

We continue with the error bound for $e_{G,1}$ in (5.18). This estimation is more or less a subsequence of calculations, where we at each step either use the Banach algebra property of $H^s(\mathbb{R})$ in Lemma 4.2, or the non-increasingness of the solution of (5.5) given in Lemma 5.6. We obtain

$$\begin{aligned}
\|e_{G,1}\|_{H^s} &\leq \int_0^{\Delta t}\int_0^s \left\| \Phi_A^{(\Delta t - s)} \left(dB\left(\Phi_A^{(s-\sigma)}(u(\sigma))\right)\left[\Phi_A^{(s-\sigma)}(B(u(\sigma)))\right]\right)\right\|_{H^s} d\sigma ds \\
&\leq \int_0^{\Delta t}\int_0^s \left\| dB\left(\Phi_A^{(s-\sigma)}(u(\sigma))\right)\left[\Phi_A^{(s-\sigma)}(B(u(\sigma)))\right]\right\|_{H^s} d\sigma ds \\
&\leq \int_0^{\Delta t}\int_0^s \left\| \left((\Phi_A^{(s-\sigma)}(u(\sigma)))(\Phi_A^{(s-\sigma)}(B(u(\sigma))))\right)_x\right\|_{H^s} d\sigma ds \\
&\leq \int_0^{\Delta t}\int_0^s \left\| (\Phi_A^{(s-\sigma)}(u(\sigma)))(\Phi_A^{(s-\sigma)}(B(u(\sigma))))\right\|_{H^{s+1}} d\sigma ds \\
&\leq C\int_0^{\Delta t}\int_0^s \left\| (\Phi_A^{(s-\sigma)}(u(\sigma)))\right\|_{H^{s+1}} \left\|\Phi_A^{(s-\sigma)}(B(u(\sigma)))\right\|_{H^{s+1}} d\sigma ds \\
&\leq C\int_0^{\Delta t}\int_0^s \|u(\sigma)\|_{H^{s+1}} \|B(u(\sigma))\|_{H^{s+1}}\, d\sigma ds.
\end{aligned}$$

Using the definition of B in (5.6), gives

$$\begin{aligned}
\|e_{G,1}\|_{H^s} &\leq C\int_0^{\Delta t}\int_0^s \|u(\sigma)\|_{H^{s+1}} \|u(\sigma)u_x(\sigma)\|_{H^{s+1}}\, d\sigma ds \\
&\leq C\int_0^{\Delta t}\int_0^s \|u(\sigma)\|_{H^{s+1}} \|u(\sigma)\|_{H^{s+1}} \|u_x(\sigma)\|_{H^{s+1}}\, d\sigma ds \\
&\leq C\int_0^{\Delta t}\int_0^s \|u(\sigma)\|_{H^{s+1}} \|u(\sigma)\|_{H^{s+1}} \|u(\sigma)\|_{H^{s+2}}\, d\sigma ds.
\end{aligned}$$

Now, we use the assumption about Hypothesis 5.2 for $H^{s+l}(\mathbb{R})$, which gives that $\|u(\sigma)\|_{H^{s+1}} \leq \|u(\sigma)\|_{H^{s+2}} \leq \|u(\sigma)\|_{H^{s+l}} \leq R$, when $l \geq 2$, which results in

$$\|e_{G,1}\|_{H^s} \leq C\int_0^{\Delta t}\int_0^s R^3\, d\sigma ds = CR^3 \int_0^{\Delta t} s\, ds = CR^3(\Delta t)^2, \tag{5.23}$$

and we have found a bound for the second term in (5.21).

The third and last term in (5.21) is estimated similarly as the second term. The major difference is the use of the regularity result for (5.6) in Lemma 5.4. We start with

the use of Lemma 5.6, which gives

$$\begin{aligned}\|e_{G,2}\|_{H^s} &\leq (\Delta t)^2 \int_0^1 \Big\| (1-\theta)\Phi_A^{\Delta t}\Big(dB(\Phi_B^{\theta\Delta t}(v)) \Big[B(\Phi_B^{\theta\Delta t}(v)) \Big] \Big) \Big\|_{H^s} d\theta \\ &\leq (\Delta t)^2 \int_0^1 \Big\| dB(\Phi_B^{\theta\Delta t}(u_0)) \Big[B(\Phi_B^{\theta\Delta t}(u_0)) \Big] \Big\|_{H^s} d\theta,\end{aligned}$$

which by the use of (2.3) and Lemma 4.2, turns into

$$\begin{aligned}\|e_{G,2}\|_{H^s} &\leq (\Delta t)^2 \int_0^1 \Big\| \Big((\Phi_B^{\theta\Delta t}(u_0)) (B(\Phi_B^{\theta\Delta t}(u_0))) \Big)_x \Big\|_{H^s} d\theta \\ &\leq (\Delta t)^2 \int_0^1 \| (\Phi_B^{\theta\Delta t}(u_0)) (B(\Phi_B^{\theta\Delta t}(u_0))) \|_{H^{s+1}} d\theta \\ &\leq C(\Delta t)^2 \int_0^1 \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+1}} \|B(\Phi_B^{\theta\Delta t}(u_0))\|_{H^{s+1}} d\theta \\ &\leq C(\Delta t)^2 \int_0^1 \Big\|\Phi_B^{\theta\Delta t}(u_0)\Big\|_{H^{s+1}} \Big\| \Big(\Phi_B^{\theta\Delta t}(u_0)\Big) \Big(\Phi_B^{\theta\Delta t}(u_0)\Big)_x \Big\|_{H^{s+1}} d\theta \\ &\leq C(\Delta t)^2 \int_0^1 \Big\|\Phi_B^{\theta\Delta t}(u_0)\Big\|_{H^{s+1}} \Big\|\Phi_B^{\theta\Delta t}(u_0)\Big\|_{H^{s+1}} \Big\| \Big(\Phi_B^{\theta\Delta t}(u_0)\Big)_x \Big\|_{H^{s+1}} d\theta \\ &\leq C(\Delta t)^2 \int_0^1 \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+1}} \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+1}} \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+2}} d\theta.\end{aligned}$$

For a sufficiently small Δt, Lemma 5.4 ensures that $\|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+1}} \leq \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+2}} \leq \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+l}} \leq R$. Thus,

$$\begin{aligned}\|e_{G,2}\|_{H^s} &\leq C(\Delta t)^2 \int_0^1 \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+1}} \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+1}} \|\Phi_B^{\theta\Delta t}(u_0)\|_{H^{s+2}} d\theta \\ &\leq C(\Delta t)^2 \int_0^1 R^3 \, d\theta \leq C(\Delta t)^2 R^3, \end{aligned} \tag{5.24}$$

and a bound for the third term in (5.21) is found.

Hence, by combining the estimates in (5.22), (5.23) and (5.24), we obtain the following bound for the error,

$$\|u_1 - u(\Delta t)\|_{H^s} \leq c_1 \, (\Delta t)^2,$$

where c_1 depends only on $\|u_0\|_{H^{s+l}}$, and Δt sufficiently small. This proves the lemma. □

5.6.2 Global Error

To estimate the global error in $H^s(\mathbb{R})$ and obtain the correct first order convergence rate for (5.2), we use the local error estimate in Lemma 5.8. The result in the lemma relies on the fact that the initial data at each step is bounded in $H^{s+l}(\mathbb{R})$. Thus, we need to show that the operator splitting solution at each step is bounded in $H^{s+l}(\mathbb{R})$, so that the local error estimate remains valid. To do this we have to use results for subequations (5.5) and (5.6), in addition with an induction argument.

We start with the regularity result for (5.6). To be more precise, we have to use Lemma 5.3 in the estimation. The problem with this lemma is that it relies on some knowledge of $\Phi_B^{\Delta t}(w_0)$. The lemma is stated using p and q defined in (5.10). In our case, we get the estimate $\|\Phi_B^{\Delta t}(w_0)\|_{H^{s+l}} \leq e^{c\alpha t}\|w_0\|_{H^{s+l}}$, as long as $\|\Phi_B^{\Delta t}(w_0)\|_{H^s} \leq \alpha$, for $s \geq 2$ and $0 \leq t \leq \Delta t$. We want the estimate to also yield for the case $s = 1$. By doing the same calculation as in the proof of Lemma 5.3, one can show that $\|\Phi_B^{\Delta t}(w_0)\|_{H^{1+l}} \leq e^{c\alpha t}\|w_0\|_{H^{1+l}}$, as long as $\|\Phi_B^{\Delta t}(w_0)\|_{H^r} \leq \alpha$ for $r \geq 2$ and $0 \leq t \leq \Delta t$. In what follows we use Lemma 5.3 for all $s \geq 1$, and keep the special case $s = 1$ in mind.

For (5.5) it turns out that it is sufficient with the fact that A do not increase the Sobolev norm, cf. Lemma 5.6.

The sketch of the proof is the following. To prove first order convergence for (5.2) we have to show boundedness of the splitting solution u_n at each time step t_n in $H^{s+l}(\mathbb{R})$, for which we use Lemma 5.3. Moreover, since Lemma 5.3 relies on boundedness of u_n in $H^s(\mathbb{R})$ we also have to show this. To simplify the proofs, we use an induction argument and prove the three results simultaneously. We assume that Hypotheses 5.1 and 5.2 holds for $k = s$.

To simplify the notation, we use the same notational convention as in [11]; we write

$$u_k^n = \Phi_{A+B}^{(n-k)\Delta t}(u_k) = \Phi^{(n-k)\Delta t}(u_k),$$

which is the exact solution of (5.4) with starting value u_k at time t_k. With this notation we see that

$$u_n = u_n^n \qquad \text{and} \qquad u(t_n) = u_n^0.$$

We start the induction argument by assuming that

$$\begin{aligned} \|u_k\|_{H^s} &\leq R, \\ \|u_k\|_{H^{s+l}} &\leq e^{2cRk\Delta t}\|u_0\|_{H^{s+l}} \leq e^{2cRT}\|u_0\|_{H^{s+l}} = C_0, \\ \|u_k - u(t_k)\|_{H^s} &\leq \gamma\Delta t, \end{aligned}$$

holds for all $k \leq n-1$, and we check that the above inequalities hold for $k = n$. In the inequalities is c as in Lemma 5.3 and $\gamma = K(R,T)c_1(C_0)T$, where $K(R,T)$ is as in (5.9) and $c_1(C_0)$ is the constant in Lemma 5.3 for initial data bounded by C_0 in H^{s+l}.

The error between u_k and the exact solution $u(t_k)$ for $k = n$, can be expanded using a telescope sum and the triangle inequality in the following way,

$$\begin{aligned} \|u_n - u(t_n)\|_{H^s} &= \|u_n^n - u_n^0\|_{H^s} = \|u_n^n - u_n^{n-1} + u_n^{n-1} - u_n^{n-2} + u_n^{n-2} - \cdots - u_n^0\|_{H^s} \\ &= \Big\| \sum_{k=0}^{n-1} u_n^{k+1} - u_n^k \Big\|_{H^s} \leq \sum_{k=0}^{n-1} \|u_n^{k+1} - u_n^k\|_{H^s}, \end{aligned}$$

which by using the notational convention becomes

$$\begin{aligned} \|u_n - u(t_n)\|_{H^s} &\leq \sum_{k=0}^{n-1} \big\|\Phi^{(n-k-1)\Delta t}(u_{k+1}) - \Phi^{(n-k)\Delta t}(u_k)\big\|_{H^s} \\ &= \sum_{k=0}^{n-1} \big\|\Phi^{(n-k-1)\Delta t}\left(\Pi^{\Delta t}(u_k)\right) - \Phi^{(n-k-1)\Delta t}\left(\Phi^{\Delta t}(u_k)\right)\big\|_{H^s}, \end{aligned} \tag{5.25}$$

where we note that $\Phi^{(n-k)\Delta t}(u_k) = \Phi^{(n-k-1)\Delta t}\left(\Phi^{\Delta t}(u_k)\right)$. For terms with $k \leq n-2$, we get using Hypothesis 5.2,

$$\|\Pi^{\Delta t}(u_k)\|_{H^s} = \|u_{k+1}\|_{H^s} \leq R, \tag{5.26}$$

and for the exact solution

$$\|\Phi^{\Delta t}(u_k)\|_{H^s} \leq \|\Phi^{\Delta t}(u_k) - \Phi^{\Delta t}(u(t_k))\|_{H^s} + \|\Phi^{\Delta t}(u(t_k))\|_{H^s},$$

which by using the Lipschitz continuity assumption in (5.9), turns into

$$\|\Phi^{\Delta t}(u_k)\|_{H^s} \leq K(R,\Delta t)\|u_k - u(t_k)\|_{H^s} + \|u(t_{k+1})\|_{H^s} \leq K(R,\Delta t)\gamma\Delta t + \rho,$$

from the induction argument. We choose Δt such that $K(R,\Delta t)\gamma\Delta t \leq R-\rho$,

$$\|\Phi^{\Delta t}(u_k)\|_{H^s} \leq R. \tag{5.27}$$

Returning to (5.25), we get for each term using Hypothesis 5.1,

$$\left\|\Phi^{(n-k-1)\Delta t}\left(\Pi^{\Delta t}(u_k)\right) - \Phi^{(n-k-1)\Delta t}\left(\Phi^{\Delta t}(u_k)\right)\right\|_{H^s} \leq K(R,T)\|\Pi^{\Delta t}(u_k) - \Phi^{\Delta t}(u_k)\|_{H^s},$$

From the estimates in (5.26) and (5.27), Lemma 5.8 yields for $k \leq n-1$, and we obtain for each term

$$\left\|\Phi^{(n-k-1)\Delta t}\left(\Pi^{\Delta t}(u_k)\right) - \Phi^{(n-k-1)\Delta t}\left(\Phi^{\Delta t}(u_k)\right)\right\|_{H^s} \leq K(R,T)c_1(C_0)\,(\Delta t)^2.$$

Thus, summing up all term and using that $n\Delta t \leq T$,

$$\|u_n - u(t_n)\|_{H^s} \leq nK(R,T)c_1(C_0)\,(\Delta t)^2 \leq \gamma\Delta t.$$

To prove the boundedness of u_n, we choose $\gamma\Delta t \leq R-\rho$ and use Hypothesis 5.2, which gives

$$\|u_n\|_{H^s} = \|u_n - u(t_n)\|_{H^s} + \|u(t_n)\|_{H^s} \leq R - \rho + \rho \leq R.$$

Thus u_n is bounded in $H^s(\mathbb{R})$ for each split step t_n, and the estimate in Lemma 5.3 is valid for each time step t_n. To bound u_n in $H^{s+l}(\mathbb{R})$, we start with

$$\|u_n\|_{H^{s+l}} = \|\Phi_A^{\Delta t} \circ \Phi_B^{\Delta t}(u_{n-1})\|_{H^{s+l}} \leq \|\Phi_B^{\Delta t}(u_{n-1})\|_{H^{s+l}},$$

where we in the second inequality have used Lemma 5.6. Lemma 5.4 gives that there exists $\Delta t \leq \bar{t}(R)$ such that $\|\Phi_B^{\Delta t}(u_{n-1})\|_{H^{s+l}} \leq 2R$, as long as $\|u_{n-1}\|_{H^{s+l}}$ is bounded. This is in our case ensured by the induction assumption. Thus, using Lemma 5.3, we get

$$\|u_n\|_{H^{s+l}} \leq e^{2cR\Delta t}\|u_{n-1}\|_{H^{s+l}} \leq e^{2cRn\Delta t}\|u_0\|_{H^{s+l}} \leq C_0.$$

Thus, the three necessary results hold by the induction argument.

We collect the main result from the above calculations, in the following theorem.

Theorem 5.9. *Assume there exists a solution of* (5.4) *and let* $s \geq 1$ *be an integer. If Hypothesis 5.1 holds for* $k = s$ *and Hypothesis 5.2 holds for* $k = s + l$ *for* $l \geq 2$*, then there is* $\overline{\Delta t} > 0$ *such that for all* $\Delta t \leq \overline{\Delta t}$ *and* $t_n = n\Delta t \leq T$*,*

$$\|u_n - u(t_n)\|_{H^s} \leq C\,\Delta t,$$

where u_n *is the Godunov splitting solution* (5.2)*, and* $\overline{\Delta t}$ *and* C *only depends on* $\|u_0\|_{H^{s+l}}$, ρ *and* T*.*

5.7 Strang Splitting

To prove the correct convergence rate for the Strang splitting (5.3), we use the same framework as in the proof of the convergence rate for the Godunov splitting (5.2). The major difference between the two proofs is that for (5.3) we need to use the higher order midpoint rule, where we for (5.2) used the rectangle rule. In addition, a higher order series expansion of the involved terms are also necessary to obtain the results.

For (5.3) it is possible to obtain first order convergence results in $H^q(\mathbb{R})$, with p and q as in (5.10). We will not focus on these results, but refer to [11, Secs. 4,5.] for a proof. To prove the second order rates we follow the same outline as in Section 5.6, and present the local error estimate in $H^s(\mathbb{R})$, which is followed up by the global error proof in $H^s(\mathbb{R})$.

5.7.1 Local Error

As discussed before, to obtain second order convergence rate for Δt for the Strang splitting (5.3), we need to get third order convergence for the local error in $H^s(\mathbb{R})$, and to obtain this, a higher order Taylor expansion for the exact solution of (5.6) is necessary. The other ideas for the rest of the proofs, are identical as those for the Godunov splitting (5.2); that is, find the error between the operator splitting solution and the exact (Taylor expanded) solution, and bound it using numerical quadratures and properties of the space. The proof is longer due to the extra order in the Taylor expansion.

Lemma 5.10. *Define s and p by* (5.10) *and assume Hypothesis 5.2 holds for $k = p$. Then the local error of the Strang splitting in* (5.3) *is bounded in $H^s(\mathbb{R})$ by*

$$\left\|\Psi^{\Delta t}(u_0) - \Phi^{\Delta t}_{A+B}(u_0)\right\|_{H^s} \leq c_2\,(\Delta t)^3,$$

where c_2 depends only on $\|u_0\|_{H^p}$.

Proof. The second order Taylor expansion for $\Phi^{\Delta t}_B(v)$, is given as

$$\begin{aligned}\Phi^{\Delta t}_B(v) =& v + \Delta t B(v) + \frac{1}{2}(\Delta t)^2 dB(v)[B(v)] \\ &+ (\Delta t)^3 \int_0^1 \frac{(1-\theta)^2}{2}\Big(d^2B\left(\Phi^{\theta\Delta t}_B(v)\right)\left[B\left(\Phi^{\theta\Delta t}_B(v)\right), B\left(\Phi^{\theta\Delta t}_B(v)\right)\right] \\ &\qquad + dB\left(\Phi^{\theta\Delta t}_B(v)\right)\left[dB\left(\Phi^{\theta\Delta t}_B(v)\right)\left[B\left(\Phi^{\theta\Delta t}_B(v)\right)\right]\right]\Big)\,d\theta,\end{aligned}$$

and is valid for Δt sufficiently small by Lemma 5.5. For easiness in the notation, we abbreviate the integrand as

$$\begin{aligned}&d^2B\left(\Phi^{\theta\Delta t}_B(v)\right)\left[B\left(\Phi^{\theta\Delta t}_B(v)\right), B\left(\Phi^{\theta\Delta t}_B(v)\right)\right] \\ &\qquad + dB\left(\Phi^{\theta\Delta t}_B(v)\right)\left[dB\left(\Phi^{\theta\Delta t}_B(v)\right)\left[B\left(\Phi^{\theta\Delta t}_B(v)\right)\right]\right] \\ &\quad = \Big(d^2B(B,B) + dB\,dB\,B\Big)(\Phi^{\theta\Delta t}_B(v)).\end{aligned} \tag{5.28}$$

By inserting the above series expansion into (5.3), we obtain the operator splitting solution after one step,

$$u_1 = \Phi_A^{\Delta t}(u_0) + \Delta t\, \Phi_A^{\Delta t/2}\left(B(\Phi_A^{\Delta t/2}(u_0))\right) + \frac{1}{2}(\Delta t)^2 \Phi_A^{\Delta t/2}\left(dB(\Phi_A^{\Delta t/2}(u_0))[B(\Phi_A^{\Delta t/2}(u_0))]\right) + e_{S,2}, \tag{5.29}$$

where

$$e_{S,2} = (\Delta t)^3 \int_0^1 \frac{(1-\theta)^2}{2} \Phi_A^{\Delta t/2}\Big(d^2B(B,B) + dB\, dB\, B\Big)\Big(\Phi_B^{\theta\Delta t}(\Phi_A^{\Delta t/2}(u_0))\Big)\, d\theta. \tag{5.30}$$

The exact solution after one split step is given as, cf. (5.17) and (5.18),

$$u(\Delta t) = \Phi_A^{\Delta t}(u_0) + \int_0^{\Delta t} \Phi_A^{(\Delta t - s)}\Big(B(\Phi_A^s(u(s)))\Big)\, ds + \int_0^{\Delta t}\int_0^s \Phi_A^{(\Delta t-s)}\Big(G(u(\sigma))\Big)\, d\sigma\, ds, \tag{5.31}$$

where $G(u(\sigma))$ is defined for a general vector v as

$$G(v) = G_{s,\sigma}(v) = dB\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(B(v))\right]. \tag{5.32}$$

Using the integral formula in (5.15), we obtain

$$G\left(u(\sigma)\right) = G\left(\Phi_A^{\sigma}(u_0)\right) + \int_0^{\sigma} dG\left(\Phi_A^{(\sigma-\tau)}(u(\tau))\right)\left[\Phi_A^{(\sigma-\tau)}(B(u(\tau)))\right]\, d\tau,$$

where the integrand is calculated as

$$dG(v)[w] = d^2B\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(w), \Phi_A^{(s-\sigma)}(B(v))\right] + dB\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(dB(v)[w])\right]. \tag{5.33}$$

Inserting the integral formula for G into (5.31) gives

$$u(\Delta t) = \Phi_A^{\Delta t}(u_0) + \int_0^{\Delta t} \Phi_A^{(\Delta t-s)}\Big(B(\Phi_A^s(u(s)))\Big)\, ds + e_{S,3}, \tag{5.34}$$

where

$$e_{S,3} = \int_0^{\Delta t}\int_0^s \Phi_A^{(\Delta t-s)}\left(dB\left(\Phi_A^s(u_0)\right)\left[\Phi_A^{(s-\sigma)}(B(\Phi_A^s(u_0)))\right]\right)\, d\sigma ds + \int_0^{\Delta t}\int_0^s\int_0^{\sigma} dG_{s,\sigma}\Big(\Phi_A^{(\sigma-\tau)}(u(\tau))\Big)\Big[\Phi_A^{(\sigma-\tau)}(B(u(\tau)))\Big]\, d\tau\, d\sigma\, ds. \tag{5.35}$$

Taking the difference of (5.29) and (5.34) gives the local error after one step

$$u_1 - u(\Delta t) = \Delta t\, \Phi_A^{\Delta t/2}\left(B(\Phi_A^{\Delta t/2}(u_0))\right) - \int_0^{\Delta t} \Phi_A^{(\Delta t-s)}\Big(B(\Phi_A^s(u(s)))\Big)\, ds + \frac{1}{2}(\Delta t)^2 \Phi_A^{\Delta t/2}\left(dB(\Phi_A^{\Delta t/2}(u_0))[B(\Phi_A^{\Delta t/2}(u_0))]\right) + (e_{S,2} - e_{S,3}).$$

To get the above expression in a more readable format, we define

$$g(s,\sigma) = \Phi_A^{(\Delta t - s)}\Big(dB\big(\Phi_A^s(u_0)\big)\big[\Phi_A^{(s-\sigma)}B\big(\Phi_A^\sigma(u_0)\big)\big]\Big). \tag{5.36}$$

By using the expressions for $e_{S,2}$ and $e_{S,3}$ given in (5.30) and (5.35), respectively, and the definition of g in (5.36) and of f in (5.20), the local error can be simplified to

$$\begin{aligned} u_1 - u(\Delta t) = {} & \Delta t f(\Delta t/2) - \int_0^{\Delta t} f(s)\,ds \\ & + \frac{1}{2}(\Delta t)^2 g(\Delta t/2, \Delta t/2) - \int_0^{\Delta t}\int_0^s g(s,\sigma)\,d\sigma\,ds \\ & + e_{S,4} - e_{S,5}, \end{aligned} \tag{5.37}$$

where

$$e_{S,4} = (\Delta t)^3 \int_0^1 \frac{(1-\theta)^2}{2}\Phi_A^{\Delta t/2}\Big(d^2B(B,B) + dB\,dB\,B\Big)\Big(\Phi_B^{\theta\Delta t}\big(\Phi_A^{\Delta t/2}(u_0)\big)\Big)\,d\theta, \tag{5.38}$$

and

$$e_{S,5} = \int_0^{\Delta t}\int_0^s\int_0^\sigma dG_{s,\sigma}\left(\Phi_A^{(\sigma-\tau)}(u(\tau))\right)\left[\Phi_A^{(\sigma-\tau)}(B(u(\tau)))\right]\,d\tau\,d\sigma\,ds. \tag{5.39}$$

The first line in (5.37) is nothing but the error of the midpoint rule (3.7), while the second line is the error of the two-dimensional quadrature rule (3.14). Using the triangle inequality, we get

$$\begin{aligned} \|u_1 - u(\Delta t)\|_{H^s} \le {} & \int_0^{\Delta t} \|K_{M_2}(t)f''(t)\|_{H^s}\,ds \\ & + \Big\|\frac{1}{2}(\Delta t)^2 g(\Delta t/2, \Delta t/2) - \int_0^{\Delta t}\int_0^s g(s,\sigma)\,d\sigma\,ds\Big\|_{H^s} \\ & + \|e_{S,4}\|_{H^s} + \|e_{S,5}\|_{H^s}, \end{aligned} \tag{5.40}$$

where $K_{M_2}(t)$ is given in (3.8) and $f''(t)$ is the Fréchet derivative. We need to find bounds in $H^s(\mathbb{R})$ for each term above.

We start with the first term in (5.40). By using the substitution $\theta = t/\Delta t$, the integral can be transformed to

$$\int_0^{\Delta t} K_{M_2}(t)f''(t)\,ds = (\Delta t)^3 \int_0^1 K_{M_2}(\theta)f''(\theta\,\Delta t)\,d\theta.$$

The second derivative of f is found as

$$\begin{aligned} f''(s) = \Phi_A^{(\Delta t - s)}\Big(& (dA(v))^2[B(v)] - dA(v)[dB(v)[A(v)]] - d^2A(v)[B(v),A(v)] \\ & - dA(v)[dB(v)[A(v)]] + d^2B(v)[A(v)]^2 + dB(v)[dA(v)[A(v)]]\Big), \end{aligned}$$

for $v = \Phi_A^s(u_0)$. In the above equation, we put in for the operators A, B and all the respective derivatives, see (2.2), (2.4), (2.3) and (2.5). Thus,

$$\begin{aligned}
f''(s) &= \Phi_A^{(\Delta t - s)}\Big(A^2(B(v)) - 2A\big(dB(v)[A(v)]\big) \\
&\qquad + d^2B(v)\big[A(v)\big]^2 + dB(v)\big[A^2(v)\big]\Big) \\
&= \Phi_A^{(\Delta t - s)}\Big(A^2(vv_x) - 2A\big((vA(v))_x\big) + \big((A(v))^2\big)_x + \big(vA^2(v)\big)_x\Big) \\
&= \Phi_A^{(\Delta t - s)}\Big(\partial_x^{2l}(vv_x) - 2\partial_x^{l+1}(v\partial_x^l(v)) + \big((\partial_x^l(v))^2\big)_x + \big(v\partial_x^{2l}(v)\big)_x\Big).
\end{aligned}$$

Since the $\Phi_A^{(\Delta t - s)}$ do not increase the Sobolev norm by Lemma 5.6, it is sufficient to investigate the norm of the argument. Writing out the argument using Leibniz' rule gives

$$\begin{aligned}
&\partial_x^{2l}(vv_x) - 2\partial_x^l\big((v\partial_x^l(v))_x\big) + \big((\partial_x^l(v))^2\big)_x + \big(v\partial_x^{2l}(v)\big)_x \\
&\qquad = \sum_{k=0}^{2l}\binom{2l}{k}\partial_x^{2l-k}v\,\partial_x^{k+1}v - 2\sum_{k=0}^{l+1}\binom{l+1}{k}\partial_x^{l+1-k}v\,\partial_x^{l+k}v \\
&\qquad\quad + 2\partial_x^l v\,\partial_x^{l+1}v + v_x\partial_x^{2l}v + v\partial_x^{2l+1}v.
\end{aligned}$$

Writing out the start and end terms in the above sums, we obtain

$$\begin{aligned}
&\sum_{k=0}^{2l}\binom{2l}{k}\partial_x^{2l-k}v\,\partial_x^{k+1}v - 2\sum_{k=0}^{l+1}\binom{l+1}{k}\partial_x^{l+1-k}v\,\partial_x^{l+k}v + 2\partial_x^l v\,\partial_x^{l+1}v + v_x\partial_x^{2l}v + v\partial_x^{2l+1}v \\
&= \sum_{k=1}^{2l-2}\binom{2l}{k}\partial_x^{2l-k}v\,\partial_x^{k+1}v + 2lv_x\,\partial_x^{2l}v + v\,\partial_x^{2l+1}v + \partial_x^{2l}v\,v_x \\
&\qquad - 2\sum_{k=0}^{l-1}\binom{l+1}{k}\partial_x^{l+1-k}v\,\partial_x^{l+k}v - 2v\,\partial_x^{2l+1}v - 2(l-1)v_x\,\partial_x^{2l}v + 2\partial_x^l v\,\partial_x^{l+1}v \\
&\qquad\quad + v_x\partial_x^{2l}v + v\partial_x^{2l+1}v \\
&= \sum_{k=1}^{2l-2}\binom{2l}{k}\partial_x^{2l-k}v\,\partial_x^{k+1}v - 2\sum_{k=0}^{l-1}\binom{l+1}{k}\partial_x^{l+1-k}v\,\partial_x^{l+k}v + 2\partial_x^l v\,\partial_x^{l+1}v,
\end{aligned}$$

which shows that all the derivatives of order $2l+1$ and $2l$ on v are cancelled out. Hence,

$$\begin{aligned}
\|f''(s)\|_{H^s} &\le \Big\|\sum_{k=1}^{2l-2}\binom{2l}{k}\partial_x^{2l-k}v\,\partial_x^{k+1}v\Big\|_{H^s} + \Big\|2\sum_{k=0}^{l-1}\binom{l+1}{k}\partial_x^{l+1-k}v\,\partial_x^{l+k}v\Big\|_{H^s} \\
&\quad + \Big\|2\partial_x^l v\,\partial_x^{l+1}v\Big\|_{H^s} \\
&\le \sum_{k=1}^{2l-2}\binom{2l}{k}\Big\|\partial_x^{2l-k}v\,\partial_x^{k+1}v\Big\|_{H^s} + 2\sum_{k=0}^{l-1}\binom{l+1}{k}\Big\|\partial_x^{l+1-k}v\,\partial_x^{l+k}v\Big\|_{H^s} \\
&\quad + \Big\|2\partial_x^l v\,\partial_x^{l+1}v\Big\|_{H^s},
\end{aligned}$$

Using the Banach algebra property in Lemma 4.2, turns the inequality into

$$\begin{aligned}\|f''(s)\|_{H^s} \le C\sum_{k=1}^{2l-2} \|\partial_x^{2l-k} v\|_{H^s} \|\partial_x^{k+1} v\|_{H^s} + C\sum_{k=0}^{l-1} \|\partial_x^{l+1-k} v\|_{H^s} \|\partial_x^{l+k} v\|_{H^s} \\ + C\|\partial_x^l v\|_{H^s} \|\partial_x^{l+1} v\|_{H^s} \le C\|v\|_{H^p}^2,\end{aligned}$$

since $2l-1 \le p$. Thus, using $v = \Phi_A^s(u_0)$ and Lemma 5.6,

$$\|f''(s)\|_{H^s} \le C\|\Phi_A^s(u_0)\|_{H^p}^2 \le C\|u_0\|_{H^s}^2.$$

Hence, we obtain a bound for the first term in (5.40)

$$\begin{aligned}\int_0^{\Delta t} \|K_{M_2}(t) f''(t)\|_{H^s}\, ds &\le (\Delta t)^3 \int_0^1 \|K_{M_2}(\theta) f''(\theta\,\Delta t)\|_{H^s}\, d\theta \\ &\le (\Delta t)^3 \int_0^1 \|f''(\theta\,\Delta t)\|_{H^s}\, d\theta \le C\|u_0\|_{H^p}^2. \end{aligned} \tag{5.41}$$

The second term in (5.40), is the error of the two-dimensional quadrature formula (3.14). Using the error bound (3.15) with $h = \Delta t$ and $a = b = \Delta t/2$, dropping the higher order terms and use norms instead of absolute values, we get

$$\begin{aligned}\Big\|\frac{1}{2}(\Delta t)^2 g(\Delta t/2, \Delta t/2) - \int_0^{\Delta t}\int_0^s g(s,\sigma)\, d\sigma\, ds\Big\|_{H^s} \\ \le C(\Delta t)^3\Big(\max_T \|\partial g/\partial s\|_{H^s} + \max_T \|\partial g/\partial \sigma\|\Big),\end{aligned} \tag{5.42}$$

Thus, we need to find a bound for the partial derivatives of g. For notational easiness, we define

$$v(s) = \Phi_A^s(u_0) \quad \text{and} \quad w(s,\sigma) = \Phi_A^{(s-\sigma)}\Big(B(v(\sigma))\Big).$$

With the two definitions, g in (5.36) is rewritten as

$$g(s,\sigma) = \Phi_A^{(\Delta t - s)}\Big(dB(v(s))\,[w(s,\sigma)]\Big).$$

The partial derivatives of g is given as

$$\begin{aligned}\frac{\partial g}{\partial s} = \Phi_A^{(\Delta t-s)}\Big(- A(dB(v(s)))\,[w(s,\sigma)] + d^2B(v(s))\,[A(v(s)), w(s,\sigma)] \\ + dB(v(s))\,[A(w(s,\sigma))]\Big)\end{aligned} \tag{5.43}$$

and

$$\frac{\partial g}{\partial \sigma} = \Phi_A^{(\Delta t-s)}\Big(dB(v(s))\Big[\Phi_A^{(s-\sigma)}\Big(- A(B(v(s))) + dB(v(\sigma))\,[A(v(\sigma))]\Big)\Big]\Big). \tag{5.44}$$

We start with finding a bound for (5.43). Lemma 5.6 gives (we leave out the arguments in v and w), and obtain

$$\begin{aligned}\Big\|\frac{\partial g}{\partial s}\Big\|_{H^s} &= \Big\|\Phi_A^{(\Delta t-s)}\Big(-A\big(dB(v)[w]\big)+d^2B(v)[A(v),w]+dB(v)[A(w)]\Big)\Big\|_{H^s}\\ &\le \|-A\big(dB(v)[w]\big)+d^2B(v)[A(v),w]+dB(v)[A(w)]\|_{H^s}.\end{aligned}$$

Using the definition for A, B and their respective derivatives gives

$$\begin{aligned}-A\big(dB(v)[w]\big)+d^2B(v)[A(v),w]+dB(v)[A(w)]&\\ = -A((vw)_x)+(A(v)w)_x+(vA(w))_x&\\ = \big(-A(vw)+A(v)w+vA(w)\big)_x,&\end{aligned}$$

which yield

$$\begin{aligned}\Big\|\frac{\partial g}{\partial s}\Big\|_{H^s} &\le \|\big(-A(vw)+A(v)w+vA(w)\big)_x\|_{H^s} \le \|-A(vw)+A(v)w+vA(w)\|_{H^{s+1}}\\ &\le \|-\partial_x^l(vw)+\partial_x^l vw+v\partial_x^l w\|_{H^{s+1}}.\end{aligned}$$

Writing out the terms inside the norm using Leibniz' rule gives,

$$\begin{aligned}-\partial_x^l(vw)+\partial_x^l v\,w+v\,\partial_x^l w &= -\sum_{k=0}^{l}\binom{l}{k}\partial_x^{l-k}v\,\partial_x^k w+\partial_x^l v\,w+v\,\partial_x^l w\\ &= -\sum_{k=1}^{l-1}\binom{l}{k}\partial_x^{l-k}v\,\partial_x^k w.\end{aligned}$$

Thus, all the derivatives of order l cancel, and we get

$$\Big\|\frac{\partial g}{\partial s}\Big\|_{H^s} \le \Big\|\sum_{k=1}^{l-1}\binom{l}{k}\partial_x^{l-k}v\,\partial_x^k w\Big\|_{H^s} \le \sum_{k=1}^{l-1}\binom{l}{k}\|\partial_x^{l-k}v\,\partial_x^k w\|_{H^s},$$

which by using Lemma 4.2 is separated as,

$$\Big\|\frac{\partial g}{\partial s}\Big\|_{H^s} = C\sum_{k=1}^{l-1}\|\partial_x^{l-k}v\|_{H^s}\|\partial_x^k w\|_{H^s} \le C\|v(s)\|_{H^{s+l}}\|w(s,\sigma)\|_{H^{s+l}} \le C\|u_0\|_{H^p}^3, \quad (5.45)$$

where the last inequality follows from Lemma 5.6

$$\|v(s)\|_{H^{s+l}} = \|\Phi_A^s(u_0)\|_{H^{s+l}} \le \|u_0\|_{H^{s+l}},$$

and

$$\begin{aligned}\|w(s,\sigma)\|_{H^{s+l}} &= \|\Phi_A^{(s-\sigma)}\Big(B(v(\sigma))\Big)\|_{H^{s+l}} \le \|v(\sigma)\,v_x(\sigma)\|_{H^{s+l}}\\ &\le C\|v(\sigma)\|_{H^{s+l}}\|v(\sigma)\|_{H^{s+l+1}} \le C\|u_0\|_{H^{s+l}}\|u_0\|_{H^{s+l+1}} \le C\|u_0\|_{H^p}^2,\end{aligned}$$

since $s+l+1 \le s+2l-1 = p$ when $l \ge 2$.

For (5.44) we get, using Lemmas 4.2 and 5.6 and (2.3),

$$\begin{aligned}
\Big\|\frac{\partial g}{\partial \sigma}\Big\|_{H^s} &= \Big\|\Phi_A^{(\Delta t-s)}\Big(dB(v)[\Phi_A^{(s-\sigma)}(-A(B(v))+dB(v)[A(v)])]\Big)\Big\|_{H^s} \\
&\le \Big\|dB(v)\Big[\Phi_A^{(s-\sigma)}(-A(B(v))+dB(v)[A(v)])\Big]\Big\|_{H^s} \\
&= \Big\|\Big(v(\Phi_A^{(s-\sigma)}(-A(B(v))+dB(v)[A(v)]))\Big)_x\Big\|_{H^s} \\
&\le \Big\|v\Big(\Phi_A^{(s-\sigma)}(-A(B(v))+dB(v)[A(v)])\Big)\Big\|_{H^{s+1}} \\
&\le C\|v\|_{H^{s+1}}\Big\|\Phi_A^{(s-\sigma)}(-A(B(v))+dB(v)[A(v)])\Big\|_{H^{s+1}} \\
&\le C\|v\|_{H^{s+1}}\Big\|-A(B(v))+dB(v)[A(v)]\Big\|_{H^{s+1}}.
\end{aligned}$$

We write out the term using the definition of A and B and Leibniz' rule. This gives.

$$\begin{aligned}
-A(B(v))+dB(v)[A(v)] &= -\partial_x^l(vv_x)+(v\,\partial_x^l v)_x \\
&= -\sum_{k=0}^{l}\binom{l}{k}\partial_x^k v\,\partial_x^{l-k+1}v + v_x\partial_x^l v + v\partial_x^{l+1}v \\
&= -\sum_{k=1}^{l-1}\binom{l}{k}\partial_x^k v\,\partial_x^{l-k+1}v,
\end{aligned}$$

From the above calculation, we observe that the derivatives of order $l+1$ on v vanish. To be precise, we get

$$\begin{aligned}
\Big\|\frac{\partial g}{\partial \sigma}\Big\|_{H^s} &\le C\|v\|_{H^{s+1}}\Big\|\sum_{k=1}^{l-1}\binom{l}{k}\partial_x^k v\,\partial_x^{l-k+1}v\Big\|_{H^{s+1}} \\
&\le C\|v\|_{H^{s+1}}\sum_{k=1}^{l-1}\binom{l}{k}\|\partial_x^k v\,\partial_x^{l-k+1}v\|_{H^{s+1}} \\
&\le C\|v\|_{H^{s+1}}\sum_{k=1}^{l-1}\|\partial_x^k v\|_{H^{s+1}}\,\|\partial_x^{l-k+1}v\|_{H^{s+1}} \\
&\le C\|v\|_{H^{s+1}}\|v\|_{H^{s+l}}^2 \le C\|v\|_{H^{s+l}}^3 \le C\|u_0\|_{H^p}^3. \qquad (5.46)
\end{aligned}$$

Returning to (5.42), we get by using (5.45) and (5.46),

$$\Big\|\frac{1}{2}(\Delta t)^2 g(\Delta t/2,\Delta t/2)-\int_0^{\Delta t}\int_0^s g(s,\sigma)\,d\sigma\,ds\Big\|_{H^s} \le C\,(\Delta t)^3\|u_0\|_{H^p}^3 \le C\,(\Delta t)^3. \quad (5.47)$$

For the third term in (5.40) we get, using (5.38),

$$\begin{aligned}
\|e_{S,4}\|_{H^s} &\le (\Delta t)^3\int_0^1\Big\|\Phi_A^{\Delta t/2}\Big(d^2B(B,B)+dB\,dB\,B\Big)\Big(\Phi_B^{\theta\Delta t}(\Phi_A^{\Delta t/2}(u_0))\Big)\Big\|_{H^s}\,d\theta \\
&\le (\Delta t)^3\int_0^1\Big\|\Big(d^2B(B,B)+dB\,dB\,B\Big)\Big(\Phi_B^{\theta\Delta t}(\Phi_A^{\Delta t/2}(u_0))\Big)\Big\|_{H^s}\,d\theta,
\end{aligned}$$

where we remember that the integrand is defined in (5.28). By the triangle inequality,

$$\|(d^2B(B,B) + dB\, dB\, B)(w)\|_{H^s} \leq \|d^2B(B,B)(w)\|_{H^s} + \|dB\, dB\, B(w)\|_{H^s},$$

where we have redefined

$$w = \Phi_B^{\theta\Delta t}(\Phi_A^{\Delta t/2}(u_0)).$$

For the first term we get, using (2.5) and Lemma 4.2

$$\begin{aligned}\|d^2B(B,B)(w)\|_{H^s} &\leq \|\left(B(w)\, B(w)\right)_x\|_{H^s} \leq \|B(w)\, B(w)\|_{H^{s+1}} \\ &\leq C\|B(w)\|^2_{H^{s+1}} \leq C\|ww_x\|^2_{H^{s+1}} \\ &\leq C\|w\|^2_{H^{s+1}}\|w\|^2_{H^{s+2}} \leq C\|w\|^4_{H^{s+3}}.\end{aligned}$$

Using Lemma 5.3 and Δt sufficiently small, we get

$$\|d^2B(B,B)(w)\|_{H^s} \leq C\|u_0\|^4_{H^p}. \tag{5.48}$$

The bound for the second term is found similarly,

$$\begin{aligned}\|dB\, dB\, B(w)\|_{H^s} &= \|\left(w\, dB(w)[B(w)]\right)_x\|_{H^s} \leq \|w\, dB(w)[B(w)]\|_{H^{s+1}} \\ &\leq C\|w\|_{H^{s+1}}\|\left(wB(w)\right)_x\|_{H^{s+1}} \leq C\|w\|_{H^{s+1}}\|w\|_{H^{s+2}}\|ww_x\|_{H^{s+2}} \\ &\leq C\|w\|_{H^{s+1}}\|w\|^2_{H^{s+2}}\|w\|_{H^{s+3}} \leq C\|w\|^4_{H^{s+3}} \leq C\|u_0\|^4_{H^p} \qquad (5.49)\end{aligned}$$

where the last inequality follows from Lemma 5.3 and Δt sufficiently small. Hence, by using (5.48) and (5.49) we get

$$\|e_{S,4}\|_{H^s} \leq C(\Delta t)^3\|u_0\|^4_{H^p} \leq C(\Delta t)^3. \tag{5.50}$$

For the fourth term in (5.40), we get using (5.39)

$$\|e_{S,5}\|_{H^s} \leq \int_0^{\Delta t}\int_0^s\int_0^\sigma \left\|dG_{s,\sigma}\left(\Phi_A^{(\sigma-\tau)}(u(\tau))\right)\left[\Phi_A^{(\sigma-\tau)}(B(u(\tau)))\right]\right\|_{H^s} d\tau\, d\sigma\, ds,$$

where G is defined in (5.32) and $dG(v)[w]$ was found in (5.33). For the integrand we obtain,

$$\begin{aligned}\|dG_{s,\sigma}(v)[w]\|_{H^s} \leq &\left\|d^2B\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(w), \Phi_A^{(s-\sigma)}(B(v))\right]\right\|_{H^s} \\ &+ \left\|dB\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(dB(v)[w])\right]\right\|_{H^s}\end{aligned}$$

where we have redefined

$$v = \Phi_A^{(\sigma-\tau)}(u(\tau)) \quad \text{and} \quad w = \Phi_A^{(\sigma-\tau)}(B(u(\tau))).$$

Using the same lemmas and techniques as before, the estimation of the first term goes as follows

$$\begin{aligned}
&\left\| d^2B\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(w), \Phi_A^{(s-\sigma)}(B(v))\right]\right\|_{H^s} \\
&\quad = \left\| \left((\Phi_A^{(s-\sigma)}(w))(\Phi_A^{(s-\sigma)}(B(v))) \right)_x \right\|_{H^s} \leq \|(\Phi_A^{(s-\sigma)}(w))(\Phi_A^{(s-\sigma)}(B(v)))\|_{H^{s+1}} \\
&\quad \leq C\|\Phi_A^{(s-\sigma)}(w)\|_{H^{s+1}} \|\Phi_A^{(s-\sigma)}(B(v))\|_{H^{s+1}} \leq C\|w\|_{H^{s+1}} \|B(v)\|_{H^{s+1}} \\
&\quad = C\|w\|_{H^{s+1}} \|vv_x\|_{H^{s+1}} \leq C\|w\|_{H^{s+1}} \|v\|_{H^{s+1}} \|v\|_{H^{s+2}}.
\end{aligned}$$

By Lemma 5.3 and Δt sufficiently small, the above is bounded by

$$\left\| d^2B\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(w), \Phi_A^{(s-\sigma)}(B(v))\right]\right\|_{H^s} \leq C\|u_0\|_{H^p}^3. \tag{5.51}$$

The second term is estimated by,

$$\begin{aligned}
\left\| dB\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(dB(v)[w])\right]\right\|_{H^s} &\leq \left\| \left((\Phi_A^{(s-\sigma)}(v))(\Phi_A^{(s-\sigma)}(dB(v)[w])) \right) \right\|_{H^{s+1}} \\
&\leq C\|v\|_{H^{s+1}} \|dB(v)[w]\|_{H^{s+1}} \\
&= C\|v\|_{H^{s+1}} \|(vw)_x\|_{H^{s+1}} \\
&\leq C\|v\|_{H^{s+1}} \|v\|_{H^{s+2}} \|w\|_{H^{s+2}},
\end{aligned}$$

and again by Lemma 5.3 this is bounded as

$$\left\| dB\left(\Phi_A^{(s-\sigma)}(v)\right)\left[\Phi_A^{(s-\sigma)}(dB(v)[w])\right]\right\|_{H^s} \leq C\|u_0\|_{H^p}^3. \tag{5.52}$$

Thus, we get, using (5.51) and (5.52),

$$\|e_{S,5}\|_{H^s} \leq \int_0^{\Delta t} \int_0^s \int_0^\sigma C(\|u_0\|_{H^p}^3 + \|u_0\|_{H^p}^4)\, d\tau\, d\sigma\, ds \leq C(\Delta t)^3. \tag{5.53}$$

We are about to conclude the proof. Using (5.41), (5.47), (5.50) and (5.53) we get for the local error in (5.40),

$$\|u_1 - u(\Delta t)\|_{H^s} \leq c_2(\Delta t)^3,$$

where c_2 only depends on $\|u_0\|_{H^p}$. This completes the proof. □

5.7.2 Global Error

To estimate the global error, and obtain the correct second order convergence rate for the Strang splitting (5.3), we use the estimate for the local error in Lemma 5.10. As was the case for the Godunov splitting, the result in the lemma relies on that the solution at each step is bounded in $H^p(\mathbb{R})$. To be precise, we need to show that the operator splitting solution for the Strang splitting at each step is bounded. This is done the same way as we did in the proof of Theorem 5.9, and therefore we drop the proof.

Theorem 5.11 (Global error in $H^s(\mathbb{R})$)**.** *Assume there exists a solution of* (5.4) *and define s and p by* (5.10)*. If Hypothesis 5.1 holds for $k = s$ and Hypothesis 5.2 holds for $k = p$, then there is a $\overline{\Delta t} > 0$ such that for all $\Delta t \leq \overline{\Delta t}$ and $t_n = n\Delta t \leq T$,*

$$\|u_n - u(t_n)\|_{H^s} \leq C\,(\Delta t)^2,$$

where u_n is the Strang splitting solution (5.3)*, and $\overline{\Delta t}$ and C only depends on $\|u_0\|_{H^p}, \rho$ and T.*

Proof. We get, following the same outline as in the proof of Theorem 5.9

$$\begin{aligned}\|u_n - u(t_n)\|_{H^s} &\leq \sum_{k=0}^{n-1} \left\|\Phi^{(n-k-1)\Delta t}(u_{k+1}) - \Phi^{(n-k)\Delta t}(u_k)\right\|_{H^s} \\ &= \sum_{k=0}^{n-1} \|\Phi^{(n-k-1)\Delta t}\left(\Psi^{\Delta t}(u_k)\right) - \Phi^{(n-k-1)\Delta t}\left(\Phi^{\Delta t}(u_k)\right)\|_{H^s}.\end{aligned}$$

Using the Lipschitz continuity (5.9) in Hypothesis 5.1, yields

$$\|u_n - u(t_n)\|_{H^s} \leq \sum_{k=0}^{n-1} K(R,T)\,\|\Psi^{\Delta t}(u_k) - \Phi^{\Delta t}(u_k)\|_{H^s}.$$

Because of the bound for $\|u_n\|_{H^p}$ at each step, Lemma 5.10 yields for every $k \leq n-1$ and we get, using $n\Delta t \leq T$,

$$\begin{aligned}\|u_n - u(t_n)\|_{H^s} &\leq \sum_{k=0}^{n-1} K(R,T)c_3(C_0)(\Delta t)^2 \leq K(R,T)c_3(C_0)n(\Delta t)^3 \\ &\leq K(R,T)c_3(C_0)T(\Delta t)^2 \leq C\,(\Delta t)^2,\end{aligned}$$

which proves the theorem. □

We state the convergence result for $H^q(\mathbb{R})$ and refer to [11] for a proof, which is very similar as those presented so far.

Theorem 5.12 (Global error in $H^q(\mathbb{R})$)**.** *Assume there exists a solution of* (5.4) *and define q and p by* (5.10)*. If Hypothesis 5.1 holds for $k = q$ and Hypothesis 5.2 holds for $k = p$, then there is a $\overline{\Delta t} > 0$ such that for all $\Delta t \leq \overline{\Delta t}$ and $t_n = n\Delta t \leq T$,*

$$\|u_n - u(t_n)\|_{H^q} \leq C\,\Delta t,$$

where u_n is the Strang splitting solution (5.3)*, and $\overline{\Delta t}$ and C only depends on $\|u_0\|_{H^p}, \rho$ and T.*

5.8 Comments

From the convergence proofs we observe that there are two differences between the Godunov splitting (5.2) and the Strang splitting (5.3). The first difference is what kind of quadrature rule that is used to get Peano kernel bounds, and the second difference is the order of the series expansions. This gives naturally simpler proofs for (5.2) compared with those for (5.3). The framework used in this text should be applicable to other partial differential equations, as long as regularity results as those in Sections 5.4 and 5.5 are present.

For (5.2) similar convergence results are obtained by switching the order of the operators A and B, as long as regularity results for (5.5) similar to those for (5.6) exist. For (5.3) there is not possible to switch the order of A and B, by the complexity of (5.6).

For the viscous Burgers' equation (5.7), Theorem 5.9 yields first order convergence for (5.2) in $H^s(\mathbb{R})$, for initial data in $H^{s+2}(\mathbb{R})$, for $s \geq 1$. Using the framework in [13], one can show first order convergence in $H^{s+1}(\mathbb{R})$ for $s \geq 1$ with initial data in $H^{s+3}(\mathbb{R})$. Thus, Theorem 5.9 prove correct convergence for the case $s = 1$, and in that sense it is a improvement, but for $s \geq 2$ there is no improvements.

For the KdV equation (5.8), Theorem 5.9 yields first order convergence in $H^s(\mathbb{R})$ for initial data in $H^{s+3}(\mathbb{R})$, for $s \geq 1$. Compared with [13, Thm. 2.4.], where first order convergence is obtained in $H^s(\mathbb{R})$ for $s \geq 2$ with initial data in $H^{s+3}(\mathbb{R})$, Theorem 5.9 yields convergence in $s = 1$, and this yield also a small improvements compared to [13, Thm. 2.4.]. Thus, the results in this text is a slightly improvement of a allready known result.

In [13, Thm. 3.5.], it is proved second order convergence for the Strang splitting (5.3) in $H^s(\mathbb{R})$ for $s \geq 8$, for initial data in $H^{s+9}(\mathbb{R})$ for the KdV equation. From Theorem 5.11 we observe that second order convergence is obtained in $H^s(\mathbb{R})$ for $s \geq 1$, with initial data in $H^{s+5}(\mathbb{R})$. Hence, Theorem 5.11 is a significantly improvement of the result in [13] for the KdV equation.

For the whole class of equations in (5.4), Theorems 5.9 and 5.11 yield general results for the operator splitting methods, and we have not found similar results in the literature.

As mentioned above, the results in [13] are obtained using another new framework, and it is natural to compare the framework from [13] and that in [11], which is used in this text. The approach in [13] introduces a two dimensional extension in time which is defined such that the evolution corresponding to each time variable is governed by one of the split operators (in our context, the operators in (5.5) and (5.6)). This extension, together with a bootstrap argument, are used to show convergence results for (5.8). This approach is complicated and gives long and tedious estimations of the subequations from the splitting. Comparing the proofs in this text with those in [13], it turns out that those in this text is easier and more elegant. Moreover, since the theoretical results using [11] are equal or better for (5.2) and (5.3) compared with those in [13], the framework in [11] should be favourable.

6 Numerical Experimentation

In this section we numerically investigate the operator splitting method of Godunov and Strang types, given in (5.2) and (5.3), respectively. Since we have derived theoretical results for these two splitting methods for the split step size Δt, we naturally have the numerical convergence rates for Δt as the main focus throughout this entire section. However, we will also investigate other aspects of the splitting methods. The theoretical results are valid for several equations, but we only study two equations in details; that is, we use the viscous Burgers' equation (5.7) and the Korteweg–de Vries (KdV) equation (5.8) as test equations.

In the numerical experimentation for the two equations, we consider the CPU runtimes and the accuracy of the two errors, in addition with the convergence rates for Δt and Δx. We also compare the two splitting methods, to find the one which is the best from a computational point of view. The outline of the testing for the two equations is as follows: We test the operator splitting method and a comparison method on a test problem which have an exact solution. The exact solution yield a safe environment to study all the different aspects of the two operator splitting solutions and the comparison method. We implement different solvers for the subequations from the splitting process, and test them to find a combination which works best for the operator splitting methods for the exact case. When all the testing for the exact case is done, we are hopefully left with numerical methods which we can use to construct some interesting non-exact solutions. We use the non-exact test problems to study if the numerical methods model the solution in the way we expect from the physical interpretation of the equations. In addition, we check the numerical convergence rates for Δt for the non-exact test cases for the operator splitting methods. Discussions and conclusions follows after the testing, for both equations.

Recall that when we apply the operator splitting method on (5.7), we obtain the two subequations

$$v_t + vv_x = 0, \tag{6.1}$$

$$w_t = w_{xx}, \tag{6.2}$$

which are solved subsequently for small time steps Δt using either (5.2) or (5.3). For (5.8) we obtain the two subequations,

$$\begin{aligned} v_t + vv_x &= 0, \\ w_t + w_{xxx} &= 0. \end{aligned} \tag{6.3}$$

We observe that both equations have the inviscid Burgers' equation (6.1) as the first subequation, while (5.7) have the diffusion equation (6.2) as the second subequation, and (5.8) the Airy equation (6.3).

When we compare (5.7) and (5.8) there is no doubts that (5.8) is a much more interesting equation, both from a physical and a numerical point of view. Both equations are transport equations and models wave translation, and in that sense two equations which

are somewhat very similar. However, the difference in the order of the highest derivative is the crucial difference, which results in the very different wave types. The waves which (5.7) describes, is just like standard waves without any special characteristics. On the other hand, the solitons which (5.8) models have many fascinating and interesting characteristics which the numerical methods should manage to preserve. The most difficult property which a numerical method should be able to model, is the two-soliton interaction problem. It is well-known that methods which work well for one-soliton solutions, necessarily not work well for the two-soliton interaction problem. Hopefully will our methods be able to approximate this phenomenon. Thus, we study (5.8) in more details than (5.7).

The observant reader will note that we in this section interchange the sign for the some of the terms in (5.7) and (5.8). For the nonlinear term uu_x and u_{xxx} this is just a matter of the wave travels to either left or right. For the diffusive term u_{xx} we are not able to change the sign, since this leads to an illposed problem.

This section is divided as follows: We introduce different numerical solution methods for the subequations in (6.1), (6.2) and (6.3). This is followed up by applying the two abovementioned operator splitting methods to (5.7) and (5.8).

The numerical methods are implemented using *Matlab 7.10.0.499 (R2010a)* 64-bit version, running on the operating system *Ubuntu 10.10 – Maverick Meerkat*. All the figures in this section are produced using built-in plotting functions in Matlab.

6.1 The Inviscid Burgers' Equation

The inviscid Burgers' equation (6.1) is a conservation law in it simplest form. The development of conservation laws is out of the scope of this text, but we give a brief introduction since this is necessary for understanding how the numerical methods are constructed. We refer to [23, Ch.2 & Ch.11.] for a brilliant introduction to conservation laws in general form, both from a mathematical and physical view. See also [21, Ch. 2.] and [14, Ch. 1.].

A conservation law (in one spatial dimension) is in its most general form given as

$$\frac{d}{dt}\int_{x_1}^{x_2} u(x,t)\,dx = f\left(u(x_1,t)\right) - f\left(u(x_2,t)\right), \tag{6.4}$$

where $[x_1, x_2]$ is an intervall on the spatial axis. By imposing regularity assumptions on u and f, the integral can be reduced to the partial differential equation

$$u_t + f(u)_x = 0, \tag{6.5}$$

where $f(u)$ is the so-called flux function. For (6.1) is $f(u) = u^2/2$.

All smooth solutions to the conservation law satisfies both (6.4) and (6.5), while solutions with discontinuities do not fulfill (6.5) in classical sense, but they satisfy (6.4). Therefore, from a numerical point of view, it is natural to develop a method which fulfill the integral form instead of the differential equation. Methods grounded on this idea are called *finite volume methods*.

To form a numerical method, divide $[x_1, x_2]$ into subintervals, which are the so-called *grid cells* (or finite volumes). We put $x_i = i\Delta x$ and $t_n = n\Delta t$, and consider the ith grid cell $(x_{i-1/2}, x_{i+1/2})$. Integrating (6.4) in time from t_n to t_{n+1}, yield

$$\int_{x_{i-1/2}}^{x_{i+1/2}} u(x, t_{n+1})\, dx - \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, t_n)\, dx = \left(\int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}, t))\, dt - \int_{t_n}^{t_{n+1}} f(u(x_{i-1/2}, t))\, dt \right).$$

We define the ith cell average as

$$u_i^n = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, t_n)\, dx,$$

where Δx is the step size on the spatial axis. Using this turns the above equation into

$$u_i^{n+1} = u_i^n - \frac{1}{\Delta x} \left(\int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}, t))\, dt - \int_{t_n}^{t_{n+1}} f(u(x_{i-1/2}, t))\, dt \right).$$

Introducing the flux average over each cell,

$$F_{i\pm1/2}^n \approx \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} f\left(u(x_{i\pm1/2}, t)\right)\, dt,$$

where Δt is the step size in time. Thus, an approximation to the conservation law (6.4) is given as

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} \left(F_{i+1/2}^n - F_{i-1/2}^n \right). \tag{6.6}$$

To form a numerical method the above expressions are replaced by numerical approximations. Dependent on how this is done, leads to different numerical methods, which are presented next. In what follows, capital letters are used to indicate the numerical methods, for instance $U_i^n = u(x_i, t_n)$.

6.1.1 The Lax–Friedrichs Method

The Lax[29]–Friedrichs[30] formula is first order accurate, and is based on central differencing the terms in (6.6). This gives

$$U_i^{n+1} = \frac{1}{2}(U_{i-1}^n + U_{i+1}^n) - \frac{\Delta t}{2\Delta x} \left(f(U_{i+1}^n) - f(U_{i-1}^n) \right).$$

The Lax–Friedrichs method is dissipative, which results in that discontinuities in the solutions are smoothed out. The amount of smoothing is dependent on the ratio between Δx and Δt. In what follows, this method is labeled LxF.

[29] Peter David Lax, 1 May 1926 – , American mathematican. One of present days most known mathematicans. Have made important contributions to pure and applied mathematics, i.e. solutions to partial differential equations. Received the Abel prize in 2005.

[30] Kurt Otto Friedrichs, 28 September 28 1901 – 31 December 1982, German mathematician. Known for the Lax–Friedrichs method. Was the co-founder of the Courant Institute at New York University and recipient of the National Medal of Science.

6.1.2 The Lax–Wendroff Method

The two-step Lax–Wendroff[31] method is an example of a two-step method, which is of second order. The method uses the first step, which is very similar to a Lax-Friedrichs step, to approximate the cell average. This approximation is then used to perform a full step. The method is given as

$$U_{i+1/2}^{n+1/2} = \frac{1}{2}(U_i^n + U_{i+1}^n) - \frac{\Delta t}{2\Delta x}\left(f(U_{i+1}^n) - f(U_i^n)\right),$$
$$U_i^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}\left(f(U_{i+1/2}^{n+1/2}) - f(U_{i-1/2}^{n+1/2})\right).$$

The method approximates the discontinuities with steeper slopes than LxF, but it introduces oscillations near the discontinuities. We label this method as LxW.

6.1.3 The MacCormack's Method

Another two-step method is the MacCormack[32] method. This method uses first forward differencing, followed up by backward differencing to achieve second order accuracy,

$$U_i^* = U_i^n - \frac{\Delta t}{\Delta x}\left(f(U_{i+1}^n) - f(U_i^n)\right),$$
$$U_i^{n+1} = \frac{1}{2}(U_i^n + U_i^*) - \frac{\Delta t}{2\Delta x}\left(f(U_i^*) - f(U_{i-1}^*)\right).$$

The method is similar to LxW, and it introduces oscillations near discontinuities. This method is labeled as McC, in what follows.

6.1.4 The Nessyahu–Tadmor Method

The Nessyahu[33]–Tadmor[34] (NT) method is a high resolution method, which is based on central schemes, staggered grids and limiters. This method is grounded on the *reconstruct-evolve-average* (REA) algorithm, see [23]. We present this method very shortly, and refer to [23, Ch. 10.] and [12, App. A.] for a deeper presentation. The idea with the method is to combine the best features of first and second order methods. First order methods are known to approximate shocks well, while second order methods are better for smooth solutions. The NT method use so-called limiters to control the

[31] Burton Wendroff, 10 March 1930 – , American mathematician. Known for his contributions to the development of numerical methods for the solution of hyperbolic partial differential equations. The Lax–Wendroff method is named after Peter Lax and him. Lax was his supervisor during his Ph.D..

[32] Robert W. MacCormack. Introduced the method in "*The Effect of viscosity in hypervelocity impact cratering*" in 1969.

[33] Haim Nessyahu, 21 June 1964 – April 1994, Israeli mathematician. Submitted his Ph.D. thesis in 1994 and was scheduled to begin a post-doctoral position at UCLA in Fall 1994. Died while trekking in Nepal in 1994.

[34] Eitan Tadmor, born 1954, Israeli mathematician. One of the most active and influential mathematicians in the area of numerical analysis, general theory of applied PDEs, and scientific computing. Founder of the spectral viscous method.

amount of the two different methods, in different regions of the solution. In particular, in regions near discontinuities a first order scheme is used, while in smooth regions a second order scheme is used. The method approximates the solution in the center of each grid cell, and is given as

$$U_i^{n+1/2} = U_i^n - \frac{\Delta t}{2\Delta x}\varphi\left(f(U_i^n) - f(U_{i-1}^n), f(U_{i+1}^n) - f(U_i^n)\right),$$
$$U_{i+1/2}^{n+1} = \frac{1}{2}(U_i^n + U_{i+1}^n) - \frac{\Delta t}{\Delta x}(g_{i+1}^n - g_i^n),$$

where

$$g_i^n = f(U_i^{n+1/2}) + \frac{\Delta x}{8\Delta t}\varphi\left(U_i^n - U_{i-1}^n, U_{i+1}^n - U_i^n\right),$$

where φ is the limiter function. There exists several limiters, but we consider only four limiters; *minmod, MC, Van-Leer*[35] and *superbee*. The minmod limiter is given as

$$minmod(a,b) = \frac{1}{2}\left(\mathrm{sgn}(a) + \mathrm{sgn}(b)\right)\min(|a|,|b|),$$

and the MC limiter is given as

$$MC(a,b) = minmod\left(\frac{a+b}{2}, 2\,minmod(a,b)\right).$$

The Van-Leer limiter is given as

$$vanleer(a,b) = \frac{ab(\mathrm{sgn}(a) + \mathrm{sgn}(b))}{a+b},$$

and at last the superbee limiter,

$$superbee(a,b) = maxmod(minmod(a,2b), minmod(2a,b)),$$

where

$$maxmod(a,b) = \frac{1}{2}\left(\mathrm{sgn}(a) + \mathrm{sgn}(b)\right)\max(|a|,|b|),$$

In what follows, we label this method as NT*, where * indicates the limiter function in use (Mm=minmod, Mc=MC, Vl=Van-Leer, Sb=superbee). For example, NTMc uses the MC limiter function.

We mention briefly the stability conditions for the numerical methods considered so far. LxF, LxW and McC are all stable under the CFL restriction

$$\frac{\Delta t}{\Delta x}\max_u |f'(u)| \leq 1. \tag{6.7}$$

[35]Bram van Leer, Dutch mathematician. Have made substantial contributions to computational fluid dynamics, fluid dynamics, and numerical analysis.

while the NT method is stable under

$$\frac{\Delta t}{\Delta x}\max_u |f'(u)| \leq \frac{1}{2}. \tag{6.8}$$

When these methods are implemented, we add a pessimistic factor in the range $(0, 1]$ for LxF, LxW and McC, and in the range $(0, 1/2]$ for NT*, which ensures that (6.7) and (6.8) are fulfilled. In what follows, this pessimistic factor is refered to only as the CFL factor. Moreover, we implement the methods such that when solving the conservation law numerically, the time steps are chosen using the CFL restriction. In such a way only a necessary amount of time steps are performed.

To illustrate the methods, we solve the following problem

$$u_t + uu_x = 0, \quad u(x,0) = \begin{cases} 1 & \text{if } x \leq 0.5, \\ 0 & \text{if } x \geq 0.5, \end{cases} \tag{6.9}$$

which has the exact solution

$$u(x,t) = \begin{cases} 1 & \text{if } x \leq 0.5(1+t), \\ 0 & \text{if } x \geq 0.5(1+t). \end{cases}$$

From Figure 6.1, we observe the dissipation introduced by LxF, and the oscillations from LxW and McC. NTMc approximates the discontinuity rather good, without oscillations.

6.1.5 The Spectral Viscosity Method

The last numerical method for (6.5), is grounded on a total different approach than the methods considered so far. The spectral viscosity (SV) method was first described in [28], and investigated further in [24]. In [8], a full method involving discontinuity detection and postprocessing is studied. The paper yields a very good numerical method, though a bit complicated. See also [29] and the references given therein. We present the method for periodic boundary conditions, see [29] for a treatment of more general boundary conditions.

The foundation for the spectral viscosity method is the Fourier transform. Consider the truncated Fourier expansion of the solution $u(x,t)$, which reads

$$P_N u(x,t) = \sum_{k=-N}^{N} \hat{u}(k,t) e^{ikx},$$

where P_N is the Fourier series operator. Inserting the series expansion into the conservation law (6.5) gives

$$(P_N u)_t + (f(P_N u))_x = 0. \tag{6.10}$$

By a standard spectral method, the above equation could have been transformed to the Fourier space. However, it is well-known that a standard spectral methods introduces

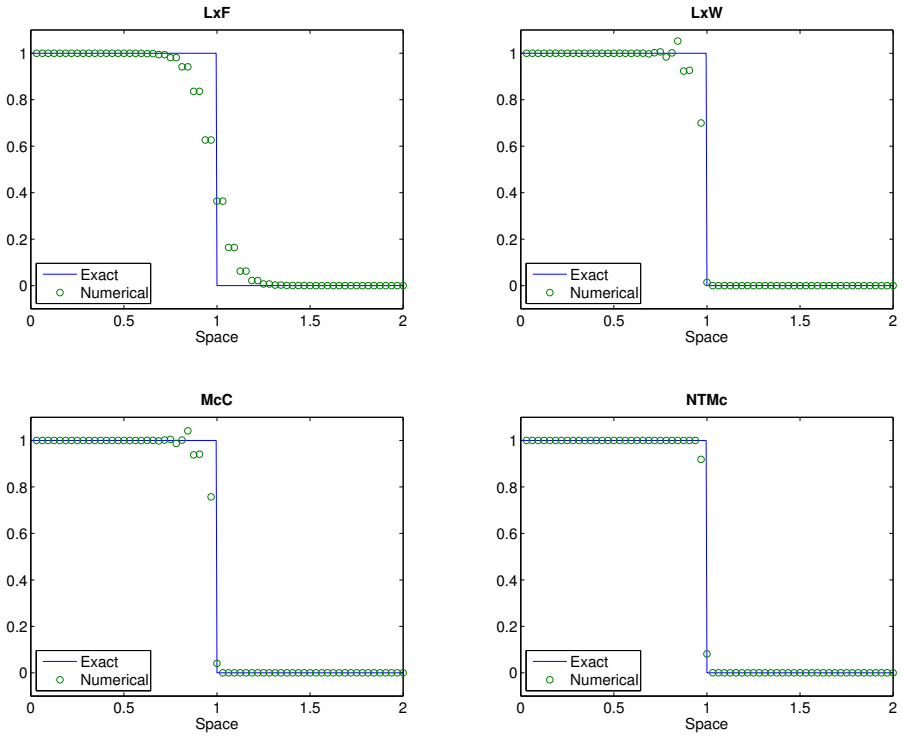


Figure 6.1: The numerical solution at time $t = 1.0$ for (6.9), using four different numerical methods. The spatial axis is discretized with 64 nodes. The dissipative behaviour for LxF and the oscillations for LxW and McC are good illustrated. The more intelligent NTMc approximates the discontinuity well, without introducing oscillations.

the Gibbs[36] phenomenon once discontinuities are about to appear in the solution. That is, local oscillations appear near the discontinuities in the series solution, which results in an unstable numerical method.

It turns out that the Gibbs oscillations appears in the high frequency domain of the solution. The elegant idea with the SV method is to modify (6.10), and introduce viscosity (or diffusion) which is carried out in the high frequency domain. This result in smoothing of the spectrum of the solutions, which hopefully reduce the oscillations. The modified version of (6.10) is given as

$$(P_N u)_t + (f(P_N u))_x = \kappa_N \left(Q_N(x,t) * (P_N u)_x\right)_x ,$$

[36] Josiah Willard Gibbs, 11 February 1839 – 28 April 1903, American physicist and mathematician. Made contributions to the vector analysis in mathematics, and studied thermodynamics. Gibbs' phase rule, Gibbs' free energy and Gibbs phenomenon, in addition with other rules and laws, are named after him.

where κ_N is the viscosity coefficient and is defined as

$$\kappa_N = \frac{1}{\sqrt{N}}.$$

Transforming into the Fourier space, and "removing" the spatial derivatives by the properties of the Fourier transform, gives a system of ordinary differential equations,

$$(\hat{u}_k)_t + 2\pi i k \hat{f}_k = -\kappa_N (2\pi k)^2 \widehat{Q}_{N_k} \hat{u}_k, \tag{6.11}$$

where $\hat{f}_k = (1/2)\hat{u}_k^2$. $\widehat{Q}_{N_k}$ is constructed so that $\widehat{Q}_{N_k} = 0$ for $|k| \ll N$ and $\widehat{Q}_{N_k} = 1$ for $|k| \approx N$. In [28], good results was obtained when $\widehat{Q}_{N_k}$ was smoothly varied between 0 and 1. We adopt this idea, and use

$$\widehat{Q}_{N_k} = \begin{cases} 0 & \text{for } k \leq \sqrt{N}, \\ \frac{1}{2}\left(1 + \tanh\left(\frac{1}{2}(k - \sqrt{N})\right)\right) & \text{otherwise,} \end{cases}$$

which is equal to the expression used in [15]. The system in (6.11) is solved using a standard forward in time, explicit scheme,

$$\widehat{U}_k^{n+1} = \widehat{U}_k^n - \Delta t (2\pi i k \hat{f}_k^n + \kappa_N (2\pi k)^2 \widehat{Q}_{N_k} \widehat{U}_k^n). \tag{6.12}$$

When the SV method is implemented, the famous *Fast Fourier Transform* (FFT) algorithm is used to transform the solution into the Fourier space.

To illustrate the SV method, we solve

$$u_t + u u_x = 0, \quad u_0(x) = -\sin(x), \quad x \in [-\pi, \pi], \tag{6.13}$$

to $t = 1.5$, using a standard spectral method and the SV method. At this point in time, a shock is about to be constructed, so this is a good example to illustrate the differences in the methods. The numerical solutions are given in Figure 6.2.

From Figure 6.2, we observe that the standard spectral method introduces oscillations, while the SV method reduce these oscillations. From the spectrum plots of the two solution methods, we see that the SV method reduces the high frequency modes. We note that for smooth solutions of (6.13), both methods approximate the solution without introducing oscillations.

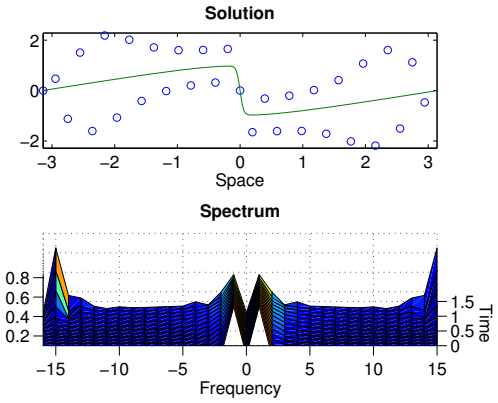
6.2 The Diffusion Equation

To solve (6.2) numerically, we use the well-known Crank[37]–Nicolson[38] difference scheme. This scheme uses central differences for the spatial derivatives in two points in time, and a forward difference formula for the time derivative. This yield

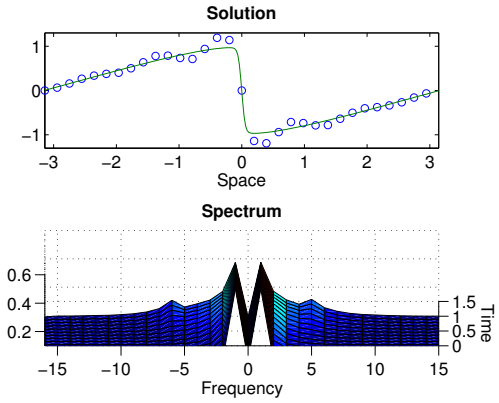
$$\frac{U_i^{n+1} - U_i^n}{\Delta t} = \frac{\kappa}{2\,(\Delta x)^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}),$$

[37] John Crank, 6 February 1916 – 3 October 2006, British physicist. Known for his work on numerical solutions of partial differential equations.

[38] Phyllis Nicolson, 21 September 1917 – 6 October 1968, British mathematician. Known for her work on the Crank–Nicolson scheme.



(a) Standard spectral method



(b) SV method

Figure 6.2: The numerical solution of (6.13) at $t = 1.5$ using a standard spectral method and the SV method. The green lines are the reference solution. The spatial axis is discretized with 32 nodes. The lower images show the spectrum evolutions in time. The standard spectral method introduces oscillations near the arising discontinuity, while SV introduces small oscillations, which is a result of the damping of the high frequency modes in the spectrum.

which can be rewritten as

$$-rU_{i-1}^{n+1} + (1+2r)U_i^{n+1} - rU_{i+1}^{n+1} = rU_{i-1}^{n+1} + (1-2r)U_i^n + rU_{i+1}^n,$$

where $r = \Delta t/(\Delta x)^2$. This method is implicit and stable when Δt is chosen such that $\Delta t \approx \Delta x/\kappa$. For a more detailed discussion about this method and other methods for (6.2) (and more general parabolic problems), see [22, Ch. 9.].

6.3 The Airy Equation

For (6.3), we introduce two methods: A Crank–Nicolson like implicit difference scheme and a spectral method.

6.3.1 The Difference Scheme

Using a central difference for the spatial derivative, and a forward difference in time gives

$$\begin{aligned}\frac{U_i^{n+1} - U_i^n}{\Delta t} = \frac{\kappa}{4\,(\Delta x)^3}(&U_{i-2}^{n+1} - 2U_{i+1}^{n+1} + 2U_{i-1}^{n+1} - U_{i-2}^{n+1} \\ &+ U_{i-2}^n - 2U_{i+1}^n + 2U_{i-1}^n - U_{i-2}^n),\end{aligned}$$

which is solved for U^{n+1},

$$\begin{aligned}-rU_{i-2}^{n+1} - 2rU_{i+1}^{n+1} + U_i^{n+1} - 2rU_{i-1}^{n+1} + rU_{i-2}^{n+1} = \\ rU_{i-2}^n - 2rU_{i+1}^n + U_i^n + 2rU_{i-1}^n - rU_{i-2}^n),\end{aligned} \tag{6.14}$$

where $r = (\kappa\Delta t)/(4(\Delta x)^3)$. This method is implicit, so we need to solve the linear system above at each time step. We do this by using an iteration method with an accuracy of 1e-06. We label this method as Diff.

6.3.2 The Spectral Method

We impose (6.3) with periodic boundary conditions. Transforming the equation into the Fourier space gives $\hat{u}_t = -i\kappa(2\pi\xi)^3\hat{u}$, which is solved to yield,

$$\hat{u}(\xi, t) = \hat{u}(\xi, 0)e^{-i\kappa(2\pi\xi)^3 t}.$$

In the discrete case, the following equations is the equivalence,

$$\hat{u}_k^n = \hat{u}_k^0\, e^{-\kappa i(2\pi k)^3\, n\Delta t}.$$

In the implementation FFT is used to calculate the discrete transform. We label this method as Spec.

6.4 The Viscous Burgers' Equation

To investigate the numerical splitting methods for the viscous Burgers' equation (5.7), we use an exact test problem and a non-exact test problem. We present the testing procedures and the results for the two problems in Sections 6.4.2 and 6.4.3. In Section 6.4.4 we discuss the results, while in Section 6.4.5 we give short conclusions.

To label the numerical operator splitting methods, we use the label for (6.1), since we always use the Crank–Nicolson method for (6.2). We let N_x be the number of nodes on the spatial axis, while N_t is the number of time steps used to solve to the end point in time.

6.4.1 A Full Difference Scheme

We compare the operator splitting methods for (5.7) with the following explicit full difference scheme, which is based on central differences for the spatial derivatives, and a forward difference for the time derivative,

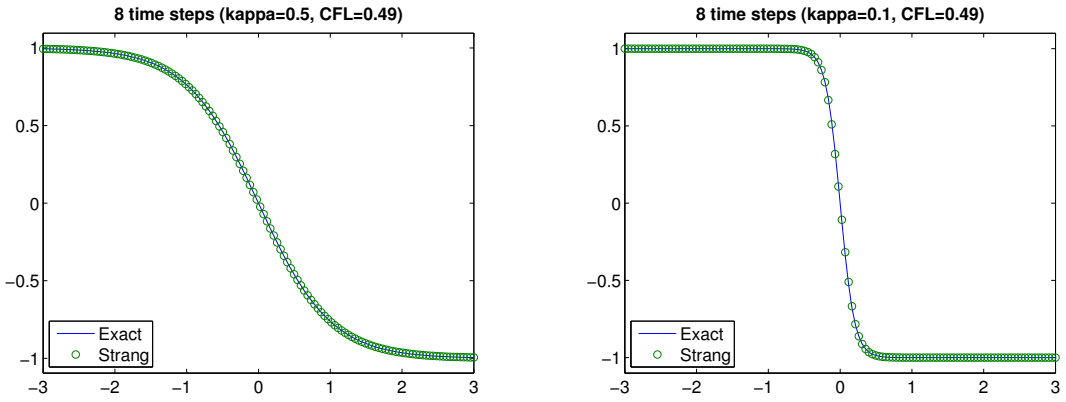
$$U_i^{n+1} = U_i^n - \frac{\Delta t}{2\,\Delta x}\left(f(U_{i+1}^n) - f(U_{i-1}^n)\right) + \frac{\kappa\,\Delta t}{(\Delta x)^2}\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right). \tag{6.15}$$

Since the method is explicit, we have to choose Δt and Δx carefully such that the method is stable. We label this method as DeSc.

6.4.2 The Exact Test Problem

As the exact test problem we use the following setup, which is time-independent and therefore is very suitable for testing purposes,

$$u_t + uu_x = \kappa u_{xx}, \qquad x \in [-3, 3], \tag{6.16}$$



(a) $t = 1.0, \kappa = 0.5$ (b) $t = 1.0, \kappa = 0.1$

Figure 6.3: The exact and numerical solutions of (6.16) using the Strang splitting (5.3) using 8 time steps and two values for κ. The spatial axis is discretized with 128 nodes. Observe that the steepness is dependent on κ.

where the initial condition at $t = 0$ is given from the exact solution

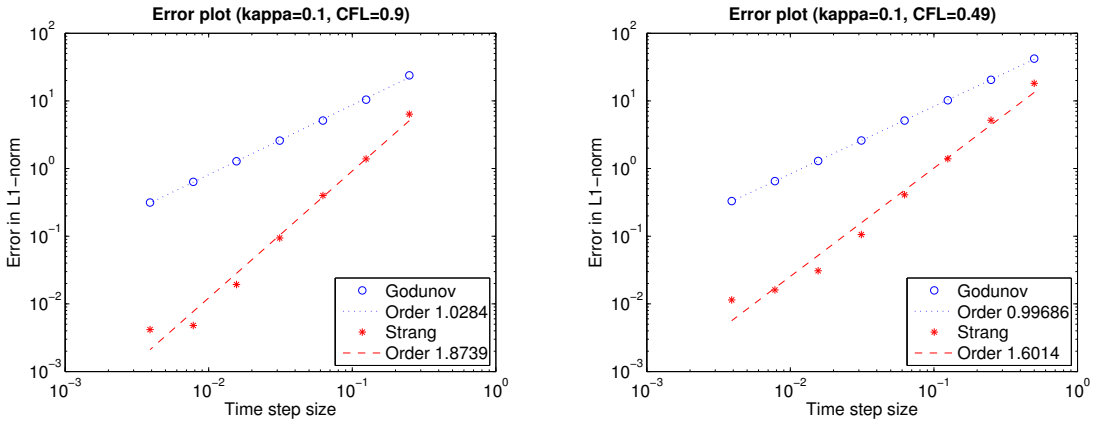
$$u(x,t) = -\tanh\left(\frac{x}{2\kappa}\right).$$

A plot of two different solutions for two values of κ is given in Figure 6.3.

We start the study with an investigation of DeSc given in (6.15). Since DeSc is explicit, we have to choose Δt very small compared to Δx, and the magnitude of κ plays also a role for the stability. We find that DeSc works well for (6.16) and it manages to produce correct solutions for all κ, as long as Δx and Δt are small enough.

We study the two operator splitting methods in more details, and start with the convergence rates for Δt. Test problem (6.16) is solved using $N_t = 2^1, 2^2, \ldots, 2^8$ to the end point $t = 1.0$. We let $N_x = 1024$ which gives $\Delta x = 0.0059$ for all the calculations. Due to the dissipative behaviour of LxF, we put $N_x = 4096$ when using LxF, that is $\Delta x = 0.0015$. We also vary the CFL pessimistic factor and κ to test if these two parameters have impact on the convergence rates for Δt. The errors are computed using the discrete L^1-, L^2- and L^∞-norms. We use these three norms to see if they are correlated in some way, or give different convergence rates. We use standard linear regression on logarithmic scales to obtain the numerical convergence rates.

The convergence rates for Δt for the Godunov splitting (5.2) and Strang splitting (5.3) are given in Table 6.1. During the testing we observe that some test cases for LxW and McC introduce oscillations. These oscillations result in *outliers* which give a regression line which do not reflect the overall linear trend for the remaining error points. Therefore, we remove the outliers before a new (and more correct) regression line is obtained. In the table these cases are marked with a star (*). For the NT schemes, we find that the CFL factor do not have significant influence on the convergence rates,



(a) LxW (b) NTMm

Figure 6.4: Numerical convergence rates using L^1-norm for Δt for the exact test problem (6.16). The rates are found using standard linear regression.

and to save typing place we present the convergence results for a CFL close to 0.5. A standard error plot is provided in Figure 6.4. We observe that all methods converge, with different numerical convergence rates. As a overall small conclusion (5.3) gives higher convergence rates, compared with (5.2).

To compare the size of the errors for the two splitting methods, we solve (6.16) to $t = 1.0$ for three values of κ, where we use $N_t = 128$ and $N_x = 2048$. For LxF, LxW, and McC we let the CFL factor be 0.9, while for the NT methods 0.49. The estimated errors in the three norms are given in Tables 6.2, 6.3 and 6.4. From the tables we observe that (5.3) is more accurate than (5.2). We observe also that LxF compared with all the other methods are less accurate in all norms. There is not much difference in the errors for the other methods.

The convergence rates for Δx are found by solving (6.16) using $N_t = 1024$ for different N_x. The numerical results for Δx are given in Table 6.5. From the table we observe that all methods converge, and the convergence rates seem to be dependent on κ. There is no trend between the norms, and overall they give different convergence rates.

To calculate the CPU runtimes, we solve (6.16) for $\kappa = 0.1$, $N_x = 1024$ to $t = 1.0$, to an accuracy to 1e-01 in the L^1-norm. To get more reliable results, we find an average time using 10 CPU runs. We also check the running times for DeSc. Due to stability problems with (6.15) we have to perform a large amount of time steps. The average runtimes are given in Table 6.6. We have given an error vs. CPU runtimes plot in Figure (6.5). To produce this plot we use $N_x = 2048$ and solve (6.15) for $N_t = 2^2, 2^3, \ldots, 2^8$. Due to the problems with dissipation using LxF, we leave it out in the CPU runtime calculations.

From Table 6.6 we observe that all of the splitting methods are faster than DeSc. We also observe that the (5.3) needs less steps to achieve same accuracy as (5.2), which naturally results in a faster runtime. The NT schemes are also a bit slower compared with the simpler LxW and McC methods. From Figure 6.5 we observe that all methods

			Godunov			*Strang*		
Method	κ	*CFL*	L^1	L^2	L^∞	L^1	L^2	L^∞
LxF	0.5	0.9	0.889	0.888	0.880	0.524	0.556	0.586
LxF	0.5	0.5	0.815	0.814	0.809	0.369	0.409	0.450
LxF	0.5	0.1	0.554	0.563	0.554	0.138	0.166	0.192
LxF	0.1	0.9	0.879	0.844	0.803	0.823	0.825	0.885
LxF	0.1	0.5	0.782	0.756	0.723	0.623	0.635	0.641
LxF	0.1	0.1	0.503	0.486	0.464	0.284	0.293	0.300
LxF	0.01	0.9	0.789	0.661	0.553	0.915	0.832	0.778
LxF	0.01	0.5	0.688	0.574	0.481	0.718	0.654	0.613
LxF	0.01	0.1	0.383	0.260	0.233	0.300	0.260	0.226
LxW	0.5	0.9	0.980	0.992	0.988	0.952	1.008	0.885
LxW	0.5	0.5	0.980	0.992	0.988	0.952	1.008	0.887
LxW	0.5	0.1	0.980	0.992	0.988	0.951	1.008	0.888
LxW	0.1	0.9	1.028*	1.012*	1.138*	1.874*	1.912*	2.067*
LxW	0.1	0.5	1.039*	1.036*	1.194*	1.866*	1.927*	2.088*
LxW	0.1	0.1	1.010*	0.993*	0.982*	1.776*	1.845*	1.892*
LxW	0.01	0.9	1.162*	1.132*	1.196*	1.508*	1.724*	1.939*
LxW	0.01	0.5	1.179*	1.141*	1.205*	1.775*	1.885*	2.040*
LxW	0.01	0.1	1.312*	1.271*	1.303*	2.313*	2.200*	2.177*
McC	0.5	0.9	0.980	0.991	0.987	0.944	1.005	0.886
McC	0.5	0.5	0.979	0.991	0.987	0.944	1.005	0.888
McC	0.5	0.1	0.979	0.991	0.987	0.944	1.006	0.888
McC	0.1	0.9	1.048	1.019	1.169	1.779	1.745	1.796
McC	0.1	0.5	1.056	1.028	1.186	1.797	1.771	1.817
McC	0.1	0.1	1.133	1.120	1.271	1.823	1.821	1.880
McC	0.01	0.9	0.956	0.833	0.817	1.347	1.235	1.110
McC	0.01	0.5	1.041	0.989	0.871	1.353	1.241	1.126
McC	0.01	0.1	1.019*	0.923*	1.049*	1.201*	1.265*	1.254*
NTMm	0.5	0.49	0.979	0.991	0.883	0.980	0.992	0.988
NTMm	0.1	0.49	0.997	0.960	0.966	1.601	1.636	1.761
NTMm	0.01	0.49	0.902	0.800	0.709	1.151	1.099	1.061
NTMc	0.5	0.49	0.949	1.006	0.884	0.980	0.992	0.988
NTMc	0.1	0.49	1.003	0.965	0.974	1.994	2.019	2.126
NTMc	0.01	0.49	0.950	0.837	0.733	1.530	1.439	1.327
NTVl	0.5	0.49	0.980	0.992	0.988	0.949	1.006	0.884
NTVl	0.1	0.49	1.002	0.965	0.973	1.974	2.003	2.109
NTVl	0.01	0.49	0.941	0.831	0.729	1.397	1.340	1.300
NTSb	0.5	0.49	0.956	1.010	0.884	0.981	0.993	0.988
NTSb	0.1	0.49	1.009	0.969	0.977	1.716	1.723	1.827
NTSb	0.01	0.49	1.012	0.880	1.137	1.281	1.201	1.137

Table 6.1: Numerical convergence rates for Δt for (6.16). (*) indicates that outliers are removed.

	Godunov			*Strang*		
Method	L^1	L^2	L^∞	L^1	L^2	L^∞
LxF	1.431e+00	4.167e-02	1.753e-03	6.689e-01	1.927e-02	8.019e-02
LxW	7.928e-01	2.279e-02	9.635e-04	2.793e-02	7.495e-04	6.922e-05
McC	7.932e-01	2.280e-02	9.640e-04	2.838e-02	7.540e-04	6.922e-05
NTMm	7.933e-01	2.280e-02	9.637e-04	2.841e-02	7.565e-04	7.080e-05
NTMc	7.930e-01	2.279e-02	9.636e-04	2.814e-02	7.528e-04	7.055e-05
NTVl	7.930e-01	2.279e-02	9.636e-04	2.813e-02	7.527e-04	7.063e-05
NTSb	7.927e-01	2.279e-02	9.634e-04	2.787e-02	7.491e-04	7.071e-05

Table 6.2: Estimated errors using L^1-, L^2- and L^∞-norm for exact test problem (6.16) with $\kappa = 0.5$, $N_t = 128$ and $N_x = 2048$ at $t = 1.0$.

	Godunov			*Strang*		
Method	L^1	L^2	L^∞	L^1	L^2	L^∞
LxF	2.433e+00	1.503e-01	1.338e-02	1.181e+00	7.224e-02	6.387e-03
LxW	1.283e+00	8.060e-02	7.300e-03	9.656e-03	8.044e-04	9.443e-05
McC	1.286e+00	8.084e-02	7.322e-03	1.366e-02	1.022e-03	1.194e-04
NTMm	1.288e+00	8.090e-02	7.320e-03	1.543e-02	1.018e-03	1.052e-04
NTMc	1.285e+00	8.073e-02	7.308e-03	1.214e-02	8.815e-04	9.812e-05
NTVl	1.285e+00	8.073e-02	7.308e-03	1.216e-02	8.817e-04	9.807e-05
NTSb	1.282e+00	8.056e-02	7.296e-03	9.070e-03	7.719e-04	9.213e-05

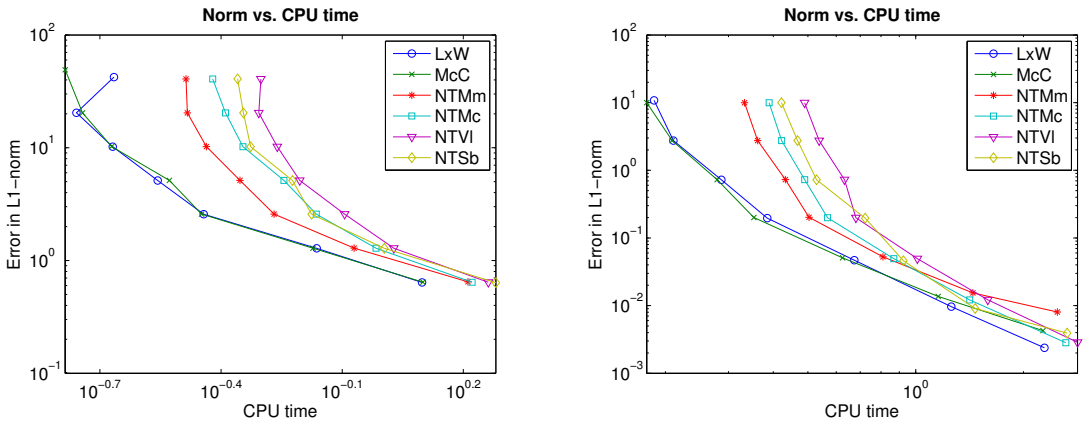
Table 6.3: Estimated errors using L^1-, L^2- and L^∞-norm for exact test problem (6.16) with $\kappa = 0.1$, $N_t = 128$ and $N_x = 2048$ at $t = 1.0$.

	Godunov			*Strang*		
Method	L^1	L^2	L^∞	L^1	L^2	L^∞
LxF	2.430e+00	4.444e-01	1.170e-01	1.341e+00	2.491e-01	6.660e-02
LxW	1.337e+00	2.592e-01	7.413e-02	9.935e-02	2.575e-02	9.276e-03
McC	1.372e+00	2.651e-01	7.592e-02	1.389e-01	3.206e-02	1.161e-02
NTMm	1.364e+00	2.629e-01	7.491e-02	1.270e-01	2.876e-02	9.858e-02
NTMc	1.395e+00	2.675e-01	7.557e-02	1.640e-01	3.362e-02	1.087e-02
NTVl	1.366e+00	2.632e-01	7.500e-02	1.295e-01	2.898e-02	9.941e-03
NTSb	1.332e+00	2.583e-01	7.329e-02	9.218e-02	2.471e-02	9.117e-03

Table 6.4: Estimated errors using L^1-, L^2- and L^∞-norm for exact test problem (6.16) with $\kappa = 0.01$, $N_t = 128$ and $N_x = 2048$ at $t = 1.0$.

		Godunov			*Strang*		
Method	κ	L^1	L^2	L^∞	L^1	L^2	L^∞
LxF	0.5	0.254	0.719	1.159	0.257	0.721	1.159
LxW	0.5	0.478	1.025	1.486	0.836	1.322	1.657
McC	0.5	0.422	0.994	1.524	0.780	1.335	1.734
NTMm	0.5	0.742	1.226	1.710	0.775	1.258	1.737
NTMc	0.5	1.315	1.859	2.332	1.806	2.273	2.493
NTVl	0.5	1.308	1.847	2.338	1.772	2.261	2.476
NTSb	0.5	1.110	1.655	2.176	1.446	1.948	2.276
LxF	0.1	0.203	0.447	0.682	0.205	0.449	0.684
LxW	0.1	1.174	1.584	1.911	0.992	1.388	1.724
McC	0.1	0.878	1.171	1.400	1.042	1.373	1.590
NTMm	0.1	0.273	0.569	0.844	0.279	0.574	0.850
NTMc	0.1	1.080	1.335	1.528	1.138	1.414	1.606
NTVl	0.1	0.889	1.170	1.436	0.939	1.224	1.495
NTSb	0.1	0.611	1.112	1.584	0.554	1.112	1.584
LxF	0.01	0.205	0.212	0.208	0.206	0.214	0.208
LxW	0.01	1.772	1.840	1.905	1.634	1.736	1.860
McC	0.01	0.715	0.698	0.712	0.951	0.975	0.948
NTMm	0.01	0.739	0.743	0.792	0.770	0.774	0.822
NTMc	0.01	0.988	1.076	1.161	1.160	1.279	1.354
NTVl	0.01	0.877	0.966	1.069	0.999	1.096	1.212
NTSb	0.01	0.557	0.775	0.935	0.509	0.722	0.879

Table 6.5: Numerical convergence rates for Δx for exact test problem (6.16).



(a) Godunov splitting

(b) Strang splitting

Figure 6.5: Plot of the error vs. CPU runtimes using logarithmic scales for (6.16).

	Godunov		*Strang*	
Method	*Time steps*	*Runtime*	*Time steps*	*Runtime*
Difference scheme*	5799	3.292	-	-
LxW	82	0.325	11	0.102
McC	82	0.331	11	0.102
NTMm	84	0.390	11	0.150
NTMc	83	0.418	11	0.172
NTVl	83	0.420	11	0.172
NTMc	82	0.428	11	0.181

Table 6.6: The CPU runtimes in seconds for computing the numerical solutions of (6.16). The runtimes are the average after 10 calculations with each method. (* The difference scheme is *not* an operator splitting method of Godunov type, but is placed there for typing reasons.).

gives smaller error for greater CPU times.

6.4.3 The Non-Exact Test Problem

Since all the methods worked well for (6.16), we use all of the numerical methods as on the non-exact test problem. We use the following setup,

$$u_t + uu_x = \kappa u_{xx}, \quad u_0 = u(x,0) = -\sin(\pi x) \cdot \chi_{[-1,1]}(x), \quad x \in [-1.5, 1.5], \tag{6.17}$$

where χ is the characteristic function, and we solve to $t = 1.0$.

We start with a qualitatively discussion on the evolution of the initial data. The initial data is a wave which travels to the right, which shape is dependent on the viscosity coefficient κ. If $\kappa = 0$ the wave will break relatively fast, which results in a creation of a shock. When $\kappa \neq 0$, as the wave begins to break, the viscosity term u_{xx} grows much faster than u_x, and at some point κu_{xx} begins to play a role for the solution. This term keeps the solution smooth for all times. If κ is small, the wave should travel to the right and get steeper and steeper, like a surface water wave breaking against a beach. For larger values for κ, the wave will be smoother, like deep water waves traveling on the surface, cf. Figures 1.1 and 6.6.

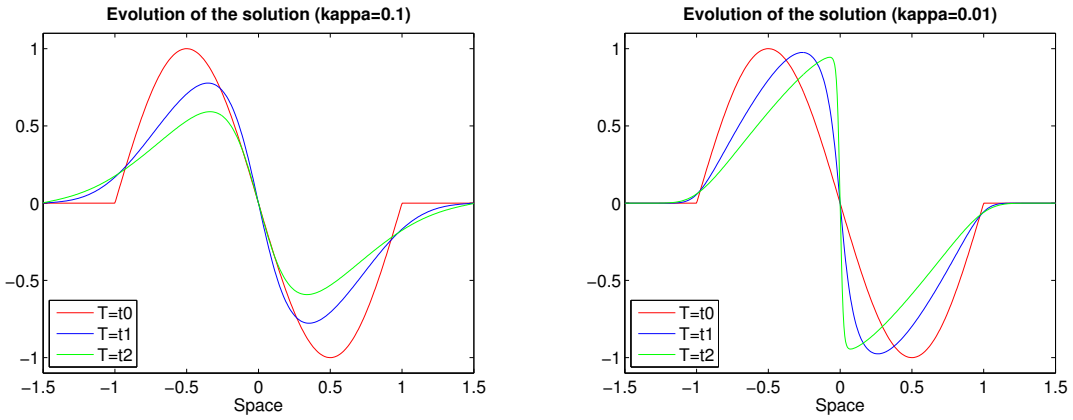
The numerical convergence rates for Δt are found similar as for (6.16) (we put $N_x = 1024$ for all method but LxF where $N_x = 4096$). Since no exact solution exist, we use a reference solution, which is a numerical calculated solution using a very fine grid, to find the estimated error. The results are given in Table 6.7, and we observe that (5.2) obtain numerical convergence results which is correct with the theoretical results. On the other hand, (5.3) obtains higher convergence rates, but not as high as the theoretical results.

For DeSc we find that the method works well for (6.17), and that it produces correct solutions.

The last thing to check is the robustness of the numerical methods and the two splitting approaches, when discontinuities are about to arise in the solutions. For (6.17) this happens when $\kappa \to 0$. At the limit, we are left with (6.1), which develops a shock

			Godunov			*Strang*		
Method	κ	*CFL*	L^1	L^2	L^∞	L^1	L^2	L^∞
LxF	0.1	0.9	0.962	0.927	0.898	0.734	0.773	0.800
LxF	0.1	0.5	0.904	0.614	0.641	0.578	0.614	0.641
LxF	0.1	0.1	0.622	0.607	0.590	0.265	0.289	0.307
LxF	0.01	0.9	0.765	0.746	0.709	0.791	0.948	1.000
LxF	0.01	0.5	0.669	0.682	0.607	0.650	0.796	0.816
LxF	0.01	0.1	0.333	0.418	0.366	0.270	0.371	0.370
LxW	0.1	0.9	1.013*	1.009*	0.991*	1.382*	1.305*	0.851*
LxW	0.1	0.5	1.014*	1.009*	0.991*	1.378*	1.305*	0.852*
LxW	0.1	0.1	1.014*	1.010*	0.991*	1.375*	1.305*	0.852*
LxW	0.01	0.9	1.068*	1.068*	1.123*	1.218*	1.445*	1.671*
LxW	0.01	0.5	1.311*	1.528*	1.733*	1.311*	1.528*	1.733*
LxW	0.01	0.1	1.069*	1.078*	1.178*	1.652*	1.760*	1.885*
McC	0.1	0.9	1.036	1.034	1.145	1.512	1.491	1.265
McC	0.1	0.5	1.044	1.042	1.167	1.528	1.514	1.305
McC	0.1	0.1	1.065	1.070	1.203	1.562	1.560	1.346
McC	0.01	0.9	1.068	0.971	0.942	1.646	1.586	1.529
McC	0.01	0.5	1.057	0.967	0.943	1.649	1.586	1.539
McC	0.01	0.1	1.437	1.286	1.156	1.655	1.593	1.563
NTMm	0.1	0.49	1.040	1.021	0.989	1.492	1.406	0.924
NTMm	0.01	0.49	0.969	0.884	0.832	1.501	1.522	1.538
NTMc	0.1	0.49	1.041	1.022	0.991	1.544	1.439	0.996
NTMc	0.01	0.49	0.977	0.893	0.842	1.754	1.736	1.743
NTVl	0.1	0.49	1.041	1.022	0.991	1.543	1.427	0.953
NTVl	0.01	0.49	0.977	0.892	0.840	1.729	1.713	1.723
NTSb	0.1	0.49	1.041	1.024	0.992	1.532	1.424	0.982
NTSb	0.01	0.49	0.986	0.900	0.847	1.627	1.625	1.581

Table 6.7: Numerical convergence rates for Δt for the operator splitting approximations of Godunov and Strang types for the non-exact test problem (6.17). A star (*) indicates that outliers has been removed.



(a) $\kappa = 0.1$ (b) $\kappa = 0.01$

Figure 6.6: The evolution of the solution of for the non-exact test problem (6.17) for different values of κ, where $t_0 < t_1 < t_2$. Observe that the solution never breaks, and develops a shock.

at $x = 0$. Thus, the solution becomes a sawtooth function in the limit. For the operator splitting methods we let $N_x = 256$ and $N_t = 128$, while we for DeSc have to put $N_t = 2900$ due to stability problems. We run several calculations where we let $\kappa \to 0$ in (6.17).

DeSc produces solutions which follows the reference solution for $\kappa = 10^{-1}, 10^{-2}, \ldots, 10^{-4}$. When $\kappa = 10^{-5}$ oscillations start to appear in the solution. The oscillations are not amplified for smaller values of κ.

LxF are stable for both splitting methods and produces solutions which follows the reference solution for $\kappa = 10^{-1}, 10^{-2}, \ldots, 10^{-9}$. However, with our grid it produces solutions which suffers with dissipation, which results in that the solutions are smoothed versions of the reference solution. If we redefine the grid in such a way that $3\Delta x \sim \Delta t$, then most of the dissipation dissapears from the solutions. The CFL pessimistic factor has also an influence on the solutions. Small values give more dissipation. Thus, the grid need to be chosen such that there is a strong relation between Δx and Δt.

LxW are stable down to $\kappa = 10^{-3}$, where oscillations appear in the solutions for both splitting methods. These oscillations increases when κ gets smaller. If we redefine N_x and N_t, we are not able to find a combination of the two step sizes which removes the oscillations for small κ. Thus, the LxW collapses when $\kappa \to 0$ for both splitting approaches.

McC are stable for both splitting methods for $\kappa = 10^{-1}, 10^{-2}, \ldots, 10^{-9}$. For very small values of κ small oscillations occur in the solutions, but they dissapear if Δx are put to be smaller. Thus, the McC approximates the solutions well.

The NT based schemes works very well for all values of κ and for both splitting methods. No oscillations appear, and the shock are approximated well. There are small differences in the solutions dependent on which limiter which is in use.

6.4.4 Discussions

The major task with the numerical testing for the viscous Burgers' equation (5.7) was to check the numerical convergence rates for Δt when applying the Godunov splitting (5.2) and Strang splitting (5.3) methods, using different numerical methods for the subequations in (6.1) and (6.2). Furthermore, we checked the convergence rates for Δx, CPU runtimes, accuracy of the errors, and at last the ability to approximate shocks using a non-exact test problem.

The numerical convergence rates for Δt for the exact test problem (6.16) are given in Table 6.1, while for the non-exact test problem (6.17) in Table 6.7. When we compare the numerical results with the theoretical results in Section 5, we observe that (5.2) produces numerical convergence rates for Δt which follow the theory very well. On the other hand, the numerical convergence rates for (5.3) are higher than for (5.2), but not as high as the theoretical result. For some of the numerical methods for the subequations, a convergence rate for Δt which follows the theoretical result well is obtained. From Table 6.1 we observe that for (5.3) the coefficient κ has major impact on the convergence rate. For $\kappa = 0.5$, which gives a very smooth solution, (5.2) is as good as (5.3), and this is the case for all the numerical methods we tested. This seems a bit strange and we check this special case in more details. We find that if we put $N_x = 4096$ or even higher, we get convergence rates for (5.3) which is similar to those in Table 6.1 for other values of κ. This happens for all the numerical methods. For smaller values of κ there are major differences for the convergence rates between the two splitting methods. Thus, it looks like (5.3) is sensitive for the number of nodes on the spatial axis, and that we need more nodes for higher values of κ to achieve convergence rates which follows the theory.

Let us comment on the behaviour of the different numerical methods for (6.1) and (6.2). We solve (6.2) with the implicit Crank–Nicolson method, which is known to work very well. We find that this method works well on our test cases. LxF suffers with dissipation, which is dependent on the ratio between Δx and Δt, and this behaviour has impact on the numerical results for the splitting methods. For all cases when LxF is used as solver for (6.1) in the two splitting methods, the convergence rates for Δt is not significantly different. The CFL pessimistic factor has an impact on the rates, and when it is small, which induces several substeps with LxF, the convergence rate for Δt are lower. This is an indication on that the dissipation has an severe impact on the convergence rates for Δt for the two splitting methods. The only way to overcome the dissipative behaviour, is to define the grid such that Δx is approximately the same as Δt. For LxW, McC and NT* we observe that there are small differences when the CFL pessimistic factor is varied, expect for LxW when $\kappa = 0.01$. For this special case the convergence rate for Δt makes a "jump", and gets in fact higher than we expect from the theory.

From the numerical testing it seems like the L^1-, L^2- and L^∞-norms are well correlated for the test problems, and the three different norms yield equal results for the convergence rates for Δt.

There is a major difference between the two splittings when accuracy is considered. When we solve the test problems for the same amount of split steps, (5.3) turns out to

be much more accurate than (5.2), cf. Tables 6.2, 6.3 and 6.4. This happens for all the different numerical methods for the subequations, in all the three norms, and there are no significant differences in the size of the errors between the methods. We also observe that when the solutions are steeper, the error decreases. Thus, from this point of view (5.3) should be favourable.

The convergence rates for Δx are given in Table 6.5. From the table we observe that all methods converge, but the convergence rates seem to be dependent on κ. LxF have significantly lower convergence rates compared with the other methods, and this is a result of the dissipation.

The full difference scheme in (6.15) works well on both the exact and non-exact test problem. Since the method is explicit, we have to choose Δx and Δt in relation with each other to get a stable method. From this point of view, and implicit scheme is preferable, since it usually are less restrictive on the two step sizes. Moreover, both operator splitting methods are faster compared to (6.15), cf. Table 6.6 and Figure 6.5. A crucial point is that (5.3) needs much fewer steps to achieve the same accuracy as (5.2). This naturally results in faster runtimes, and (5.3) should be favourable from this point of view.

All the numerical methods, except LxW, approximate arising shocks for the non-exact test problem well. The high resolution NT schemes seem to be better in this compared with LxF and McC. There was no difference in the two splitting methods on this point. Thus, from this point of view the NT schemes should be favourable.

From the tests on (5.7), we can give some small conclusions based on the results. It seems like it is easier to obtain numerical results for (5.2), which follows the theoretical results in a good way, compared with (5.3). However, there are major differences in the accuracy of the two splitting methods; (5.3) produces very accurate solutions in fewer steps than (5.2). As a result, one can perform relatively few time steps using (5.3), and get good reults. Thus, (5.3) should be chosen before (5.2). The different numerical methods work well, but LxF suffers with dissipation, which makes it less usable. Except from this, it is a matter of taste which numerical method we use to solve (6.1), in the two splitting methods.

6.4.5 Conclusions

We have solved the viscous Burgers' equation (5.7) numerically using the operator splitting approaches of Godunov and Strang type, given in (5.2) and (5.3), respectively. As initial data we have used an initial condition which has an exact solution, and one which gives a non-exact solution. The main focus was to investigate the numerical convergence rates for Δt numerically, and compare with the theoretical results in Theorems 5.9 and 5.11. In addition, several other aspects with the splitting approaches and the numerical methods was investigated.

We found that that the numerical convergence rates for Δt followed the theoretical result for the Godunov splitting well, while for the Strang splitting it was in some way dependent on the number of nodes on the spatial axis. However, the numerical results

followed overall the theoretical result for the Strang splitting. Hence, the Strang splitting should in general be favourable from this point of view.

All the numerical methods for the subequations worked well for both splitting approaches. The Lax–Friedrichs method for (6.1) suffers with dissipation, which results in a careful chosing of the step sizes Δx and Δt, to prevent the solutions from being to smeared out. Furthermore, there were minor differences between the Lax–Wendroff, MacCormach and the Nessyahu–Tadmor method, and all worked very well on the test problems.

The operator splitting method is a successful numerical method for the viscous Burgers' equation.

6.5 The Korteweg–de Vries Equation

For the KdV equation (5.8) we use two exact test problems as test environments for the two operator splitting methods and a comparison method which we introduce below. The first test case is a one-soliton problem, while the second is a two-soliton interaction problem. Thus, these two problems are well-suited to study the numerical methods' ability to model two important phenomena which (5.8) describes. We expect that methods that work well on the one-soliton problem, are not necessarily able to approximate the two-soliton problem well. After the exact test cases we impose (5.8) with initial data which give non-exact solutions, and we use them to check the convergence rate for Δt. The validity of the solutions are checked using three densitites which (5.8) should conserve. We introduce them below.

At the end of this section, we add a source term to (5.8), and try the operator splitting method on the resulting equation. This is followed up by discussions and conclusions on the numerical testing for (5.8).

The numerical operator splitting methods are labeled as follows: The method for (6.3) is named first, separated with a hyphen and followed by the method for (6.1). For example; *Diff-SV* denotes the method for which (6.3) is solved with the difference method (6.14), while (6.1) is solved with the spectral viscosity method (6.12). We let N_x be the number of nodes on the spatial axis, while N_t is the number of time steps used to solve to the end point in time. We use periodic boundary conditions for all the test cases.

6.5.1 Conserved Densities

The KdV equation (5.8) is a conservation law, which results in that it conserves some densitites. It can be shown that (5.8) have an infinite number of conserved densities, but in this study we only consider the three most basic densities, and refer to [4, Ch. 5.] for more examples. To derive the conserved densitites for (5.8), we have to go back to the conservation law in (6.5), from which we see that if $f(u) = \kappa u_{xx} - (1/2)u^2$, we can

write (5.8) in conservation form as

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2} + \kappa u_{xx}\right) = 0, \tag{6.18}$$

and we emphasize the relation with (6.1). We use periodic boundary conditions with period $P = [p_0, p_1]$ for (6.18), and we obtain the first density by assuming that we are allowed to interchange differentiation and integration,

$$\frac{d}{dt}\left(\int_{p_0}^{p_1} u\,dx\right) = \int_{p_0}^{p_1} u_t\,dx = -\int_{p_0}^{p_1}\left(\frac{u^2}{2} + \kappa u_{xx}\right)_x dx = -\left[-\frac{1}{2}u^2 + \kappa u_{xx}\right]_{p_0}^{p_1} = 0,$$

from the periodic boundary conditions. Thus, the area under the solution should be conserved,

$$\int_P u\,dx = C_1,$$

where C_1 is a constant. By multiplying (6.18) by u we get

$$\frac{\partial}{\partial t}\left(\frac{u^2}{2}\right) + \frac{\partial}{\partial x}\left(\frac{u^3}{3} + \kappa u u_{xx} - \frac{u_x^2}{2}\right) = 0,$$

and by doing the same calculation as above, we obtain the second conserved density

$$\int_P u^2\,dx = C_2.$$

We follow the same approach to find the third and last density. By adding $(-1/2\kappa)u^2$ times (6.18) to u_x times the derivative wrt. x of (6.18), we get

$$(-1/2\kappa)u^2(u_t + uu_x + \kappa u_{xxx}) + u_x(u_{xt} + u_x^2 + uu_{xx} + \kappa u_{xxxx}) = 0,$$

which can be rewritten as

$$\frac{\partial}{\partial t}\left(-\frac{u^3}{6\kappa} + \frac{u_x^2}{2}\right) + \frac{\partial}{\partial x}\left(-\frac{u^4}{8\kappa} - \frac{u^2 u_{xx}}{2} + u u_x^2 + \kappa u_x u_{xxx} - \frac{\kappa u_{xx}^2}{2}\right) = 0,$$

from which the third density is given as

$$\int_P \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = C_3.$$

We use these three densities to check the correctness of the numerical solutions.

6.5.2 A Full Difference Scheme

As a comparison method we use the following full implicit difference scheme,

$$\begin{aligned} r_2 U_{i+2}^{n+1} + \left(r_1(U_i^n + U_{i+1}^n) - 2r_2\right) U_{i+1}^{n+1} + U_i^{n+1} + \\ \left(-r_1(U_i^n + U_{i+1}^n) + 2r_2\right) U_{i-1}^{n+1} - r_2 U_{i-2}^{n+1} = U_i^n, \end{aligned} \tag{6.19}$$

where

$$r_1 = \frac{\Delta t}{6(\Delta x)} \quad \text{and} \quad r_2 = \frac{\kappa \Delta t}{2(\Delta x)^3}.$$

This scheme is unconditionally stable according to linear analysis, and it is of first order for Δt and second order for Δx. It was originally introduced in [9], and in what follows we label (6.19) as DiSc.

6.5.3 The One-Soliton Exact Test Problem

The one-soliton test problem is given as follows

$$u_t + uu_x + \kappa u_{xxx} = 0, \quad x \in [-\pi, \pi], \tag{6.20}$$

where the initial data at $t = 0$ is given from the exact solution,

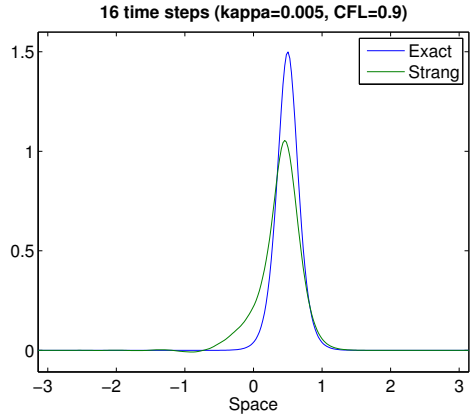
$$u(x,t) = 3c \operatorname{sech}^2\left(\sqrt{\frac{c}{4\kappa}}(x - ct)\right). \tag{6.21}$$

The constant c is the wave speed, and κ determines the sharpness of the soliton. In the experimentation we put $c = 0.5$ and $\kappa = 0.005$, and solve to $t = 1.0$.

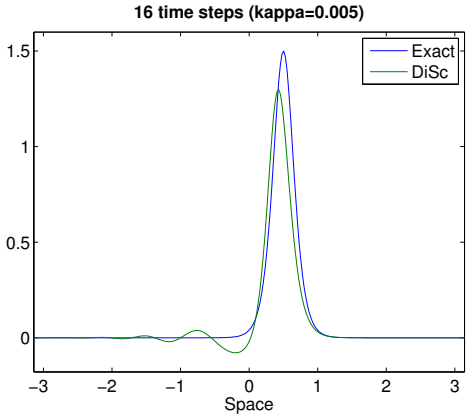
Let us first make some comments on the behaviour of the numerical operator splitting methods applied to (6.20). Diff-LxF and Spec-LxF do not manage to approximate the solution in a good manner. From the testing on the viscous Burgers' equation in Section 6.4, we know that LxF suffers with dissipation. For (6.20) this behaviour is in some way amplified out of bounds, cf. Figure 6.7. There is no difference if we use (5.2) or (5.3) as the operator splitting method. Thus, Diff-LxF and Spec-LxF are not suited for (6.20), and we drop these two methods in the remaining of the numerical study of the KdV equation.

Diff-LxW, Spec-LxW, Diff-McC and Spec-McC introduce oscillations for small values of N_t. For the NT based schemes we experience the same behaviour. The different limiters do not have significantly influence on the oscillations. Spec-SV also introduces oscillations, but the frequency is much lower compared with the other methods. We observe the same behaviour for Diff-SV, but this method gives the smallest oscillations. The oscillations for all the numerical methods are reduced when N_t is increased. We also observe that the amplitudes of the oscillations are smaller for (5.3), compared with (5.2). See Figure 6.7 for plots for four methods.

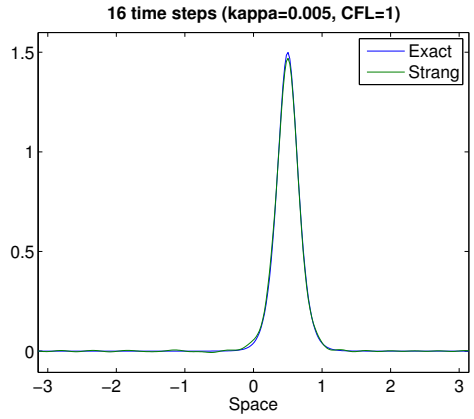
DiSc also introduces oscillations for small N_t. These oscillations is different from the oscillations for the other methods, and appear as low frequency waves behind the soliton, cf. Figure 6.7. We also observe that DiSc has some small problems approximating the



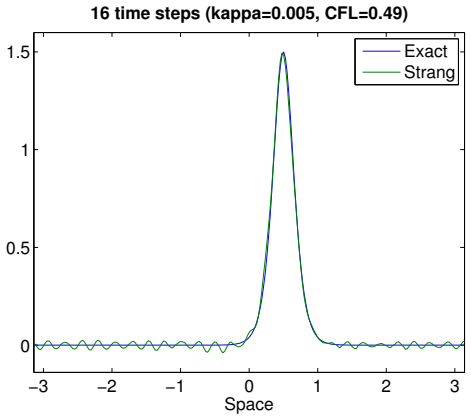
(a) Diff-LxF



(b) DiSc



(c) Diff-SV



(d) Spec-NTMc

Figure 6.7: Numerical solution to (6.20) using four different numerical operator splitting methods, where $N_x = 256$ and $N_t = 16$. Diff-LxF is very dissipative while the other methods approximate the soliton well, but oscillations become a part of the solutions.

		Godunov			*Strang*		
Method	*CFL*	L^1	L^2	L^∞	L^1	L^2	L^∞
Diff-LxW	0.90	1.082	1.068	1.038	1.757	1.680	1.573
Spec-LxW	0.90	1.044	0.986	0.935	1.299	1.141	1.068
Diff-McC	0.90	1.057	1.065	1.042	1.775	1.697	1.586
Spec-McC	0.90	1.027	0.975	0.928	1.276	1.132	1.050
Diff-NTMm	0.49	1.048	1.050	1.027	1.625	1.588	1.503
Spec-NTMm	0.49	1.015	0.972	0.924	1.300	1.149	1.081
Diff-NTMc	0.49	1.055	1.061	1.043	1.621	1.559	1.470
Spec-NTMc	0.49	1.044	0.986	0.935	1.299	1.141	1.068
Diff-NTVl	0.49	1.053	1.060	1.041	1.622	1.560	1.469
Spec-NTVl	0.49	1.331	1.189	0.943	1.331	1.189	1.109
Diff-NTSb	0.49	1.057	1.065	1.042	1.552	1.449	1.373
Spec-NTSb	0.49	1.032	1.002	0.947	1.347	1.229	1.133
Diff-SV	-	1.063	1.060	1.023	1.407	1.262	1.469
Spec-SV	-	1.059	1.048	0.996	1.537	1.441	1.359

Table 6.8: Numerical convergence rates for Δt for the operator splitting solutions of Godunov and Strang type for (6.20).

top shape of the soliton. However, both the oscillations and the approximation problem decrease when we increase N_t, and as a result DiSc works very well for (6.20).

Let us analyze the operator splitting methods in more details. To obtain the numerical converence rates for Δt, we follow the same outline as in Section 6.4. During the testing we find out that the CFL condition do not affect the convergence rates significantly, so we use 0.9 for LxW and McC, and 0.49 for the NT based methods in the remaining of this study. The numerical convergence rates for Δt are given in Table 6.8. We see that for (5.2) we get overall convergence rates which follows the theoretical results. For (5.3), we get better convergence result than for (5.2), but not as high as the theoretical result. The three norms are well correlated, and give approximately equal convergence rates for Δt.

The numerical convergence rates for Δx are given in Table 6.9. In the calculations we let $N_x = 2^5, 2^6, \ldots, 2^{10}$, to avoid large outliers in the regression analysis. From the table we observe that both splitting methods and all the combination of numerical methods converge for Δx. However, there is no correlation between the three norms; all produce different convergence rates for Δx.

To compare the size of the errors we solve (6.20) using $N_x = 256$ and $N_t = 128$, and the results are given in Table 6.10. We observe that there are small differences between the two splitting approaches, and that all the numerical methods approximates the solution to the same order. However, by using a finer grid we find that (5.3) gives smaller error compared with (5.2), and therefore is more accurate. The approximated L^1 errors are relatively large, but we find out that this happens because of the oscillations introduced by the numerical methods, in addition with approximation problems with the

	Godunov			*Strang*		
Method	L^1	L^2	L^∞	L^1	L^2	L^∞
Diff-LxW	0.672	1.290	1.714	0.972	1.516	1.945
Spec-LxW	0.473	0.941	1.357	0.502	0.963	1.381
Diff-McM	0.640	1.264	1.673	1.026	1.556	1.985
Spec-McM	0.495	0.937	1.352	0.534	0.959	1.375
Diff-NTMm	0.519	0.952	1.416	0.521	0.957	1.451
Spec-NTMm	0.440	0.836	1.313	0.439	0.842	1.337
Diff-NTMc	0.974	1.617	2.132	1.109	1.675	2.124
Spec-NTMC	0.703	1.183	1.699	0.730	1.209	1.732
Diff-NTVl	0.962	1.569	2.096	1.084	1.639	2.083
Spec-NTVl	0.698	1.158	1.670	0.721	1.181	1.703
Diff-NTSb	0.265	0.796	1.293	0.313	0.796	1.264
Spec-NTSb	0.375	0.917	1.381	0.410	0.933	1.412
Diff-SV	0.556	1.070	1.517	0.579	1.051	1.492
Spec-SV	0.740	1.294	1.732	0.832	1.370	1.833

Table 6.9: Numerical convergence rates for Δx for the operator splitting solutions of Godunov and Strang type for (6.20).

shape of the top of the soliton. By making N_x larger, these problems dissappear, but the CPU runtimes increase radically. DiSc manages to approximate the exact solution with the same accuracy as the operator splitting methods.

To calculate the CPU runtimes we put $N_x = 512$ and see how many time steps which are necessary to achieve an accuracy of 1e-01 in the L^1-norm for (6.20). The average CPU runtimes for the methods are given in Table 6.11. We observe the major differences between (5.2) and (5.3), which is that (5.3) needs radically fewer steps than (5.2) to achieve the same accuracy, which results in faster CPU runtimes. There is a difference in which solver we use for (6.3); Spec results in better runtimes than Diff. We also see that there is a small difference in the NT based method, dependent on the limiter in use. The relative long runtimes for Diff-SV and Spec-SV is a result of how the system of equations for SV in (6.12) is solved. DiSc is nearly as fast as (5.3), though it uses many steps to get the right accuracy.

From the norm vs. CPU runtimes in Figure 6.8, we observe that the error decreases for larger CPU times, which is what we expect. The major differences in the CPU runtimes for methods based on Diff and Spec, and the two SV methods, are good illustrated in the plots. We also observe that the Strang splitting gives longer runtimes, but it is more accurate.

For the conserved densities we get, using the exact solution of (6.20) in (6.21),

$$\int_{-\pi}^{\pi} u\,dx = 0.6000, \quad \int_{-\pi}^{\pi} u^2\,dx = 0.6000, \quad \int_{-\pi}^{\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = -36.0431.$$

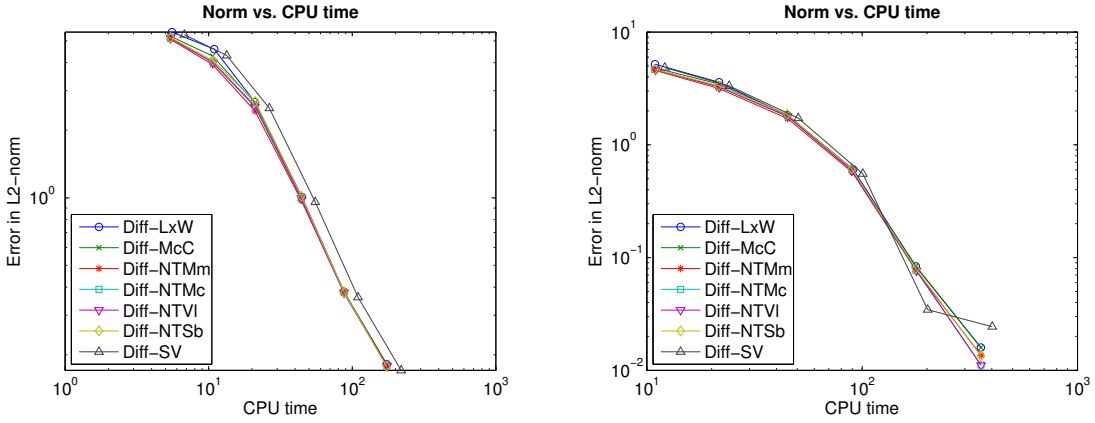
The relative errors for the conserved densities are given in Table 6.12, where we have

	Godunov			*Strang*		
Method	L^1	L^2	L^∞	L^1	L^2	L^∞
DiSC*	1.534e+00	2.505e-01	8.275e-02	-	-	-
Diff-LxW	5.470e-01	6.267e-02	2.140e-02	3.269e-02	3.707e-02	1.227e-02
Spec-LxW	1.433e+00	1.965e-01	5.636e-02	1.264e+00	1.790e-01	5.042e-02
Diff-McC	5.138e-01	5.701e-02	1.970e-02	2.526e-01	3.036e-02	9.855e-03
Spec-McC	1.446e+00	1.939e-01	5.554e-02	1.283e+00	1.768e-01	5.013e-02
Diff-NTMm	1.401e+00	1.915e-01	6.306e-02	1.413e+00	1.889e-01	5.244e-02
Spec-NTMm	1.576e+00	2.730e-01	8.950e-02	1.581e+00	2.637e-01	7.906e-02
Diff-NTMc	6.037e-01	5.671e-02	1.779e-02	5.327e-01	6.070e-02	2.112e-02
Spec-NTMc	7.973e-01	1.201e-01	3.710e-02	6.891e-01	1.043e-01	2.674e-02
Diff-NTVl	6.130e-01	5.658e-02	1.538e-02	5.394e-01	6.086e-02	2.104e-02
Spec-NTVl	8.473e-01	1.362e-01	4.323e-02	7.878e-01	1.223e-01	3.248e-02
Diff-NTSb	1.626e+00	2.176e-01	6.366e-02	1.424e+00	2.211e-01	7.481e-02
Spec-NTSb	9.986e-01	1.266e-01	3.998e-02	8.520e-01	1.179e-01	3.441e-02
Diff-SV	1.046e+00	1.235e-01	4.514e-02	1.287e+00	1.428e-01	5.374e-02
Spec-SV	1.414e+00	1.752e-01	4.826e-02	1.531e+00	1.780e-01	4.881e-02

Table 6.10: Estimated errors using L^1-, L^2- and L^∞-norm for (6.20) with $N_t = 128$ and $N_x = 256$. (* DiSc is *not* an operator splitting method of Godunov type, but is placed there for typing reasons.)

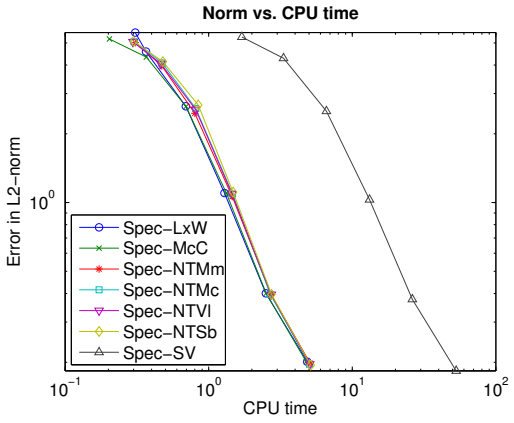
	Godunov		*Strang*	
Method	*Time steps*	*Runtime*	*Time steps*	*Runtime*
DiSc*	405	4.598	-	-
Diff-LxW	135	37.912	31	17.482
Spec-LxW	441	12.211	55	3.040
Diff-McC	134	37.066	31	17.658
Spec-McC	446	12.343	55	3.141
Diff-NTMm	122	32.641	31	18.017
Spec-NTMm	210	5.829	44	2.519
Diff-NTMc	127	34.664	31	17.789
Spec-NTMc	173	4.854	41	2.396
Diff-NTVl	119	34.261	31	17.781
Spec-NTVL	171	4.852	41	2.387
Diff-NTSb	140	36.895	39	22.384
Spec-NTSb	168	4.737	41	2.401
Diff-SV	123	68.990	18	14.601
Spec-SV	126	41.184	23	8.276

Table 6.11: The CPU runtimes (in seconds) for computing the numerical solutions of (6.20). The runtimes are the average after 10 calculations with each method. (* DiSc is *not* an operator splitting method of Godunov type, but is placed there for typing reasons.)

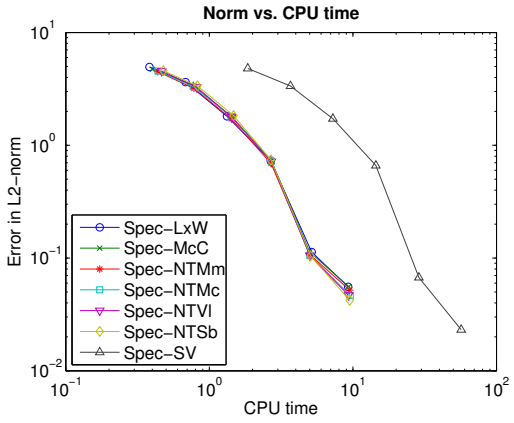


(a) Diff-* (Godunov)

(b) Diff-* (Strang)



(c) Spec-* (Godunov)



(d) Spec-* (Strang)

Figure 6.8: Plot of the error vs. CPU runtimes using logarithmic schales for (6.20). Observe the major differences in the runtimes between the Godunov splitting and Strang splitting, in additon with between the different numerical methods.

Method	$\int u\,dx$	$\int u^2\,dx$	$\int(-u^3/3\kappa + u_x^2)\,dx$
DiSc	2.5215 %	3.7141 %	6.1519 %
Diff-LxW	0.0014/0.0021 %	0.0270/0.0262 %	0.0712/0.0602 %
Spec-LxW	0.0020/0.0005 %	0.0278/0.0275 %	0.0772/0.0531 %
Diff-McC	0.0014/0.0025 %	0.0268/0.0260 %	0.0711/0.0601 %
Spec-McC	0.0021/0.0005 %	0.0270/0.0266 %	0.0768/0.0527 %
Diff-NTMm	0.0012/0.0080 %	0.3473/0.3507 %	0.6010/0.5961 %
Spec-NTMm	0.0019/0.0001 %	0.3463/0.3442 %	0.6030/0.5778 %
Diff-NTMc	0.0003/0.0085 %	0.0111/0.0029 %	0.0440/0.0078 %
Spec-NTMc	0.0017/0.0003 %	0.0148/0.0147 %	0.0523/0.0283 %
Diff-NTVl	0.0008/0.0088 %	0.0264/0.0175 %	0.0683/0.0415 %
Spec-NTVl	0.0017/0.0003 %	0.0296/0.0294 %	0.0768/0.0528 %
Diff-NTSb	0.0010/0.0085 %	0.3201/0.3591 %	0.5025/0.5786 %
Spec-NTSb	0.0014/0.0009 %	0.3127/0.3105 %	0.4906/0.5137 %
Diff-SV	0.0013/0.0018 %	0.1589/0.1576 %	0.2810/0.2660 %
Spec-SV	0.0023/0.0002 %	0.1461/0.1414 %	0.1442/0.1392 %

Table 6.12: Relative errors for the conserved densities for (6.20) for DiSc and the methods based on Godunov/Strang splittings, using $N_x = 512$ and $N_t = 128$.

used $N_x = 512$ and $N_t = 128$ in the calculations. We observe that all the operator splitting methods conserve the densitites very well. On the other hand, DiSc have larger relative errors, but by increasing N_t these errors decrease.

6.5.4 The Two-Soliton Exact Test Problem

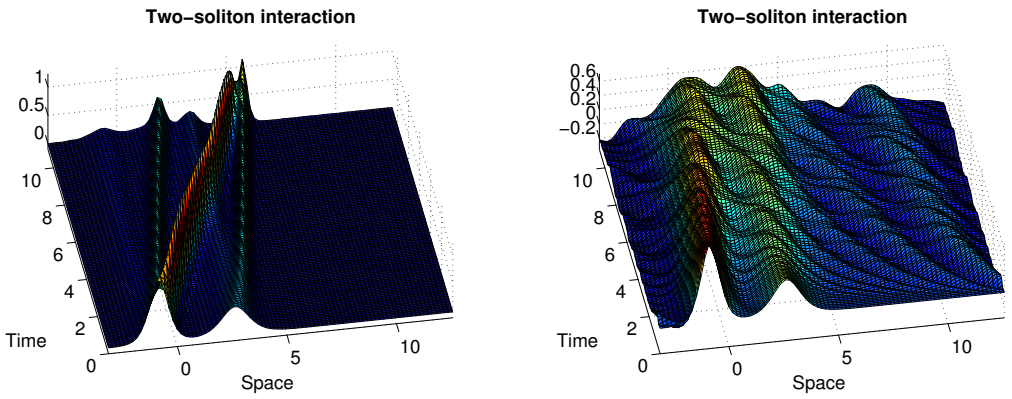
Physically, two solitons which have different shapes moves with different velocities, which is a result of the dependence between the height of the soliton and the velocity. A higher and steeper soliton travels faster than a lower and smoother soliton. If the two solitons travel along a surface, the higher soliton will overtake the lower soliton, and in the interaction they will not merge and not change shape. After the interaction, the two solitons should have the same shapes and velocities as before the interaction, cf. Figure 1.1.

Inspired by [15], we use the following exact test problem for the two-soliton interaction phenomenon,

$$u_t + uu_x + \kappa u_{xxx} = 0, \tag{6.22}$$

where the initial data at $t = 0$ is given from the exact solution,

$$u(x,t) = 2\frac{k_1^2 e^{\theta_1} + k_2^2 e^{\theta_2} + 2(k_2 - k_1)e^{\theta_1+\theta_2} + a(k_2^2 e^{\theta_1} + k_1^2 e^{\theta_2})e^{\theta_1+\theta_2}}{(1 + e^{\theta_1} + e^{\theta_2} + ae^{\theta_1+\theta_2})^2},$$



(a) Diff-LxW (b) Spec-LxW

Figure 6.9: Numerical solution to (6.22) using Diff-LaxW and Spec-LxW. None of the methods approximate the two-soliton solution in a good manner.

where

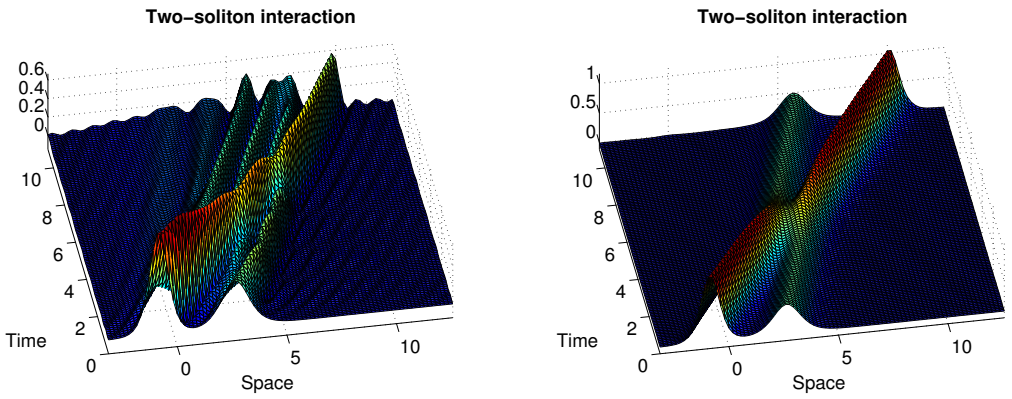
$$k_1 = \frac{3}{2}, \qquad k_2 = 1, \qquad a = \left(\frac{k_1 - k_2}{k_1 + k_2}\right)^2 = \frac{1}{25}$$
$$\theta_1 = k_1 \frac{x}{6\sqrt{\kappa}} - k_1^3 \frac{t}{6^{3/2}\sqrt{\kappa}} + 3, \quad \theta_2 = k_2 \frac{x}{6\sqrt{\kappa}} - k_2^3 \frac{t}{6^{3/2}\sqrt{\kappa}} - 3.$$

We use $\kappa = 0.005$ and let x be in the intervall $[-\pi, 4\pi]$ and solve to $t = 11$, when the interaction between the two solitons is totally over.

Let us comment on the behaviour of the different methods applied to (6.22). DiSc introduces no oscillations, but three more solitons arise in the solution and the two initial solitons travel with wrong velocities. Diff-LxW introduces no oscillations, but also here three more solitons are introduced, and the velocities are wrong for the initial solitons. Spec-LxW introduces oscillations and do not conserve the shape of the two solutions. In addition, the velocity of the two initial solitons are totally wrong. The results are independent on whether we use (5.2) or (5.3) as the splitting method. For Diff-McC and Spec-McC we get similar results. Thus, DiSc, Diff-LxW, Spec-LxW, Diff-McC and Spec-McC are *not* well suited methods for the interaction phenomenon. A plot of the solutions for Diff-LxW and Spec-LxW is given in Figure 6.9.

For Diff-NTMm, Diff-NTMc, Diff-NTVl and Diff-NTSb we get similar results as for Diff-LxW and Diff-McC. For Spec-NTMm, Spec-NTMc, Spec-NTVl and Spec-NTSb we get similar results as for Spec-LxW and Spec-McC. Thus, there is no difference using the more intelligent NT based methods on the two-soliton interaction problem.

We are left with two methods; Diff-SV and Spec-SV, and it turns out that Spec-SV is the best method for the interaction phenomenon. It manages to travel the solitons with correct velocities, and the interaction part is approximated in a good manner. We observe that (5.3) approximates the solution slightly better than (5.2). Diff-SV



(a) Diff-SV (b) Spec-SV

Figure 6.10: Numerical solution to (6.22) using Diff-SV and Spec-SV. Diff-SV is a useless method, while Spec-SV approximates the interaction very well.

	Godunov			*Strang*		
Method	L^1	L^2	L^∞	L^1	L^2	L^∞
Spec-SV	1.016	1.087	1.044	1.049	1.150	1.220

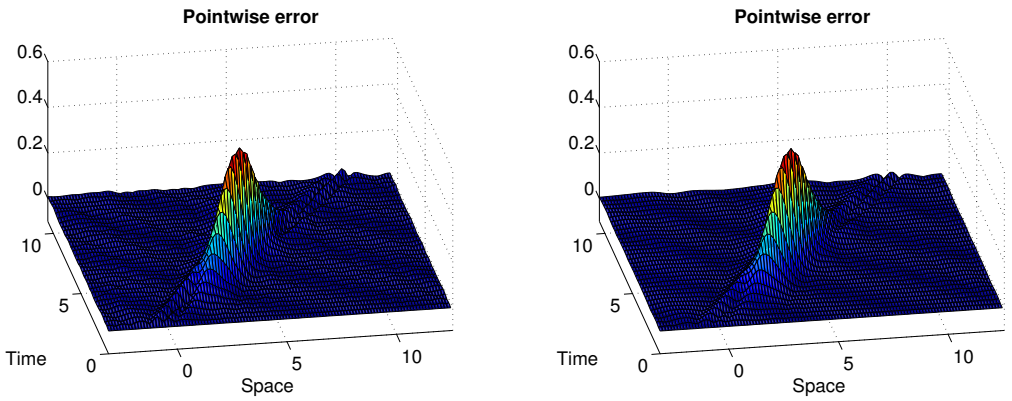
Table 6.13: Numerical convergence rates for Δt for the operator splitting solutions of Godunov and Strang type for (6.22).

introduces several solitons and oscillations, and it loses the soliton structure totally. This happens for both splitting methods. Thus, it turns out to be not well suited for this phenomenon, cf. Figure 6.10.

Hence, we are left with only *one* method which manages to approximate the two-soliton interaction phenomenon. To take the investigation a step further, we calculate the pointwise errors (in absolute values) for (5.2) and (5.3), and a plot is given in Figure 6.11. From the figure we observe that the largest errors occur when the two solitons interact, for both splitting methods, and that (5.3) is a bit more accurate than (5.2). However, after the interaction, the errors decrease and the numerical solutions are good approximations of the exact solution.

The numerical convergence rates for Δt for Spec-SV for (6.22) are given in Table 6.13. We have used $N_t = 2^4, 2^5, \ldots, 2^9$ in the calculations. From the table we observe that both splitting approaches converge, but there is no major difference in the convergence rates. We also see that the three norms are well correlated, and gives approximately equal answers.

There is only one test left for Spec-SV; the conserved densities. For the exact solution



(a) Godunov (b) Strang

Figure 6.11: Pointwise absolute error for the numerical solutions of (6.22) using Godunov and Strang splittings. The largest error is in the part of the solution where the two solitons interact. The Strang splitting is a bit more accurate than the Godunov splitting.

of (6.22), we find that,

$$\int_{-\pi}^{4\pi} u\,dx = 2.1078, \quad \int_{-\pi}^{4\pi} u^2\,dx = 1.2256, \quad \int_{-\pi}^{4\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = -61.0285.$$

We use $N_x = 1024$ and $N_t = 512$ in the calculations, and get for (5.2),

$$\int_{-\pi}^{4\pi} u\,dx = 2.1079, \quad \int_{-\pi}^{4\pi} u^2\,dx = 1.2258, \quad \int_{-\pi}^{4\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = -59.9223,$$

and for (5.3)

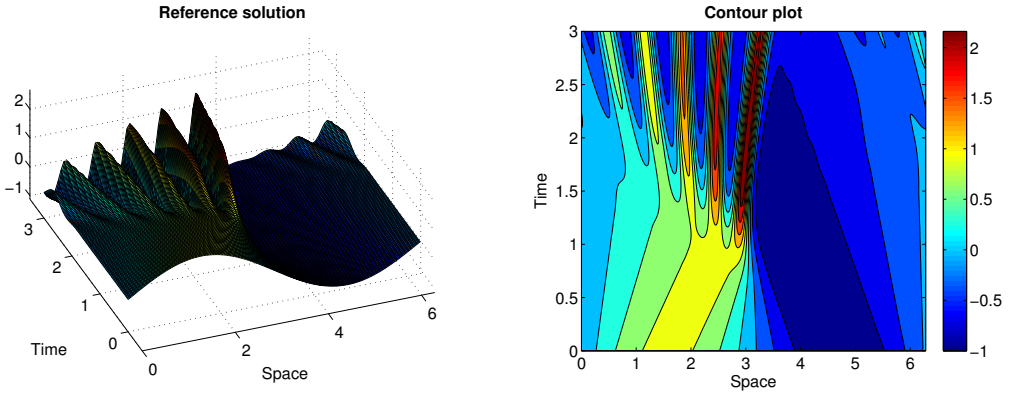
$$\int_{-\pi}^{4\pi} u\,dx = 2.1079, \quad \int_{-\pi}^{4\pi} u^2\,dx = 1.2258, \quad \int_{-\pi}^{4\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = -60.0402.$$

We see that all the numerical densities have an error of less than two percent, so we can conclude that Spec-SV conserves the three densities very well for both splitting methods.

Hence, based on our testing it turns out that the Spec-SV method is a well suited numerical method for the KdV equation, for both splitting methods in (5.2) and (5.3). All the other methods which worked well on the one-soliton problem (6.20), did not manage to solve (6.22) in a good manner.

6.5.5 The Non-Exact Test Problems

We impose the KdV equation (5.8) with initial data for which we do not have an exact solution. By studying the mechanism in each term of (5.8), we can qualitatively describe



(a) Numerical solution

(b) Contour plot

Figure 6.12: Numerical solution to (6.23). Observe the steepening of the initial sine wave, right before the dispersive term kicks in and prevents the creation of a shock.

how the solutions should behave. We use small values for κ in (5.8), and this results in that in the beginning of the movement, the nonlinear part is the dominant one, and as a result it steepen the initial data on the road to a shock creation. At some time, the third derivative gets larger and becomes the dominant part of the equation, and prevents the nonlinear part from creating a shock. Hence, the solution will be smooth for all time. The shape of the solutions is another question. From the discussion in [4], we know that the solution, after some time, should behave like solitons. That is, the initial data should in some way be deformed to waves with sech^2 shapes, and the number of solitons is dependent on the initial data and κ.

From the numerical testing on exact datas in the previous subsections, we found that Spec-SV was overall the best numerical method for both splitting methods. Therefore we study this method for both splitting methods in details for the non-exact test problems. To check the validity of the solutions, we use the three conserved densities. If they vary much with the initial data, the numerical solutions may not be correct. The other numerical methods are not leaved out, but we only test them briefly on the test cases in this section.

The first test case is the following,

$$u_t + uu_x + \kappa u_{xxx} = 0, \quad x \in [0, 2\pi], \quad u(x, 0) = \sin(x), \tag{6.23}$$

where we put $\kappa = 0.005$. A solution plot is given in Figure 6.12. We observe that what is expected happens. The initial wave is transported to the right and a shock is about to be constructed. After a short time, solitons with different heights and velocities appears in the solution. The steepest solitons travel to the right, while the smallest travel to the left. We also observe that the velocities are different for the waves.

		Godunov			*Strang*		
Test case	*Method*	L^1	L^2	L^∞	L^1	L^2	L^∞
1: (6.23)	Spec-SV	0.784	0.747	0.763	0.954	0.946	0.959
2: (6.24)	Spec-SV	1.428	1.250	1.082	1.856	1.817	1.773
3: (6.25)	Spec-SV	0.901	0.751	0.591	0.949	0.837	0.812

Table 6.14: Numerical convergence rates for Δt for the operator splitting solutions of Godunov and Strang type for three non-exact test problems.

The conserved densities for the initial condition in (6.23) are

$$\int_0^{2\pi} u\,dx = 0, \quad \int_0^{2\pi} u^2\,dx = 3.1416, \quad \int_0^{2\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = 3.1369,$$

while we get for the numerical solution at $t = 3.0$,

$$\int_0^{2\pi} u\,dx = 0, \quad \int_0^{2\pi} u^2\,dx = 3.1198, \quad \int_0^{2\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = 2.8909.$$

Thus, the densities are approximately conservered. The numerical convergence rates for Δt are given in Table 6.14. From the table we observe that (5.3) gives slightly better convergence rates, compared with (5.2). However, the rates are not as high as the theoretical results. The three norms are well correlated.

We find that the number of solitons in the solution of (6.23) is dependent on κ. For $\kappa > 0.005$ we get fewer solitons, and for $\kappa < 0.005$ more solitons. For very small values of κ we experience that we have to have a very fine grid to manage to catch all the very steep solitons in the solution. This naturally gives huge CPU runtimes for Spec-SV.

For the other methods, we find that Diff-LxW and Spec-LxW manage to approximate the solution well for (5.2) and (5.3). The solution conserves the densities well, but the third density is somewhat harder to conserve due to small oscillations in the solution, which results in variation in the derivative. However, by using a very fine grid it is approximately conserved. We experience the same behaviour for Diff-McC and Spec-McC.

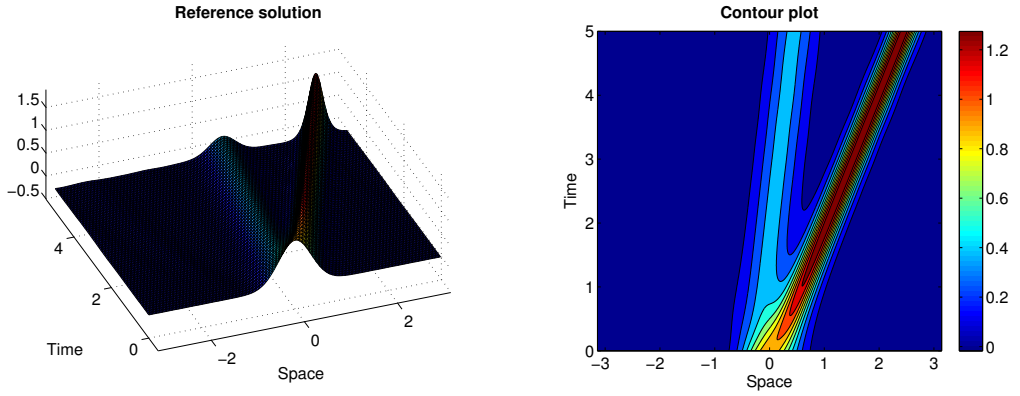
For Diff-NT* and Spec-NT* we get nearly same behaviour, but the small oscillations introduced by the abovementioned methods do not appear in the solution. Thus, it seems like the NT* based methods work slightly better. They also conserve the densities well.

Diff-SV and DiSc manage to approximate the shape of the solution well, but they conserve only the two first densities very well. The third density is too large. However, the number of solitons in the solution is correct.

We conclude that all the methods work very well on this test case, even though they have some problems with conserving the third density. The shape of the solution is correct for all methods.

In **the second test case** we use a gaussian as the initial data,

$$u_t + uu_x + \kappa u_{xxx} = 0, \quad x \in [-\pi, \pi], \quad u(x,0) = e^{-4x^2}, \tag{6.24}$$



(a) Numerical solution (b) Contour plot

Figure 6.13: Numerical solution to (6.24). The initial data is separated into two solitons.

which we solve for $\kappa = 0.005$. The initial wave is a compressed version of a soliton. As in the previous test case, the solution should in some form be deformed to sech^2 functions. The numerical solution is given in Figure 6.13. We observe that the gaussian is transformed into two solitons, each with different height and speed.

The conserved densities for the initial condition is

$$\int_{-\pi}^{\pi} u\,dx = 0.8862, \quad \int_{-\pi}^{\pi} u^2\,dx = 0.6267, \quad \int_{-\pi}^{\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = -31.6046,$$

while we at the end point $t = 5.0$ find for numerical solution

$$\int_{-\pi}^{\pi} u\,dx = 0.8862, \quad \int_{-\pi}^{\pi} u^2\,dx = 0.5620, \quad \int_{-\pi}^{\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = -31.6371.$$

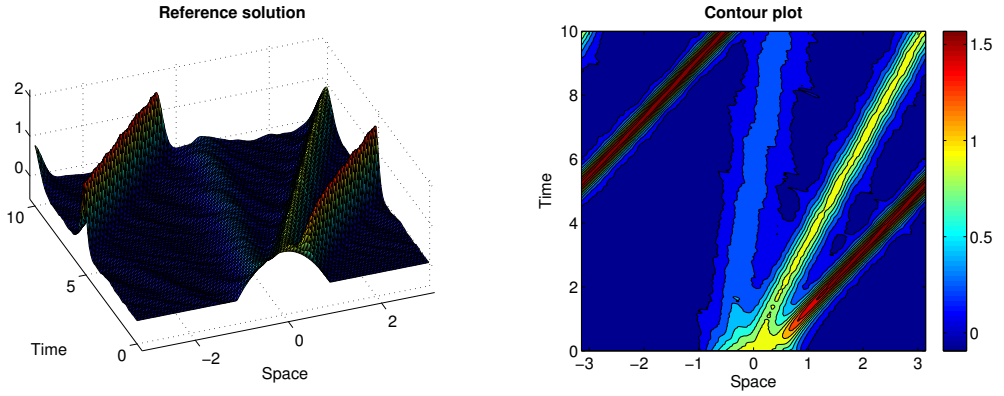
Hence all the conserved densities are conserved, and the solution is correct. The numerical convergenc rates for Δt are given in Table 6.14, and we observe that (5.3) gives better convergence rates than (5.2).

All the other numerical methods work well on (6.24), and there is slightly differences between them. All manages to approximate the shape of the solution well, and conserves the three densities. DiSc is the only method which suffers with the densitites, at most is the error about 11 percent for the third density.

The third test case is the following,

$$u_t + uu_x + \kappa u_{xxx} = 0, \quad x \in [-\pi, \pi], \quad u(x,0) = (1-x^2)\chi_{[-1,1]}, \tag{6.25}$$

where $\kappa = 0.005$ which we solve to $t = 10$. The initial data is a parabola, which is to bold to be a soliton. As with the other non-exact test problems, we expect that several solitons become a part of the solution as the time passes by. A plot of the solution is



(a) Numerical solution (b) Contour plot

Figure 6.14: Numerical solution to (6.25). The initial data is separated into three solitons, each with different shapes and velocities.

given in Figure 6.14, from which we observe that three solitons arise in the solution, all which have sech^2 shapes.

The conserved densities for the initial condition in (6.25) are found as,

$$\int_{-\pi}^{\pi} u\,dx = 1.3333, \quad \int_{-\pi}^{\pi} u^2\,dx = 1.0667, \quad \int_{-\pi}^{\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = -58.3033,$$

while we get for the numerical solution at $t = 10.0$,

$$\int_{-\pi}^{\pi} u\,dx = 1.3276, \quad \int_{-\pi}^{\pi} u^2\,dx = 1.0616, \quad \int_{-\pi}^{\pi} \left(-\frac{u^3}{3\kappa} + u_x^2\right) dx = -58.1694.$$

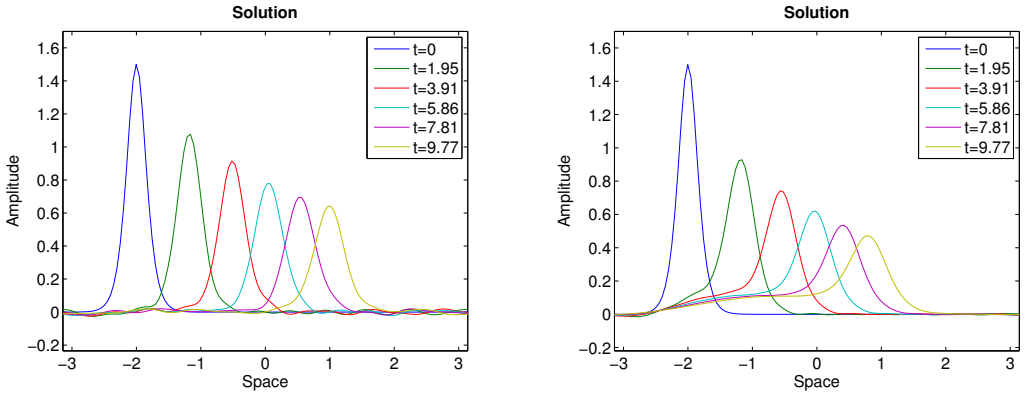
Thus, the densities are conservered. The numerical convergence rates for Δt are given in Table 6.14, from which we observe that there is no major difference between the two splitting methods.

When DiSc is applied to (6.25) it manages to transform the initial parabola into the three solitons, but the velocity of the steepest soliton is wrong. In addition, the three densities are not conservered. Thus, DiSc works not well on this problem. Diff-LxW, Spec-LxW, Diff-McC and Spec-McC introduce small oscillations in the solution, and in that sense they are not well suited for (6.25). Diff-NT* and Spec-NT* manages to approximate the shape of the solution very well, and in addition they conserve the three densities. The same is true for Diff-SV.

6.5.6 The Korteweg–de Vries Equation with a Source Term

We extend the numerical solvers to yield for the KdV equation with a small source term added. This gives the following equation,

$$u_t + uu_x + \kappa u_{xxx} = \varepsilon F(x, u, u_{xx}), \quad x \in [-\pi, \pi], \tag{6.26}$$



(a) $F(x, u, u_{xx}) = a_1 u + a_2 u^2 + a_3 u^3$ (b) $F(x, u, u_{xx}) = \varepsilon u_{xx}$

Figure 6.15: Numerical solution to (6.26) using two different sources $F(x, u, u_{xx})$. By inspecting the length between the tops of the soliton at each time, one find that the velocity is reduced, which is a result of the energy loss induced by the source.

where $F(x, u, u_{xx})$ is a smooth function and ε is a constant. If ε is small, the equation yield the so-called Perturbed KdV equation. In the previous subsections, we found that the operator splitting method applied to the standard KdV equation (5.8), was a method which worked overall very well. Therefore, it is tempting to apply the splitting approach once more for (6.26) to form a numerical solution. We split the equation, and get

$$v_t + vv_x = 0, \tag{6.27}$$

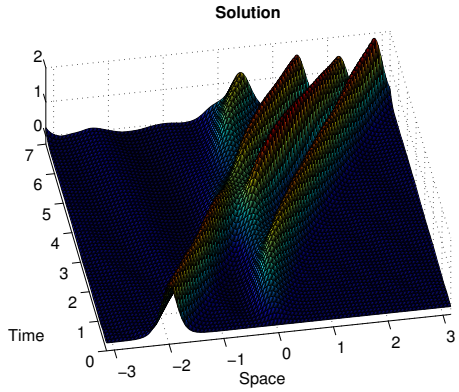
$$w_t + \kappa w_{xxx} = 0, \tag{6.28}$$

$$\zeta_t = \varepsilon F(x, \zeta, \zeta_{xx}), \tag{6.29}$$

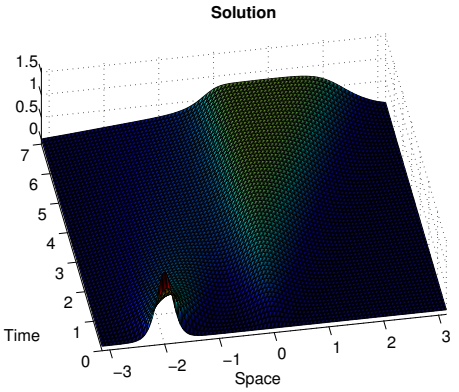
as the split equations. We concatenate the subsolutions using the Godunov splitting (5.2), such that we solve the three subequations a full time step Δt to form a full splitting step. For (6.27) and (6.28) we use SV and Spec, respectively. For (6.29), we form a method which is dependent on the form of the source. We construct four small test problems, where we put $\kappa = 0.005$ and let x be in the intervall $[-\pi, \pi]$. The constant ε is varied between the examples. As initial condition for the problems we use (6.21) (with $c = 0.5$), and shift it two units to the left.

The first test case is inspired by [16]. We put $F(x, u, u_{xx}) = a_1 u + a_2 u^2 + a_3 u^3$ in (6.29) and $\varepsilon = 0.01$. To solve (6.29) we use an explicit forward difference in time. For the coefficients we use $a_1 = 2$, $a_2 = -20$ and $a_3 = 4$. The added source results in that the soliton reduces its energy because of the negative sign in front of a_2. Thus, the velocity is reduced in addition with the height. If we use only positive coefficients for $F(x, u, u_{xx})$, the velocity and height increase. A plot of the solution is given in Figure 6.15. We observe that the height of the soliton is reduced and that the velocity decreases.

In the second test case we put $F(x, u, u_{xx}) = u_{xx}$. With this source (6.26) turns into the Kortweg–de Vries Burgers' equation. To solve (6.29) we use the Crank–Nicolson



(a) $F(x,u,u_{xx}) = 3c\,\mathrm{sech}^2\left(\sqrt{\frac{c}{4\kappa}}x\right)$



(b) $F(x,u,u_{xx}) = 3c\,\mathrm{sech}^2\left(\sqrt{\frac{c}{4\kappa}}x\right) + \frac{1}{10}u_{xx}$

Figure 6.16: Numerical solution to (6.26) using two different sources $F(x,u,u_{xx})$. The major difference between (a) and (b) is the appearence of the diffusion term in (b), which result in a smoothing of the solution.

scheme. The solution for this test case, should be close related to those for (6.20), but because of the source term the soliton is smoothed. As a result of this smoothing, the velocity should go down. We put $\varepsilon = 0.01$ in the calculations. A plot of solution is given in Figure 6.15, and we observe that the solution is a smoothed version of those in Figure 6.7.

For the third test case we use

$$F(x,u,u_{xx}) = 3c\,\mathrm{sech}^2\left(\sqrt{\frac{c}{4\kappa}}x\right),$$

where we put c and κ as above. We use $\varepsilon = 1.0$. Thus, the source produces solitons similar to the initial data around $x = 0$. These new solitons should travel to the right. A plot of the solution is given in Figure 6.16.

In the fourth test case we combine the second and third test case, and use

$$F(x,u,u_{xx}) = 3c\,\mathrm{sech}^2\left(\sqrt{\frac{c}{4\kappa}}x\right) + \frac{1}{10}u_{xx},$$

where we put c and κ as above, and we use $\varepsilon = 1.0$. The solution should behave similar as in the third test case, but the solitons should be smeared out. A plot is given in Figure 6.16.

We have not been much formal and precise in the testing for (6.26), and we have not analytically analyzed the behaviour of the equation. We have only extended our already existing solvers for the KdV equation, to yield for an added source term. However, by using physical interpretation of $F(x,u,u_{xx})$ we have been able to predict the behaviour of the numerical solutions of (6.26), and the solutions followed our intuition in a good way. We have only used rather simple expression for the source term, and more "exotic"

sources could have been tested, but they should in some sense have some relation with the physics of the equation. As a small conclusion, it looks like our extended operator splitting method for (6.26) works well, but the results should be controlled in more details.

6.5.7 Discussions

The main purpose for the numerical experiments for the KdV equation (5.8) was to study the numerical convergence rates for the split step size Δt when applying the Godunov splitting (5.2) and the Strang splitting (5.3) methods, and compare the numerical results with the theoretical results in Section 5. In this investigation we used an exact one-soliton test case, a two-soliton exact test case and several non-exact test cases.

For the one-soliton test problem (6.20) we experienced that all methods but Diff-LxF and Spec-LxF worked very well. The convergence rates for Δt are given in Table 6.8, from which we observe that (5.2) gets overall numerical results which follows the theory very well, while (5.3) gets higher convergence rates, but not as high as the theoretical results. From our numerical results it seems like (5.3) is sensitive for which numerical method that is used for solving the subequations, specially those for (6.3). The methods which use Diff as the numerical method for (6.3) obtain higher convergence rates compared with the methods with Spec. Same behaviour is not present for (5.2). We should not expect that the numerical results fit perfectly with the theory, and our numerical results is an indication on that (5.3) gives a higher convergence rate than (5.2), which is the important essence of the theoretical results. In addition, the L^1-, L^2- and L^∞-norm are well-correlated and give approximately the same convergence rates for Δt.

The numerical behaviour of the two splitting methods in (5.2) and (5.3) is in some way different. The behaviour is best seen when few time steps are used to create the solutions. Oscillations appear in all methods, though with different amplitudes and frequencies, and Diff-SV and Spec-SV turns out to be the best method from this point of view. Moreover, the oscillations is reduced when the number of time steps is increased. However, there is a clear tendency that the oscillations with (5.3) have lower amplitudes compared with the solutions using (5.2). Thus, since (5.3) performs two "half-steps" with (6.3), in each split step, it looks like the numerical methods for (6.3) reduces in some way these oscillations.

The Strang splitting (5.3) involves three function calls in each split step, while the Godunov splitting (5.2) involves two. Thus, when the same amount of steps are performed, (5.3) yield naturally higher CPU runtimes. However, the important point is the accuracy of the splittings, and we experienced that (5.3) needs significantly fewer time steps to give the same accuracy as (5.2), cf. Table 6.11. Thus, from this point of view is (5.3) favourable. When we consider the numerical methods for the subequations, we observe that there is a major difference if we solve (6.3) with Diff or Spec, and the latter is much faster. The differences between LxF, McC and NT* for the (6.1) are small. The slowest methods are the two methods which use SV for the conservation law. This is a result of how the linear system in (6.12) is solved. A natural improvement would have been to solve the system with a Crank–Nicolson method, which would have given

a linear system to solve at each time step. With this implicit scheme we could probably have taken longer time steps, which would have resulted in faster CPU runtimes. When in addition fast iteration methods exist for solving linear system, this would probably have given a much faster method.

Since (5.8) is a conservation law and therefore conserve some densities, we tested how well (5.2) and (5.3) in use with the different numerical methods conserved three densitites. The conclusion is that both the splitting methods in combination with all the numerical methods for the subequations conserve the densities very well for the one-soliton test problem.

DiSc works very well for the one-soliton test problem, and manages to conserve the three densities, though with a relative error larger than the operator splitting solutions. We experience that this difference method is stable for both rough and fine grids, and in addition is it relatively fast. Hence, DiSc works as well as the operator splitting methods for the one-soliton test problem.

For the two-soliton test problem (6.22) all but Spec-SV collapse, in the sense that they not manage to give a correct solution. Spec-SV also conserves the densities for this problem well, for both splitting methods. The convergence rates for Δt is not different for the two splitting methods for Spec-SV, cf. Table 6.13. The pointwise error is highest in the interaction of the solitons, cf. Figure 6.11. The interesting with Spec-SV is that even though the error is relatively high in the middle of the domain, the method is able to regain and reduce the error after the interaction. This is a property of the method which we not have seen using other methods, that is, if first an error is introduced it remains in the solution for all times after that point. This means that Spec-SV manages to remember the shape and movement of the two solitons throughout the interaction. Thus, we have found a numerical method which manages to approximate the one-soliton and two-soliton phenomena in a good manner.

For the non-exact test cases Spec-SV in combination with the two splitting methods work well, based on the conserved densities and the mechanism in each term of (5.8). From the collapse on the two-soliton exact test problem, we thought that the other methods were not able to approximate solutions which involved more than one soliton. We were wrong! The other operator splitting methods and DiSc manage to produce correct non-exact solutions which conserve the three densitites, for the non-exact cases. Moreover, the success for the other methods seem to be dependent on the initial data, specially for DiSc, in addition with a strong dependence on Δx and Δt. For instance, to conserve the densities using some of the methods we have to use a much finer grid than with Spec-SV. On the other hand, the CPU runtime for Spec-SV is very high compared with the other methods, which means that we can use finer grids for the other methods and obtain same CPU runtimes. However, by our testing Spec-SV should be favourable since it works well for all the test problems.

As a small conclusion on the numerical convergence rates for Δt, which overall have been the main purpose to study throughout all the different test problem, it seemes like it it difficult to obtain numerical convergence rates for Δt for (5.3) for (5.8) which is as high as the theoretical result. The numerical results for (5.3) seems to be way more

sensitive for the initial data, the grid, and the underlying numerical solvers, than the more simple (5.2).

The non-exact test cases for (5.8) with a source term is more or less a tryout of the operator splitting method for equations with more than two spatial split terms. From our results, which is interpreted using physical intuition, it seems like it is possible to use the operator splitting method for equations with more than two spatial terms. We concatenated the split equations in the most simple way using the Godunov method, since there is no intuitive way of combining "half" steps to form a method similar to Strang splitting for two spatial terms. All the four test cases make sense, and the solution seem to be correct from our intuition, but we have no exact cases to check the numerical methods with. From an analytical point of view, it should be possible to get theoretical convergence result for the Godunov type of an operator splitting method for equations with more than two spatial terms. On the other hand, operator splitting methods which give high order convergence for Δt for these types of equations are harder to find. An interesting study would have been to study the operator splitting method applied on very complicated equations involving many terms, each which as a subequation in a splitting method is relatively easy to solve.

6.5.8 Conclusions

We have solved the KdV equation (5.8) numerically using the operator splitting method of Godunov and Strang type, given in (5.2) and (5.3), respectively. As initial data, we have used an exact one-soliton solution, an exact two-soliton interaction solution and several non-exact test cases. The main focus was to investigate the numerical convergence rates for Δt numerically, and compare them with the theoretical results in Section 5. In addition, several other aspects with the splitting approaches and the numerical methods was investigated, and at last the two operator splitting methods were compared with a full implicit difference scheme (DiSc).

We found that the numerical convergence rates for Δt followed the theoretical results for the Godunov splitting well for the one-soliton exact test problem, and all the numerical methods worked well. In addition, all the numerical methods conserved three densities, which the KdV equation conserve. The Strang splitting gave better rates than the Godunov splitting, but not as high as the theoretical result. However, the Strang splitting produced more accurate solutions, and should because of this generally be more favourable for this type of problems.

For the two-soliton exact test problem, all numerical methods but a numerical method which use a spectral method for the Airy equation in combination with a spectral viscous method for the inviscid Burgers' equation (Spec-SV) collapsed, in the sense that they not gave correct solutions. Spec-SV was also able to conserve the three densities very well. Hence, Spec-SV seems to be a good numerical method for the KdV equation, using both the Godunov and Strang splitting methods.

We tested how well the operator splitting methods and DiSc solved the KdV equation for three non-exact test cases, which we used the conservered densities and physical intuition as check points for the correctness of the solutions. Even though all of the

methods but Spec-SV collapsed for the two-soliton exact test problem, several of the different methods were able to produce solutions which was intuitively correct using the check points. DiSc also produced sensible solutions.

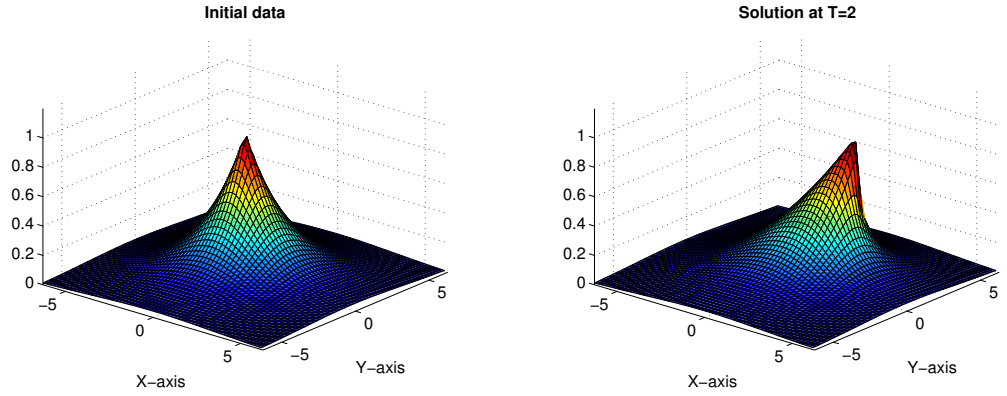
At last, as a try-out we tested the operator splitting method for the KdV equation with a source term, by splitting it into three equations, and concatenate them using the Godunov method. The numerical solutions followed our intuition, but should be investigated in more details, specially from an analytical view point.

Based on our testing, the operator splitting methods of Godunov and Strang types are successful methods for the KdV equation, and Spec-SV is the best overall method for the equation, both for the Godunov and Strang operator splitting methods.

Appendices

A Two Dimensional Examples

We have in this text only considered the viscous Burgers' equation and the Korteweg–de Vries equation in one spatial dimension, and applied the operator splitting method to split the original equations into two subequations, which we solved for small time steps. In these cases the two spatial terms were derivatives in the same dimension. The splitting method can also be used on equations which involves terms which have derivatives in several dimensions, and we use this very small section to illustrate the operator splitting method on two dimensional problems.



(a) Initial data (b) Solution

Figure A.1: Initial data and numerical solution of the two dimensional inviscid Burgers' equation (A.1) using the operator splitting method of Godunov type.

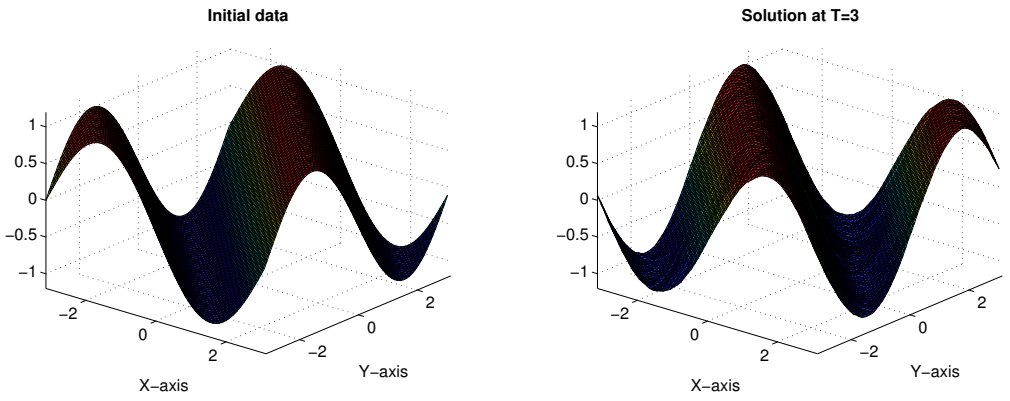
As the first example, we extend the inviscid Burgers' equation to two dimensions. To do this, consider the inviscid Burgers' equation

$$u_t + uu_\xi = 0,$$

and put $\xi = x + y$, which yield the two dimensional inviscid Burgers' equation,

$$u_t + uu_x + uu_y = 0, \quad \Omega = [-2\pi, 2\pi] \times [-2\pi, 2\pi], \tag{A.1}$$

by a change of variables. Since we started with the one dimensional equation, we can use the intuition to interpret how the solutions will behave. As initial data we use a two dimensional guassian centered at origo. Since the coefficients in front of the two nonlinear parts is one, the initial data will be transported in the direction [1,1], and the initial data should be steepened through this movement. From the plot of the solution in Figure A.1 we observe that this happens. Thus, the operator splitting method works as we expect on this example.



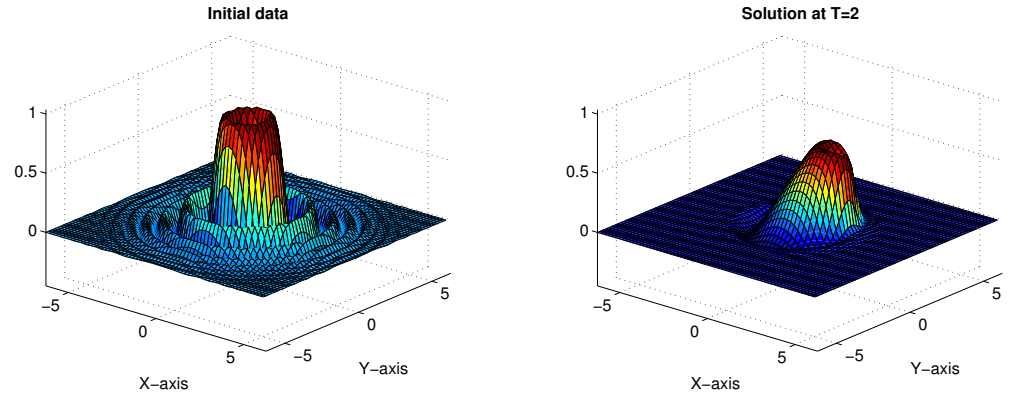
(a) Initial data (b) Solution

Figure A.2: Initial data and numerical solution of the two dimensional Airy equation (A.2) using the operator splitting method of Godunov type.

As the second example, we use a two dimensional Airy equation,

$$u_t - 2u_{xxx} + u_{yyy} = 0, \quad \Omega = [-\pi, \pi] \times [-\pi, \pi], \tag{A.2}$$

which has exact solution $u(x, y, t) = \sin(x + y - t)$. Thus, the solution should be a two dimensional wave. A solution plot is given in Figure A.2, and we see that the solution behaves as we expect. Thus, the operator splitting method works for this problem.



(a) Initial data (b) Solution

Figure A.3: Initial data and numerical solution of the two dimensional viscous Burgers' equation (A.3) using the operator splitting method of Godunov type.

For the third example, we extend the one dimensional viscous Burgers' equation, $u_t + uu_\xi - \kappa u_{\xi\xi} = 0$, to two dimensions. This is done by using $\xi = x + y$ and a change

of variables, which gives

$$u_t + uu_x + uu_y - \kappa(u_{xx} + u_{yy}) = 0, \quad [-2\pi, 2\pi] \times [-2\pi, 2\pi]. \tag{A.3}$$

This example is close related to (A.1), the only difference is the added diffusion which will have an smoothing effect on the solutions. We use

$$u_0(x, y) = \sin(x^2 + y^2)\, e^{-\sqrt{x^2+y^2}},$$

as the initial data and put $\kappa = 0.05$. The initial data should be traveled in the $[1, 1]$ direction and during this movement smoothed slightly out. From Figure A.3 we see that the small initial waves far from origin are smoothed out and are gone at $t = 2.0$, and we are left with the highest of the initial waves which are a tilted and smoothed version of the initial wave.

These examples are good illustrations of the robustness of the operator splitting method, and that the method can be applied to several different equation with great success.

B Small Biographies

The small biographies throughout the text, are based on the following references.

Stefan Banach
- http://en.wikipedia.org/wiki/Stefan_Banach, date: 18 March 2011.

Joseph Valentin Boussinesq
- http://en.wikipedia.org/wiki/Joseph_Valentin_Boussinesq, date: 15 May 2011.

Johannes Martinus Burgers
- http://en.wikipedia.org/wiki/Johannes_Martinus_Burgers, date: 15 May 2011.

Augustin Louis Cauchy
- http://snl.no/Augustin_Louis_Cauchy, date: 15 May 2011.
- Page 61 in [25].

John Crank
- http://en.wikipedia.org/wiki/John_Crank, date: 19 May 2011.

Jean-Marie Duhamel
- http://en.wikipedia.org/wiki/Jean-Marie_Duhamel, date: 18 March 2011.
- http://www.britannica.com/EBchecked/topic/173205/Jean-Marie-Constant-Duhamel, date: 18 March 2011.

Euclid
- http://snl.no/Evklid/gresk_matematiker, date: 23 March 2011.
- http://en.wikipedia.org/wiki/Euclid, date: 23 March 2011.
- Page 110 in [25].

Jean Baptiste Joseph Fourier
- http://snl.no/Jean_Baptiste_Joseph_Fourier, date: 15 May 2011.
- Page 133 in [25].
- http://en.wikipedia.org/wiki/Joseph_Fourier, date: 15 May 2011.

Maurice René Fréchet
- http://en.wikipedia.org/wiki/Maurice_Rene_Frechet, date: 18 March 2011.

Kurt Otto Friedrichs
- http://en.wikipedia.org/wiki/Kurt_O._Friedrichs, date: 19 May 2011.

Josiah Willard Gibbs
- http://snl.no/Josiah_Willard_Gibbs, date: 19 May 2011.
- http://en.wikipedia.org/wiki/Willard_Gibbs, date: 19 May 2011.

Sergei Konstantinovich Godunov
- http://en.wikipedia.org/wiki/Sergei_K._Godunov, date: 15 May 2011.

David Hilbert
- http://snl.no/David_Hilbert, date: 18 March 2011.
- Page 171 in [25].

- http://en.wikipedia.org/wiki/David_Hilbert, date: 18 March 2011.

Diedrik Korteweg
- http://en.wikipedia.org/wiki/Diederik_Korteweg, date: 15 May 2011.

Peter David Lax
- http://snl.no/Peter_D._Lax, date: 19 May 2011.
- http://www.abelprisen.no/no/prisvinnere/2005/, date: 19 May 2011.
- http://en.wikipedia.org/wiki/Peter_Lax, date: 19 May 2011.

Bram Van Leer
- http://en.wikipedia.org/wiki/Bram_van_Leer, date: 19 May 2011.

Gottfried Wilhelm Leibniz
- http://snl.no/Gottfried_Wilhelm_Leibniz, date: 16 May 2011.
- Pages 273-275 in [25].
- http://en.wikipedia.org/wiki/Leibniz, date: 16 May 2011.

Marius Sophus Lie
- http://snl.no/Sophus_Lie, date: 21 May 2011.
- Page 277 in [25].
- http://en.wikipedia.org/wiki/Sophus_Lie, date: 21 May 2011.

Claude-Louis Navier
- http://en.wikipedia.org/wiki/Claude-Louis_Navier, date: 15 May 2011.

Haim Nessyahu
- http://www.cscamm.umd.edu/people/faculty/tadmor/students/HaimNessyahu.htm, date 19 May 2011.

Phyllis Nicolson
- http://en.wikipedia.org/wiki/Phyllis_Nicolson, date: 19 May 2011.

Marc-Antoine Parseval
- http://en.wikipedia.org/wiki/Parseval, date: 15 May 2011.

Michel Plancherel
- http://en.wikipedia.org/wiki/Plancherel, date: 15 May 2011.

Giuseppe Peano
- http://snl.no/Giuseppe_Peano, date: 15 May 2011.
- Page 345 in [25].
- http://en.wikipedia.org/wiki/Giuseppe_Peano, date: 15 May 2011.

John William Strutt Rayleigh
- http://snl.no/John_William_Strutt_Rayleigh, date: 15 May 2011.
- http://en.wikipedia.org/wiki/John_William_Strutt,_3rd_Baron_Rayleigh, date: 15 May 2011.

John Scott Russel
- http://en.wikipedia.org/wiki/John_Scott_Russell, date: 15 May 2011.

Erwin Schrödinger
- http://en.wikipedia.org/wiki/Erwin_Schrodinger, date: 15 May 2011.
- http://snl.no/Erwin_Schrödinger, date: 15 May 2011.

Laurent Schwartz
- http://snl.no/Laurent_Schwartz, date 15 May 2011.
- http://en.wikipedia.org/wiki/Laurent_Schwartz, date 15 May 2011.

Karl Hermann Amandus Schwarz
- http://en.wikipedia.org/wiki/Hermann_Amandus_Schwarz, date: 15 May 2011.

Sergei Lvovich Sobolev
- http://en.wikipedia.org/wiki/Sergei_Sobolev, date: 18 March 2011.

Sir George Gabriel Stokes
- http://en.wikipedia.org/wiki/George_Gabriel_Stokes, date: 15 May 2011.
- http://snl.no/George_Gabriel%2C_Sir_Stokes, date: 15 May 2011.

William Gilbert Strang
- http://www-math.mit.edu/∼gs/, date: 15 May 2011.
- http://en.wikipedia.org/wiki/Gilbert_Strang, date: 15 May 2011.

Eitan Tadmor
- *Computational Methods in Applied Mathematics,* Vol.4, No.3, pp. 265–270, 2004.
- http://www.cscamm.umd.edu/people/faculty/tadmor/, date: 19 May 2011.

Brook Taylor
- http://snl.no/Brook_Taylor, date: 15 May 2011.
- Page 433 in [25]
- http://en.wikipedia.org/wiki/Brook_taylor, date: 15 May 2011.

Gustav de Vries
- http://en.wikipedia.org/wiki/Gustav_de_Vries, date: 15 May 2011.

Burton Wendroff
- http://www.math.unm.edu/∼bbw/, date: 19 May 2011.
- http://en.wikipedia.org/wiki/Burton_Wendroff, date: 19 May 2011.

William Henry Young
- http://en.wikipedia.org/wiki/William_Henry_Young, date: 16 May 2011.

References

[1] Robert A. Adams. *Calculus: A complete course. Fifth edition.* Pearson Education Canada Inc., Toronto, Ontario, 2003.

[2] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces. Second edition.* Adademic Press, 2003.

[3] A. Ambrosetti and G. Prodi. *A Primer of Nonlinear Analysis.* Cambridge University Press, Cambridge, 1993.

[4] P. G. Drazin & R. S. Johnson. *Solitons: an introduction.* Cambridge University Press, Cambridge, 1989.

[5] Lawrence C. Evans. *Partial Differential Equations. Second Edition.* American Mathematical Society, Providence, Rhode Island, 2010.

[6] Hans G. Feichtinger and Tobias Werther. *Robustness of regular sampling in Sobolev algebras.* Sampling, wavelets, and tomography, John J. Benedetto, Ahmed I. Zayed, editors, pp. 83–113, Applied and Numerical Harmonic Analysis, Birkhäuser Boston, 2004.

[7] C. Gasquet and P. Witomski (Translated by R. Ryan). *Fourier Analysis and Applications. Filtering, Numerical Computation, Wavelets.* Springer-Verlag New York, Inc. 1999.

[8] Anne Gelb, Eitan Tadmor. *Enhanced spectral viscosity approximations for conservation laws*, Applied Numerical Mathematics 33, pp. 3–21, 2000.

[9] Katuhiko Goda. *On Stability of Some Finite Difference Schemes for the Korteweg–de Vries Equation.*, J. Phys. Soc. Japan, Vol. 35, no. 1, pp. 229–236, 1975.

[10] E. Hairer, S. P. Nørsett, G. Wanner. *Solving Ordinary Differential Equations I, Nonstiff Problems, Second Revised Edition.* Springer-Verlag Berlin Heidelberg, 1993.

[11] Helge Holden, Christian Lubich, and Nils Henrik Risebro. *Operator splitting for partial differential equations with Burgers nonlinearity.* arXiv:1102.4218v1, http://arxiv.org/abs/1102.4218v1.

[12] Helge Holden, Kenneth H. Karlsen, Knut-Andreas Lie, Nils Henrik Risebro. *Splitting Methods for Partial Differential Equations with Rough Solutions. Analysis and Matlab programs.* European Mathematical Society, 2010.

[13] Helge Holden, Kenneth H. Karlsen, Nils Henrik Risebro, and Terence Tao. *Operator splitting for the KdV equation.* Math. Comp. 80, pp. 821–846, 2011.

[14] Helge Holden, Nils Henrik Risebro. *Front Tracking for Hyperbolic Conservation Laws*, Springer-Verlag New York, Inc., 2002.

[15] Helge Holden, Kenneth Hvistendahl Karlsen, and Nils Henrik Risebro. *Operator Splitting Method for Generalized Korteweg–De Vries equations*, J. Comp. Physics 153, pp. 203–222, 1999.

[16] P. C Jain, Rama Shankar and Dheeraj Bhardwaj. *Numerical solution of the Korteweg–de Vries (KdV) equation.* Chaos Solitons Fractals 8, no. 6, pp. 943–951, 1997.

[17] A. Jeffrey and T. Kakutani. *Weak nonlinear dispersive waves: A discussion centered around the Korteweg–de Vries equation.* SIAM Review, Vol. 14, No. 4, pp. 582–643, 1972.

[18] Kenneth Hvistendahl Karlsen, Nils Henrik Risebro. *An operator splitting method for nonlinear convection-diffusion equations*. Numer. Math. 77, No. 3, pp. 365–382, 1997.

[19] V. Lakshmikantham and S. Leela. *Nonlinear Differential Equations in Abstract Spaces*. Pergamon Press, Oxford, 1981.

[20] Giovanni Leoni. *A First Course in Sobolev Spaces*, American Mathematical Society, Providence, Rhode Island, 2009.

[21] Randall J. LeVeque. *Numerical Methods for Conservation Laws. Second Edition* Birkhäuser Verlag, Basel, Switzerland, 1992.

[22] Randall J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations. Steady-State and Time-Dependent Problems.* Society for Industrial and Applied Mathematics, Philadelphia, 2007.

[23] Randall J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002.

[24] Yvon Maday, Eitan Tadmor. *Analysis of the Spectral Vanishing Viscosity Method for Periodic Conservation Laws*, SIAM J. Numer. Anal., Vol 26, Issue 4, pp. 854–870, 1989.

[25] *Matematikkleksikon.* Kunnskapsforlaget, H. Aschehoug & Co A/S og Gyldendal ASA, 2. opplag, 2006. (In Norwegian).

[26] R. S. Pathak. *A Course in Distribution Theory and Applications*. Alpha Science International Ltd., UK, 2001.

[27] J. Stoer, R. Bulirsch. *Introduction to Numerical Analysis*, Springer-Verlag New York Inc., 1980.

[28] Eitan Tadmor. *Convergence of Spectral Methods for Nonlinear Conservation Laws*, SIAM J. Numer. Anal., Vol 26, Issue 1, pp. 30–44, 1989.

[29] Eitan Tadmor. *Super Viscosity and Spectral Approximations of Nonlinear Conservation Laws*, Numerical Methods for Fluid Dynamics IV, M. J. Baines and K. W. Morton, eds, pp. 69–82, Clarendon Press, Oxford, 1993.